European Congress of Mathematics

Portorož, 20–26 June 2021

Edited by
Ademir Hujdurović
Klavdija Kutnar
Dragan Marušič
Štefko Miklavič
Tomaž Pisanski
Primož Šparl

EMS PRESS

EMS
PRESS

European Congress of Mathematics

Portorož, 20–26 June 2021

Edited by
Ademir Hujdurović
Klavdija Kutnar
Dragan Marušič
Štefko Miklavič
Tomaž Pisanski
Primož Šparl

EMS PRESS

Editors

Ademir Hujdurović
Faculty of Mathematics, Natural Sciences
and Information Technologies
University of Primorska
Glagoljaška 8
6000 Koper, Slovenia
Email: ademir.hujdurovic@upr.si

Klavdija Kutnar
Faculty of Mathematics, Natural Sciences
and Information Technologies
University of Primorska
Glagoljaška 8
6000 Koper, Slovenia
Email: klavdija.kutnar@upr.si

Dragan Marušič
Faculty of Mathematics, Natural Sciences
and Information Technologies
University of Primorska
Glagoljaška 8
6000 Koper, Slovenia
Email: dragan.marusic@upr.si

Štefko Miklavič
Faculty of Mathematics, Natural Sciences
and Information Technologies
University of Primorska
Glagoljaška 8
6000 Koper, Slovenia
Email: stefko.miklavic@upr.si

Tomaž Pisanski
Faculty of Mathematics, Natural Sciences
and Information Technologies
University of Primorska
Glagoljaška 8
6000 Koper, Slovenia
Email: tomaz.pisanski@upr.si

Primož Šparl
University of Ljubljana
Faculty of Education
Kardeljeva ploščad 16
1000 Ljubljana, Slovenia
Email: primoz.sparl@pef.uni-lj.si

2020 Mathematics Subject Classification: 00B25

ISBN 978-3-98547-051-8, eISBN 978-3-98547-551-3, DOI 10.4171/8ECM

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Published by EMS Press, an imprint of the

European Mathematical Society – EMS – Publishing House GmbH
Institut für Mathematik
Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin, Germany

https://ems.press

Cover illustration based on a photo by Antonio Marano licensed under CC BY 2.0
Typesetting using the authors’ LaTeX sources: Ashry Abdalla, Roisin, Belgium
Printed in Germany
♾ Printed on acid free paper

Preface

The Eighth European Congress of Mathematics (8ECM) was special in many ways. It was the first time that an ECM was entrusted to Slovenia, a relatively small European country of two million inhabitants. In addition, in contrast to previous ECMs that were organized in major cities, this ECM was planned to take place in a small town, Piran, on the shore of the Adriatic coast. Although the municipality of Piran has less than twenty thousand inhabitants, it has a suitable congress infrastructure. Local organization was focused in the small but ambitious University of Primorska in the nearby town of Koper. The main challenge was to break the stereotypes about Slovenia and its mathematics, not only abroad but also locally. “We are too small for such a big project. Forget it!” said a leading Slovenian mathematician. After winning the bid for the 8ECM, the tide changed and all Slovenian institutions practicing mathematics enthusiastically offered their support for the success of the congress.

The second challenge was to increase interest in the congress among the general international mathematical community. Looking at the attendance details of the past seven ECMs, we noticed that the number of participants never exceeded the participation at the first ECM in Paris, which had about 1,500 attendees. Only in the seventh ECM in Berlin, one of the key centers of European mathematics, did the trend turn and the participation surpassed 1,000 attendees. Before we decided to put in a bid for hosting the congress, we wanted to understand why this event, one of the most important international mathematical events, does not attract more participants. By informally interviewing various mathematicians from different countries, including some of the organizers of previous congresses, we identified certain issues, three of which are mentioned below.

First, even the first-rate mathematicians who were actively involved in previous congresses as speakers, prize winners, scientific or prize committee members, etc., in general, rarely find time to attend later congresses. Similarly, many officers of the European Mathematical Society (EMS), belonging to various EMS committees, consider their participation at the ECM of lesser importance. By not being a strong positive role model, they also fail to reach out to the younger generations of European and world mathematicians. For instance, the absence of members of scientific committees who select plenary and invited speakers, and similarly members of prize committees, unfortunately, sends a very negative message not only to the speakers and prize winners themselves but also to the general mathematical community, that it is prestige and not mathematical content that counts at the congress. This is a challenge that the leadership of the EMS should address for future congresses.

Second, many excellent mathematicians tend to avoid worldwide and European mathematics congresses. They find these meetings too big and too broad; they prefer smaller, specialized meetings, which they find much more attractive and productive. We addressed this challenge by increasing the weight of the bottom-up approach, expanding the number of minisymposia and minisymposia speakers. In addition, we allowed each minisymposium to select a special speaker and gave these speakers the opportunity to present their contribution within these proceedings.

The third issue, and by far the most damaging, was beyond our control. The outbreak of the pandemic in the final stages of our preparations at the end of 2019 forced us to rapidly adapt to the developing situation. The numbers of infections were rising steeply, worldwide. By March of 2020, it became clear that our original plan needed drastic changes. All decisions were made hand in hand with the executive committee of the EMS. The option of canceling the congress was never on the table. Postponing the ECM for a couple of months seemed too risky. Eventually, we decided to postpone it for a year. By then, the tools for online conferences were sufficiently developed and most mathematicians had adjusted to giving their presentations via the Internet. Although we were prepared for a live congress, we knew very well that several countries still prohibited their scientists from traveling abroad at that time. This is why we opted for a hybrid approach.

Structure of the 8ECM

There were 62 minisymposia with 902 talks, and 95 talks were delivered in special sessions.

Plenary Speakers. Peter Bühlmann, Xavier Cabré, Franc Forstnerič, Alice Guionnet, Gitta Kutyniok, Monika Ludwig, János Pach, Alfio Quarteroni, Karl-Theodor Sturm, Umberto Zannier.

Invited Speakers. Andrej Bauer, Yves Benoist, Robert Berman, Martin Burger, Albert Cohen, Marius Crainic, Mirjam Dür, Alison Etheridge, Rupert Frank, Aleksey Kostenko, Emmanuel Kowalski, Daniel Kressner, Daniela Kühn, Eugenia Malinnikova, Domenico Marinucci, Eva Miranda, Richard Nickl, Burak Özbağcı, Ilaria Perugia, Gabriel Peyré, Yuri Prokhorov, Alexander A. Razborov, Aner Shalev, Špela Špenko, László Székelyhidi, Anna-Karin Tornberg, Nick Trefethen (FRS), Stuart White.

EMS Prize Lectures. Karim Adiprasito, Ana Caraiani, Alexander Efimov, Kaisa Matomäki, Joaquim Serra, Simion Filip, Alexandr Logunov, Phan Thành Nam, Jack Thorne, Maryna Viazovska.

Abel Lecture. László Lovász.

Felix Klein Prize Lecture. Arnulf Jentzen.

Otto Neugebauer Prize Lecture. Karine Chemla.

Hirzebruch Lecture. Martin Hairer.

Public Lectures. Bojan Mohar, Andrei Okounkov, Stanislav Smirnov, Kathryn Hess, Robin Wilson.

Scholarships were awarded following the regulations of the open call.[1] Out of 274 applications from 64 countries around the world, 105 scholarships were awarded.

The internal satellite event "Optimization in Insurance" was held in Portorož on 23 June 2021.[2] Due to COVID-19, five of the fifteen external satellite events were canceled or postponed.

Four Open Panels and Society Meetings

A highlight of the 8ECM was an open live interview with Jean-Pierre Bourguignon, one of the most influential contemporary European mathematicians who, among other things, served as the second President of the EMS and as President of the European Research Council and left a huge impact on the prestigious Institut des Hautes Études Scientifiques as its Director for 19 years. The interview was conducted by Günter Ziegler, a prominent mathematician who currently serves as the President of the Freie Universität Berlin. The event was chaired by Maria J. Esteban, the Chair of the 8ECM Scientific Committee, and was broadcast live with open access. The interview took place on the last day of the congress, on Friday, 25 June 2021, and was followed by the closing ceremony.

There were eight accompanying events, a career day, and a student competition "Best of 8ECM". There were sixteen 8ECM exhibitors. There were 1,771 participants who completed registration. Participants came from seventy-seven countries, and there were nineteen countries with more than twenty registered participants: Italy, Slovenia, Germany, UK, Spain, USA, France, Russia, Poland, Czech Republic, Croatia, Hungary, Austria, Ukraine, Switzerland, China, Canada, Belgium, and Romania. There were 1,058 contributions in total.

These proceedings covered forty presentations coming from plenary speakers (7), invited speakers (14), EMS prize winners (6), public lecturers (2), and minisymposia keynote speakers (11).

The 8ECM program was broadcast using the Zoom Webinar platform: one Zoom Webinar license for 3,000 participants (used for plenary talks, public talks, the opening, the interview with Jean-Pierre Bourguignon, and the closing ceremony), eight Zoom

[1] See https://www.8ecm.si/about-8ecm/8ecm-scholarships.

[2] See https://conferences.famnit.upr.si/event/20.

Webinar licenses for 1,000 participants (used for invited talks and prize talks), and forty-two Zoom Webinar licenses for 500 participants (used for minisymposia, round tables, exhibitors, etc.).

We certainly hope that the next ECMs will be held live, perhaps with certain key talks and other events broadcast over the Internet, and all talks recorded for posterity. We also hope that the next ECMs will be attended more widely by members of EMS committees and also by members of ECM committees. The bottom-up approach could be significantly extended through engagement by national mathematical societies.

Tomaž Pisanski, 8ECM Organizing Committee Chair
Dragan Marušič, 8ECM Local Scientific Committee Chair
Klavdija Kutnar, 8ECM Organizing Committee Deputy Chair
Ademir Hujdurović, 8ECM Organizing Committee Member

Contents

Plenary lectures

Invited lectures

EMS prize lectures

Public lectures

Minisymposia keynote lectures

8ECM Committees

Scientific committee
Maria J. Esteban (president), CNRS and Paris Dauphine University
Aad van der Vaart, TU Delft
Alexander Kuznetsov, Steklov Mathematical Institute
Barbara Kaltenbacher, University of Klagenfurt
Halil Mete Soner, Princeton University
Joachim Weickert, Saarland University
Luigi Ambrosio, Scuola Normale Superiore di Pisa
Oliver Riordan, University of Oxford
Pavel Pudlák, Czech Academy of Sciences
Péter Pál Pálfy, Hungarian Academy of Sciences
Stefaan Vaes, KU Leuven
Tadeusz Januszkiewicz, Polish Academy of Sciences
Tomaz Pisanski, University of Primorska
Tudor Stefan Ratiu, Shanghai Jiao Tong University
Valeria Simoncini, University of Bologna
Zeev Rudnick, Tel Aviv University

Local scientific committee
Dragan Marušič (chair), University of Primorska
Tomaž Pisanski, University of Primorska
Klavdija Kutnar, University of Primorska
Josip Globevnik, University of Ljubljana
Matej Brešar, University of Maribor and University of Ljubljana
Marko Tadić, University of Zagreb
Annalisa Buffa, École Polytechnique Fédérale de Lausanne

EMS prize committee
Martin Bridson (chair), University of Oxford
Eduardo Casas, University of Cantabria
Fabrizio Catanese, Bayreuth University
Maria Chudnovsky, Princeton University
Monique Dauge, University of Rennes
Herbert Edelsbrunner, Institute of Science and Technology Austria
Per Christian Hansen, Technical University of Denmark
David Kazhdan, University of Jerusalem
Pekka Koskela, University of Jyväskylä

Jan Philip Solovej, University of Copenhagen
Domokos Szász, Budapest University of Technology and Economics
Ragnar Winther, University of Oslo
Cem Yalçın Yıldırım, Boğaziçi University

Organizing committee
Tomaž Pisanski (chair), University of Primorska
Klavdija Kutnar (deputy chair), University of Primorska
Aleš Oven (secretary), University of Primorska
Ademir Hujdurović, University of Primorska
Alen Orbanić, Abelium d.o.o.
Boštjan Kuzman, University of Ljubljana
Daniel Eremita, University of Maribor
Dean Crnković, University of Rijeka
Gregor Dolinar, University of Ljubljana
Jasna Prezelj, University of Primorska and University of Ljubljana
Martin Milanič, University of Primorska
Nino Bašić, University of Primorska
Rok Požar, University of Primorska
Russ Woodroofe, University of Primorska
Susan Cook, University of Primorska
Ted Dobson, University of Primorska
Vida Groznik, University of Primorska
Vito Vitrih, University of Primorska
Dragan Stevanović, Serbian Academy of Science and Arts
Michael Drmota, Vienna University of Technology
Norbert Seifter, University of Leoben
Pablo Spiga, University of Milano-Bicocca
Tamás Szőnyi, Eötvös Loránd University and University of Primorska
Wacław Marzantowicz (special advisor), Adam Mickiewicz University
Tine Šukljan (IT support team leader), Innorenew CoE and University of Primorska

List of sponsors

Main sponsor
The Slovenian Insurance Association

Gold sponsor
Elsevier

Sponsors
Actual I.T., d.d.
Adacta d.o.o.
CREAplus d.o.o.
Dinit d.o.o.

Sponsor for the minisymposium Number Theory (MS-ID 63)
Journal de Théorie des Nombres de Bordeaux

Sponsor for the minisymposia Number Theory (MS-ID 63) and Arithmetic and Geometry of Algebraic Surfaces (MS-ID 45)
Dipartimento di Matematica e Informatica, Università della Calabria

Sponsor for the minisymposium Arithmetic and Geometry of Algebraic Surfaces (MS-ID 45)
Università degli Studi di Genova

Donors
E-misija d.o.o. – Računalniška oprema
Hidria d.o.o.
Polydron Ltd.
SIQ Ljubljana
SORBIT VALJI d.o.o.
The Slovenian Association of Actuaries
ZBS The Bank Association of Slovenia – GIZ
Zometool Inc.

Donors to the EMS prizes fund
EMS Press
Foundation Compositio Mathematica

Plenary lectures

DOI 10.4171/8ECM/13

Regularity of stable solutions to reaction-diffusion elliptic equations

Xavier Cabré

Abstract. The boundedness of stable solutions to semilinear (or reaction-diffusion) elliptic PDEs has been studied since the 1970s. In dimensions 10 and higher, there exist stable energy solutions which are unbounded (or singular). This note describes, for non-expert readers, a recent work in collaboration with Figalli, Ros-Oton, and Serra, where we prove that stable solutions are smooth up to the optimal dimension 9. This solves an open problem posed by Brezis in the mid-nineties concerning the regularity of extremal solutions to Gelfand-type problems. We also describe, briefly, a famous analogue question in differential geometry: the regularity of stable minimal surfaces.

1. Hilbert's 19th problem and the principle of least action

In many physical phenomena and geometric problems, observable states try to minimize a certain functional. When we describe the possible states by functions u of one or several real variables, the functional is a real valued function A acting on such functions. In classical mechanics, A is called the *action* and is given by the integral of a Lagrangian. A simple example is the motion of a particle under gravitation, in which its position is given by $u = u(x)$ (where $x = t \in \mathbb{R}$ is time) and the action is the difference of kinetic and potential energies. In geometry, two important examples are geodesics (curves in a Riemannian manifold that are critical points of the length functional) and minimal surfaces (hypersurfaces of Euclidean space that are critical points of the area functional).

Hilbert's 19th problem asks whether minimizers of elliptic functionals are always analytic. When the functional is given by $A(u) = \int_\Omega L(\nabla u(x))dx$ for some domain $\Omega \subset \mathbb{R}^n$ and convex function $L : \mathbb{R}^n \to \mathbb{R}$ (here $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}$), the problem was solved independently in the late 1950s by Ennio De Giorgi and John Forbes Nash, Jr.

2020 Mathematics Subject Classification. Primary 35B65; Secondary 35B35.
Keywords. Semilinear elliptic equations, stable solutions, extremal solutions, regularity, a priori estimates.

Our work [5] takes up on the same question for the Lagrangian $L(\nabla u, u) = \frac{1}{2}|\nabla u|^2 - F(u)$, which depends also on the variable u.

The *principle of least action* in mechanics states that observable states should be not only critical points of the action but absolute minimizers of it. In the previous setting, when $L = L(\nabla u)$ is a convex function defined in $\mathbb{R}^n$, we have that A is convex. As a consequence, any critical point of A (if it exists) is an absolute minimizer. Thus, the principle of least action is, here, fulfilled. However, the principle is violated in some real life situations, described next, in which we observe states that are only local minimizers (minimizers among small perturbations), even in situations when an absolute minimizer exists. Still, the observed state being a local minimizer, it is therefore a *stable state*, in the sense that the second variation of the functional is nonnegative definite when computed at such state.

In these minimization problems, the competitors among which one looks for critical points are functions, or surfaces, with prescribed given boundary values—the end points of the trajectory of a mechanical particle, or of a geodesic, or a given wire from which soap films or (minimal) surfaces are spanned. The key point is that the functional in many of these variational problems is not convex. Thus, it may admit critical points which are not absolute minimizers (but are only local minimizers) and even unstable critical points (which, consequently, are not even local minimizers).

2. Stable minimal surfaces

An instructive example is that of catenoids: a soap film or minimal surface formed between two coaxial parallel circular rings. In the nice paper "*In situ observation of a soap-film catenoid—a simple educational physics experiment*" by Ito and Sato [15], catenoids are experimentally produced in a lab and videotaped while the distance between the two circular wires is continuously increased. Note that, besides catenoids, there always exists another critical point of the area: the two flat disks spanned by the wires. For each small enough distance, two catenoids exist and the one with a thicker neck is an absolute minimizer of the area (clearly the disks have much larger area). As the wires separate, there is a distance h_0 at which both states (the thick-neck catenoid and the two disks) have the same area. Right after it, the two disks become the absolute minimizer, while the catenoid is only a local minimizer. Still, for an interval of distances $h > h_0$ the videotaped surface is the catenoid—not the absolute minimizing disks. Such catenoid is a *stable minimal surface*—stability understood as defined in the previous section. Up to these distances, the unobserved thinner neck catenoid always existed and was an unstable minimal surface (this fits with the idea that a functional with two local minima should have a third unstable critical point). Finally, there is a second larger distance h_c at which the local minimizer (the thick catenoid) and the unstable critical point (the thin one) get together to produce an inflection point. Right after it, the two-disks is the only critical point. In the experiment [15], the unstable catenoid is photographed for a short instant, right before the distance h_c. Quickly after such instant, the thin neck collapses and the catenoid film succeeds to transform itself into the two disks.

The regularity theory of minimal surfaces has been the source of many important progresses in the area of PDEs. In the 1960s the Italian school proved that the Simons cone

$$x_1^2 + \cdots + x_m^2 = x_{m+1}^2 + \cdots + x_{2m}^2$$

is an absolute minimizer of area (for its own boundary values in any ball) if $2m \geq 8$, while it is not even stable in dimensions 2, 4, and 6.[1] Thus, minimizing minimal surfaces of dimension n may have (conical) singularities when $n \geq 7$. At the same time, a sequence of outstanding contributions by different authors (J. Simons' being a prominent one) established that n-dimensional (absolute) minimizing minimal surfaces in $\mathbb{R}^{n+1}$ are always smooth when $n \leq 6$.

It is a long-standing open problem to extend this regularity result to the larger class of *stable minimal surfaces*. It is only known to be true for surfaces of dimension 2, 3, or 4. See [7, 8] for more details on these issues.

3. Stable solutions to reaction-diffusion elliptic equations

The paper [5] takes on the analogue question (the regularity of *stable solutions*) for equations of the form $-\Delta u = f(u)$, where Δ is the Laplacian. They are called semilinear or reaction-diffusion elliptic equations and arise in many physical and biological situations. In the following combustion problem, a similar phenomenon to that of catenoids occurs. It concerns the thermal self-ignition of a chemically active mixture of gases in a container. The model was introduced by Frank-Kamenetskii in the 1930s but became popular within the mathematical community when Barenblatt wrote Chapter 15 of the volume [12], edited by Gelfand in 1963. Here $x \in \Omega \subset \mathbb{R}^n$ denotes points in the container Ω and $u = u(x)$ is the temperature at the point. The action functional is the difference of kinetic and potential energies:

$$A(u) = \int_\Omega \left(\frac{1}{2} \left| \nabla u(x) \right|^2 - F\big(u(x)\big) \right) dx,$$

where $F : \mathbb{R} \to \mathbb{R}$ is a given function, which Barenblatt chose to be $F(u) = \lambda e^u$, with λ a positive constant, from Arrhenius law in chemical kinetics. For convenience,

[1] This different behavior can be roughly understood noticing that the Jacobian for area in spherical coordinates, $r^{2m-2} dr$, becomes smaller at the origin as the dimension $2m$ increases. Note that for $2m = 2$, the minimizer clearly avoids the origin: for the boundary values of the cone, it is given by two parallel lines (and not by the "cross" passing through the origin).

we will impose vanishing boundary conditions: the temperature is kept at $u = 0$ on the boundary $\partial\Omega$. Making a first variation $u + \varepsilon v$ and integrating by parts, one easily sees that critical points of A satisfy the *reaction-diffusion equation*

$$-\Delta u = f(u) \quad \text{in } \Omega \subset \mathbb{R}^n, \tag{3.1}$$

where $f = F'$. In the case (among others) of the so-called *Gelfand problem*, $-\Delta u = \lambda e^u$, the situation is similar to the one of catenoids. For a certain range $\lambda \in (0, \lambda^*)$ of parameters, there exists a stable solution u_λ—that is, a solution at which the functional A has a nonnegative definite second variation. Such stable solution is not an absolute minimizer since A is unbounded by below (note that the potential density $F(u) = e^u$ grows faster at infinity than the quadratic kinetic density $|\nabla u(x)|^2$). For some parameters $\lambda \in (0, \lambda^*)$, there might also exist (this will depend on the container Ω) unstable solutions of the same problem. For $\lambda = \lambda^*$, the limit of the functions u_λ is an L^1 weak stable solution, called the extremal solution. When $\lambda > \lambda^*$, no solution exists—in the same way that catenoids did not exist for distances h between the wires larger than h_c.

To better understand the problem, let us consider the nonlinear heat equation

$$v_t - \Delta v = f(v), \tag{3.2}$$

where $v = v(x,t)$ and t is time. Now, a stable solution $u = u(x)$ of (3.1) can be understood as a stationary solution of (3.2) which is stable in the sense of Lyapunov—note that a simple computation shows that the action functional $A(v(\cdot,t))$ is non-increasing in the time t. The problem is nonlinear due to the sources of heat, $f(v(x,t))$ or $f(u(x))$: the production of heat depends nonlinearly on the actual temperature. As described in [13, 14], equation (3.2) describes the evolution of an initially uniform temperature $v(\cdot, 0) \equiv 0$ which diffuses in space and increases in the container due to the heat release given by the reaction term $f(v)$—note that in Gelfand's problem the initial heat source $\lambda f(0) = \lambda e^0 = \lambda$ is already positive. The parameters λ for which there exists a stable solution of (3.1) correspond to ignition failure (the reactive component undergoes partial oxidation and results in establishing a stationary temperature profile equal to the stable solution). Instead, $\lambda > \lambda^*$ (when there exists no stationary solution) means successful auto-ignition in the combustion process.

Since we will turn now to regularity issues, let us recall that Fourier invented his omnipresent Fourier series to understand the linear heat equation. On the other hand, the regularity theory for the stationary linear Poisson equation $-\Delta u = g(x)$ is at the center of PDE theory and also propitiated the development of many tools, such as the theory of singular integrals in harmonic analysis. In particular, the Lebesgue-integrability requirements for $g = g(x)$ which are needed to guarantee the boundedness of the potential function u are well known. This is relevant since, for our nonlinear equation (3.1), $-\Delta u = f(u)$, the obstruction for regularity is the possibility that u becomes unbounded somewhere, that is, u blows-up at some points (and hence u is singular). As we will see next, this singular behavior can be produced by the strength of some reaction terms $f(u)$. A technical detail for experts is that, since our problem is variational, in what follows we consider only (singular) solutions for which each term of the action functional is integrable (that is, energy solutions).

When $n \geq 3$, $\Omega = B_1$ is the unit ball,

$$u = \log \frac{1}{|x|^2}, \qquad f(u) = 2(n-2)e^u,$$

then a simple computation shows that we are in the presence of a singular solution of (3.1) vanishing on ∂B_1. As the Simons cone in minimal surfaces theory, this explicit solution turns out to be stable in high dimensions, precisely when $n \geq 10$. On the other hand, in the 1970s Crandall and Rabinowitz [9] established that if

$$f(u) = e^u \quad \text{or} \quad f(u) = (1+u)^p \quad \text{with } p > 1,$$

then stable solutions in any smooth bounded domain Ω are bounded (and hence smooth and analytic, by classical elliptic regularity theory) when $n \leq 9$. These results were the main reason for Haim Brezis to raise the following question in the 1990s (which we cite almost literally from a later reference).

Brezis ([1, Open problem 1]). *Is there something "sacred" about dimension 10? More precisely, is it possible in "low" dimensions to construct some f (and some Ω) for which a singular stable solution exists? Alternatively, can one prove in "low" dimensions that every stable solution is smooth for every f and every Ω?*

Other open questions on stable solutions were posed by Brezis and Vázquez [2].

The last twenty five years have produced a large literature on Gelfand-type problems. See the monograph [10] for an extensive list of results and references. For a certain type of nonlinearities f, some of these works are related to micro-electro-mechanical systems (MEMS); see [11].

The main developments proving that stable solutions to (3.1) are smooth (no matter what the nonlinearity f is) were made

- by Nedev [16] in 2000, when $n \leq 3$ (and f is convex);
- by Cabré and Capella [4] in 2006, when $\Omega = B_1$ (u is radially symmetric) and $n \leq 9$;
- by Cabré [3] in 2010, when $n \leq 4$ (and Ω is convex).

Note that the 2006 result in the radially symmetric case, [4], accomplished the optimal dimension $n \leq 9$ for every nonlinearity f. This gave hope for the result to be true also in the general nonradial case, though no certainty was assured—note that Brezis' statement above leaves both the affirmative and negative answers as possible ones. Since 2010, after [3], the regularity result was only known up to dimension $n = 4$. Two attempts in higher dimensions (recorded in [5]) gave only very partial answers.

The work [5] finally solves the open problem, by establishing the regularity of stable solutions to (3.1) in the interior of any open set Ω in the optimal dimensions $n \leq 9$ under the only requirement for the nonlinearity f to be nonnegative. Furthermore, adding the vanishing boundary condition $u = 0$ on $\partial\Omega$, the article proves regularity up to the boundary when Ω is of class C^3 and $n \leq 9$, assuming now f to be nonnegative, nondecreasing, and convex. Both results come along with new universal Hölder-continuity estimates which have a very weak norm (the L^1-norm) of the solution on their right-hand sides. They read, respectively, as

$$\|u\|_{C^\alpha(\overline{B}_{1/2})} \leq C\,\|u\|_{L^1(B_1)}, \quad \|u\|_{C^\alpha(\overline{\Omega})} \leq C_\Omega\,\|u\|_{L^1(\Omega)},$$

where $\alpha > 0$ and C are dimensional constants, while C_Ω depends only on Ω. These estimates are rather surprising (because of their universality) for a nonlinear problem, specially since they make no reference to the reaction nonlinearity f. The stability of the solution u is crucial for their validity. For the expert reader, [5] also establishes another open problem from [2]: an a priori $H^1 = W^{1,2}$ estimate for stable solutions in all dimensions n.

Whether the nonnegativeness of f is a needed requirement for interior regularity remains as an open question. It is only known to be unnecessary for $n \leq 4$, as well as for $n \leq 9$ in the radial case.

The proofs in the article are too technical to be described here. Let us only say that a key point is to use the stability property under two different types of small perturbations of the solution u (one in the radial direction, the other in the normal direction to the level sets):

$$u\big(x + \varepsilon|x|^{(2-n)/2}\zeta(x)x\big), \quad u\left(x + \varepsilon\eta(x)\frac{\nabla u(x)}{|\nabla u(x)|}\right),$$

where ζ and η are cut-off functions.

After [5], an analogue result for equations involving the p-Laplacian has been proved by Cabré, Miraglio, and Sanchón [6]. It is optimal in terms of dimensions for $p > 2$, but not for $p < 2$—this case remains as an open problem. On the other hand, for the recently very active area of fractional Laplacians, an optimal result for $(-\Delta)^s u = f(u)$ is largely open—even in the radial case. The optimal dimensions for regularity have only been accomplished in a 2014 work of Ros-Oton [17] for the Gelfand nonlinearity $f(u) = \lambda e^u$ in symmetric convex domains—but for any fraction $s \in (0, 1)$ of the Laplacian.

Funding. Research supported by grants MTM2017-84214-C2-1-P and RED2018-102650-T funded by MCIN/AEI/10.13039/501100011033 and by “ERDF A way of making Europe.”

References

[1] H. Brezis, Is there failure of the inverse function theorem? In *Morse Theory, Minimax Theory and Their Applications to Nonlinear Differential Equations*, pp. 23–33, New Stud. Adv. Math. 1, Int. Press, Somerville, MA, 2003 Zbl 1200.35144 MR 2056500

[2] H. Brezis and J. L. Vázquez, Blow-up solutions of some nonlinear elliptic problems. *Rev. Mat. Univ. Complut. Madrid* **10** (1997), no. 2, 443–469 Zbl 0894.35038 MR 1605678

[3] X. Cabré, Regularity of minimizers of semilinear elliptic problems up to dimension 4. *Comm. Pure Appl. Math.* **63** (2010), no. 10, 1362–1380 Zbl 1198.35094 MR 2681476

[4] X. Cabré and A. Capella, Regularity of radial minimizers and extremal solutions of semilinear elliptic equations. *J. Funct. Anal.* **238** (2006), no. 2, 709–733 Zbl 1130.35050 MR 2253739

[5] X. Cabré, A. Figalli, X. Ros-Oton, and J. Serra, Stable solutions to semilinear elliptic equations are smooth up to dimension 9. *Acta Math.* **224** (2020), no. 2, 187–252 Zbl 1467.35172 MR 4117051

[6] X. Cabré, P. Miraglio, and M. Sanchón, Optimal regularity of stable solutions to nonlinear equations involving the p-Laplacian. *Adv. Calc. Var.* **15** (2022), no. 4, 749–785 Zbl 1500.35073 MR 4489602

[7] O. Chodosh and C. Li, Stable minimal hypersurfaces in $\mathbf{R}^4$. arXiv:2108.11462

[8] T. H. Colding and W. P. Minicozzi II, In search of stable geometric structures. *Notices Amer. Math. Soc.* **66** (2019), no. 11, 1785–1791 Zbl 1465.53005 MR 3971084

[9] M. G. Crandall and P. H. Rabinowitz, Some continuation and variational methods for positive solutions of nonlinear elliptic eigenvalue problems. *Arch. Ration. Mech. Anal.* **58** (1975), no. 3, 207–218 Zbl 0309.35057 MR 382848

[10] L. Dupaigne, *Stable solutions of elliptic partial differential equations*. Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math. 143, Chapman & Hall/CRC, Boca Raton, FL, 2011 Zbl 1228.35004 MR 2779463

[11] P. Esposito, N. Ghoussoub, and Y. Guo, *Mathematical analysis of partial differential equations modeling electrostatic MEMS*. Courant Lect. Notes Math. 20, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2010 Zbl 1223.35003 MR 2604963

[12] I. M. Gelfand, Some problems in the theory of quasilinear equations. *Amer. Math. Soc. Transl. Ser. 2* **29** (1963), 295–381 Zbl 0127.04901 MR 0153960

[13] P. V. Gordon and V. Moroz, Gelfand-type problem for two-phase porous media. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **470** (2014), no. 2163, 20130573 Zbl 1348.35095 MR 3159562

[14] P. V. Gordon, V. Moroz, and F. Nazarov, Gelfand-type problem for turbulent jets. *J. Differential Equations* **269** (2020), no. 7, 5959–5996 Zbl 1447.35138 MR 4104947

[15] M. Ito and T. Sato, In situ observation of a soap-film catenoid—a simple educational physics experiment. *European J. Phys.* **31** (2010), no. 2, 357–365 MR 2595581

[16] G. Nedev, Regularity of the extremal solution of semilinear elliptic equations. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (2000), no. 11, 997–1002 Zbl 0955.35029 MR 1779693

[17] X. Ros-Oton, Regularity for the fractional Gelfand problem up to dimension 7. *J. Math. Anal. Appl.* **419** (2014), no. 1, 10–19 Zbl 1294.35190 MR 3217131

Xavier Cabré
Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluis Companys 23,
08010 Barcelona; Departament de Matemàtiques and IMTech, Universitat Politècnica
de Catalunya, Diagonal 647, 08028 Barcelona; and Centre de Recerca Matemàtica,
Edifici C, Campus Bellaterra, 08193 Bellaterra, Spain; xavier.cabre@upc.edu

DOI 10.4171/8ECM/08

Minimal surfaces in Euclidean spaces by way of complex analysis

Franc Forstnerič

Abstract. This is an expanded version of my plenary lecture at the 8th European Congress of Mathematics in Portorož on 23 June 2021. The main part of the paper is a survey of recent applications of complex-analytic techniques to the theory of conformal minimal surfaces in Euclidean spaces. New results concern approximation, interpolation, and general position properties of minimal surfaces, existence of minimal surfaces with a given Gauss map, and the Calabi–Yau problem for minimal surfaces. To be accessible to a wide audience, the article includes a self-contained elementary introduction to the theory of minimal surfaces in Euclidean spaces.

1. Minimal surfaces: A link between mathematics, science, engineering, and art

Minimal surfaces are among the most beautiful and aesthetically pleasing geometric objects. These are surfaces in space which locally minimize area, in the sense that any small enough piece of the surface has the smallest area among surfaces with the same boundary. From the physical viewpoint, these are surfaces minimizing tension, hence in equilibrium position. They appear in a variety of applications to engineering, biology, architecture, and others.

The subject has a luminous history, going back to 1744 when Leonhard Euler [32] showed that pieces of the surface now called *catenoid* (see Example 2.7) have smallest area among all surfaces of rotation in the 3-dimensional Euclidean space $\mathbb{R}^3$. The catenoid derives it name from *catenary*, the curve that an idealized hanging chain assumes under its own weight when supported only at its ends. The model catenary is the graph of the hyperbolic cosine function $y = \cosh x$, and a catenoid is obtained by rotating this curve around the x-axis in the (x, y, z)-space. Topologically, a catenoid is a cylinder, and as a conformal surface it is the puncture plane $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. From

2020 Mathematics Subject Classification. Primary 53A10; Secondary 32H02.
Keywords. Minimal surface, conformal harmonic map, Calabi–Yau problem.

the mathematical viewpoint, the catenoid is one of the most paradigmatic examples of minimal surfaces, and it appears in several important classification results and in proofs of major theorems.

The subject of minimal surfaces was put on solid footing by Joseph–Louis Lagrange who developed the calculus of variations during 1760–61, thereby reducing the problem of finding stationary points of functionals to a second-order partial differential equation, now called Lagrange's equation. His work was published in 1762 by Accademia delle scienze di Torino [51,52] and is available in his collected works [53]. In [51], Lagrange applied his new method to a variety of problems in physics, dynamics, and geometry. In particular, he derived the *equation of minimal graphs*. The term *minimal surface* has since been used for a surface which is a stationary point of the area functional. The question whether a domain in a minimal surface truly minimizes the area among nearby surfaces with the same boundary can be analyzed by considering the second variation of area. It was later shown that a minimal graph in $\mathbb{R}^3$ over a compact convex domain in $\mathbb{R}^2$ is an absolute area minimizer, and hence small enough pieces of any minimal surface are area minimizers.

In 1776, Jean Baptiste Meusnier [66] discovered that domains in a surface in $\mathbb{R}^3$ are minimal in the sense of Lagrange if and only if the surface has vanishing mean curvature at every point. He also described the second known minimal surface, the *helicoid*; see Example 2.8. It is obtained by a line in 3-space rotating at a constant rate as it moves at a constant speed along the axis of rotation, which is perpendicular to the rotating line. Helicoid is the geometric shape of a device known as *Archimedes' screw* (or the water screw, screw pump, or Egyptian screw), named after Greek philosopher and mathematician Archimedes who described it around 234 BC on the occasion of his visit to Egypt. There is evidence that this device had been used in ancient Egypt much earlier. The helicoid is sometimes called "double spiral staircase"—each of the two half-lines sweeps out a spiral staircase, and these two staircases only meet along the axis of rotation. Therefore, its physical model is a convenient device for letting people ascend and descend a staircase without the two crowds meeting in-between. From a different field, DNA molecules assume the shape of a helicoid.

Topologically and conformally the helicoid is the plane. Its name derives from helix—for every point on the helicoid, there is a helix (a spiral curve) contained in the helicoid which passes through that point. The helicoid plays a major role in the classification of properly embedded minimal surfaces in $\mathbb{R}^3$; see the survey paper [28] by Tobias H. Colding and William P. Minicozzi.

Minimal surfaces appear naturally in the physical world. Laws of physics imply that a soap film spanned by a given frame (i.e., a closed Jordan curve) is a minimal surface. The reason is that this shape minimizes the surface tension and puts it in equilibrium position. Soap films, bubbles, and surface tension were studied by the Belgian physicist Joseph Plateau in the 19th century. Based on his experiments, Karl Weierstrass formulated in 1873 the *Plateau problem*, conjecturing that any closed Jordan curve in $\mathbb{R}^3$ spans a minimal surface (in fact, a minimal disc). This was confirmed by Tibor Radó [71, 72] (1930) and Jesse Douglas [31] (1931). For his work on the Plateau problem, Douglas received one of the first two Fields Medals at the International Congress of Mathematicians in Oslo in 1936. Half a century later, it was shown that the disc of smallest area with given boundary curve (the Douglas–Morrey solution of the Plateau problem) has no branch points; see the monograph by Anthony Tromba [77]. Furthermore, if the curve lies in the boundary of a convex domain in $\mathbb{R}^3$, then the solution is embedded according to William H. Meeks and Shing Tung Yau [63, 64].

Minimal surfaces are also studied in more general Riemannian manifolds of dimension at least three. Holomorphic curves in complex Euclidean spaces $\mathbb{C}^n$ for $n > 1$, or in any complex Kähler manifold of complex dimension at least two, are special but important examples of minimal surfaces. As pointed out by Colding and Minicozzi [28], there are several fields where minimal surfaces are actively used in understanding physical phenomena. In particular, they come up in the study of compound polymers, protein folding, etc. They also play a prominent role in art, especially in architecture.

The connection between minimal surfaces in Euclidean spaces and complex analysis has been known since mid-19th century. The basic fact is that a conformal immersion $X : M \to \mathbb{R}^n$ from a Riemann surface M parameterizes a minimal surface if and only if the map X is harmonic (see Theorem 2.1); equivalently, the complex derivative $\partial X/\partial z$ in any local holomorphic coordinate z on M is holomorphic. Furthermore, the immersion X is conformal if and only if $\partial X/\partial z$ assumes values in the null quadric $\mathbb{A} \subset \mathbb{C}^n$, given by the equation $z_1^2 + z_2^2 + \cdots + z_n^2 = 0$ (see (2.23)), and $\partial X/\partial z \neq 0$ if X is an immersion. This leads to the *Enneper–Weierstrass representation* of any conformally immersed minimal surface $M \to \mathbb{R}^n$ as the real part of the integral of a holomorphic map $f : M \to \mathbb{A}_* = \mathbb{A} \setminus \{0\} \subset \mathbb{C}^n$ (see Theorem 2.6). The period vanishing conditions on f along closed curves in M ensure that the integral is well defined. The formula is most concrete in dimension $n = 3$ (see (2.25)) due to an explicit 2-sheeted parameterization of the null quadric $\mathbb{A} \subset \mathbb{C}^3$ by $\mathbb{C}^2$.

This connection between minimal surfaces and holomorphic maps was used by Bernhard Riemann around 1860 in his construction of properly embedded minimal surfaces in $\mathbb{R}^3$, now called *Riemann's minimal examples* [73] (see the paper [60] by William H. Meeks and Joaquín Pérez), and in numerous further works by other authors. It was popularized again in modern times by Robert Osserman [69].

Despite the long and illustrious history of the subject, the author in collaboration with Antonio Alarcón, Francisco J. López, and others obtained in the last decade a string of new results by exploiting the Enneper–Weierstrass representation. The main point in our approach is that the punctured null quadric $\mathbb{A}_*$ is a complex homogeneous manifold, hence an *Oka manifold*, a notion introduced in [34] and treated in [35, Chapter 5]. This implies that holomorphic maps from any open Riemann surface (and, more generally, from any Stein manifold, that is, a closed complex submanifold of a complex Euclidean space $\mathbb{C}^N$) to $\mathbb{A}_*$ satisfy the Runge–Mergelyan approximation theorem and the Weierstrass interpolation theorem in the absence of topological obstructions. Together with methods of convexity theory, this gave rise to many new constructions of conformal minimal surfaces with interesting properties; see Theorem 3.1. By using parametric versions of these results, it was possible to determine the rough topological shape (i.e., the weak or strong homotopy type) of the space of nonflat conformal minimal immersions from any given open Riemann surface into $\mathbb{R}^n$ (see Theorem 3.2). It was also shown that every natural candidate is the Gauss map of a conformal minimal surface in $\mathbb{R}^n$ (see Theorem 3.3).

Another complex analytic technique, which has recently had a major impact on the field, is an adaptation of the classical Riemann–Hilbert boundary value problem to conformal minimal surfaces and holomorphic null curves in Euclidean spaces. This led to an essentially optimal solution of the *Calabi–Yau problem for minimal surfaces*, originating in conjectures of Eugenio Calabi from 1965; see Theorems 3.5 and 3.6. This technique was also used in the construction of complete proper minimal surfaces in minimally convex domains of $\mathbb{R}^n$ (see [16, Chapter 8]).

The recent results presented in Section 3 are carefully explained in the monograph [16] published in March 2021. The corresponding developments on non-orientable minimal surfaces are described in the AMS Memoir [15] from 2020. It is needless to say that both of these publications contain many other results not mentioned here.

In 2021, the author and David Kalaj [38] obtained an optimal Schwarz–Pick lemma for conformal minimal discs in the ball of $\mathbb{R}^n$ and introduced the notion of hyperbolicity of domains in $\mathbb{R}^n$, in analogy with Kobayashi hyperbolicity of complex manifolds. This new topic is currently being developed, and it is too early to include it here.

2. An elementary introduction to minimal surfaces

To make the article accessible to a wide audience including advanced undergraduate students of Mathematics, we present in this section a self-contained introduction to the theory of minimal surfaces in Euclidean spaces. We assume familiarity with elementary calculus, topology, and rudiments of complex analysis; however, no a priori knowledge of differential geometry is expected. We shall use the fact that metric-related quantities such as length, area, and curvature of curves and surfaces in a Euclidean space $\mathbb{R}^n$ are invariant under translations and orthogonal maps of $\mathbb{R}^n$; these are the isometries of the Euclidean metric, also called *rigid motions*. For simplicity of presentation, we focus on minimal surfaces parameterized by plane domains, although the same methods apply on an arbitrary open Riemann surface. More complete treatment is available in a number of texts; see [16, 20, 26, 30, 55, 58, 59, 68, 69], among others. For the theory of non-orientable minimal surfaces, see [15].

2.1. Conformal maps and conformal structures on surfaces

From the physical viewpoint, the most natural parameterization of a minimal surface is by a *conformal map* (from a plane domain, or a conformal surface). A conformal parameterization minimizes the total energy of the map and makes the tension uniformly spread over the surfaces. We give a brief introduction to the subject of conformal maps, referring to [16, Sections 1.8–1.9] for more details and further references.

Let D be a domain in $\mathbb{R}^2$ with coordinates (u, v). A $\mathscr{C}^1$ map $X : D \to \mathbb{R}^n$ $(n \geq 2)$ is an *immersion* if the partial derivatives $X_u = \partial X/\partial u$ and $X_v = \partial X/\partial v$ are linearly independent at every point of D. An immersion is said to be *conformal* if its differential dX_p at any point $p \in D$ preserves angles. It is elementary to see (cf. [16, Lemma 1.8.4]) that an immersion X is conformal if and only if

$$|X_u| = |X_v| \quad \text{and} \quad X_u \cdot X_v = 0. \tag{2.1}$$

Here, $\mathbf{x} \cdot \mathbf{y}$ denotes the Euclidean inner product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$ is the Euclidean length of $\mathbf{x}$. A smooth map $X : D \to \mathbb{R}^n$ (of class $\mathscr{C}^1$, not necessarily an immersion) is called conformal if (2.1) holds at each point. It clearly follows that X has rank zero at non-immersion points.

Let M be a topological surface. A *conformal structure* on M is given by an atlas $\mathcal{U} = \{(U_i, \phi_i)\}_{i \in I}$ with charts $\phi_i : U_i \xrightarrow{\cong} V_i \subset \mathbb{R}^2$ whose transition maps

$$\phi_{i,j} = \phi_i \circ \phi_j^{-1} : \phi_j(U_i \cap U_j) \to \phi_i(U_i \cap U_j)$$

are conformal diffeomorphisms of plane domains. Identifying $\mathbb{R}^2$ with the complex plane $\mathbb{C}$, each map $\phi_{i,j}$ is biholomorphic or anti-biholomorphic. A surface M endowed with a conformal structure (more precisely, with an equivalence class of conformal structures) is a *conformal surface*. If M is orientable, then by choosing the charts ϕ_i in a conformal atlas to preserve orientation, the transition maps $\phi_{i,j}$ are biholomorphic; hence, $\mathcal{U}$ is a complex atlas and $(M, \mathcal{U})$ is a *Riemann surface*. A connected non-orientable conformal surface M admits a two-sheeted conformal covering $\widetilde{M} \to M$ by a Riemann surface $\widetilde{M}$.

Assume now that g is a *Riemannian metric* on a smooth surface M, i.e., a smoothly varying family of scalar products g_p on tangent spaces T_pM, $p \in M$. In any local coordinate (u, v) on M, the metric g has an expression

$$g = E\,du^2 + 2F\,du\,dv + G\,dv^2,$$

where the coefficient functions E, F, G satisfy $EG - F^2 > 0$. A local chart (u, v) is said to be *isothermal* for g if the above expression simplifies to

$$g = \lambda(u, v)(du^2 + dv^2) = \lambda |dz|^2, \quad z = u + \mathrm{i}v$$

for some positive function λ. An important result, first observed by Carl Friedrich Gauss, is that in a neighborhood of any point of M there exist smooth isothermal coordinates. One way to obtain such coordinates is from solutions of the classical *Beltrami equation*. We refer to [16, Sections 1.8–1.9] for a more precise statement and references. Since the transition map between any pair of isothermal charts is a conformal diffeomorphism, we thus obtain a conformal atlas on M consisting of isothermal charts. The upshot is that every Riemannian metric on a smooth surface determines a conformal structure. Furthermore, a pair of Riemannian metrics g, $\tilde{g}$ on M determine the same conformal structure if and only if $\tilde{g} = \mu g$ for a smooth positive function μ on M.

Denote by $\mathbf{x} = (x_1, \ldots, x_n)$ the Euclidean coordinates on $\mathbb{R}^n$ and by

$$ds^2 = dx_1^2 + \cdots + dx_n^2$$

the Euclidean metric. If $X = (X_1, \ldots, X_n) : M \to \mathbb{R}^n$ is a smooth immersion, then

$$g = X^*(ds^2) = (dX_1)^2 + \cdots + (dX_n)^2$$

is a Riemannian metric on M, called the *first fundamental form*. By the definition of g, the map $X : (M, g) \to (\mathbb{R}^n, ds^2)$ is an isometric immersion. By what has been said, g determines a conformal structure on M (assuming now that M is a surface), and in this structure the map X is a conformal immersion. More precisely, $X(u, v)$ is conformal in any isothermal local coordinate (u, v) on M.

This shows that any immersion $X : M \to \mathbb{R}^n$ from a smooth surface determines a unique conformal structure on M which makes X a conformal immersion. If in addition M is oriented, we get the structure of a Riemann surface. Results of conformality theory imply that if D is a domain in $\mathbb{R}^2$ and $X : D \to \mathbb{R}^n$ is an immersion, then there is a diffeomorphism $\phi : D' \to D$ from another domain $D' \subset \mathbb{R}^2$ such that the immersion $X \circ \phi : D' \to \mathbb{R}^n$ is conformal. In particular, if D is the disc, then we may take $D' = D$.

The same arguments and conclusions apply to immersions of a smooth surface M into an arbitrary Riemannian manifold $(N, \tilde{g})$ in place of $(\mathbb{R}^n, ds^2)$.

2.2. First variation of area and energy

Assume that $D \subset \mathbb{R}^2_{(u,v)}$ is a bounded domain with piecewise smooth boundary and $X : \overline{D} \to \mathbb{R}^n$ is a smooth immersion. Precomposing X with a diffeomorphism from another such domain in $\mathbb{R}^2$, we may assume that X is conformal; see (2.1). We consider the *area functional*

$$\mathrm{Area}(X) = \int_D |X_u \times X_v| \, du \, dv = \int_D \sqrt{|X_u|^2 |X_v|^2 - |X_u \cdot X_v|^2} \, du \, dv \qquad (2.2)$$

and the *Dirichlet energy functional*

$$\mathcal{D}(X) = \frac{1}{2} \int_D |\nabla X|^2 \, du \, dv = \frac{1}{2} \int_D \big(|X_u|^2 + |X_v|^2\big) \, du \, dv. \qquad (2.3)$$

We have elementary inequalities

$$|\mathbf{x}|^2 |\mathbf{y}|^2 - |\mathbf{x} \cdot \mathbf{y}|^2 \le |\mathbf{x}|^2 |\mathbf{y}|^2 \le \frac{1}{4}\big(|\mathbf{x}|^2 + |\mathbf{y}|^2\big)^2, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

which are equalities if and only if $\mathbf{x}, \mathbf{y}$ is a conformal frame, i.e., $|\mathbf{x}| = |\mathbf{y}|$ and $\mathbf{x} \cdot \mathbf{y} = 0$. Applying this to the vectors $\mathbf{x} = X_u$ and $\mathbf{y} = X_v$ gives $\mathrm{Area}(X) \le \mathcal{D}(X)$, with equality if and only if X is conformal. Hence, these two functionals have the same critical points on the set of conformal immersions.

It is elementary to find critical points of these functionals. The calculation is simpler for the Dirichlet functional $\mathcal{D}$, but the expression for the first variation is the same for both functionals at a conformal map X. Assuming that $G : \overline{D} \to \mathbb{R}^n$ is a smooth map vanishing on bD, the first variation of $\mathcal{D}$ at X in direction G equals

$$\frac{d}{dt}\Big|_{t=0} \mathcal{D}(X + tG) = \int_D (X_u \cdot G_u + X_v \cdot G_v) \, du \, dv = -\int_D \Delta X \cdot G \, du \, dv, \quad (2.4)$$

where $\Delta X = X_{uu} + X_{vv}$ is the Laplace of X. (We integrated by parts and used $G|_{bD} = 0$.) The right-hand side of (2.4) vanishes for all G if and only if $\Delta X = 0$. This proves the following theorem.

Theorem 2.1. *Let D be a relatively compact domain in $\mathbb{R}^2$ with piecewise smooth boundary. A smooth conformal immersion $X : \overline{D} \to \mathbb{R}^n$ $(n \ge 3)$ is a stationary point of the area functional* (2.2) *if and only if X is harmonic: $\Delta X = 0$.*

For completeness, we also calculate the first variation of area at a conformal immersion X. Let $G : \overline{D} \to \mathbb{R}^n$ be as above. Consider the expression under the integral (2.2) for the map $X_t = X + tG$, $t \in \mathbb{R}$. Taking into account (2.1), we obtain

$$|X_u + tG_u|^2 \cdot |X_v + tG_v|^2 = |X_u|^4 + 2t\,(X_u \cdot G_u + X_v \cdot G_v)\,|X_u|^2 + O(t^2),$$
$$\big|(X_u + tG_u) \cdot (X_v + tG_v)\big|^2 = O(t^2).$$

It follows that

$$\frac{d}{dt}\Big|_{t=0} \Big(|X_u + tG_u|^2 |X_v + tG_v|^2 - \big|(X_u + tG_u) \cdot (X_v + tG_v)\big|^2\Big)$$
$$= 2|X_u|^2 (X_u \cdot G_u + X_v \cdot G_v)$$

and therefore

$$\frac{d}{dt}\Big|_{t=0} \mathrm{Area}(X+tG) = \int_D (X_u\cdot G_u + X_v\cdot G_v)\,du\,dv = -\int_D \Delta X\cdot G\,du\,dv.$$

(We integrated by parts and used that $G|_{bD}=0$. The factor $2|X_u|^2$ also appears in the denominator when differentiating the expression for $\mathrm{Area}(X+tG)$ at $t=0$, so this term cancels.) Comparing with (2.4), we see that

$$\frac{d}{dt}\Big|_{t=0} \mathrm{Area}(X+tG) = \frac{d}{dt}\Big|_{t=0} \mathcal{D}(X+tG) = -\int_D \Delta X\cdot G\,du\,dv$$

if X is a conformal immersion.

The same result holds on any compact domain with piecewise smooth boundary in a conformal surface M. A conformal diffeomorphism changes the Laplacian by a multiplicative factor, so there is a well-defined notion of a harmonic function on M.

2.3. Characterization of minimality by vanishing mean curvature

In this section, we prove a result due to Meusnier [66] which characterizes minimal surfaces in terms of vanishing mean curvature; see Theorem 2.3.

To explain the notion of curvature of a smooth plane curve $C\subset\mathbb{R}^2$ at a point $p\in C$, we apply a rigid change of coordinates in $\mathbb{R}^2$ taking p to $(0,0)$ and the tangent line T_pC to the x-axis, so locally near $(0,0)$ the curve is the graph $y=f(x)$ of a smooth function on an interval around $0\in\mathbb{R}$, with $f(0)=f'(0)=0$. Therefore,

$$y=f(x)=\frac{1}{2}f''(0)x^2+o(x^2). \tag{2.5}$$

Let us find the circle which agrees with this graph to the second order at $(0,0)$. Clearly, such a circle has center on the y-axis, so it is of the form $x^2+(y-r)^2=r^2$ for some $r\in\mathbb{R}\setminus\{0\}$, unless $f''(0)=0$ when the x-axis (a circle of infinite radius) does the job. Solving the equation on y near $(0,0)$ gives

$$\begin{aligned} y &= r-\sqrt{r^2-x^2} = r-r\sqrt{1-\frac{x^2}{r^2}} \\ &= r-r\Big(1-\frac{x^2}{2r^2}+o(x^2)\Big) = \frac{1}{2r}x^2+o(x^2). \end{aligned}$$

A comparison with (2.5) shows that for $f''(0)\neq 0$ the number $r=1/f''(0)\in\mathbb{R}\setminus\{0\}$ is the unique number for which the circle agrees with the curve (2.5) to the second order at $(0,0)$. This best fitting circle is called the *osculating circle*. The number

$$\kappa = f''(0) = 1/r \tag{2.6}$$

is the *signed curvature* of the curve (2.5) at $(0,0)$, its absolute value $|\kappa| = |f''(0)| \geq 0$ is the *curvature*, and $|r| = 1/|\kappa| = 1/|f''(0)|$ is the *curvature radius*. If $f''(0) = 0$, then the curvature is zero and the curvature radius is $+\infty$.

Consider now a smooth surface $S \subset \mathbb{R}^3$. Let (x, y, z) be coordinates on $\mathbb{R}^3$. Fix a point $p \in S$. A rigid change of coordinates gives $p = (0,0,0)$ and $T_pS = \{z = 0\} = \mathbb{R}^2 \times \{0\}$. Then, S is locally near the origin of a graph of the form

$$z = f(x, y) = \frac{1}{2}\big(f_{xx}(0)x^2 + 2f_{xy}(0,0)xy + f_{yy}(0)y^2\big) + o(x^2 + y^2). \tag{2.7}$$

The symmetric matrix

$$A = \begin{pmatrix} f_{xx}(0,0) & f_{xy}(0,0) \\ f_{xy}(0,0) & f_{yy}(0,0) \end{pmatrix} \tag{2.8}$$

is called the *Hessian matrix* of f at $(0,0)$. Given a unit vector $v = (v_1, v_2)$ in the (x, y)-plane, let Σ_v be the 2-plane through $0 \in \mathbb{R}^3$ spanned by v and the z-axis. The intersection $C_v := S \cap \Sigma_v$ is then a planar curve contained in S, given by

$$z = f(v_1 t, v_2 t) = \frac{1}{2}(Av \cdot v)t^2 + o(t^2) \tag{2.9}$$

for $t \in \mathbb{R}$ near 0. Since $|v| = 1$, the parameters (t, z) on Σ_v are Euclidean parameters, i.e., the Euclidean metric ds^2 on $\mathbb{R}^3$ restricted to the plane Σ_v is given by $dt^2 + dz^2$. From our discussion of curves and the formula (2.6), we infer that the number

$$\kappa_v = Av \cdot v = f_{xx}(0)v_1^2 + 2f_{xy}(0,0)v_1v_2 + f_{yy}(0)v_2^2$$

is the signed curvature of the curve C_v at the point $(0,0)$.

On the unit circle $|v|^2 = v_1^2 + v_2^2 = 1$ the quadratic form $v \mapsto Av \cdot v$ reaches its maximum κ_1 and minimum κ_2; these are the *principal curvatures* of the surface (2.7) at $(0,0)$. Since A is symmetric, κ_1 and κ_2 are its eigenvalues. The real numbers

$$H = \kappa_1 + \kappa_2 = \operatorname{trace} A, \qquad K = \kappa_1\kappa_2 = \det A \tag{2.10}$$

are, respectively, the *mean curvature* and the *Gaussian curvature* of S at $(0,0,0)$.

Note that the trace of A (2.8) equals the Laplacian $\Delta f(0,0)$. On the other hand, the trace of a matrix is the sum of its eigenvalues. This implies

$$\Delta f(0,0) = \kappa_1 + \kappa_2 = H. \tag{2.11}$$

Lemma 2.2. *Let D be a domain in $\mathbb{R}^2$. If $X : D \to \mathbb{R}^n$ is a smooth conformal immersion, then for every $p \in D$ the vector $\Delta X(p)$ is orthogonal to the plane $dX_p(\mathbb{R}^2) \subset \mathbb{R}^n$. Equivalently, the following identities hold on D:*

$$\Delta X \cdot X_u = 0, \qquad \Delta X \cdot X_v = 0. \tag{2.12}$$

Proof. Recall from (2.1) that X is conformal if and only if $X_u \cdot X_u = X_v \cdot X_v$ and $X_u \cdot X_v = 0$. Differentiating the first identity on u and the second one on v yields

$$X_{uu} \cdot X_u = X_{uv} \cdot X_v = -X_{vv} \cdot X_u,$$

whence $\Delta X \cdot X_u = (X_{uu} + X_{vv}) \cdot X_u = 0$. Likewise, differentiating the first identity on v and the second one on u gives $\Delta X \cdot X_v = 0$. ■

We can now prove the following result due to Meusnier [66].

Theorem 2.3. *A smooth conformal immersion $X = (x, y, z) : D \to \mathbb{R}^3$ from a domain $D \subset \mathbb{R}^2$ parameterizes a surface with vanishing mean curvature function if and only if the map X is harmonic, $\Delta X = (\Delta x, \Delta y, \Delta z) = 0$.*

Proof. Fix a point $p_0 \in D$; by a translation of coordinates we may assume that $p_0 = (0, 0) \in \mathbb{R}^2$. Since the differential $dX_{(0,0)} : \mathbb{R}^2 \to \mathbb{R}^3$ is a conformal linear map, we may assume up to a rigid motion on $\mathbb{R}^3$ that $X(0, 0) = (0, 0, 0)$ and

$$dX_{(0,0)}(\xi_1, \xi_2) = \mu(\xi_1, \xi_2, 0) \quad \text{for all } \xi = (\xi_1, \xi_2) \in \mathbb{R}^2$$

for some $\mu > 0$. Equivalently, at $(u, v) = (0, 0)$ the following hold:

$$x_u = y_v = \mu > 0, \quad x_v = y_u = 0, \quad z_u = z_v = 0. \tag{2.13}$$

Note that

$$\mu = |X_u| = |X_v| = \frac{1}{\sqrt{2}}|\nabla X|. \tag{2.14}$$

The implicit function theorem shows that there is a neighborhood $U \subset D$ of the origin such that the surface $S = X(U)$ is a graph $z = f(x, y)$ with $df_{(0,0)} = 0$, so f is of the form (2.7). Since the immersion X is conformal, (2.12) shows that ΔX is orthogonal to the (x, y)-plane $\mathbb{R}^2 \times \{0\}$ at the origin, which means that

$$\Delta x = \Delta y = 0 \quad \text{at } (0, 0). \tag{2.15}$$

We now calculate $\Delta z(0, 0)$. Differentiation of $z(u, v) = f(x(u, v), y(u, v))$ gives

$$\begin{aligned} z_u &= f_x x_u + f_y y_u, \quad z_v = f_x x_v + f_y y_v, \\ z_{uu} &= (f_x x_u + f_y y_u)_u \\ &= f_{xx} x_u^2 + f_{xy} x_u y_u + f_x x_{uu} + f_{yx} x_u y_u + f_{yy} y_u^2 + f_y y_{uu}. \end{aligned}$$

At the point $(0, 0)$, taking into account (2.13) and $f_x = f_y = 0$ we get $z_{uu} = \mu^2 f_{xx}$. A similar calculation gives $z_{vv} = \mu^2 f_{yy}$ at $(0, 0)$, so we conclude that

$$\Delta z(0, 0) = \mu^2 \Delta f(0, 0) = \mu^2 H, \tag{2.16}$$

where H is the mean curvature of S at the origin (see (2.11)). Denoting by $\mathbf{N} = (0,0,1)$ the unit normal vector to S at $0 \in \mathbb{R}^3$, it follows from (2.14), (2.15), and (2.16) that

$$\Delta X = \frac{1}{2}|\nabla X|^2 H\mathbf{N} \tag{2.17}$$

holds at $(0,0) \in D$. In particular, $\Delta X = 0$ if and only if $H = 0$. This formula is clearly independent of the choice of a Euclidean coordinate system. ■

Combining Theorems 2.1 and 2.3 gives the following corollary.

Corollary 2.4. *Let D be a relatively compact domain in $\mathbb{R}^2$ with piecewise smooth boundary. A smooth conformal immersion $X : \overline{D} \to \mathbb{R}^3$ is a stationary point of the area functional if and only if the immersed surface $S = X(D)$ has vanishing mean curvature at every point.*

Although we used conformal parameterizations, neither curvature nor area depends on the choice of parameterization. This motivates the following definition.

Definition 2.5. A smooth surface in $\mathbb{R}^3$ is a *minimal surface* if and only if its mean curvature vanishes at every point.

Every point in a minimal surface is a saddle point, and the surface is equally curved in both principal directions but in the opposite normal directions. Furthermore, the Gaussian curvature $K = \kappa_1\kappa_2 = -\kappa_1^2 \le 0$ is nonpositive at every point. The integral

$$\mathrm{TC}(S) = \int_S K \cdot dA \in [-\infty, 0] \tag{2.18}$$

of the Gaussian curvature function with respect to the surface area on S is called the *total Gaussian curvature*. This number equals zero if and only if S is a piece of a plane.

The results presented in this section easily extend to surfaces in $\mathbb{R}^n$ for any $n \ge 3$ which are parameterized by conformal immersions $X : M \to \mathbb{R}^n$ from any open Riemann surface M. (By the maximum principle for harmonic maps, there are no compact minimal surfaces in $\mathbb{R}^n$.) There is a sphere S^{n-3} of unit normal vectors to the surface at a given point, and one must consider the mean curvature of the surface in any given normal direction. This gives the mean curvature vector field $\mathbf{H}$ along the surface, which is orthogonal to it at every point. For surfaces in $\mathbb{R}^3$ we have $\mathbf{H} = H\mathbf{N}$, where H is the mean curvature function (2.10) and $\mathbf{N}$ is a unit normal vector field to the surface. The formula (2.17) can then be written in the form

$$\frac{2}{|\nabla X|^2}\Delta X = \Delta_g X = \mathbf{H},$$

where $\Delta_g X$ denotes the intrinsic Laplacian of the map X with respect to the induced metric $g = X^*ds^2$ on the surface M (cf. [16, Lemma 2.1.2]). The formula (2.4) for the first variation of area still holds. It shows that the mean curvature vector field $\mathbf{H}$ is the negative gradient of the area functional, and the surface is a minimal surface if and only if $\mathbf{H} = 0$. We refer to [16,55,69] or any other standard source for the details.

2.4. The Enneper–Weierstrass representation

In this section we explain the Enneper–Weierstrass formula, which provides a connection between holomorphic maps $D \to \mathbb{C}^n$ with special properties from domains $D \subset \mathbb{C}$ and conformal minimal immersions $D \to \mathbb{R}^n$ for $n \geq 3$. The same connection holds more generally for maps from any open Riemann surface.

Let $z = x + \mathrm{i}y$ be a complex coordinate on $\mathbb{C}$. Let us recall the following basic operators of complex analysis, also called *Wirtinger derivatives*:

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - \mathrm{i}\frac{\partial}{\partial y}\right), \quad \frac{\partial}{\partial \bar z} = \frac{1}{2}\left(\frac{\partial}{\partial x} + \mathrm{i}\frac{\partial}{\partial y}\right).$$

The differential of a function $F(z)$ can be written in the form

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = \frac{\partial F}{\partial z}dz + \frac{\partial F}{\partial \bar z}d\bar z,$$

where $dz = dx + \mathrm{i}dy$ and $d\bar z = dx - \mathrm{i}dy$. Note that $\frac{\partial F}{\partial z}dz$ is the $\mathbb{C}$-linear part and $\frac{\partial F}{\partial \bar z}d\bar z$ is the $\mathbb{C}$-antilinear part of dF. In particular, $\partial F/\partial \bar z = 0$ holds for holomorphic functions, and $\partial F/\partial z = 0$ holds for antiholomorphic ones. In terms of these operators, the Laplacian equals

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = 4\frac{\partial}{\partial \bar z}\frac{\partial}{\partial z} = 4\frac{\partial}{\partial z}\frac{\partial}{\partial \bar z}.$$

Hence, a function $F : D \to \mathbb{R}$ is harmonic if and only if $\partial F/\partial z$ is holomorphic.

It follows that a smooth map $X = (X_1, X_2, \ldots, X_n) : D \to \mathbb{R}^n$ is a harmonic immersion if and only if the map $f = (f_1, f_2, \ldots, f_n) : D \to \mathbb{C}^n$ with components $f_j = \partial X_j/\partial z$ is holomorphic and the component functions f_j have no common zero. Furthermore, conformality of X is equivalent to the following nullity condition:

$$f_1^2 + f_2^2 + \cdots + f_n^2 = 0. \tag{2.19}$$

Indeed, we have that $4f_j^2 = (X_{j,x} - \mathrm{i}X_{j,y})^2 = (X_{j,x})^2 - (X_{j,y})^2 - 2\mathrm{i}X_{j,x}X_{j,y}$, and hence

$$4\sum_{j=1}^{n} f_j^2 = |X_x|^2 - |X_y|^2 - 2\mathrm{i}X_x \cdot X_y.$$

Comparing with the conformality conditions (2.1) proves the claim.

Since we know by Theorem 2.1 that a conformal immersion is harmonic if and only if it parameterizes a minimal surface, this gives the following result.

Theorem 2.6 (The Enneper–Weierstrass representation). *Let D be a connected domain in $\mathbb{C}$. For every smooth conformal minimal immersion $X = (X_1, X_2, \dots, X_n) : D \to \mathbb{R}^n$, the map $f = (f_1, f_2, \dots, f_n) = \partial X/\partial z : D \to \mathbb{C}^n \setminus \{0\}$ is holomorphic and satisfies the nullity conditions* (2.19)*. Conversely, a holomorphic map $f : D \to \mathbb{C}^n \setminus \{0\}$ satisfying* (2.19) *and the period vanishing conditions*

$$\Re \oint_C f\, dz = 0 \quad \textit{for every closed curve } C \subset D \tag{2.20}$$

determines a conformal minimal immersion $X : D \to \mathbb{R}^n$ given by

$$X(z) = c + 2\Re \int_{z_0}^{z} f(\zeta)\, d\zeta, \quad z \in D \tag{2.21}$$

for any base point $z_0 \in D$ and vector $c \in \mathbb{R}^n$.

Conditions (2.20) guarantee that the integral in (2.21) is well defined, that is, independent of the path of integration. The imaginary components

$$\Im \oint_C f\, dz = \mathfrak{p}(C) \in \mathbb{R}^n \tag{2.22}$$

of the periods define the *flux homomorphism* $\mathfrak{p} : H_1(D, \mathbb{Z}) \to \mathbb{R}^n$ on the first homology group of D. Indeed, by Green's formula the period $\oint_C f\, dz$ only depends on the homology class $[C] \in H_1(D, \mathbb{Z})$ of a closed path $C \subset D$.

Remark (The first homology group). If D is a domain in $\mathbb{R}^2 \cong \mathbb{C}$, then its first homology group $H_1(D, \mathbb{Z})$ is a free abelian group $\mathbb{Z}^\ell$ ($\ell \in \{0, 1, 2, \dots, \infty\}$) with finitely or countably many generators. If D is bounded, connected, and its boundary bD consists of l_1 Jordan curves $\Gamma_1, \dots, \Gamma_{l_1}$ and l_2 isolated points (punctures) $p_1, \dots, p_{l_2}$, then the group $H_1(D, \mathbb{Z})$ has $\ell = l_1 + l_2 - 1$ generators which are represented by loops in D based at any given point $p_0 \in D$, each surrounding one of the holes of D. (By a *hole*, we mean a compact connected component of the complement $\mathbb{C} \setminus D$. A hole which is an isolated point of $\mathbb{C} \setminus D$ is called a *puncture*.) Indeed, if Γ_1 is the outer boundary curve of D, then every other boundary curve $\Gamma_2, \dots, \Gamma_{l_1}$ of D is contained in the bounded component of $\mathbb{C} \setminus \Gamma_1$, so it bounds a hole of D. Likewise, each of the points $p_1, \dots, p_{l_2}$ is a hole (a puncture). Every hole contributes one generator to $H_1(D, \mathbb{Z})$. The same loops then generate the fundamental group $\pi_1(D, p_0)$ as a free nonabelian group, and group $H_1(D, \mathbb{Z})$ is the abelianization of $\pi_1(D, p_0)$. A similar description of the homology group $H_1(D, \mathbb{Z})$ holds for every surface, except that its genus enters the picture as well; see [16, Section 1.4]. For basics on homology and cohomology, see J. P. May [56].

It is clear from Theorem 2.6 that the following quadric complex hypersurface in $\mathbb{C}^n$ plays a special role in the theory of minimal surfaces in $\mathbb{R}^n$:

$$\mathbb{A} = \mathbb{A}^{n-1} = \big\{(z_1,\ldots,z_n) \in \mathbb{C}^n : z_1^2 + z_2^2 + \cdots + z_n^2 = 0\big\}. \tag{2.23}$$

This is called the *null quadric* in $\mathbb{C}^n$, and $\mathbb{A}_* = \mathbb{A} \setminus \{0\}$ is the *punctured null quadric*. Note that $\mathbb{A}$ is a complex cone with the only singular point at 0. Theorem 2.6 says that we get all conformal minimal surfaces $D \to \mathbb{R}^n$ as integrals of holomorphic maps $f : D \to \mathbb{A}_* \subset \mathbb{C}^n$ satisfying the period vanishing conditions (2.20).

The Enneper–Weierstrass representation in $\mathbb{R}^3$. In dimension $n = 3$, the null quadric $\mathbb{A}$ admits a 2-sheeted quadratic parameterization $\phi : \mathbb{C}^2 \to \mathbb{A}$ given by

$$\phi(z,w) = \big(z^2 - w^2, \mathrm{i}(z^2 + w^2), 2zw\big). \tag{2.24}$$

This map is branched at $0 \in \mathbb{C}^2$, and $\phi : \mathbb{C}^2 \setminus \{0\} \to \mathbb{A}_*$ is a 2-sheeted holomorphic covering map. It follows that every conformal minimal immersion $X = (X_1, X_2, X_3) : D \to \mathbb{R}^3$ can be written in the following form (see [69] or [16, pp. 107–108]):

$$X(z) = X(z_0) + 2\Re \int_{z_0}^{z} \left(\frac{1}{2}\left(\frac{1}{\mathfrak{g}} - \mathfrak{g}\right), \frac{\mathrm{i}}{2}\left(\frac{1}{\mathfrak{g}} + \mathfrak{g}\right), 1 \right) \partial X_3. \tag{2.25}$$

Here, $\partial X = \frac{\partial X}{\partial z} dz = (\partial X_1, \partial X_2, \partial X_3)$, and

$$\mathfrak{g} = \frac{\partial X_3}{\partial X_1 - \mathrm{i}\, \partial X_2} : D \to \mathbb{CP}^1 = \mathbb{C} \cup \{\infty\} \tag{2.26}$$

is a holomorphic map to the Riemann sphere (a meromorphic function on D), called the *complex Gauss map* of X. Identifying $\mathbb{CP}^1$ with the unit 2-sphere $S^2 \subset \mathbb{R}^3$ by the stereographic projection from the point $(0,0,1) \in S^2$, $\mathfrak{g}$ corresponds to the classical Gauss map $\mathbf{N} = X_x \times X_y / |X_x \times X_y| : D \to S^2$ of X.

Many important quantities and properties of a minimal surface are determined by its Gauss map. In particular, we have that

$$g = X^* ds^2 = 2\big(|\partial X_1|^2 + |\partial X_2|^2 + |\partial X_3|^2\big) = \frac{\big(1 + |\mathfrak{g}|^2\big)^2}{4|\mathfrak{g}|^2} |\partial X_3|^2$$
$$Kg = -\frac{4|d\mathfrak{g}|^2}{\big(1 + |\mathfrak{g}|^2\big)^2} = -\mathfrak{g}^*(\sigma^2_{\mathbb{CP}^1}).$$

Here, K is the Gauss curvature function (2.10) of the metric $X^* ds^2$ and $\sigma^2_{\mathbb{CP}^1}$ is the spherical metric on $\mathbb{CP}^1$. It follows that the total Gaussian curvature (see (2.18)) of a conformal minimal surface $X : D \to \mathbb{R}^3$ equals the negative spherical area of the image of the Gauss map $\mathfrak{g} : D \to \mathbb{CP}^1$ counted with multiplicities, where the area of the sphere $\mathbb{CP}^1 = S^2$ is 4π:

$$\mathrm{TC}(X) = -\operatorname{Area} \mathfrak{g}(D). \tag{2.27}$$

It is a recent result that every holomorphic map $D \to \mathbb{CP}^1$ is the complex Gauss map of a conformal minimal immersion $X : D \to \mathbb{R}^3$; see Theorem 3.3. Hence, the total Gaussian curvature of a minimal surface can be any number in $[-\infty, 0]$.

Example 2.7 (Catenoid). A conformal parameterization of a standard catenoid (see [16, Figure 2.1, p. 117]) is given by the map $X = (X_1, X_2, X_3) : \mathbb{R}^2 \to \mathbb{R}^3$,

$$X(u,v) = (\cos u \cdot \cosh v, \sin u \cdot \cosh v, v). \tag{2.28}$$

It is 2π-periodic in the u variable, hence infinitely-sheeted. Introducing the variable $z = \mathrm{e}^{-v+\mathrm{i}u} \in \mathbb{C}^*$, we pass to the quotient $\mathbb{C}/(2\pi\mathbb{Z}) \cong \mathbb{C}^*$ and obtain a single-sheeted parameterization $X : \mathbb{C}^* \to \mathbb{R}^3$ having the Enneper–Weierstrass representation

$$X(z) = (1,0,0) - 2\Re \int_1^z \left(\frac{1}{2}\left(\frac{1}{\zeta} - \zeta\right), \frac{\mathrm{i}}{2}\left(\frac{1}{\zeta} + \zeta\right), 1 \right) \frac{d\zeta}{\zeta}. \tag{2.29}$$

Its Gauss map is $\mathfrak{g}(z) = z$ and extends to the identity map $\mathbb{CP}^1 \to \mathbb{CP}^1$. Hence, by (2.27) the catenoid has total Gaussian curvature equal to -4π.

The catenoid is one of the most paradigmatic examples in the theory of minimal surfaces. A compendium of major results about it can be found in [16, Example 2.8.1].

Example 2.8 (Helicoid). A conformal parameterization $X : \mathbb{R}^2 \to \mathbb{R}^3$ of the standard left helicoid, shown on [16, Figure 2.2, p. 119], is

$$X(u,v) = (\sin u \cdot \sinh v, -\cos u \cdot \sinh v, u). \tag{2.30}$$

Its Weierstrass representation in the complex coordinate $z = u + \mathrm{i}v \in \mathbb{C}$ is

$$X(z) = \Re \int_0^z \left(\frac{1}{2}\left(\frac{1}{\mathrm{e}^{\mathrm{i}\zeta}} - \mathrm{e}^{\mathrm{i}\zeta}\right), \frac{\mathrm{i}}{2}\left(\frac{1}{\mathrm{e}^{\mathrm{i}\zeta}} + \mathrm{e}^{\mathrm{i}\zeta}\right), 1 \right) d\zeta.$$

Its complex Gauss map $\mathfrak{g}(z) = \mathrm{e}^{\mathrm{i}z}$ is transcendental, so the helicoid has infinite total Gaussian curvature $-\infty$. Changing the sign of the second component in (2.30) gives a right helicoid. Like the catenoid, the helicoid is a paradigmatic example satisfying various uniqueness theorems. E. Catalan [23] proved in 1842 that the helicoid and the plane are the only ruled minimal surfaces in $\mathbb{R}^3$, i.e., unions of straight lines. Much more recently, W. H. Meeks and H. Rosenberg proved in 2005 [62] that the helicoid and the plane are the only properly embedded, simply connected minimal surfaces in $\mathbb{R}^3$. Their proof uses curvature estimates of T. H. Colding and W. P. Minicozzi [27].

Remark (Branch points). Our definition of a conformal map $X : D \to \mathbb{R}^n$ of class $\mathscr{C}^1(D)$ requires that equations (2.1) hold. We have already observed that such a map has rank zero at non-immersion points. Assuming that X is harmonic at immersion points, it follows that $f = \partial X/\partial z : D \to \mathbb{C}^n$ is a continuous map with values in the null quadric $\mathbb{A}$ (2.23) which is holomorphic at immersion points of X and vanishes at non-immersion points. By a theorem of T. Radó [70] (cf. [74, Theorem 15.1.7]), such an f is holomorphic everywhere on D, and in particular its zero set consists of isolated points (assuming that X and hence f are nonconstant). This shows that the minimal surface parameterized by X has only isolated singularities. See [77] for more details.

There are interesting examples of minimal surfaces with branch points. For example, *Henneberg's surface* (see [16, Example 2.8.9]) is a complete non-orientable minimal surface with two branch points (a branched minimal Möbius strip), named after Ernst Lebrecht Henneberg [46] who first described it in his doctoral dissertation in 1875. It was the only known non-orientable minimal surface until 1981 when W. H. Meeks [57] discovered a properly immersed minimal Möbius strip in $\mathbb{R}^3$. A properly embedded minimal Möbius strip in $\mathbb{R}^4$ was found in 2017 [15, Example 6.1].

2.5. Holomorphic null curves

There is a family of holomorphic curves in $\mathbb{C}^n$ which are close relatives of conformal minimal surfaces in $\mathbb{R}^n$. A holomorphic map $Z = (Z_1, \ldots, Z_n) : D \to \mathbb{C}^n$ for $n \geq 3$ from a domain $D \subset \mathbb{C}$ satisfying the nullity condition

$$(Z_1')^2 + (Z_2')^2 + \cdots + (Z_n')^2 = 0$$

is a *holomorphic null curve* in $\mathbb{C}^n$. Its complex derivative $f = Z'$ assumes values in the null quadric $\mathbb{A}$ (2.23), and we have $\oint_C f dz = \oint_C dZ = 0$ for any closed curve $C \subset D$. Conversely, a holomorphic map $f : D \to \mathbb{A}$ satisfying the period vanishing conditions

$$\oint_C f dz = 0 \quad \text{for every closed curve } C \subset D \tag{2.31}$$

integrates to a holomorphic null curve

$$Z(z) = c + \int_{z_0}^{z} f(\zeta) d\zeta, \quad z \in D, \tag{2.32}$$

where $z_0 \in D$ is any given base point and $c \in \mathbb{C}^n$. Indeed, conditions (2.31) guarantee that the integral in (2.32) is independent of the choice of a path of integration. These period conditions are trivial on a simply connected domain D.

If $Z = X + \mathrm{i}Y : D \to \mathbb{C}^n$ is an immersed holomorphic null curve, then its real part $X = \Re Z : D \to \mathbb{R}^n$ and imaginary part $Y = \Im Z : D \to \mathbb{R}^n$ are conformal minimal surfaces which are harmonic conjugates of each other. Indeed, denoting the complex variable in $\mathbb{C}$ by $z = x + \mathrm{i}y$, the Cauchy–Riemann equations imply

$$f = Z' = Z_x = X_x + \mathrm{i}Y_x = X_x - \mathrm{i}X_y = 2\frac{\partial X}{\partial z}.$$

Since $f = Z' : D \to \mathbb{A}_*^{n-1}$ satisfies the nullity condition (2.19), X is a conformal minimal immersion. In the same way we find that $f = Z' = Y_y + \mathrm{i}Y_x = 2\mathrm{i}Y_z$, so Y is a conformal minimal immersion. Being harmonic conjugates, X and Y are called *conjugate minimal surfaces*. Conformal minimal surfaces in the 1-parameter family

$$X^t = \Re(\mathrm{e}^{\mathrm{i}t}Z) : D \to \mathbb{R}^n, \quad t \in \mathbb{R},$$

are called *associated minimal surfaces* of the holomorphic null curve Z.

Conversely, if $X : D \to \mathbb{R}^n$ is a conformal minimal surface and the holomorphic map $f = 2\frac{\partial X}{\partial z} : D \to \mathbb{A}^{n-1}$ satisfies period vanishing conditions (2.31), then f integrates to a holomorphic null curve $Z : D \to \mathbb{C}^n$ (2.32) with $\Re Z = X$. In general, the imaginary parts of the periods (2.32) determine the flux homomorphism $H_1(M, \mathbb{Z}) \to \mathbb{R}$ of the minimal surface X (see (2.22)); hence, X is the real part of a holomorphic null curve if and only if it has vanishing flux. The periods (2.31) always vanish on a simply connected domain D, and hence every conformal minimal immersion $D \to \mathbb{R}^n$ is the real part of a holomorphic null curve $D \to \mathbb{C}^n$.

The relationship between conformal minimal surfaces and holomorphic null curves extends to maps having (isolated) branch points.

Example 2.9 (Helicatenoid). Consider the holomorphic immersion $Z : \mathbb{C} \to \mathbb{C}^3$,

$$Z(z) = (\cos z, \sin z, -\mathrm{i}z) \in \mathbb{C}^3, \quad z = x + \mathrm{i}y \in \mathbb{C}. \tag{2.33}$$

We have that

$$Z'(z) = (-\sin z, \cos z, -\mathrm{i}), \quad \sin^2 z + \cos^2 z + (-\mathrm{i})^2 = 0.$$

Hence, Z is a holomorphic null curve. Consider the 1-parameter family of its associated minimal surfaces in $\mathbb{R}^3$ for $t \in [0, 2\pi]$:

$$X^t(z) = \Re\big(\mathrm{e}^{\mathrm{i}t}Z(z)\big) = \cos t \begin{pmatrix} \cos x \cdot \cosh y \\ \sin x \cdot \cosh y \\ y \end{pmatrix} + \sin t \begin{pmatrix} \sin x \cdot \sinh y \\ -\cos x \cdot \sinh y \\ x \end{pmatrix}. \tag{2.34}$$

At $t = 0$ and $t = \pi$ we have a catenoid (see Example 2.7), while at $t = \pm\pi/2$ we have a helicoid (see Example 2.8). Hence, these are conjugate minimal surfaces in $\mathbb{R}^3$. The holomorphic null curve (2.33) is called *helicatenoid*.

3. A survey of new results

This section is a survey of recent results in the theory of minimal surfaces in Euclidean spaces, which were discussed in my lecture at the 8ECM. A detailed presentation is available in the monograph [16] and, for non-orientable surfaces, in the AMS Memoir [15] by Alarcón, the author, and López.

3.1. Approximation, interpolation, and general position theorems

Holomorphic approximation is a central topic in complex analysis. Holomorphic functions and maps with interesting properties are often constructed inductively, exhausting the manifold by an increasing sequence of compact sets such that one can approximate holomorphic functions uniformly on each one by holomorphic functions on M. The quintessential example is Runge's theorem from 1885 [75] on approximation of holomorphic functions on a compact set $K \subset \mathbb{C}$ with connected complement by holomorphic polynomials. A major extension is Mergelyan's theorem [65] from 1951.

In order to generalize Runge's theorem, we need the following concept. Denote by $\mathcal{O}(M)$ the algebra of holomorphic functions on a complex manifold M. Given a compact set K in M, its $\mathcal{O}(M)$-convex hull (or holomorphic hull) is the set

$$\widehat{K} = \big\{ z \in M : \big|f(z)\big| \le \sup_K |f| \text{ for all } f \in \mathcal{O}(M) \big\}.$$

If $K = \widehat{K}$, then K is said to be *holomorphically convex*, or $\mathcal{O}(M)$-convex, or a *Runge compact*. If M is the complex plane or, more generally, an open Riemann surface, then the hull $\widehat{K}$ is the union of K and all relatively compact connected components of $M \setminus K$ (the holes of K in M). There is no topological characterization of the hull in higher-dimensional complex manifolds.

Holomorphically convex sets are the natural sets for holomorphic approximation. Runge's theorem was extended to open Riemann surfaces by H. Behnke and K. Stein [21] in 1949, who proved that any holomorphic function on a neighborhood of a Runge compact K in open Riemann surface M can be approximated uniformly on K by holomorphic functions on M. A related result on higher-dimensional complex manifolds is the Oka–Weil theorem which pertains to Runge compacts in $\mathbb{C}^n$ and, more generally, in any *Stein manifold* (a closed complex submanifold of a Euclidean space $\mathbb{C}^n$). A recent survey of holomorphic approximation theory can be found in [33].

We have seen in Section 2.4 that every conformal minimal immersion $M \to \mathbb{R}^n$ from an open Riemann surface M is the integral of a holomorphic map $f : M \to \mathbb{A}_* \subset \mathbb{C}^n$ into the punctured null quadric $\mathbb{A}_*$; furthermore, f must satisfy the period vanishing conditions (2.20). Hence, a Runge-type approximation theorem for conformal minimal surfaces in $\mathbb{R}^n$ (or holomorphic null curves in $\mathbb{C}^n$) reduces to the approximation problem for holomorphic maps $f : M \to \mathbb{A}_*$ satisfying the period vanishing conditions (2.20) (or (2.31) when considering null curves). This is a nonlinear approximation problem. The first part, ignoring the period conditions, fits within Oka theory. In particular, the manifold $\mathbb{A}_*$ is easily seen to be a homogeneous space of the complex orthogonal group $O_n(\mathbb{C})$. Runge-type approximation theorems for holomorphic maps from Stein manifolds to complex homogeneous manifolds were proved by Hans Grauert [41] (1957) and Grauert and Kerner [42] (1963). More generally, a complex manifold Y is said to be an *Oka manifold* if and only if approximation results of this type hold for holomorphic maps $M \to Y$ from any Stein manifold in the absence of topological obstructions. Oka theory also includes interpolation theorems for holomorphic maps, generalizing classical theorems of K. Weierstrass [78] and H. Cartan [22]. For the theory of Oka manifolds, see [35].

The second part of the problem, ensuring the period vanishing conditions (2.20) or (2.31) for holomorphic maps to $\mathbb{A}_*$, can be treated by using sprays of holomorphic maps together with elements of convexity theory. More precisely, Gromov's 1-dimensional convex integration lemma from [43] is useful in this regard. The main techniques underlying all subsequent developments were established in [5] (2014). Their application led to the following result, which is a summary of several individual theorems. Parts (i), (ii), and (iv) are due to Alarcón, the author, and López [5, 12, 15] (the special case of (i) for $n = 3$ was obtained beforehand in [19]), while (iii) was proved by Alarcón and Castro–Infantes [2, 3]. Related results for conformal minimal surfaces of finite total curvature were given by Alarcón and López [18].

Main Theorem 3.1. *Let K be a compact set with piecewise smooth boundary and without holes (a Runge compact) in an open Riemann surface M. Then:*

(i) *Every conformal minimal immersion $X : K \to \mathbb{R}^n$ ($n \geq 3$) can be approximated uniformly on K by proper conformal minimal immersions $\widetilde{X} : M \to \mathbb{R}^n$.*

(ii) *The approximating map $\widetilde{X}$ can be chosen to have only simple double points if $n = 4$, and to be an embedding if $n \geq 5$.*

(iii) *In addition, one can prescribe the values of $\widetilde{X}$ on any closed discrete subset of M (Weierstrass-type interpolation).*

(iv) *The analogous results hold for non-orientable minimal surfaces in $\mathbb{R}^n$ and for holomorphic null curves in $\mathbb{C}^n$, $n \geq 3$.*

The proof of Theorem 3.1 is fairly complex, and we shall only outline the main idea. Fix a nowhere vanishing holomorphic 1-form θ on the open Riemann surface M. (Such a 1-form always exists; see [44].) By Enneper–Weierstrass (Theorem 2.6), it suffices to prove the Runge approximation theorem for holomorphic maps $f : M \to \mathbb{A}_*$ satisfying the period vanishing conditions (2.20).

Consider an inductive step. Assume that $K \subset L$ are connected Runge compacts with piecewise smooth boundaries in M, $X : K \to \mathbb{R}^n$ is a conformal minimal surface, and $f = 2\partial X/\theta : K \to \mathbb{A}_*$. We wish to approximate X by a conformal minimal immersion $\widetilde{X} : L \to \mathbb{R}^n$. We may assume that $f(K)$ is not contained in a complex ray $\mathbb{C}^*\mathbf{z}$ of the null quadric $\mathbb{A}_*$, for otherwise the result is trivial. There are two main cases to consider, the noncritical case and the critical case.

The noncritical case. There is no change of topology from K to L. It is well known that there are closed curves $C_1, \dots, C_\ell$ in K forming a basis of $H_1(K, \mathbb{Z})$ whose union $C = \bigcup_{j=1}^{\ell} C_j$ is a Runge compact. Let $\mathbb{B}^n$ denote the unit ball of $\mathbb{C}^n$. By using flows of holomorphic vector fields on $\mathbb{C}^n$ tangent to $\mathbb{A}$, we construct a smooth map

$$F : K \times \mathbb{B}^{n\ell} \to \mathbb{A}_*, \quad F(\cdot, 0) = f = 2\partial X/\theta,$$

which is holomorphic on $\overset{\circ}{K} \times \mathbb{B}^n$, such that the associated period map

$$\mathbb{B}^{n\ell} \ni t \mapsto \left(\int_{C_j} F(\cdot, t)\theta \right)_{j=1}^{\ell} \in \mathbb{C}^{n\ell}$$

is biholomorphic onto its image. Such a *period dominating spray* can be found of the form

$$F(p, t) = \phi^1_{g_1(p)t_1} \circ \phi^2_{g_2(p)t_2} \circ \cdots \circ \phi^{n\ell}_{g_{n\ell}(p)t_{n\ell}}\big(f(p)\big) \in \mathbb{A}_*, \quad p \in K, \tag{3.1}$$

where each ϕ^j is the flow of a holomorphic vector field tangent to $\mathbb{A}$ and $g_j \in \mathcal{O}(M)$. We first construct smooth functions g_i on C which give a period dominating spray; this can be done since the convex hull of $\mathbb{A}$ equals $\mathbb{C}^n$. As C is Runge in M, we can approximate the g_i's by holomorphic functions on M, thereby obtaining a holomorphic period dominating spray F as above.

In the next key step, we use that $\mathbb{A}_*$ is an Oka manifold, so we can approximate F by a holomorphic map $\widetilde{F} : M \times \mathbb{B}^{n\ell} \to \mathbb{A}_*$. (There is no topological obstruction since $\mathbb{A}_*$ is connected.) If the approximation is close enough, the implicit function theorem furnishes a parameter value $\tilde{t} \in \mathbb{B}^{n\ell}$ close to 0 such that the map $\tilde{f} = F(\cdot, \tilde{t}) : M \to \mathbb{A}_*$ has vanishing real periods on the curves $C_1, \dots, C_\ell$. Hence, fixing a point $p_0 \in K$, the map $\widetilde{X} : L \to \mathbb{R}^n$ given by

$$\widetilde{X}(p) = X(p_0) + \Re \int_{p_0}^{p} \tilde{f}\theta, \quad p \in L,$$

is a conformal minimal immersion which approximates $X : K \to \mathbb{R}^n$ on K.

The critical case. Assume now that E is an embedded smooth arc in $L \setminus \mathring{K}$ attached with its endpoints to K such that $K \cup E$ is a deformation retract of L. (Thus, L has the same topology as $K \cup E$. This situation arises when passing a critical point of index 1 of a strongly subharmonic Morse exhaustion function on M.) Let $a, b \in bK$ denote the endpoints of E. We extend f smoothly across E to a map $f : K \cup E \to \mathbb{A}_*$ such that

$$\Re \int_E f\theta = X(b) - X(a) \in \mathbb{R}^n.$$

This is possible since the convex hull of $\mathbb{A}_*$ equals $\mathbb{C}^n$. We then proceed as in the noncritical case: embed f into a period dominating spray of smooth maps $K \cup E \to \mathbb{A}_*$ which are holomorphic on $\mathring{K} = K \setminus bK$, approximate it by a holomorphic spray on L by Mergelyan's theorem, and pick a parameter value for which the map in the spray has vanishing real periods on $K \cup E$, and hence on L. The Enneper–Weierstrass formula gives a conformal minimal surface $\widetilde{X} : L \to \mathbb{R}^n$ approximating X on K.

The proof of the basic approximation theorem (i) (without the properness condition) is then completed by induction on a suitable exhaustion of M by Runge compacts, alternatively using the above two cases. Critical points of index 2 do not arise.

Interpolation (part (iii)) is easily built into the same inductive construction. Indeed, in each of the two cases considered above, we can arrange that none of the points $p_j \in M$ at which we wish to interpolate lies on the boundary of K or L. By choosing the functions g_i in the spray F (3.1) to vanish at those points p_j which lie in the interior of K, we ensure that the spray F is fixed at these points (independent of the parameter t), and hence the approximating map $\widetilde{X}$ will agree with X at these points. For each of the finitely many points $p_j \in \mathring{L} \setminus K$ we choose a smooth embedded arc $E_j \subset L \setminus \mathring{K}$ with one endpoint p_j and the other endpoint $q_j \in bK$ such that $E_j \setminus \{q_j\} \subset L \setminus K$ and these arcs are pairwise disjoint. The set $S = K \cup \bigcup_j E_j$ is then a Runge compact. We extend the map $f : K \to \mathbb{A}_*$ smoothly to S such that for each j, $\int_{E_j} f\theta$ has the correct value which ensures that the integral assumes the prescribed value at p_j. It remains to apply the same method as above with a spray which is period dominating also on each of the arcs E_j and to use Mergelyan approximation on the set S.

Properness of the approximating conformal minimal immersion $\widetilde{X} : M \to \mathbb{R}^n$ (part (ii) of the theorem) requires considerable additional work. The main point is to prove a relative version of the approximation theorem in part (i) in which all but two components of the given map X extend to harmonic functions on all of M. One can keep these components fixed while approximating the remaining two components such that the resulting map $\widetilde{X}$ is a conformal minimal immersion. This requires a more precise version of the Oka principle. This result is then used in an inductive scheme which is designed so that $|\widetilde{X}(z)|$ tends to infinity as the point $z \in M$ goes to the ideal boundary of M (i.e., it exists in any compact subset).

Finally, the general position theorem in part (ii) uses the same technique together with the transversality theorem. The details of proof are considerably more involved from the technical viewpoint, and we shall not deal with this subject here.

3.2. Topological structure of spaces of minimal surfaces

Assume that M is an open Riemann surface. Fix a nowhere vanishing holomorphic 1-form θ on M. Let $n \geq 3$. An immersion $M \to \mathbb{R}^n$ is said to be *nonflat* if its image is not contained in an affine 2-plane. We introduce the following notations:

- $\mathscr{O}(M, \mathbb{A}_*)$ and $\mathscr{C}(M, \mathbb{A}_*)$ denote spaces of holomorphic and continuous maps $M \to \mathbb{A}_*$, respectively;
- $\mathrm{CMI}(M, \mathbb{R}^n)$ denotes the space of conformal minimal immersions $M \to \mathbb{R}^n$;
- $\mathrm{CMI}_{\mathrm{nf}}(M, \mathbb{R}^n)$ is the subspace of $\mathrm{CMI}(M, \mathbb{R}^n)$ consisting of nonflat immersions;
- $\mathrm{NC}(M, \mathbb{C}^n)$ is the space of holomorphic null immersions $M \to \mathbb{C}^n$;
- $\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n)$ is the subspace of $\mathrm{NC}(M, \mathbb{C}^n)$ consisting of nonflat immersions.

Consider the commutative diagram

$$\begin{tikzcd}
\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \arrow[r, "\phi"] \arrow[d, "\Re"'] & \mathscr{O}(M, \mathbb{A}_*) \arrow[r, hook, "\tau"] & \mathscr{C}(M, \mathbb{A}_*) \\
\Re\,\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \arrow[r, hook, "\iota"] & \mathrm{CMI}_{\mathrm{nf}}(M, \mathbb{R}^n) \arrow[u, "\psi"'] &
\end{tikzcd}$$

where

- the maps $\phi : \mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \to \mathscr{O}(M, \mathbb{A}_*)$ and $\psi : \mathrm{CMI}_{\mathrm{nf}}(M, \mathbb{C}^n) \to \mathscr{O}(M, \mathbb{A}_*)$ are given by $Z \mapsto \partial Z/\theta$ and $X \mapsto 2\partial X/\theta$, respectively;
- the map $\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \to \Re\,\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n)$ is the projection $Z = X + \mathrm{i}Y \mapsto X$;
- the maps $\iota : \Re\,\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \hookrightarrow \mathrm{CMI}_{\mathrm{nf}}(M, \mathbb{R}^n)$ and $\tau : \mathscr{O}(M, \mathbb{A}_*) \hookrightarrow \mathscr{C}(M, \mathbb{A}_*)$ are the natural inclusions.

Recall that a continuous map $\phi : X \to Y$ between topological spaces is said to be a *weak homotopy equivalence* if it induces a bijection of path components of the two spaces and, for each integer $k \in \mathbb{N}$, an isomorphism $\pi_k(\phi) : \pi_k(X) \xrightarrow{\cong} \pi_k(Y)$ of their kth homotopy groups. The map ϕ is a *homotopy equivalence* if there is a continuous map $\psi : Y \to X$ such that $\psi \circ \phi : X \to X$ is homotopic to the identity on X and $\phi \circ \psi : Y \to Y$ is homotopic to the identity on Y. These notions indicate that the spaces X and Y have the same rough topological shape.

Since $\mathbb{A}_*$ is an Oka manifold, the inclusion $\tau : \mathscr{O}(M, \mathbb{A}_*) \hookrightarrow \mathscr{C}(M, \mathbb{A}_*)$ is a weak homotopy equivalence by the Oka–Grauert principle (see [35, Corollary 5.5.6]), and by Lárusson [54] it is a homotopy equivalence if M is of finite topological type; i.e., if the homology group $H_1(M, \mathbb{Z})$ is a finitely generated abelian group.

The real-part projection map $\Re : \mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n) \to \Re\,\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n)$ is evidently a homotopy equivalence.

It turns out that all other maps in the above diagram are also weak homotopy equivalences. The first part of the following theorem was proved by the author and Lárusson in [39], and the second part was proved by Alarcón, the author, and López in [14]. Validity of statement (a) for $\mathrm{CMI}(M, \mathbb{R}^n)$ and $\mathrm{NC}(M, \mathbb{C}^n)$ remains an open problem.

Main Theorem 3.2. *Let M be an open Riemann surface.*

(a) *Each of the maps ι, ϕ, ψ in the above diagram is a weak homotopy equivalence, and a homotopy equivalence if M is of finite topological type.*

(b) *The map $\tau \circ \psi : \mathrm{CMI}(M, \mathbb{R}^n) \to \mathscr{C}(M, \mathbb{A}_*)$ induces a bijection of path components of the two spaces. Hence,*

$$\pi_0\big(\mathrm{CMI}(M, \mathbb{R}^n)\big) = \begin{cases} \mathbb{Z}_2^\ell, & n = 3,\ H_1(M, \mathbb{Z}) = \mathbb{Z}^\ell; \\ 0, & n > 3. \end{cases}$$

It follows that each of the spaces $\mathrm{NC}_{\mathrm{nf}}(M, \mathbb{C}^n)$ and $\mathrm{CMI}_{\mathrm{nf}}(M, \mathbb{C}^n)$ is weakly homotopy equivalent to the space $\mathscr{C}(M, \mathbb{A}_*)$ of continuous maps $M \to \mathbb{A}_*$, and is homotopy equivalent to $\mathscr{C}(M, \mathbb{A}_*)$ if the surface M has finite topological type.

The group $\mathbb{Z}_2 = \{0, 1\}$, which appears in part (b), is the fundamental group of the punctured null quadric $\mathbb{A}_* \subset \mathbb{C}^3$; see (2.24) and note that $\mathbb{C}^2 \setminus \{0\}$ is simply connected. If $X \in \mathrm{CMI}(M, \mathbb{R}^3)$, then $\partial X / \partial z : M \to \mathbb{A}_*$ maps every generator of the homology group $H_1(M, \mathbb{Z})$ either to the generator of $\pi_1(\mathbb{A}_*)$ or to the trivial element. This gives 2^ℓ choices, each one determining a connected component of $\mathrm{CMI}(M, \mathbb{R}^3)$. The null quadric $\mathbb{A}_* \subset \mathbb{C}^n$ for $n > 3$ is simply connected.

These results are proved by using the parametric versions of techniques discussed in Section 3.1. Each of the maps in question satisfies the parametric h-principle, which implies that it is a weak homotopy equivalence.

3.3. The Gauss map of a conformal minimal surface

The Gauss map is of major importance in the theory of minimal surfaces. We have already seen that the Gauss map of a conformal minimal immersion $X : M \to \mathbb{R}^3$ is a holomorphic map $\mathfrak{g} : M \to \mathbb{CP}^1$ (2.26), which coincides with the classical Gauss map $M \to S^2$ under the stereographic projection from S^2 onto $\mathbb{CP}^1$. In general, for any dimension $n \geq 3$ one defines the *generalized Gauss map* of a conformal minimal immersion $X = (X_1, X_2, \dots, X_n) : M \to \mathbb{R}^n$ as the Kodaira-type holomorphic map

$$\mathcal{G} = [\partial X_1 : \partial X_2 : \cdots : \partial X_n] : M \to Q^{n-2} \subset \mathbb{CP}^{n-1}, \tag{3.2}$$

where

$$Q = Q^{n-2} = \left\{ [z_1 : \cdots : z_n] \in \mathbb{CP}^{n-1} : \sum_{j=1}^{n} z_j^2 = 0 \right\}$$

is the projectivization of the punctured null quadric $\mathbb{A}_*$, a smooth quadric complex hypersurface in $\mathbb{CP}^{n-1}$. A recent discovery is the following converse result from [14] (see also [16, Theorem 5.4.1]), which shows that every natural candidate is the Gauss map of a conformal minimal surface.

Main Theorem 3.3. *Assume that* $n \ge 3$.

(i) *For every holomorphic map* $\mathcal{G} : M \to Q^{n-2}$ *from an open Riemann surface there exists a conformal minimal immersion* $X : M \to \mathbb{R}^n$ *with the Gauss map* $\mathcal{G}$.

(ii) *If* M *is a compact bordered Riemann surface and* $\mathcal{G} : M \to Q^{n-2}$ *is a map of class* $\mathscr{A}^{r-1}(M, Q^{n-2})$ *for some* $r \in \mathbb{N}$*, then there is a conformal minimal immersion* $X : M \to \mathbb{R}^n$ *of class* $\mathscr{C}^r(M, \mathbb{R}^n)$ *with the Gauss map* $\mathcal{G}$.

Here, $\mathscr{A}^{r-1}(M, Q^{n-2})$ denotes the space of maps $M \to Q^{n-2}$ of class $\mathscr{C}^{r-1}$ which are holomorphic in the interior $M \setminus bM$ of M.

Furthermore, the following assertions hold true in both cases in the above theorem.

(i) The conformal minimal immersion X can be chosen to have vanishing flux. In particular, every holomorphic map $\mathcal{G} : M \to Q^{n-2}$ is the Gauss map of a holomorphic null curve $M \to \mathbb{C}^n$.

(ii) If $\mathcal{G}(M)$ is not contained in any projective hyperplane of $\mathbb{CP}^{n-1}$, then X can be chosen with arbitrary flux, to have prescribed values on a given closed discrete subset Λ of M, to be an immersion with simple double points if $n = 4$, and to be an injective immersion if $n \ge 5$ and the prescription of values on Λ is injective.

When $n = 3$, the quadric Q^1 is an embedded rational curve in $\mathbb{CP}^2$ parameterized by the biholomorphic map

$$\mathbb{CP}^1 \ni t \xmapsto{\tau} \left[\frac{1}{2}\left(\frac{1}{t} - t\right) : \frac{\mathrm{i}}{2}\left(\frac{1}{t} + t\right) : 1 \right] = \left[1 - t^2 : \mathrm{i}(1 + t^2) : 2t\right] \in Q^1. \tag{3.3}$$

Writing $(1 - t^2, \mathrm{i}(1 + t^2), 2t) = (a, b, c)$, we easily find that

$$t = \frac{c}{a - \mathrm{i}\,b} = \frac{b - \mathrm{i}\,a}{\mathrm{i}\,c} \in \mathbb{CP}^1.$$

Suppose that $X = (X_1, X_2, X_3) : M \to \mathbb{R}^3$ is a conformal minimal immersion, and write $2\partial X = 2(\partial X_1, \partial X_2, \partial X_3) = (\phi_1, \phi_2, \phi_3)$. In view of the above formula for $t = t(a, b, c)$ it is natural to consider the holomorphic map

$$\mathfrak{g} = \frac{\phi_3}{\phi_1 - \mathrm{i}\,\phi_2} = \frac{\partial X_3}{\partial X_1 - \mathrm{i}\,\partial X_2} : M \to \mathbb{CP}^1.$$

This is the complex Gauss map (2.26) of X, which appears in the Enneper–Weierstrass representation (2.25). The generalized Gauss map $\mathscr{G} : M \to Q^1 \subset \mathbb{CP}^2$ (3.2) of X is then expressed by $\mathscr{G} = \tau \circ \mathfrak{g}$, where $\tau : \mathbb{CP}^1 \to Q^1$ is given by (3.3).

Let us say a few words about the proof of Theorem 3.3. The first step is to lift the given map $\mathscr{G} : M \to Q$ to a holomorphic map $G : M \to \mathbb{A}_*$. Note that the natural projection $\mathbb{A}_* \to Q$ sending $(z_1, \dots, z_n)$ to $[z_1 : \cdots : z_n]$ is a holomorphic fibre bundle with fibre $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. The existence of a continuous lifting follows by noting that the homotopy type of M is a wedge of circles, and every oriented $\mathbb{C}^*$-bundle over a circle is trivial. Further, since $\mathbb{C}^*$ is an Oka manifold, every continuous lifting is homotopic to a holomorphic lifting according to the Oka principle [35, Corollary 5.5.11].

In the second and main step of the proof, the holomorphic map $G : M \to \mathbb{A}_*$ is multiplied by a nowhere vanishing holomorphic function $h : M \to \mathbb{C}^*$ such that the product $f = hG : M \to \mathbb{A}_*$ has vanishing periods along closed curves in M (see (2.31)), and hence it integrates to a holomorphic null immersion $Z : M \to \mathbb{C}^n$. Its real part $X = \Re Z : M \to \mathbb{R}^n$ is then a conformal minimal immersion having the Gauss map $\mathscr{G}$. The construction of such a multiplier h follows the idea of proof of Theorem 3.1, but the details are fairly nontrivial and we refer to the cited works.

There are many results in the literature relating the behavior of a minimal surface to properties of its Gauss map. A particularly interesting question is how many hyperplanes in a general position in $\mathbb{CP}^{n-1}$ can be omitted by the Gauss map of a complete conformal minimal surface of finite total curvature. A discussion of this topic can be found in [16, Chapter 5] and in several other sources.

3.4. The Calabi–Yau problem

A smooth immersion $X : M \to \mathbb{R}^n$ is said to be complete if X^*ds^2 is a complete metric on M. Equivalently, for every divergent path $\gamma : [0, 1) \to M$ (i.e., such that $\gamma(t)$ leaves every compact set in M as $t \to 1$) the image path $X \circ \gamma : [0, 1) \to \mathbb{R}^n$ has infinite Euclidean length. Clearly, if X is proper, then it is complete since any such path $X \circ \gamma(t)$ diverges to infinity as $t \to 1$. The converse is not true; it is easy to construct complete immersions (and embeddings if $n \geq 3$) with bounded image $X(M) \subset \mathbb{R}^n$.

It is however not so easy to find complete bounded immersions with additional properties, such as conformal minimal or, in case when the target is a complex

Euclidean space $\mathbb{C}^n$, holomorphic. The following conjecture was posed by Eugenio Calabi in 1965, [50, p. 170]. Calabi's conjecture was also promoted by Shiing-Shen Chern [24, p. 212].

Conjecture 3.4. *Every complete minimal hypersurface in* $\mathbb{R}^n$ $(n \geq 3)$ *is unbounded. Furthermore, every complete nonflat minimal hypersurface in* $\mathbb{R}^n$ $(n \geq 3)$ *has an unbounded projection to every* $(n-2)$*-dimensional affine subspace.*

A particular reason which may have led Calabi to propose these conjectures was the theorem of Chern and R. Osserman [25] from that time. Their result says in particular that if $X : M \to \mathbb{R}^n$ $(n \geq 3)$ is a complete conformal minimal surface of finite total Gaussian curvature $\mathrm{TC}(X) > -\infty$, then M is the complement of finitely many points $p_1, \ldots, p_m$ in a compact Riemann surface R, the holomorphic 1-form ∂X has an effective pole at each point p_j, and X is proper. (The first statement holds even without the completeness assumption on X, due to a result of Huber [47] from 1957.) The Chern–Osserman theorem says that such an X is complete if and only if ∂X has an effective pole at each puncture p_j. The asymptotic behavior of X at the punctures was described by M. Jorge and W. Meeks [48] in 1983.

It turns out that, at least in dimension $n = 3$, Calabi's conjecture is both right and wrong, depending on whether the minimal surface is embedded or merely immersed. (This point was not specified in the original question.) In dimension $n = 3$, the answer is radically different for these two cases, as we now explain.

The first counterexample to Calabi's conjecture in the immersed case was given by L. P. de M. Jorge and F. Xavier in 1980 [49], who constructed a complete nonflat conformal minimal immersion $\mathbb{D} \to \mathbb{R}^3$ from the disc with the range contained in a slab between two parallel planes.

In 1982, S.-T. Yau pointed out in [80, Problem 91] that the question whether there are complete bounded minimal surfaces in $\mathbb{R}^3$ remained open despite Jorge–Xavier's example. This became known as the *Calabi–Yau problem for minimal surfaces*.

The problem was resolved for immersed surfaces by N. Nadirashvili [67] who in 1996 constructed a complete conformal minimal immersion $\mathbb{D} \to \mathbb{R}^3$ with the image contained in a ball. Many subsequent results followed, showing similar results for topologically more general surfaces; see [16, Section 7.1] for a survey and references. However, the conformal type of the examples could not be controlled by the methods developed in those papers, except for the disc. The reason is that the increase of the intrinsic radius of a surface was achieved by applying Runge's theorem on pieces of a suitable labyrinth in the surface, chosen such that any divergent path avoiding most pieces has infinite length, while crossing a piece of the labyrinth increases the length by a prescribed amount. However, Runge's theorem does not allow to control the map everywhere, and hence small pieces of the surface had to be cut away in order to keep the image bounded. This surgery changes the conformal type of the surface, and only its topological type can be controlled by this method.

After Nadirashvili's paper, Yau revisited the Calabi–Yau conjectures in his 2000 millennium lecture and proposed several new questions (see [81, p. 360] or [82, p. 241]). He asked in particular: What is the geometry of complete bounded minimal surfaces in $\mathbb{R}^3$? Can they be embedded? What can be said about the asymptotic behavior of these surfaces near their ends?

Concerning Calabi's conjecture for embedded surfaces, Colding and Minicozzi showed in 2008 [29] that every complete embedded minimal surface in $\mathbb{R}^3$ of finite topological type is proper in $\mathbb{R}^3$. Their result was extended to surfaces of finite genus and countably many ends by W. H. Meeks, J. Pérez, and A. Ros in 2018, [61]. Hence,

> *Calabi's conjecture holds true for embedded minimal surfaces of finite genus and countably many ends in $\mathbb{R}^3$.*

Against this background, we have the following result for immersed surfaces.

Main Theorem 3.5. *Every open Riemann surface of finite genus and at most countably many ends, none of which are point ends, is the conformal structure of a complete bounded immersed minimal surface in $\mathbb{R}^3$.*

By the uniformization theorem of Z.-X. He and O. Schramm [45, Theorem 0.2] (1993) solving Koebe's conjecture, every open Riemann surface of finite genus and at most countably many ends is conformally equivalent to a domain of the form

$$M = R \setminus \bigcup_i D_i, \tag{3.4}$$

where R is a compact Riemann surface without boundary and $\{D_i\}_i$ is a finite or countable family of pairwise disjoint compact geometric discs or points in R. (A *geometric disc* in R is a compact subset whose preimage in the universal holomorphic covering space of R, which is one of the surfaces $\mathbb{CP}^1$, $\mathbb{C}$, or $\mathbb{D}$, is a family of pairwise disjoint round discs or points.) Such an M is called a *circled domain* in R. Hence, Theorem 3.5 is a corollary to the following more precise result, which includes information about the boundary behavior of surfaces.

Main Theorem 3.6. *Assume that M is a circled domain of the form* (3.4). *For any $n \geq 3$ there exists a continuous map $X : \bar{M} \to \mathbb{R}^n$ such that $X : M \to \mathbb{R}^n$ is a complete conformal minimal immersion and $X : bM \to \mathbb{R}^n$ is a topological embedding. If $n \geq 5$, then there is a topological embedding $X : \bar{M} \to \mathbb{R}^n$ such that $X : M \to \mathbb{R}^n$ is a complete embedded minimal surface.*

This means that the image $X(M)$ is a complete immersed minimal surface whose boundary $X(bM)$ consists of pairwise disjoint Jordan curves. The control of conformal structures on complete minimal surfaces in Theorems 3.5 and 3.6 is one of the main new aspects of these results; the other one is that the surfaces in Theorem 3.6 have Jordan boundaries. These answer the aforementioned questions by Yau.

For surfaces M of type (3.4) with finitely many boundary components, Theorem 3.6 was proved in [4]. This covers all finite bordered Riemann surfaces in view of the uniformization theorem [76, Theorem 8.1] due to E. L. Stout. In this case, we actually showed that any conformal minimal immersion $\bar{M} \to \mathbb{R}^n$ can be approximated uniformly on $\bar{M}$ by a map X as in the theorem. The general case for countably many ends was obtained in [10]; an approximation theorem also holds in that case.

The situation regarding point ends remains elusive and does not have a clear-cut answer. On the one hand, a bounded conformal minimal surface cannot be complete at an isolated point end (a puncture) since a bounded harmonic function extends across a puncture. On the other hand, it was shown in [10, Theorem 5.1] that an analogue of Theorem 3.6 holds for connected domains of the form

$$M = R \setminus \Big(E \cup \bigcup_i D_i \Big),$$

where E is a compact set in a compact Riemann surface R and $D_i \subset R \setminus E$ are pairwise disjoint geometric discs such that the distance to E is infinite within M. In particular, there are complete bounded conformal minimal surfaces in $\mathbb{R}^3$ with point ends which are limits of disc ends.

Our construction uses an adaptation of the Riemann–Hilbert boundary value problem to holomorphic null curves and conformal minimal surfaces, together with a method of exposing boundary points of such surfaces. This technique is explained in detail in [16, Chapter 6]. The modifications which we use provide a good control of the position of the whole surface in the ambient space, thereby keeping it bounded. The main technical lemma of independent interest (see [16, Lemma 7.3.1]) enables one to make the intrinsic radius of a conformal bordered minimal surface in $\mathbb{R}^n$ as large as desired by a deformation of the surface which is uniformly as small as desired. One uses this lemma in an inductive process which converges to a bounded complete limit surface. This lemma also allows the construction of complete minimal surfaces with other interesting geometric properties. In particular, every bordered Riemann surface admits a complete proper conformal minimal immersion into any convex domain in $\mathbb{R}^n$ (embedding if $n \geq 5$) and, more generally, into any minimally convex domain (see [16, Section 8.3]). A smoothly bounded domain in $\mathbb{R}^3$ is minimally convex if and only if the boundary has nonnegative mean curvature at each point.

We give a brief description of the modifications which lead to proof of the above results. A complete presentation of this technique is given in [16, Chapter 6], and Theorem 3.6 is proved in [16, Chapter 7]. Illustrations can be found in my lecture [36].

Each step consists of two substeps. In the first substep, we choose a large but finite number of roughly equidistributed points on the boundary of the surface and change the surface so that it grows long spikes (tentacles) at these points, which however remain uniformly close to the attachment points. (Imagine the picture of a corona virus.) The effect of this modification is that curves in the surface which terminate near one of the exposed boundary points get elongated by a prescribed amount. See [16, Section 6.7].

In the second substep, we perform a Riemann–Hilbert type modification which increases the intrinsic radius along each of the boundary arcs between a pair of exposed points, without destroying the effect of substep 1. To each boundary arc between a pair of exposed points we attach a 3-dimensional cylinder, consisting of a 1-parameter family of conformal minimal discs centered at points of the given arc. The boundaries of these discs form a 2-dimensional cylinder, a product of the arc with a circle, and their radii shrink to zero near the exposed endpoints of the arc. Is it then possible to modify the surface by pushing each arc very near the corresponding 2-dimensional cylinder, with the modification tempering out near the exposed endpoints and away from the arcs. So, the modification in substep 2 is big very close to the boundary (except near the exposed points), and it is arbitrarily small outside a given neighborhood of the boundary. The new conformal minimal surface is contained in an arbitrarily small neighborhood of the union of the surface from substep 1 and the 3-dimensional cylinders that have been attached to the arcs in substep 2. The metric effect of the modification in substep 2 is that the length of any path in the surface terminating at an interior point of one of the boundary arcs increases almost by the radius of the disc that was attached at this point. (For curves terminating near the exposed points a desired elongation was already achieved in substep 1.) For technical reasons, we actually work with ∂-derivatives of these conformal minimal surfaces, including the boundary discs, so the entire picture concerns families of holomorphic maps with values in the punctured null quadric $\mathbb{A}_*$. In order to control the period conditions, we work with sprays of such configurations, like in the proof of Theorem 3.1. Special attention is paid to avoid introducing branch points to our surfaces in the process. As said before, this provides the main modification lemma, and its inductive application leads to the proof of Theorem 3.6.

By this method, the Calabi–Yau property has been established in several geometries: for holomorphic curves in complex manifolds [6], holomorphic null curves in $\mathbb{C}^n$ and conformal minimal surfaces in $\mathbb{R}^n$ for $n \geq 3$ [4, 7, 10], holomorphic Legendrian curves in complex contact manifolds [8, 13], and superminimal surfaces in self-dual or anti-self-dual Einstein 4-manifolds [37]. For a survey and further references, see [16, Section 7.4]. An axiomatic approach to the Calabi–Yau problem was proposed in [11].

The analogue of the Calabi–Yau problem for complex submanifolds in $\mathbb{C}^n$, which is known as *Paul Yang's problem* who raised it in 1977 [79], has also received a lot of recent attention. In particular, J. Globevnik showed [40] that for any pair of integers $1 \leq k < n$, the ball of $\mathbb{C}^n$ admits holomorphic foliations by complete k-dimensional proper complex subvarieties, most of which are without singularities (submanifolds). Another construction using a different technique was given by Alarcón et al. [17], and it was also shown that there are nonsingular holomorphic foliations of the ball having complete leaves (Alarcón [1]). Furthermore, there are nonsingular holomorphic foliations of the ball whose leaves are complete properly embedded discs [9]. The techniques in these papers do not apply to more general minimal surfaces, and they do not provide control of complex structures of examples.

In conclusion, I propose the following conjecture. Although I am fully aware of the lack of technical tools to solve it in this generality, I believe that it is true.

Conjecture 3.7. *The Calabi–Yau property holds for bordered minimal surfaces in any smooth Riemannian manifold* (N, g) *with* $\dim N \geq 3$. *Explicitly, for every bordered Riemann surface,* M, *and conformal minimal immersion* $X : \overline{M} \to N$ *it is possible to approximate* X *uniformly on* M *by complete conformal minimal immersions* $M \to N$.

Acknowledgments. I wish to thank Antonio Alarcón and Francisco J. López for having introduced me to this beautiful subject back in 2011. I also thank the mentioned colleagues, as well as Barbara Drinovec Drnovšek and Finnur Lárusson, for the continuing collaboration on this subject.

Funding. This research was supported by Program P1-0291 from ARRS, Republic of Slovenia.

References

[1] A. Alarcón, Complete complex hypersurfaces in the ball come in foliations. *J. Differential Geom.*, to appear

[2] A. Alarcón and I. Castro-Infantes, Complete minimal surfaces densely lying in arbitrary domains of $\mathbb{R}^n$. *Geom. Topol.* **22** (2018), no. 1, 571–590 Zbl 1378.53070 MR 3720350

[3] A. Alarcón and I. Castro-Infantes, Interpolation by conformal minimal surfaces and directed holomorphic curves. *Anal. PDE* **12** (2019), no. 2, 561–604 Zbl 1402.53005 MR 3861901

[4] A. Alarcón, B. Drinovec Drnovšek, F. Forstnerič, and F. J. López, Every bordered Riemann surface is a complete conformal minimal surface bounded by Jordan curves. *Proc. Lond. Math. Soc. (3)* **111** (2015), no. 4, 851–886 Zbl 1344.53008 MR 3407187

[5] A. Alarcón and F. Forstnerič, Null curves and directed immersions of open Riemann surfaces. *Invent. Math.* **196** (2014), no. 3, 733–771 Zbl 1297.32009 MR 3211044

[6] A. Alarcón and F. Forstnerič, Every bordered Riemann surface is a complete proper curve in a ball. *Math. Ann.* **357** (2013), no. 3, 1049–1070 Zbl 1288.32014 MR 3118624

[7] A. Alarcón and F. Forstnerič, The Calabi–Yau problem, null curves, and Bryant surfaces. *Math. Ann.* **363** (2015), no. 3-4, 913–951 Zbl 1343.53053 MR 3412347

[8] A. Alarcón and F. Forstnerič, Darboux charts around holomorphic Legendrian curves and applications. *Int. Math. Res. Not. IMRN* **2019** (2019), no. 3, 893–922 Zbl 1429.53088 MR 3910477

[9] A. Alarcón and F. Forstnerič, A foliation of the ball by complete holomorphic discs. *Math. Z.* **296** (2020), no. 1-2, 169–174 Zbl 1447.32037 MR 4140736

[10] A. Alarcón and F. Forstnerič, The Calabi–Yau problem for Riemann surfaces with finite genus and countably many ends. *Rev. Mat. Iberoam.* **37** (2021), no. 4, 1399–1412 Zbl 1469.53014 MR 4269402

[11] A. Alarcón, F. Forstnerič, and F. Lárusson, Holomorphic Legendrian curves in $\mathbb{CP}^3$ and superminimal surfaces in $\mathbb{S}^4$. *Geom. Topol.* **25** (2021), no. 7, 3507–3553 Zbl 07483973 MR 4372635

[12] A. Alarcón, F. Forstnerič, and F. J. López, Embedded minimal surfaces in $\mathbb{R}^n$. *Math. Z.* **283** (2016), no. 1-2, 1–24 Zbl 1360.53020 MR 3489056

[13] A. Alarcón, F. Forstnerič, and F. J. López, Holomorphic Legendrian curves. *Compos. Math.* **153** (2017), no. 9, 1945–1986 Zbl 1373.53106 MR 3705282

[14] A. Alarcón, F. Forstnerič, and F. J. López, Every meromorphic function is the Gauss map of a conformal minimal surface. *J. Geom. Anal.* **29** (2019), no. 4, 3011–3038 Zbl 1432.30030 MR 4015426

[15] A. Alarcón, F. Forstnerič, and F. J. López, New complex analytic methods in the study of non-orientable minimal surfaces in $\mathbb{R}^n$. *Mem. Amer. Math. Soc.* **264** (2020), no. 1283, vi+77 Zbl 07213239 MR 4078111

[16] A. Alarcón, F. Forstnerič, and F. J. López, *Minimal Surfaces from a Complex Analytic Viewpoint*. Springer Monogr. Math., Springer, Cham, 2021 Zbl 07332812 MR 4237295

[17] A. Alarcón, J. Globevnik, and F. J. López, A construction of complete complex hypersurfaces in the ball with control on the topology. *J. Reine Angew. Math.* **751** (2019), 289–308 Zbl 1423.32018 MR 3956697

[18] A. Alarcón and F. J. López, Algebraic approximation and the Mittag–Leffler theorem for minimal surfaces. *Anal. PDE*, to appear

[19] A. Alarcón and F. J. López, Minimal surfaces in $\mathbb{R}^3$ properly projecting into $\mathbb{R}^2$. *J. Differential Geom.* **90** (2012), no. 3, 351–381 Zbl 1252.53005 MR 2916039

[20] J. L. M. Barbosa and A. G. Colares, *Minimal Surfaces in* $\mathbf{R}^3$. *Translated from the Portuguese*. Lecture Notes in Math. 1195, Springer, Berlin, 1986 Zbl 0609.53001 MR 853728

[21] H. Behnke and K. Stein, Entwicklung analytischer Funktionen auf Riemannschen Flächen. *Math. Ann.* **120** (1949), 430–461 Zbl 0038.23502 MR 29997

[22] H. Cartan, Variétés analytiques complexes et cohomologie. In *Colloque sur les fonctions de plusieurs variables, tenu à Bruxelles, 1953*, pp. 41–55, Georges Thone, Liège; Masson & Cie, Paris, 1953 Zbl 0053.05301 MR 0064154

[23] E. Catalan, Sur les surfaces réglées dont l'aire est un minimum. *J. Math. Pure Appl.* **7** (1842), 203–211

[24] S.-S. Chern, The geometry of G-structures. *Bull. Amer. Math. Soc.* **72** (1966), 167–219 Zbl 0136.17804 MR 192436

[25] S.-S. Chern and R. Osserman, Complete minimal surfaces in euclidean n-space. *J. Analyse Math.* **19** (1967), 15–34 Zbl 0172.22802 MR 226514

[26] T. H. Colding and W. P. Minicozzi II, *Minimal Surfaces*. Courant Lect. Notes Math. 4, New York University, Courant Institute of Mathematical Sciences, New York, 1999 Zbl 0987.49025 MR 1683966

[27] T. H. Colding and W. P. Minicozzi II, The space of embedded minimal surfaces of fixed genus in a 3-manifold. IV. Locally simply connected. *Ann. of Math. (2)* **160** (2004), no. 2, 573–615 Zbl 1076.53069 MR 2123933

[28] T. H. Colding and W. P. Minicozzi II, Shapes of embedded minimal surfaces. *Proc. Natl. Acad. Sci. USA* **103** (2006), no. 30, 11106–11111 Zbl 1175.53008 MR 2242650

[29] T. H. Colding and W. P. Minicozzi II, The Calabi–Yau conjectures for embedded surfaces. *Ann. of Math. (2)* **167** (2008), no. 1, 211–243 Zbl 1142.53012 MR 2373154

[30] T. H. Colding and W. P. Minicozzi II, *A course in Minimal Surfaces*. Grad. Stud. Math. 121, American Mathematical Society, Providence, RI, 2011 Zbl 1242.53007 MR 2780140

[31] J. Douglas, Solution of the problem of Plateau. *Trans. Amer. Math. Soc.* **33** (1931), no. 1, 263–321 Zbl 57.0601.01 MR 1501590

[32] L. Euler, *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes: sive solutio problematis isoperimetrici latissimo sensu accepti (1744)*. Opera Omnia (1) 24, Harvard University Press, Cambridge, MA, 1969

[33] J. E. Fornæss, F. Forstnerič, and E. F. Wold, Holomorphic approximation: the legacy of Weierstrass, Runge, Oka–Weil, and Mergelyan. In *Advancements in Complex Analysis—from Theory to Practice*, pp. 133–192, Springer, Cham, 2020 Zbl 07216541 MR 4264040

[34] F. Forstnerič, Oka manifolds. *C. R. Math. Acad. Sci. Paris* **347** (2009), no. 17-18, 1017–1020 Zbl 1175.32005 MR 2554568

[35] F. Forstnerič, *Stein Manifolds and Holomorphic Mappings. The Homotopy Principle in Complex Analysis*. 2nd edn., Ergeb. Math. Grenzgeb. (3) 56, Springer, Cham, 2017 Zbl 1382.32001 MR 3700709

[36] F. Forstnerič, Minimal surfaces from a complex analytic viewpoint. 2021, https://8ecm.si/system/admin/abstracts/presentations/000/000/663/original/8ECM2021.pdf?1626190740

[37] F. Forstnerič, The Calabi–Yau property of superminimal surfaces in self-dual Einstein four-manifolds. *J. Geom. Anal.* **31** (2021), no. 5, 4754–4780 Zbl 1466.53072 MR 4244884

[38] F. Forstnerič and D. Kalaj, Hyperbolicity theory for minimal surfaces in Euclidean spaces. 2021, arXiv:2102.12403

[39] F. Forstnerič and F. Lárusson, The parametric h-principle for minimal surfaces in $\mathbb{R}^n$ and null curves in $\mathbb{C}^n$. *Comm. Anal. Geom.* **27** (2019), no. 1, 1–45 Zbl 1414.53009 MR 3951019

[40] J. Globevnik, A complete complex hypersurface in the ball of $\mathbb{C}^N$. *Ann. of Math. (2)* **182** (2015), no. 3, 1067–1091 Zbl 1333.32018 MR 3418534

[41] H. Grauert, Holomorphe Funktionen mit Werten in komplexen Lieschen Gruppen. *Math. Ann.* **133** (1957), 450–472 Zbl 0080.29202 MR 98198

[42] H. Grauert and H. Kerner, Approximation von holomorphen Schnittflächen in Faserbündeln mit homogener Faser. *Arch. Math. (Basel)* **14** (1963), 328–333 Zbl 0113.29102 MR 153871

[43] M. L. Gromov, Convex integration of differential relations. I. *Izv. Akad. Nauk SSSR Ser. Mat.* **37** (1973), 329–343 Zbl 0281.58004 MR 0413206

[44] R. C. Gunning and R. Narasimhan, Immersion of open Riemann surfaces. *Math. Ann.* **174** (1967), 103–108 Zbl 0179.11402 MR 223560

[45] Z.-X. He and O. Schramm, Fixed points, Koebe uniformization and circle packings. *Ann. of Math. (2)* **137** (1993), no. 2, 369–406 Zbl 0777.30002 MR 1207210

[46] L. Henneberg, Ueber diejenige Minimalfläche, welche die Neil'sche Parabel zur ebenen geodätischen Linie hat. *Wolf Z.* **21** (1876), 17–21 Zbl 08.0528.01

[47] A. Huber, On subharmonic functions and differential geometry in the large. *Comment. Math. Helv.* **32** (1957), 13–72 Zbl 0080.15001 MR 94452

[48] L. P. Jorge and W. H. Meeks III, The topology of complete minimal surfaces of finite total Gaussian curvature. *Topology* **22** (1983), no. 2, 203–221 Zbl 0517.53008 MR 683761

[49] L. P. d. M. Jorge and F. Xavier, A complete minimal surface in $\mathbf{R}^3$ between two parallel planes. *Ann. of Math. (2)* **112** (1980), no. 1, 203–206 Zbl 0455.53004 MR 584079

[50] S. Kobayashi and J. Eells Jr., *Proceedings of the United States-Japan Seminar in Differential Geometry, Kyoto, Japan, 1965*. Nippon Hyoronsha, Tokyo, 1966

[51] J.-L. Lagrange, Application de la méthode exposée dans le mémoire précédent à la solution des problèmes de dynamique différents. In *Accademia delle scienze di Torino (1760–1761)*, pp. 365–468, Oeuvres de Lagrange 1, Gauthier–Villars, Paris, 1867

[52] J.-L. Lagrange, Essai d'une nouvelle méthode pour déterminer les maxima et les minima des formules intégrales indéfinies. In *Accademia delle scienze di Torino (1760–1761)*, pp. 335–362, Oeuvres de Lagrange 1, Gauthier–Villars, Paris, 1867

[53] J.-L. Lagrange, *Oeuvres. Tome 1. Publiées par les soins de J.-A. Serret. Avec une notice sur la vie et les ouvrages de J.-L. Lagrange par J.-B. J. Delambre. Nachdruck der Ausgabe Paris 1867*. Georg Olms Verlag, Hildesheim, 1973 MR 0439546

[54] F. Lárusson, Absolute neighbourhood retracts and spaces of holomorphic maps from Stein manifolds to Oka manifolds. *Proc. Amer. Math. Soc.* **143** (2015), no. 3, 1159–1167 Zbl 1318.32014 MR 3293731

[55] H. B. Lawson Jr., *Lectures on Minimal Submanifolds. Vol. I.* 2nd edn., Mathematics Lecture Series 9, Publish or Perish, Wilmington, DE, 1980 Zbl 0434.53006 MR 576752

[56] J. P. May, *A Concise Course in Algebraic Topology*. Chicago Lect. Math., University of Chicago Press, Chicago, IL, 1999 Zbl 0923.55001 MR 1702278

[57] W. H. Meeks III, The classification of complete minimal surfaces in $\mathbf{R}^3$ with total curvature greater than -8π. *Duke Math. J.* **48** (1981), no. 3, 523–535 Zbl 0472.53010 MR 630583

[58] W. H. Meeks III and J. Pérez, The classical theory of minimal surfaces. *Bull. Amer. Math. Soc. (N.S.)* **48** (2011), no. 3, 325–407 Zbl 1232.53003 MR 2801776

[59] W. H. Meeks III and J. Pérez, *A Survey on Classical Minimal Surface Theory*. Univ. Lecture Ser. 60, American Mathematical Society, Providence, RI, 2012 Zbl 1262.53002 MR 3012474

[60] W. H. Meeks III and J. Pérez, The Riemann minimal examples. In *The Legacy of Bernhard Riemann After One Hundred and Fifty Years. Vol. II*, pp. 417–457, Adv. Lect. Math. (ALM) 35, Int. Press, Somerville, MA, 2016 Zbl 1359.53010 MR 3525901

[61] W. H. Meeks III, J. Pérez, and A. Ros, The embedded Calabi–Yau conjecture for finite genus. *Duke Math. J.* **170** (2021), no. 13, 2891–2956 Zbl 07433912 MR 4312191

[62] W. H. Meeks III and H. Rosenberg, The uniqueness of the helicoid. *Ann. of Math. (2)* **161** (2005), no. 2, 727–758 Zbl 1102.53005 MR 2153399

[63] W. H. Meeks III and S. T. Yau, The classical Plateau problem and the topology of three-dimensional manifolds. The embedding of the solution given by Douglas-Morrey and an analytic proof of Dehn's lemma. *Topology* **21** (1982), no. 4, 409–442 Zbl 0489.57002 MR 670745

[64] W. W. Meeks III and S. T. Yau, The existence of embedded minimal surfaces and the problem of uniqueness. *Math. Z.* **179** (1982), no. 2, 151–168 Zbl 0479.49026 MR 645492

[65] S. N. Mergelyan, On the representation of functions by series of polynomials on closed sets. *Doklady Akad. Nauk SSSR (N.S.)* **78** (1951), 405–408 MR 0041929

[66] J. B. Meusnier, Mémoire sur la courbure des surfaces (1776). *Mem. Math. Phys. Acad. Sci. Paris* **10** (1785), 477–510

[67] N. Nadirashvili, Hadamard's and Calabi–Yau's conjectures on negatively curved and minimal surfaces. *Invent. Math.* **126** (1996), no. 3, 457–465 Zbl 0881.53053 MR 1419004

[68] J. C. C. Nitsche, *Lectures on Minimal Surfaces. Vol. 1. Introduction, Fundamentals, Geometry and Basic Boundary Value Problems. Translated from the German by Jerry M. Feinberg. With a German foreword*. Cambridge University Press, Cambridge, 1989 Zbl 0688.53001 MR 1015936

[69] R. Osserman, *A Survey of Minimal Surfaces*. 2nd edn., Dover Publications, New York, 1986 MR 852409

[70] T. Radó, Über eine nicht fortsetzbare Riemannsche Mannigfaltigkeit. *Math. Z.* **20** (1924), no. 1, 1–6 Zbl 50.0255.02 MR 1544659

[71] T. Radó, On Plateau's problem. *Ann. of Math. (2)* **31** (1930), no. 3, 457–469 MR 1502955

[72] T. Radó, The problem of the least area and the problem of Plateau. *Math. Z.* **32** (1930), no. 1, 763–796 Zbl 56.0436.01 MR 1545197

[73] B. Riemann, Ueber die Fläche vom kleinsten Inhalt bei gegebener Begrenzung. Bearbeitet von K. Hattendorf. *Gött. Abh.* **13** (1868), 3–52 Zbl 01.0218.01

[74] W. Rudin, *Function Theory in the Unit Ball of* $\mathbb{C}^n$. *Reprint of the 1980 edition*. Classics Math., Springer, Berlin, 2008 Zbl 1139.32001 MR 2446682

[75] C. Runge, Zur Theorie der Eindeutigen Analytischen Functionen. *Acta Math.* **6** (1885), no. 1, 229–244 Zbl 17.0379.01 MR 1554664

[76] E. L. Stout, Bounded holomorphic functions on finite Reimann surfaces. *Trans. Amer. Math. Soc.* **120** (1965), 255–285 Zbl 0154.32903 MR 183882

[77] A. Tromba, *A Theory of Branched Minimal Surfaces*. Springer Monogr. Math., Springer, Heidelberg, 2012 Zbl 1247.53002 MR 2920587

[78] K. Weierstrass, Ueber die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen. *Berl. Ber.* **1885** (1885), 633–640, 789–806 Zbl 17.0384.02

[79] P. Yang, Curvatures of complex submanifolds of $\mathbf{C}^n$. *J. Differential Geometry* **12** (1977), no. 4, 499–511 Zbl 0409.53043 MR 512921

[80] S. T. Yau, Problem section. In *Seminar on Differential Geometry*, pp. 669–706, Ann. of Math. Stud. 102, Princeton Univ. Press, Princeton, NJ, 1982 Zbl 0479.53001 MR 645762

[81] S.-T. Yau, Review of geometry and analysis. In *Mathematics: Frontiers and Perspectives*, pp. 353–401, Amer. Math. Soc., Providence, RI, 2000 Zbl 0969.53001 MR 1754787

[82] S.-T. Yau, Review of geometry and analysis. Kodaira's issue. *Asian J. Math.* **4** (2000), no. 1, 235–278 Zbl 1031.53004 MR 1803723

Franc Forstnerič
Department of Mathematics, Faculty of Mathematics and Physics, University of Ljubljana;
and Institute of Mathematics, Physics and Mechanics, Jadranska 19, 1000 Ljubljana, Slovenia;
franc.forstneric@fmf.uni-lj.si

DOI 10.4171/8ECM/34

Bernoulli random matrices

Alice Guionnet

Abstract. Random matrix theory has become a field on its own with a breadth of new results, techniques, and ideas in the last thirty years. In these proceedings, I illustrate some of these advances by describing what we now know about the spectrum and the eigenvectors of Bernoulli matrices.

1. Introduction

Jacques (or Jakob) Bernoulli (1654–1705) was a renowned Swiss mathematician who made important contributions to probability theory and partial differential equations. He was the first to discover the number e. But his most famous result is, at least for probabilists, the first proof of the law of large numbers. To this end, he analyzed the concept of the Bernoulli law, which is the simplest non-trivial distribution you can think of, being the sum of two Dirac masses. It is the distribution of a random variable b which can only take two values 0 and 1. We denote

$$p = \mathbb{P}(b = 1) = 1 - \mathbb{P}(b = 0).$$

A very common example is a coin that, once thrown, falls either on head (modeled by the state 1) or tail (modeled by 0). Even if one would expect in general the probability of each event to be equal to $1/2$, it may well be rather $p \in (0, 1)$ if the coin is rigged. In Ars Conjectandi, Bernoulli showed that if one throws such a coin independently a number n of times, then, with large probability, one should see approximately pn heads if n is large enough. To state this law of large numbers more precisely, he showed that if $b_1, \dots, b_n$ denotes the outcome of n-independent Bernoulli trials, then for any $a < p < b$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} b_i \in [a, b]\right) = 1.$$

2020 Mathematics Subject Classification. Primary 60B20; Secondary 05C80.
Keywords. Random matrices, random graphs.

But how close can we choose a, b to p so that this result remains true? Few years later, A. de Moivre (1667–1754) quantified the size of the error and proved the first central limit theorem, namely that a, b can be at a distance of about $1/\sqrt{n}$ of p in the sense that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{np(1-p)}}\sum_{i=1}^{n}(b_i - p) \in [a,b]\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}\,dx.$$

This was the first occurrence of the central limit theorem and the start of modern probability theory and statistics. Implicitly, we so far assumed that p does not depend on n and belongs to (0, 1). Later on, we shall also be interested in the case where p depends on n. Then, it can be checked that the central limit theorem still holds as long as pn goes to infinity. If pn goes to a finite constant c, then it cannot hold since $\sum_{i=1}^{n} b_i$ is an integer so that the above random variable is discrete. In fact, it converges towards the Poisson distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{np(1-p)}}\sum_{i=1}^{n}(b_i - p) \in [a,b]\right) = \sum_{k\in c+\sqrt{c}[a,b]} \frac{1}{k!}c^k e^{-c}.$$

We will see later that this transition between such continuous and discrete limits is also key to describing the spectrum of Bernoulli random matrices. The last concept which is central in probability theory and important in these notes is entropy. It was introduced by Ludwig Boltzmann (1844–1906) and Claude Shannon (1916–2001) in physics and information theory, respectively, as a way to measure disorder. For again n-independent Bernoulli trials with parameter p, it is defined for any $q \in [0,1]$ by

$$\lim_{\varepsilon\downarrow 0}\lim_{n\to\infty}\frac{1}{n}\ln \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} b_i \in [q-\varepsilon, q+\varepsilon]\right) = -S_p(q),$$

where $S_p(q) = \frac{q}{p}\ln\frac{q}{p} + \frac{1-q}{1-p}\ln\frac{1-q}{1-p}$ is the entropy or rate function.

In this survey, I will discuss Bernoulli random matrices. A Bernoulli random matrix is an $n \times n$ symmetric matrix with independent Bernoulli entries (modulo the symmetry constraint) whose size n is going to infinity. I will discuss the law of large numbers, the fluctuations, and the entropy for their spectrum and eigenvectors. There are many motivations to study random matrices. The first goes back to Wishart who considered random matrices to study correlations in large data sets. Such questions are very modern, with the need to analyze larger and larger data sets and machine learning. The second comes from physics and works of Wigner and Dyson. They proposed to model the Hamiltonian of excited nuclei by random matrices, an idea which turned out to be quite successful as indeed real nuclei turned out to have energy levels distributed like the eigenvalues of random matrices. But Bernoulli matrices

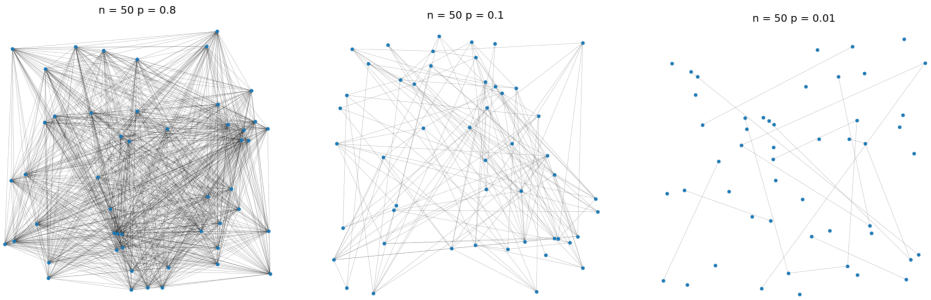


Figure 1. Courtesy of D. Coulette.

are special among all other random matrices because they describe the adjacency matrix of an Erdős–Rényi graph $G(n, p)$. Indeed, the latter is just a graph built on n (labeled) vertices, with an edge drawn independently between each couple of vertices with probability p. Studying the eigenvalues of the adjacency matrix of a graph gives valuable geometric information, such as the size of its boundary (expanders) or the number of specific configurations, such as triangles, that it contains. One can also be interested in the combinatorial properties of such matrices and for instance focus on the probability that the matrix is singular; see e.g. [68]. My viewpoint will be to investigate the properties of the eigenvalues and eigenvectors of Bernoulli random matrices, as a particularly nice and well-documented example of random matrices.

To simplify, I will restrict myself to symmetric Bernoulli matrices $\mathbf{B}_n$ throughout these notes:

$$\mathbf{B}_n(i, j) = \mathbf{B}_n(j, i),$$

and assume that $(B_n(i, j),\ i \le j)$ follows a Bernoulli law with parameter p. Also, I will take $\mathbf{B}_n(i, i)$ random, but could take it equal to zero without changing much the statements of most of the results.

My goal is to understand the spectrum of $\mathbf{B}_n$ as well as the properties of its eigenvectors as n goes to infinity. One can easily guess that these properties should depend on the parameter p. Indeed, thinking about the Erdős–Rényi graph, one sees that the average degree of a vertex is pn. The graph will be very dense if pn goes to infinity fast enough but sparse if it is finite.

Indeed, it is well known since the breakthrough paper of Erdős and Rényi (see Figure 1) that if $np < 1$, $G(n, p)$ will almost surely have no connected component of size greater than $O(\ln n)$; if $np = 1$, there is a giant connected component but it is of size of order $n^{2/3}$; if np goes to a constant $c > 1$, it will have a unique giant component but lots of small components, and isolated vertices will continue to exist until $np < (1 - \varepsilon) \ln n$; whereas if $np > (1 + \varepsilon) \ln n$ the graph will almost surely be connected. Here ε is some positive real number as small as wished. In the case where

np is of order c, the finite size connected components will create small diagonal blocks in the Bernoulli matrix, with entries equal either to zero or one and therefore finitely many possible eigenvalues. Hence, we expect the spectrum to accumulate at these possible values. But should there be other possible eigenvalues? Similarly, we see that the eigenvectors related with these eigenvalues are localized on a few vertices. But should we also have delocalized eigenvectors? On the contrary, in the case where $np > (1+\varepsilon)\ln n$, we may expect eigenvectors to be delocalized and the spectrum to be nicely continuous. In this case, a whole theory has been developed to show that the spectrum and the eigenvalues of Bernoulli matrices have the same properties as those of a random matrix with Gaussian entries. The latter is well known to be much easier to study, for instance, because the joint law of its eigenvalues is rather simple and independent of the eigenvectors. Conversely, Bernoulli matrices resemble more heavy-tailed matrices when pn is of order one, in the sense that it has mostly very tiny entries but a few large entries. Understanding the transition between these two behaviors is at the heart of random matrix theory.

In this survey, I will start discussing the asymptotic behavior of the spectrum in both sparse and dense cases. Then, I will consider its fluctuations, both local and global, as well as the properties of its eigenvectors. Finally, I will discuss the large deviations of the spectrum, for instance how to estimate the probability that the second eigenvalue of Bernoulli matrices takes an unexpected value.

2. Law of large numbers

In this section, we shall see that the limiting distribution of the spectrum differs a lot according to whether pn goes to infinity or not.

A first remark should be made about the matrix $\mathbf{B}_n$: its entries are not centered. It will be more convenient to center them and renormalize the matrix properly. To this end, we make the decomposition

$$\mathbf{B}_n = \sqrt{np(1-p)}\mathbf{X}_n + p\mathbb{1},$$

where $\mathbb{1}$ is a matrix whose entries are all equal to one, whereas the entries of $\mathbf{X}_n$ are centered and renormalized to have covariance $1/n$:

$$\mathbf{X}_n(i,j) = \frac{\mathbf{B}_n(i,j) - p}{\sqrt{np(1-p)}}.$$

The matrix $\mathbb{1}$ has one non-trivial eigenvalue which equals n, and flat eigenvector

$$\mathbf{1} = (1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n}).$$

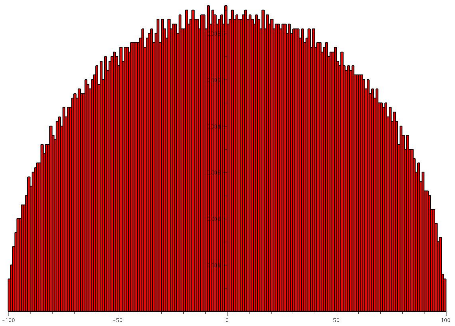


Figure 2.

Conversely, the spectrum of $\mathbf{X}_n$ has eigenvalues mostly of order one in the sense that $\mathbb{E}[\mathrm{Tr}(\mathbf{X}_n^2)] = \mathbb{E}[\sum \lambda_i^2] = n$. Therefore, the above decomposition shows that $\mathbf{B}_n$ has a very large eigenvalue of order n, and the rest is roughly given by the eigenvalues of $\mathbf{X}_n$ taken on $\mathbf{1}^{\perp}$. Moreover, by Weyl's interlacing properties, the eigenvalues $(\lambda_i^B)_{1\le i\le n}$ of $\mathbf{B}_n/\sqrt{np(1-p)}$ and $(\lambda_i^X)_{1\le i\le n}$ of $\mathbf{X}_n$ are interlaced:

$$\lambda_n^X \le \lambda_n^B \le \lambda_{n-1}^X \cdots \le \lambda_1^X \le \lambda_1^B.$$

Therefore, it is in general not difficult to retrieve the properties of the eigenvalues of $\mathbf{B}_n/\sqrt{np(1-p)}$ from those of $\mathbf{X}_n$. Hereafter, we will therefore concentrate mostly on $\mathbf{X}_n$.

2.1. Dense case

The first result describes the asymptotic distribution of the spectrum in the dense case and shows that the limit is described by the famous semi-circle law; see Figure 2.

Theorem 2.1. *Assume that pn goes to infinity as n goes to infinity. Then, almost surely, for any $a < b$*

$$\lim_{n\to\infty} \frac{1}{n}\#\{i : \lambda_i^B \in \sqrt{np(1-p)}[a,b]\} = \lim_{n\to\infty} \frac{1}{n}\#\{i : \lambda_i^X \in [a,b]\} = \sigma\big([a,b]\big),$$

where σ is the semi-circle law given by

$$\sigma(dx) = \frac{1}{2\pi}\sqrt{4-x^2}1_{|x|\le 2}dx. \tag{2.1}$$

The semi-circle law is ubiquitous to random matrix theory as it describes the asymptotic behavior of random matrices with Gaussian entries, but in fact any random matrix with independent centered entries $(a_{ij})_{i,j}$ such that $\mathbb{E}[|\sqrt{n}a_{ij}|^{2+\varepsilon}]$ is

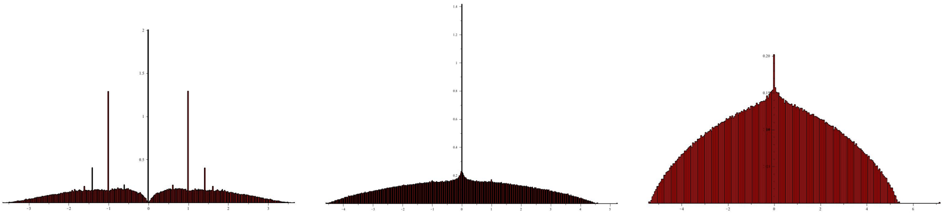


Figure 3. Simulation for $c = 1, 2, 3$ (courtesy of J. Salez).

uniformly bounded for some $\varepsilon > 0$. Such a convergence was proved first by Wigner in the case where p is independent of n based on the computation of the moments $\mathbb{E}[\operatorname{Tr}\mathbf{X}_n^k]$. Indeed, one can expand the trace of moments of matrices in terms of the entries, and observe that the indices which contribute to the first order of this expansion can be described by rooted trees, whereas $\sigma(x^k)$ is equal to the Catalan numbers which enumerate them.

2.2. Sparse case

On the other hand, the limiting distribution of the spectrum is very different when pn is of order one. Namely, we have the following theorem; see [52, 70].

Theorem 2.2. *Assume that pn goes to $c \in (0, +\infty)$ as n goes to infinity. Then, almost surely, for any $a < b$*

$$\lim_{n\to\infty} \frac{1}{n}\#\{i : \lambda_i^B \in \sqrt{np(1-p)}[a,b]\} = \lim_{n\to\infty} \frac{1}{n}\#\{i : \lambda_i^X \in [a,b]\} = \mu_c([a,b]).$$

The limit law μ_c depends on c; some plots are shown in Figure 3.

The simulations indicate the presence of atoms. They were shown to be exactly given by totally real algebraic integers in [58] for all $c>0$; these are the roots of monic polynomials with integer coefficients. It is easy to understand that the atoms should be totally algebraic integers as finite connected components are diagonal blocks with 0 or 1 entries whose characteristic polynomials have such roots. It is a much stronger statement to show that all such roots are atoms, in particular since totally algebraic integers are dense in the real numbers. μ_c has also a continuous spectrum: it was indeed proved in [30] that μ_c has a non-trivial continuous part if and only if $c > 1$. This result is in fact hard to prove as the limit laws μ_c's are described as the solution of complicated equations [28]; see also [17, 20]. However, such description could be used in [8] to prove the existence of an absolutely continuous part for sufficiently large c. Moreover, the first-order expansion of μ_c in c going to infinity was derived in [38]. The spectrum at the origin seems to have a Dirac mass whose weight could be computed [29].

2.3. Idea of the proof

The first proof of Theorem 2.1 estimated the moments $\frac{1}{n}\operatorname{Tr}(\mathbf{X}_n)^k$ for all integer numbers k; see [69] for the first theorem and [17, 52, 70] for the sparse case. However, in order to go into more local results like the behavior of the eigenvectors or the local fluctuations, and as well to have more explicit formulas for the limit law, it is more convenient to study the resolvent. This path can be used to study the asymptotics of the spectral measure of any self-adjoint matrix $\mathbf{X}_n$ with independent entries modulo the symmetry constraint, and was generalized to study heavy-tailed matrices in [17, 20, 52] based on the ideas from [35]. The idea is to derive the asymptotics of the Stieltjes transform

$$G_n(z) = \frac{1}{n}\operatorname{Tr}(z - \mathbf{X}_n)^{-1} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{z - \lambda_i^X}$$

for a complex number z away from the real line. To this end, we use the Schur complement formula which reads

$$(z - \mathbf{X}_n)_{ii}^{-1} = \frac{1}{z - X_{ii} - \langle X_i, (z - \mathbf{X}^{(i)})^{-1} X_i\rangle}, \tag{2.2}$$

where $X_i = (X_{ij})_{j \neq i}$ and $\mathbf{X}^{(i)}$ is the associated principal minor, namely the $(N-1) \times (N-1)$ matrix obtained from $\mathbf{X}_n$ by removing the ith row and column. X_{ii} goes to zero with N and we can check (e.g. by estimating the L^2 norm of the difference) that with probability going to one

$$\langle X_i, (z - \mathbf{X}^{(i)})^{-1} X_i\rangle = \sum_{j : j \neq i} X_{ij}^2 (z - \mathbf{X}^{(i)})_{jj}^{-1} + o(1). \tag{2.3}$$

This is where the “light tail” hypothesis pn going to infinity starts to matter. Then, the entries X_{ij}^2 go to zero and have variance $1/n$ so that, since the X_{ij} are independent of $\mathbf{X}^{(i)}$, the law of large numbers (or a second moment computation) asserts that with probability going to one

$$\sum_{j:j\neq i} X_{ij}^2 (z - \mathbf{X}^{(i)})_{jj}^{-1} = \sum_{j:j\neq i} \mathbb{E}[X_{ij}^2](z - \mathbf{X}^{(i)})_{jj}^{-1} + o(1) = \frac{1}{n}\sum_{j:j\neq i}(z - \mathbf{X}^{(i)})_{jj}^{-1} + o(1).$$

But again $\mathbf{X}^{(i)}$ and $\mathbf{X}_n$ vary only by a rank two matrix (if we complete $\mathbf{X}^{(i)}$ by zero entries at the ith row and column), so that their spectrum is interlaced by Weyl’s interlacing property. As a consequence

$$\frac{1}{n}\sum_{j\neq i}(z - \mathbf{X}^{(i)})_{jj}^{-1} = \frac{1}{n}\sum_{i}(z - \mathbf{X}_n)_{jj}^{-1} + O\left(\frac{1}{\Im(z)n}\right).$$

This approximation, together with (2.2) and (2.3), implies that with high probability

$$G_n(z) = \frac{1}{n}\sum_i (z - \mathbf{X}_n)^{-1}_{jj} = \frac{1}{z - G_n(z)} + o(1). \tag{2.4}$$

After recalling that $G_n(z)$ goes to zero as N goes to infinity, we conclude that since $G_n(z)$ goes to zero as the imaginary part of z goes to infinity,

$$G_n(z) = \frac{1}{2}(z - \sqrt{z^2 - 4}) + o(1)$$

is approximately the Stieltjes transform of the semicircle law $G_\sigma(z) = \frac{1}{2}(z - \sqrt{z^2 - 4})$. Since G_n is analytic and uniformly bounded for $\Im z > \varepsilon$, Montel's theorem implies that G_n converges to this limit away from the real line, which yields the vague convergence of the empirical measure of the eigenvalues. Because $\frac{1}{n}\operatorname{Tr}(\mathbf{X}_n^2)$ is in L^1, the weak convergence follows.

On the contrary, in the heavy-tailed case where pn is of order one, the entries of X_{ij} are often very small but of order one with a positive probability. Hence, the previous law of large numbers does not hold true any more and we cannot expect such a simple equation as (2.4). In fact, $\sum_{j \neq i} X_{ij}^2 (z - \mathbf{X}^{(i)})^{-1}_{jj}$, if it converges, will a priori converge to a random variable. To study this convergence, we make the following assumption on the law μ_n of X_{ij}:

$$\lim_{n\to\infty} n\left(\int (e^{-iux^2} - 1)\, d\mu_n(x)\right) = \Phi(u) \tag{2.5}$$

with Φ such that there exists g on $\mathbb{R}^+$, with $g(y)$ bounded by Cy^κ for some $\kappa > -1$, such that for $u \in \mathbb{C}^-$,

$$\Phi(u) = \int_0^\infty g(y) e^{\frac{iy}{u}}\, dy. \tag{2.6}$$

This is satisfied by the adjacency matrix of Erdős–Rényi graph with $\Phi(u) = c(e^{iu} - 1)$ if pn goes to c and g is a Bessel function [20], but also for other cases, for instance for α stable laws with $\Phi(u) = c(iu)^{\alpha/2}$ and $g(y) = Cy^{\alpha/2-1}$ for some constants c, C. Then, it was shown in [17, 20] that $G_n(z) = \frac{1}{n}\operatorname{Tr}(z - \mathbf{X}^n)^{-1}$ converges almost surely towards G given by

$$G(z) = i\int e^{itz} e^{\rho_z(t)}\, dt, \quad z \in \mathbb{C}^+, \tag{2.7}$$

where $\rho_z : \mathbb{R}^+ \to \{x + iy; x \leq 0\}$ is the unique solution, analytic in $z \in \mathbb{C}^+$, of the non-linear equation

$$\rho_z(t) = \int_0^\infty g(y) e^{\frac{iy}{t}z + \rho_z(\frac{y}{t})}\, dy. \tag{2.8}$$

This entails the convergence of the spectral measure of $\mathbf{X}^n$, with σ replaced by a probability measure with Stieltjes transform given by (2.7). The argument to prove (2.7) and (2.8) is as follows. We first remark that G_n concentrates in the sense that it is close to its average; see Theorem 3.2. We let ρ^n be the order parameter $\rho_z^n(x) := \mathbb{E}[\frac{1}{n}\sum \Phi(x(z-\mathbf{X}^{(i)})_{jj}^{-1})]$. By (2.2) and (2.3), we find that, if $\Im z > 0$,

$$\begin{aligned}
G_n(z) \simeq \mathbb{E}\big[G_n(z)\big] &= -i\mathbb{E}\bigg[\int_0^\infty e^{itz-it\sum_{j\neq i} X_{ij}^2(z-\mathbf{X}^{(i)})_{jj}^{-1}}\,dt\bigg] + o(1)\\
&= i\int_0^\infty e^{itz}\mathbb{E}\bigg[\prod_{j\neq i}\mathbb{E}[e^{-itX_{ij}^2(z-\mathbf{X}^{(i)})_{jj}^{-1}}]\,dt\bigg] + o(1)\\
&= -i\int_0^\infty e^{itz}\mathbb{E}\bigg[\prod_{j\neq i}\bigg(1+\frac{1}{n}\Phi\big(t(z-\mathbf{X}^{(i)})_{jj}^{-1}\big)\bigg)\,dt\bigg] + o(1)\\
&= i\int_0^\infty e^{itz+\rho_z^n(t)}\,dt + o(1).
\end{aligned}$$

To conclude, we need to show the convergence of ρ^n. But ρ^n can be seen to be analytic away from the real axis, and uniformly bounded under our hypothesis. This is enough to see that it is tight and any limit point will be analytic by Montel theorem. Hence, it is enough to show that it has a unique limit point for z with large imaginary part. To this end, we get an equation for ρ^n which follows from (2.6) by

$$\begin{aligned}
\rho_z^n(t) &= \int_0^\infty g(y)\mathbb{E}\Big[e^{\frac{iy}{x(z-\mathbf{X}^{(i)})_{11}^{-1}}}\Big]\,dy\\
&\simeq \int_0^\infty g(y)\mathbb{E}\Big[e^{\frac{iy}{x}(z-\sum_{j\geq 2} X_{ij}^2(z-\mathbf{X}^{(1)})_{jj}^{-1})}\Big]\,dy + o(1)\\
&\simeq \int_0^\infty g(y)e^{\frac{iy}{x}z}e^{\rho_z^n(\frac{y}{x})}\,dy + o(1),
\end{aligned}$$

where in the second line we used (2.2) and (2.3). One can conclude by proving the uniqueness of the solutions to this equation when z is far from the real line by showing that the non-linear equation is then a contraction. The above arguments were made complete in [17, 19, 20]. Another approach to heavy-tailed matrices and sparse Bernoulli matrices based on Aldous' Poisson-weighted infinite tree was proposed in [25].

2.4. Extreme eigenvalues

The asymptotic behavior of the extreme eigenvalues also depend on c: they stick to the bulk when $pn \gg \ln n$ and then go away at distance of order $\sqrt{\ln n}$. We, more precisely, have the following result, putting together the article of Benaych-Georges, Bordenave, and Knowles [18] and that of Alt, Ducatez, and Knowles [4]; see also [65].

Theorem 2.3.
- *Assume that $pn/\ln n \to +\infty$. Then the largest eigenvalue of* $\mathbf{X}_n$ *sticks to the bulk:* $\lambda_1^X \to 2$.
- *Assume that $pn/\ln n \to 0$. Then $\lambda_1^X \simeq \sqrt{\ln n/\ln(\ln n/pn)}$.*
- *Assume that $pn \simeq C\ln n$. Then for $C > 1/(\ln 4 - 1) := C^*$ the eigenvalues stick to the bulk, whereas for $C < 1/(\ln 4 - 1)$*

$$\lambda_1^X = \frac{\alpha}{\sqrt{\alpha - 1}}, \qquad \alpha = \max \frac{1}{pn} \sum_j B_{ij}.$$

Observe that $\sum_j B_{ij}$ is the degree of vertex i: the largest eigenvalue is hence created by the largest degree in the graph. In fact, in the work of Alt, Ducatez, and Knowles [4], it is shown that all eigenvalues outside the bulk are created by vertices with large degrees when $pn \le C^* \ln n$.

3. Fluctuations

3.1. Concentration of measure

Concentration of measure has become a central tool in probability and, in particular, in random matrix theory. It allows us to prove that some quantities, such as smooth function of independent variables, are not much random. It was crucial in the previous proof of the convergence of the spectral measure. However, it generally depends on the tails of the random variables. Herbst's argument allows considering random variables with sub-Gaussian tails and more precisely random variables whose distribution satisfies log-Sobolev inequalities, which is the case for instance when their density is strictly log-concave as for Gaussian's variables. To deal with bounded variables such as the entries of Bernoulli matrices, one should rather use the theory developed by Talagrand [61]. This was done in [44], where the spectrum of random matrices was observed to be a smooth function of its entries and the associated Lipschitz norm was computed. It resulted in the following theorem [44, Theorem 1.1]. We hereafter consider a symmetric matrix $\mathbf{A}$ with independent entries above the diagonal with distribution $a_{ij}/\sqrt{n}$, where a_{ij} is distributed according to P_{ij} supported in a compact set K with width $|K|$.

Theorem 3.1. (1) *Take f convex and Lipschitz with Lipschitz norm $\|f\|_L$. Then, for any $\delta > \delta_0(n) = 8|K|\|f\|_L/n$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{A})\right) - \mathbb{E}\left[\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{A})\right)\right]\right| > \delta\|f\|_L\right) \le 4\exp\left\{-n^2\frac{\left(\delta - \delta_0(n)\right)^2}{16|K|^2}\right\}.$$

(2) *There exists a finite constant $c > 0$ such that for any $\delta > \delta_1(n) \simeq \sqrt{\delta_0(n)}$*

$$\mathbb{P}\left(\sup_{f \in \mathrm{Lip}_{\mathcal{K}}} \left|\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{A})\right) - \mathbb{E}\left[\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{A})\right)\right]\right| > \delta\|f\|_L\right)$$
$$\leq \exp\left\{-n^2\frac{\left(\delta-\delta_1(n)\right)^2}{c|K|^2}\right\}.$$

(3) *Let λ_1^A be the largest eigenvalue of* $\mathbf{A}$. *Then*

$$\mathbb{P}\left(\left|\lambda_1^A - \mathbb{E}[\lambda_1^A]\right| \geq \delta|K|\right) \leq \exp\left\{-\frac{\left(\delta - 8|K|/\sqrt{n}\right)^2 n}{16}\right\}.$$

This result is a direct application of Talagrand's beautiful theory and the computation of Lipschitz constants of functions of the spectral measure in terms of the entries; see [6, 45]. The original statement proves concentration around the median rather than the mean, but it is easy to go from one result to the other up to some error $\delta_0(n)$, $\delta_1(n)$. The second point is deducted from the first by approximating a general function by convex functions. It applies to Bernoulli matrices straightforwardly by taking $|K| = 1/\sqrt{p(1-p)}$.

Theorem 3.2. *Take f convex and Lipschitz with Lipschitz norm $\|f\|_L$. Then, for any $\delta > \delta_0(n) = 8\sqrt{\pi}|f|_L/np(1-p)$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{X}_n)\right) - \mathbb{E}\left[\frac{1}{n}\operatorname{Tr}\left(f(\mathbf{X}_n)\right)\right]\right| > \delta + \delta_0(n)\right)$$
$$\leq \exp\left\{-p(1-p)n^2\frac{(\delta)^2}{16|f|_L^2}\right\}.$$

Moreover, for any $\delta > \delta_0'(n) = O(1/\sqrt{p(1-p)n})$

$$\mathbb{P}\left(\left|\lambda_1 - \mathbb{E}[\lambda_1]\right| > \delta + \delta_0'(n)\right) \leq \exp\left\{-p(1-p)n\delta^2\right\}.$$

As we can see, the speed of the concentration deteriorates with p going to zero to be of order n when np is of order one. In fact, it can be shown that the worse concentration estimates for the empirical measure are of the order of exponential in n. Indeed, we have the following result due to Bordenave, Caputo, and Chafai [24] which is based on the Azuma–Hoeffding inequality and requires only the independence of the vectors of the random matrix.

Lemma 3.3. *Let $\|f\|_{TV}$ be the total variation norm:*

$$\|f\|_{TV} = \sup_{x_1 < \cdots < x_p} \sum_{i=2}^{p} \left|f(x_i) - f(x_{i-1})\right|.$$

Then, for any self-adjoint matrix $\mathbf{X}_n$ *with independent vectors* $((X_{ij},\ i \le j),\ 1 \le j \le n)$ *and eigenvalues* $(\lambda_i)_{1 \le i \le n}$ *for any function* f *with finite total variation norm so that* $E[|\frac{1}{n}\sum_{i=1}^n f(\lambda_i)|] < \infty$*, and any* $\delta > 0$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n f(\lambda_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(\lambda_i)\right]\right| \ge \delta \|f\|_{TV}\right) \le 2e^{-\frac{n\delta^2}{8}}.$$

In the general case, however, the extreme eigenvalues do not concentrate and can be very large for heavy-tailed entries [4,9].

3.2. Global fluctuations

It is a natural question to wonder how the empirical measure of the eigenvalues fluctuates and, in particular, whether the concentration result of Theorem 3.2 is on the optimal scale. In the case where p is of order one, this question was first answered by Jonsson [51] by estimating moments, and in the context of Gaussian matrices by Johansson [50] by using loop equations. The main point is that the central limit theorem does not require a renormalization by the famous $\sqrt{n}$ as for the classical central limit theorems.

Theorem 3.4. *Assume that* $p \in (0,1)$ *independent of* n*. Let* f *be a continuously differentiable function. Let* λ_i *be the eigenvalues of* $\mathbf{X}_n$*. Then*

$$\sum_{i=1}^n f(\lambda_i) - \mathbb{E}\left[\sum_{i=1}^n f(\lambda_i)\right]$$

converges in distribution towards a centered Gaussian variable with variance

$$V(f) = \frac{1}{2\pi^2}\int_{-2}^{2}\int_{-2}^{2}\left(\frac{f(x)-f(y)}{x-y}\right)^2 \frac{(4-xy)}{\sqrt{4-x^2}\sqrt{4-y^2}}\,dx\,dy.$$

The central limit theorem also holds if one recenters with respect to the limit rather than the expectation; see e.g. [56].

On the contrary, if pn goes to a constant c, we see that Theorem 3.3 gives the optimal speed and we have a "more" classical central limit theorem [7,20,59]:

Theorem 3.5. *Assume that* pn *goes to* $c \in (0,+\infty)$*. Let* f *be a* C_b^1 *function. Then*

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n f(\lambda_i) - \mathbb{E}\left[\sum_{i=1}^n f(\lambda_i)\right]\right)$$

converges in law towards a centered Gaussian variable with non-trivial variance.

Together with [46], we claim that at least for pn of order one, or in $[n^{\varepsilon}, n^{1-\varepsilon}]$, or p of order one, we have the following theorem.

Theorem 3.6. *Let f be a C_b^1 function. Then*

$$\sqrt{p}\left(\sum_{i=1}^{n} f(\lambda_i) - \mathbb{E}\left[\sum_{i=1}^{n} f(\lambda_i)\right]\right)$$

converges in law towards a centered Gaussian variable with non-trivial covariance.

This result should hold for any $p > 1/n$.

3.3. Local laws

An important breakthrough towards the understanding of local fluctuations and eigenvectors is to analyze the so-called local laws as foreseen in [41]. Namely, to estimate $\sum f(\lambda_i)$ for less smooth functions, in fact for functions on a mesoscopic scale $f(x) = g(N^{\alpha}(x - E))$ for some $\alpha \in (0, 1)$. Equivalently, one can look at $f(x) = (z - x)^{-1}$ with $z = E + i\eta$ with η of order $N^{-\alpha}$ (indeed the latter can serve to approximate conveniently the first). In this scale, it was proved that if pn goes to infinity, the mesoscopic distribution of the eigenvalues is still very close from the semi-circle distribution. Indeed, let us define the Stieltjes transform to be given by

$$G_n(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{z - \lambda_i}, \quad G_{\mu}(z) = \int \frac{1}{z - \lambda}\, d\mu(\lambda).$$

In [40, Theorems 2.8 and 2.10], the following result was proved, where ζ-high probability means a probability greater than or equal to $1 - e^{-v(\ln n)^{\zeta}}$ for some $v > 0$.

Theorem 3.7. *There are universal constants $C_1, C_2 > 0$ such that the following holds. Assume that*

$$pn \geq (\ln n)^{C_1 \xi}, \quad \xi = C_2 \ln\ln n.$$

Then, for $E \in [-3, 3]$ and $D = \{z = E + i\eta, 0 < \eta < 3\}$,

$$\bigcap_{z \in D}\left\{\left|G_n(z) - m_{\sigma}(z)\right| \leq (\ln n)^{C_2 \xi}\left(\min\left\{\frac{1}{pn\sqrt{\kappa_E + \eta}}, \frac{1}{\sqrt{pn}}\right\} + \frac{1}{n\eta}\right)\right\}$$

holds with ζ-high probability. Moreover, for $\eta > (\ln n)^{C\zeta} n^{-1}$

$$\#\{i : \lambda_i \in [E - \eta, E + \eta]\} = n\sigma([E - \eta, E + \eta])\left(1 + O(\ln n)^{C\zeta}\left(\frac{1}{n\eta^{\frac{3}{2}}} + \frac{1}{pn\eta}\right)\right)$$

with ζ-high probability.

The above theorem applies for any p such that pn goes to infinity much faster than any $\ln n$; see e.g. [4]. Below $\ln n$, the extreme eigenvalues were shown to be dictated by the largest degree in the graph [18].

A similar statement in the sparse case where pn goes to a finite constant is still open. Indeed, the fact that μ_c has a dense set of atoms and a continuous part makes the analysis a priori much more involved and the local law more difficult to conjecture. An easier heavy tail matrix model was studied in [17, 26, 35], namely the random matrices with alpha-stable independent entries. In this case, the entries follow the alpha-stable law $\mathbb{P}(|A_{ij}| \geq t) \simeq t^{-\alpha}/n$. When $\alpha < 2$, it was shown in [17, 35] that the empirical measure converges towards a limiting law μ_α which is different from the semi-circle law. One of the advantages of this model is that μ_α is absolutely continuous except possibly for a discrete set of atoms. Of course, one cannot expect the eigenvalues to be as rigid in the heavy-tailed case since this would contradict the central limit theorem (which holds as in Theorem 3.6; see [20]). Hence, in this case, large eigenvalues should be less rigid, creating large fluctuations. The following result was proved if the A_{ij} are α-stable variables in [26, 27]: for all $t \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(itA_{11})\big] = \exp\left(-\frac{1}{n}w_\alpha |t|^\alpha\right), \tag{3.1}$$

for some $0 < \alpha < 2$ and $w_\alpha = \pi/(\sin(\pi\alpha/2)\Gamma(\alpha))$. We put

$$\rho = \begin{cases} \frac{1}{2} & \text{if } \frac{8}{5} \leq \alpha < 2, \\ \frac{\alpha}{8-3\alpha} & \text{if } 1 < \alpha < \frac{8}{5}, \\ \frac{\alpha}{2+3\alpha} & \text{if } 0 < \alpha \leq 1. \end{cases} \tag{3.2}$$

Then, there exists a finite set $\mathcal{E}_\alpha \subset \mathbb{R}$ such that if $K \subset \mathbb{R}\backslash\mathcal{E}_\alpha$ is a compact set and $\delta > 0$, the following holds. There are constants $c_0, c_1 > 0$ such that for all integers $n \geq 1$, if $I \subset K$ is an interval of length $|I| \geq c_1 n^{-\rho}(\ln n)^2$, then

$$\big|N_I - n\mu_\alpha(I)\big| \leq \delta n|I|, \tag{3.3}$$

with probability at least $1 - 2\exp(-c_0 n\delta^2|I|^2)$. The fact that our result might not be true on a finite set of values should only be technical. This result was improved in [2, Theorems 3.4 and 3.5] in order to tackle I of size $n^{-\omega(\alpha)}$ with $\omega(\alpha) > 1/2$ (and $\Re(z)$ small enough when $\alpha < 1$). Such an optimal scale is important in the study of the local fluctuations of the spectrum.

In both light and heavy tails, the main point is to estimate the Stieltjes transform $G_n(z) = \frac{1}{n}\sum_{i=1}^n (z - \lambda_i)^{-1}$ for z going to the real axis: $z = E + i\eta$ with η of order nearly as good as n^{-1} for light tails, $n^{-\rho}$ for heavy tails. This is done by showing that G_n is characterized approximately by a closed set of equations. In the case of lights tails, one has simply a quadratic equation for G_n and needs to show that the error terms remain small as z approaches the real line. In the heavy-tailed case, the equations are much more complicated, see (2.7) and (2.8), and therefore more difficult to handle. Similar questions are completely open for other heavy-tailed matrices, including Bernoulli matrices with pn of order one.

3.4. Local fluctuations

When the average degree pn is large, one expects the eigenvalues to behave exactly as the eigenvalues of a symmetric matrix with independent Gaussian entries (so-called GOE matrices). The advantage of Gaussian matrices is that they are an integrable model of random matrices in the sense that many of their properties can be exactly computed. To start with, the joint distribution of its eigenvalues $(\lambda_i^G)_{1\le i\le n}$ is explicit:

$$d\mathbb{P}(\lambda^G) = \frac{1}{Z}\Delta(\lambda)e^{-\frac{n}{4}\sum(\lambda_i^G)^2}\prod d\lambda_i^G, \tag{3.4}$$

where

$$\Delta(\lambda) = \prod_{i<j}|\lambda_i - \lambda_j|$$

is the Vandermonde determinant. In particular, this formula does not depend on the eigenvectors. Based on this formula, Tracy and Widom could study the local fluctuations of the spectrum $(\lambda_i^G)_{1\le i\le n}$ [66, 67] and they proved that

$$\lim_{n\to\infty}\mathbb{P}\big(n^{2/3}(\lambda_1^G - 2) \le s\big) = F_1(s),$$

where F_1 is the distribution function of the Tracy–Widom law. For the eigenvalues in the bulk, it was proved [55] that, for all smooth compactly supported function,

$$\mathcal{E}_{\mathbf{G}_n}(O, E) = \mathbb{E}\big[O\big(n(\lambda_i^G - E), \dots, n(\lambda_{i+p}^G - E)\big)\big]$$

converges as n goes to infinity and the limit is described in terms of Pfaffian distributions.

The universality in the bulk was obtained after a series of works including notably [41, 42, 62] and [39, Theorem 2.5] (for $\phi \ge 2/3$) and improved in [48] (for $\phi > 0$) to finally get the following theorem.

Theorem 3.8 (Bulk universality). *Suppose that $pn > n^\phi$ with $\phi > 0$. There exists b_n going to zero so that for all smooth compactly supported function O, any $E \in (-2, 2)$,*

$$\lim_{n\to\infty}\int_{E-b_n}^{E+b_n}\frac{dE'}{2b_n}\big(\mathcal{E}_{\mathbf{G}_n}(O, E') - \mathcal{E}_{\mathbf{B}_n}(O, E')\big) = 0.$$

Moreover, the universality at the edge was obtained in [39, Theorem 2.7]; see also [60].

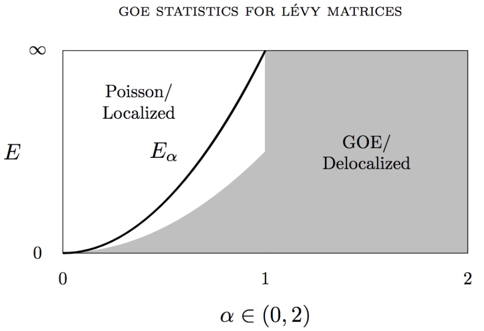


Figure 4.

Theorem 3.9 (Edge universality). *Suppose that $pn > n^\phi$, $\phi > 2/3$. Then there exists $\delta > 0$ such that*

$$\mathbb{P}\big(n^{2/3}(\lambda_2^B - 2) \le s\big) = \mathbb{P}\big(n^{2/3}(\lambda_2^G - 2) \le s + O(n^{-\delta})\big) + O(n^{-\delta}).$$

This statement was generalized to $pn > n^{1/3}$ but the largest eigenvalue then needs to be shifted by a deterministic drift of order $1/pn$ [53]. Beyond this threshold, the fluctuations of the second largest eigenvalue starts to be Gaussian.

When pn decreases below $1/3$, it was proved that universality stops to hold and fluctuations of the largest eigenvalue start to be Gaussian. The precise transition between Tracy–Widom law and Gaussian fluctuations when p is of order $n^{-2/3}$ was described in [49]. When $n^{o(1)} \ll pn \ll n^{1/3}$, the papers [47, 49] show that the fluctuations of the extreme eigenvalues are Gaussian, even if they stick to the bulk. In the case where $pn \ll \ln n$, Theorem 2.3 asserts that the eigenvalues go away from the bulk, at distance of order $\sqrt{\ln n}$. The corresponding eigenvectors are localized close to the vertices with a high degree. In an even more recent preprint [5], the same authors show that these eigenvalues follow a Poisson point process. Such questions are open for Bernoulli random matrices with pn of order $c \in (0, +\infty)$ and eigenvalues in the bulk. Indeed, as we have seen, the limiting density is a mixture of atoms and continuous density and it is not yet clear how to zoom in the spectrum in such a situation. However, such questions could be analyzed for Lévy matrices with α-stable entries in the regime where local law can be obtained on the optimal scale $n^{-1/2}$ [2]. Figure 4 depicts the expected regimes. In fact, one expects the following transition to occur (see [63]).

- If $\alpha \in [1, 2]$, all eigenvectors corresponding to finite eigenvalues are completely delocalized. Further, for any $E \in \mathbb{R}$, the local statistics of the eigenvalues near E converge to those of the GOE as N goes to infinity.
- If $\alpha \in (0, 1)$, there exists a mobility edge E_α such that for $|E| < E_\alpha$ the local statistics of the eigenvalues near E converge to those of the GOE as N goes to

infinity. But if $|E| > E_\alpha$, the local statistics of the eigenvalues near E converge to those of a Poisson point process and all eigenvectors in this region are localized. The fact that local statistics are given by those of Gaussian matrices for $\alpha \in (1,2)$ or $\alpha \in (0,1)$ and E small enough, except for E in some finite set, was proved in [2, Theorems 2.4 and 2.5].

3.5. Properties of the eigenvectors

The properties of the eigenvectors are intimately related with local laws. Indeed, by definition of the eigenvectors, if v is an eigenvector of the symmetric matrix $\mathbf{X}_n$ for the eigenvalue E and we set $\langle v, e_i\rangle = v_i$, then X_1 is the first column vector of $\mathbf{X}_n$ while $\mathbf{X}_n^{(1)}$ is the $(n-1)\times(n-1)$ principal minor of $\mathbf{X}_n$ obtained by removing the column and row vector given by X_1 and X_1^T:

$$v_1^2 = \left(1 + \left\langle X_1, (E - \mathbf{X}_n^{(1)})^{-2} X_1\right\rangle\right)^{-1},$$

where, at least in the dense cases $\langle X_1, (E - \mathbf{X}_n^{(1)})^{-2} X_1\rangle$ is close to $\frac{1}{n}\operatorname{Tr}(E - \mathbf{X}_n)^{-2}$, and so is governed by the local law. In [40, Theorem 2.16], the following theorem was proved.

Theorem 3.10 (Complete delocalization of eigenvectors). *Assume the hypotheses of Theorem 3.7 with $pn > n^\phi$ with $\phi > 0$. Let v_i be the eigenvectors of $\mathbf{B}_n$ for the eigenvalues $\lambda_n \le \lambda_{n-1} \cdots \le \lambda_1$. Then*

$$\max_{i \le n} \|v_i\|_\infty \le \frac{(\ln n)^{4\xi}}{\sqrt{n}}$$

with ξ-high probability.

This result was extended to q going to infinity logarithmically only more recently [3]. We roughly state their result:

- (Semilocalized phase) Assume that $C\sqrt{\ln n}\ln\ln n \le \sqrt{pn} \le 3\ln n$ and let w be a normalized eigenvector of $\mathbf{B}_n$ with non-trivial eigenvalue $E \ge 2 + C\zeta^{1/2}$. We let $\Lambda(\alpha) = \alpha/\sqrt{\alpha - 1}$ and $\alpha_x = \sum_y \mathbf{B}_{xy}/pn$. We let $W_{E,\delta}$ be the set of vertices such that $\Lambda(\alpha_x) \in [E-\delta, E+\delta]$. Then for each $x \in W_{E,\delta}$, there exists a normalized vector $v(x)$ supported in a ball around x and radius $c\sqrt{\ln n}$, such that the support of $v(cx)$ and $v(y)$ is distinct if $x \ne y$ and

$$\sum_{x \in W_{E,\delta}} \left\langle v(x), w\right\rangle^2 \ge 1 - C\left(\sqrt{\ln n}\,pn \ln pn + \sqrt{\ln n}\,pn \frac{1}{E-2}\right)^2 \delta^{-2}.$$

Moreover,

$$\sum_{y \in B_r(x)} \left(v(x)\right)_y^2 \le \frac{1}{(\alpha_x - 1)^{r+1}}.$$

- (Delocalized phase) For any $\nu > 0$ and $\kappa > 0$, there exists a constant $C > 0$ such that for $pn \in [C\sqrt{\ln n}, (\ln n)^{3/2}]$, if w is a normalized eigenvector for $\mathbf{B}_n$ with eigenvalue $E \in [-2+\kappa, -\kappa] \cup [\kappa, 2-\kappa]$,
$$\|w\|_\infty^2 \le n^{-1+\kappa}$$
with probability greater than $1 - n^{-\nu}$.

This question is completely open for Bernoulli random matrices with pn of order one but the understanding of Lévy matrices is again more complete. Based on [2, 26, 27], we can assert that Tarquini, Biroli, and Tarzia's conjecture [63] is partly proved. Indeed the complete delocalization is proved for $\alpha \in (1,2)$ and $\alpha \in (0,1)$ and small enough eigenvalues. A sort of localization for $\alpha \in (0,1)$ for large enough eigenvalue was derived in [26], and was shown to be not true for small enough eigenvalues in [27]: the transition and the value of the mobility edge is still an open question. In fact, even in the case where the eigenvalue statistics belong to the universality class of Gaussian matrices, the fine properties of the eigenvectors of Lévy matrices differ [1]. Let us also mention [57] which shows under quite general assumptions that eigenvectors are somehow uniformly delocalized in the sense that any subset of at least eight coordinates carries a non-negligible part of the mass of an eigenvector.

4. Rare events

It is sometimes important to estimate the probability of rare events, such as the probability that the extreme eigenvalues take unlikely values or the empirical measure of the eigenvalues shows an unlikely profile, and what kind of optimal strategy can lead to such deviations from the expected behavior. In the case of Gaussian symmetric matrices, the joint density of the eigenvalues is known (3.4). One finds by sort of Laplace's principle [15, 16] the large deviations for the empirical measure and the largest eigenvalue.

Theorem 4.1. *Let $\lambda_n^G \le \lambda_2^G \cdots \le \lambda_1^G$ be the eigenvalues of a GOE matrix. Then, the following holds.*

- *Let $E(\mu) = \frac{1}{2}\iint(\frac{x^2}{4} + \frac{y^2}{4} - \ln|x-y|)\,d\mu(x)\,d\mu(y)$ and set $\mathcal{E}(\mu) = E - \inf E$. Then $\mathcal{E}$ is a good rate function and the distribution of the empirical measure of the eigenvalues $\hat{\mu}_n = \frac{1}{n}\sum \delta_{\lambda_i^G}$ satisfies a large deviations principle (LDP) with speed n^2 with rate function $\mathcal{I}$, that is for every closed set F*
$$\limsup_{n\to\infty} \frac{1}{n^2} \ln \mathbb{P}(\hat{\mu}_n \in F) \le -\inf_F \mathcal{E},$$
whereas for any open set O
$$\limsup_{n\to\infty} \frac{1}{n^2} \ln \mathbb{P}(\hat{\mu}_n \in O) \ge -\inf_O \mathcal{E}.$$

- *Let $I_G(x) = \frac{1}{2}\int_2^x \sqrt{4-y^2}dy$ for $x \geq 2$ and $I_G(x) = +\infty$ for $x < 2$. Then I is a good rate function and the distribution of λ_1^G satisfies an LDP with speed n and good rate function I_G.*

In this case, deviations of the spectrum can be created independently from the eigenvectors which stay uniformly distributed. On the other hand, if the entries have sharp exponential decay, large deviations can be created by large entries. Assume that for some $\alpha \in (0, 2)$, there exists $a > 0$ so that for all i, j

$$\lim_{t\to\infty} 2^{-1_{i=j}} t^{-\alpha} \ln \mathbb{P}\big(|\sqrt{n}X_{ij}| \geq t\big) = -a.$$

Theorem 4.2. • *The law of the empirical measure satisfies an LDP in the speed $n^{1+\frac{\alpha}{2}}$ and good rate function which is infinite unless $\mu = \sigma \boxplus \nu$ and then equals $a \int |x|^\alpha d\nu(x)$* [22].

- *The law of the largest eigenvalue satisfies an LDP with rate $n^{\frac{\alpha}{2}}$ and GRF proportional to $(\int (x-y)^{-1} d\sigma(y))^{-\alpha}$* [10].

However, the situation is much less understood for Bernoulli matrices and again the sparse and the dense regimes lead to very different results and techniques. We discuss these questions hereafter.

4.1. Large deviations for the extreme eigenvalues

Let us first consider the dense case. In [12, 43], we considered the large deviations for the largest eigenvalue of Wigner matrices and showed that if the entries are Rademacher, then the same large deviation principle holds, whereas in general there is a transition between deviations close to two where the rate function is the Gaussian one whereas for large deviations towards large enough values the rate function is more of a heavy tail type. In a work in progress with F. Augeri, R. Ducatez and J. Husson, we prove the following theorem.

Theorem 4.3. • *Assume that $p = 1/2$. Then the law of λ_1^X satisfies an LDP in the scale n and with the same rate function I_G as for the GOE matrix.*

- *Assume that $p \in (0, 1/2)$. Then for x close enough to 2, the probability that λ_1^X is close to x is the same as in the Gaussian case. But for x large enough,*

$$\limsup_{\delta\downarrow 0} \limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\big(|\lambda_1^X - x| < \delta\big) = -I_p(x),$$

where $I_p(x) < I_G(x)$.

The case $p \in (1/2, 1)$ is under investigation. In fact, analyzing the large deviation requires to understand good strategies to create the deviations. For $p = 1/2$, it is shown that an optimal strategy is to tilt the law of the entries in order to change their expectation so that the matrix looks like a rank one deformation of Bernoulli matrix with a delocalized deformation. The eigenvectors also stay delocalized through this deformation. When $p < 1/2$ and x is large, it turns out that the optimal strategy is to create fully connected components of size $\sqrt{n}$. For $p > 1/2$, the picture is less clear and we suspect that vertices with high degree are optimal ways to create large eigenvalues.

Let us now consider the sparse case following [21]: in this case we already saw that large eigenvalues are created by vertices with large degree, namely with row or column vectors with many entries equal to one.

Theorem 4.4. *Let* $L_p = \frac{\ln n}{\ln\ln n - \ln(np)}$ *and assume that*

$$\ln(1/np) \ll \ln n \quad \text{and} \quad np \ll \sqrt{\ln n/\ln\ln n}.$$

Let λ_2 *be the second largest eigenvalue of* $\mathbf{B}_n$. *Then for any* $\delta \geq 0$,

$$\lim_{n\to\infty} \frac{-\ln P\big(\lambda_2 \geq (1+\delta)\sqrt{L_p}\big)}{\ln n} = 2\delta + \delta^2,$$

whereas

$$\lim_{n\to\infty} \frac{-\ln P\big(\lambda_2 \leq (1-\delta)\sqrt{L_p}\big)}{\ln n} = 2\delta - \delta^2.$$

4.2. Large deviations for the empirical measure

In [23, Theorem 1.6], a large deviation for the empirical measure of the eigenvalue in the sparse case was derived: we do not make precise the rate function as it is obtained by contraction from the large deviation for the empirical neighborhood distribution.

Theorem 4.5. *Assume that* pn *is fixed. Then the law of* $\hat{\mu}_n$ *satisfies a large deviation principle with speed* n.

This question is still open when $pn \gg 1$. When p is of order one, we should expect to have a large deviation with speed n^2 according to the concentration of measure, but the rate function should not be equal to the Gaussian one even when $p = 1/2$ because the Dirac at the origin should have rate function bounded above by $\ln p$ (whereas it is infinite in the Gaussian case).

4.3. Large deviations for triangle counts

The traces of Bernoulli matrices have a combinatorial interpretation. For instance, $\mathrm{Tr}(\mathbf{B}_n^3)$ is the number $T_{n,p}$ of triangles in the Erdős–Rényi graph. Observe that its expectation is of order p^3n^3. In the well-known paper [34, Theorem 4.1], the following theorem was proved.

Theorem 4.6. *Let*

$$I_p(f) = \sup_{\phi}\left\{\int_0^1\int_0^1 f(x,y)\phi(x,y)\,dx\,dy - \frac{1}{2}\iint \ln\big(pe^{2\phi(x,y)} + (1-p)\big)\,dx\,dy\right\}$$

and set $\varphi(p,t) = \in \{I_p(f), \int f(x,y)f(y,v)f(v,x)dxdydv \geq 6t\}$. *Then for each* $p \in (0,1)$,

$$\lim_{n\to\infty} \frac{1}{n^2} \ln \mathbb{P}(T_{n,p} \geq tn^3) = -\varphi(p,t).$$

This result extends to any moment $\mathrm{Tr}(\mathbf{B}_n^k)$. However, observe that it does not tell us about deviations of the empirical measure since $x \to x^k$ is unbounded so that deviations of the extreme eigenvalues matter. It is natural to wonder what happens as well when p goes to zero. This question was attacked in [33,36,37], but we state here [11, Proposition 1.19]

Theorem 4.7. *Let p go to zero with n so that* $(\ln n)^4 \ll np^2$. *Set* $v_n = n^2p^2\ln(1/p)$. *Then for* $t \geq 1$,

$$\lim_{n\to\infty} \frac{1}{v_n} \ln \mathbb{P}\big(\mathrm{Tr}(\mathbf{B}_n^d) \geq tn^d p^d\big) = -\Phi(t),$$

where $\Phi(t) = \frac{1}{2}(t-1)^{2/d}$ *if* $n^{-1} \ll p \ll n^{-1/2}$, *but* $\Phi(t) = \min\{\theta_t, \frac{1}{2}(t-1)^{2/d}\}$ *if* $p \gg n^{-1/2}$ *and* θ_t *is the solution of* $P_{C_d}(\theta_t) = t$, *where* P_{C_d} *is the independence polynomial of the d-cycle.*

4.4. The singularity probability

A well-known problem has been to estimate the probability that a matrix $\widetilde{\mathbf{B}}_n$ with all independent Bernoulli entries (hence not self-adjoint) is singular. In a breakthrough paper, Tikhomirov [64] (see also [54]) could exactly estimate it, by showing that the best strategy to achieve singularity is to have a zero column or row vector.

Theorem 4.8. *There exists a finite constant C such that if* $C\ln n/n \leq p \leq \frac{1}{2}$,

$$\mathbb{P}(\widetilde{\mathbf{B}}_n \textit{ is singular}) = \big(2 + o_n(1)\big)(1-p)^n n.$$

Such an optimal estimate is not yet known for the symmetric Bernoulli matrix $\mathbf{B}_n$ (even though it is conjectured) but the paper [32] proves that the probability that it is singular is bounded above by $e^{-O(\sqrt{n})}$. This was improved in an exponential upper bound in [31].

5. Open problems

(1) Local law for Bernoulli matrices when pn is of order one. This could be at best on the scale $\sqrt{n}$ but is tricky even to state because of the atoms of the limit law.

(2) Localization/delocalization of the eigenvectors of Bernoulli matrices for pn of order one (one would conjecture that Dirac masses yield localization but the continuous part yields delocalization, however the right criteria to express this remains to be given). Find a critical c^* such that for $np > c^*$ there exists delocalized vectors with connected support with high probability.

(3) Large deviations for the empirical measure of the eigenvalues of Bernoulli matrices (all p so that $pn \gg 1$). Even when $p = 1/2$, one does not expect to retrieve the Gaussian rate function since the entropy should be finite at δ_0 (as can be seen by requiring all entries to be equal).

(4) Precise estimate on the singularity probability in the symmetric case.

(5) In comparison, d-regular graphs which are picked uniformly at random are conjectured to be in the universality class of Gaussian random matrices for all $d \geq 3$. This was proved for d going to infinity fast enough [13, 14], and recently Huang and Yau could get the local law and the delocalization of the eigenvectors up to $d = 3$.

Acknowledgments. I thank Charles Bordenave, Christophe Garban, Jiaoyang Huang, Antti Knowles, Justin Salez, and Ofer Zeitouni for pointing out recent developments of the subject and for their very helpful comments on preliminary versions of this survey.

Funding. This work was supported in part by ERC-2019-ADG Project 884584-LDRAM.

References

[1] A. Aggarwal, P. Lopatto, and J. Marcinek, Eigenvector statistics of Lévy matrices. *Ann. Probab.* **49** (2021), no. 4, 1778–1846 Zbl 1467.60003 MR 4260468

[2] A. Aggarwal, P. Lopatto, and H.-T. Yau, GOE statistics for Lévy matrices. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 11, 3707–3800 Zbl 07445595 MR 4310816

[3] J. Alt, R. Ducatez, and A. Knowles, Delocalization transition for critical Erdős–Rényi graphs. *Comm. Math. Phys.* **388** (2021), no. 1, 507–579 Zbl 1477.15029 MR 4328063

[4] J. Alt, R. Ducatez, and A. Knowles, Extremal eigenvalues of critical Erdős–Rényi graphs. *Ann. Probab.* **49** (2021), no. 3, 1347–1401 Zbl 1467.05236 MR 4255147

[5] J. Alt, R. Ducatez, and A. Knowles, Poisson statistics and localization at the spectral edge of sparse Erdős–Rényi graphs. 2021, arXiv:2106.12519

[6] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*. Cambridge Stud. Adv. Math. 118, Cambridge University Press, Cambridge, 2010 Zbl 1184.15023 MR 2760897

[7] G. W. Anderson and O. Zeitouni, A CLT for regularized sample covariance matrices. *Ann. Statist.* **36** (2008), no. 6, 2553–2576 Zbl 1360.60047 MR 2485007

[8] A. Arras and C. Bordenave, Existence of absolutely continuous spectrum for Galton–Watson random trees. 2021, arXiv:2105.10177

[9] A. Auffinger, G. Ben Arous, and S. Péché, Poisson convergence for the largest eigenvalues of heavy tailed random matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** (2009), no. 3, 589–610 Zbl 1177.15037 MR 2548495

[10] F. Augeri, Large deviations principle for the largest eigenvalue of Wigner matrices without Gaussian tails. *Electron. J. Probab.* **21** (2016), Paper No. 32 Zbl 1338.60010 MR 3492936

[11] F. Augeri, Nonlinear large deviation bounds with applications to Wigner matrices and sparse Erdős–Rényi graphs. *Ann. Probab.* **48** (2020), no. 5, 2404–2448 Zbl 1456.60063 MR 4152647

[12] F. Augeri, A. Guionnet, and J. Husson, Large deviations for the largest eigenvalue of sub-Gaussian matrices. *Comm. Math. Phys.* **383** (2021), no. 2, 997–1050 Zbl 1479.60011 MR 4239836

[13] R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau, Bulk eigenvalue statistics for random regular graphs. *Ann. Probab.* **45** (2017), no. 6A, 3626–3663 Zbl 1379.05098 MR 3729611

[14] R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau, Edge rigidity and universality of random regular graphs of intermediate degree. *Geom. Funct. Anal.* **30** (2020), no. 3, 693–769 Zbl 1453.05117 MR 4135670

[15] G. Ben Arous, A. Dembo, and A. Guionnet, Aging of spherical spin glasses. *Probab. Theory Related Fields* **120** (2001), no. 1, 1–67 Zbl 0993.60055 MR 1856194

[16] G. Ben Arous and A. Guionnet, Large deviations for Wigner's law and Voiculescu's non-commutative entropy. *Probab. Theory Related Fields* **108** (1997), no. 4, 517–542 Zbl 0954.60029 MR 1465640

[17] G. Ben Arous and A. Guionnet, The spectrum of heavy tailed random matrices. *Comm. Math. Phys.* **278** (2008), no. 3, 715–751 Zbl 1157.60005 MR 2373441

[18] F. Benaych-Georges, C. Bordenave, and A. Knowles, Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs. *Ann. Probab.* **47** (2019), no. 3, 1653–1676 Zbl 1447.60017 MR 3945756

[19] F. Benaych-Georges and A. Guionnet, Central limit theorem for eigenvectors of heavy tailed matrices. *Electron. J. Probab.* **19** (2014), Paper No. 54 Zbl 1293.15021 MR 3227063

[20] F. Benaych-Georges, A. Guionnet, and C. Male, Central limit theorems for linear statistics of heavy tailed random matrices. *Comm. Math. Phys.* **329** (2014), no. 2, 641–686 Zbl 1294.60039 MR 3210147

[21] B. B. Bhattacharya, S. Bhattacharya, and S. Ganguly, Spectral edge in sparse random graphs: Upper and lower tail large deviations. *Ann. Probab.* **49** (2021), no. 4, 1847–1885 Zbl 1467.05237 MR 4260469

[22] C. Bordenave and P. Caputo, A large deviation principle for Wigner matrices without Gaussian tails. *Ann. Probab.* **42** (2014), no. 6, 2454–2496 Zbl 1330.60012 MR 3265172

[23] C. Bordenave and P. Caputo, Large deviations of empirical neighborhood distribution in sparse random graphs. *Probab. Theory Related Fields* **163** (2015), no. 1-2, 149–222 Zbl 1327.60067 MR 3405616

[24] C. Bordenave, P. Caputo, and D. Chafaï, Spectrum of large random reversible Markov chains: Heavy-tailed weights on the complete graph. *Ann. Probab.* **39** (2011), no. 4, 1544–1590 Zbl 1245.60008 MR 2857250

[25] C. Bordenave, P. Caputo, and D. Chafaï, Spectrum of non-Hermitian heavy tailed random matrices. *Comm. Math. Phys.* **307** (2011), no. 2, 513–560 Zbl 1235.60008 MR 2837123

[26] C. Bordenave and A. Guionnet, Localization and delocalization of eigenvectors for heavy-tailed random matrices. *Probab. Theory Related Fields* **157** (2013), no. 3-4, 885–953 Zbl 1296.15019 MR 3129806

[27] C. Bordenave and A. Guionnet, Delocalization at small energy for heavy-tailed random matrices. *Comm. Math. Phys.* **354** (2017), no. 1, 115–159 Zbl 1388.60022 MR 3656514

[28] C. Bordenave and M. Lelarge, Resolvent of large random graphs. *Random Structures Algorithms* **37** (2010), no. 3, 332–352 Zbl 1209.05222 MR 2724665

[29] C. Bordenave, M. Lelarge, and J. Salez, The rank of diluted random graphs. *Ann. Probab.* **39** (2011), no. 3, 1097–1121 Zbl 1298.05283 MR 2789584

[30] C. Bordenave, A. Sen, and B. Virág, Mean quantum percolation. *J. Eur. Math. Soc. (JEMS)* **19** (2017), no. 12, 3679–3707 Zbl 1385.60057 MR 3730511

[31] M. Campos, M. Jenssen, M. Michelen, and J. Sahasrabudhe, The singularity probability of a random symmetric matrix is exponentially small. 2021, arXiv:2105.11384

[32] M. Campos, L. Mattos, R. Morris, and N. Morrison, On the singularity of random symmetric matrices. *Duke Math. J.* **170** (2021), no. 5, 881–907 Zbl 1467.60005 MR 4255046

[33] S. Chatterjee and A. Dembo, Nonlinear large deviations. *Adv. Math.* **299** (2016), 396–450 Zbl 1356.60045 MR 3519474

[34] S. Chatterjee and S. R. S. Varadhan, The large deviation principle for the Erdős–Rényi random graph. *European J. Combin.* **32** (2011), no. 7, 1000–1017 Zbl 1230.05259 MR 2825532

[35] P. Cizeau and J. P. Bouchaud, Theory of Lévy matrices. *Phys. Rev. E* **50** (1994), no. 3, 1810–1822

[36] N. Cook and A. Dembo, Large deviations of subgraph counts for sparse Erdős–Rényi graphs. *Adv. Math.* **373** (2020), 107289, 53 Zbl 1473.60058 MR 4130460

[37] R. Eldan, Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.* **28** (2018), no. 6, 1548–1596 Zbl 1428.60045 MR 3881829

[38] N. Enriquez and L. Ménard, Spectra of large diluted but bushy random graphs. *Random Structures Algorithms* **49** (2016), no. 1, 160–184 Zbl 1348.05189 MR 3521277

[39] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, Spectral statistics of Erdős–Rényi graphs II: Eigenvalue spacing and the extreme eigenvalues. *Comm. Math. Phys.* **314** (2012), no. 3, 587–640 Zbl 1251.05162 MR 2964770

[40] L. Erdős, A. Knowles, H.-T. Yau, and J. Yin, The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** (2013), Paper No. 59 Zbl 1373.15053 MR 3068390

[41] L. Erdős, B. Schlein, and H.-T. Yau, Wegner estimate and level repulsion for Wigner random matrices. *Int. Math. Res. Not. IMRN* **2010** (2010), no. 3, 436–479 Zbl 1204.15043 MR 2587574

[42] L. Erdős, B. Schlein, and H.-T. Yau, Universality of random matrices and local relaxation flow. *Invent. Math.* **185** (2011), no. 1, 75–119 Zbl 1225.15033 MR 2810797

[43] A. Guionnet and J. Husson, Large deviations for the largest eigenvalue of Rademacher matrices. *Ann. Probab.* **48** (2020), no. 3, 1436–1465 Zbl 1444.60021 MR 4112720

[44] A. Guionnet and O. Zeitouni, Concentration of the spectral measure for large matrices. *Electron. Comm. Probab.* **5** (2000), 119–136 Zbl 0969.15010 MR 1781846

[45] A. Guionnet and O. Zeitouni, Large deviations asymptotics for spherical integrals. *J. Funct. Anal.* **188** (2002), no. 2, 461–515 Zbl 1002.60021 MR 1883414

[46] Y. He, Bulk eigenvalue fluctuations of sparse random matrices. *Ann. Appl. Probab.* **30** (2020), no. 6, 2846–2879 Zbl 1482.15037 MR 4187130

[47] Y. He and A. Knowles, Fluctuations of extreme eigenvalues of sparse Erdős–Rényi graphs. *Probab. Theory Related Fields* **180** (2021), no. 3-4, 985–1056 Zbl 1468.05155 MR 4288336

[48] J. Huang, B. Landon, and H.-T. Yau, Bulk universality of sparse random matrices. *J. Math. Phys.* **56** (2015), no. 12, Paper No. 123301 Zbl 1329.05262 MR 3429490

[49] J. Huang, B. Landon, and H.-T. Yau, Transition from Tracy-Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős–Rényi graphs. *Ann. Probab.* **48** (2020), no. 2, 916–962 Zbl 1440.05181 MR 4089498

[50] K. Johansson, On fluctuations of eigenvalues of random Hermitian matrices. *Duke Math. J.* **91** (1998), no. 1, 151–204 Zbl 1039.82504 MR 1487983

[51] D. Jonsson, Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* **12** (1982), no. 1, 1–38 Zbl 0491.62021 MR 650926

[52] O. Khorunzhy, M. Shcherbina, and V. Vengerovsky, Eigenvalue distribution of large weighted random graphs. *J. Math. Phys.* **45** (2004), no. 4, 1648–1672 Zbl 1068.05062 MR 2043849

[53] J. O. Lee and K. Schnelli, Local law and Tracy–Widom limit for sparse random matrices. *Probab. Theory Related Fields* **171** (2018), no. 1-2, 543–616 Zbl 1429.60012 MR 3800840

[54] A. E. Litvak and K. E. Tikhomirov, Singularity of sparse Bernoulli matrices. *Duke Math. J.* **171** (2022), no. 5, 1135–1233 Zbl 07513357 MR 4402560

[55] M. L. Mehta, *Random Matrices*. 3rd edn., Pure Appl. Math. (Amsterdam) 142, Elsevier/Academic Press, Amsterdam, 2004 Zbl 1107.15019 MR 2129906

[56] L. Pastur and M. Shcherbina, *Eigenvalue Distribution of Large Random Matrices*. Math. Surveys Monogr. 171, American Mathematical Society, Providence, RI, 2011 Zbl 1244.15002 MR 2808038

[57] M. Rudelson and R. Vershynin, No-gaps delocalization for general random matrices. *Geom. Funct. Anal.* **26** (2016), no. 6, 1716–1776 Zbl 1375.60027 MR 3579707

[58] J. Salez, Every totally real algebraic integer is a tree eigenvalue. *J. Combin. Theory Ser. B* **111** (2015), 249–256 Zbl 1307.05150 MR 3315609

[59] M. Shcherbina and B. Tirozzi, Central limit theorem for fluctuations of linear eigenvalue statistics of large random graphs. *J. Math. Phys.* **51** (2010), no. 2, 023523, 20 Zbl 1309.05163 MR 2605074

[60] A. Soshnikov, Universality at the edge of the spectrum in Wigner random matrices. *Comm. Math. Phys.* **207** (1999), no. 3, 697–733 Zbl 1062.82502 MR 1727234

[61] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces. *Publ. Math. Inst. Hautes Études Sci.* **81** (1995), 73–205 Zbl 0864.60013 MR 1361756

[62] T. Tao and V. Vu, The Wigner–Dyson–Mehta bulk universality conjecture for Wigner matrices. *Electron. J. Probab.* **16** (2011), no. 77, 2104–2121 Zbl 1245.15041 MR 2851058

[63] E. Tarquini, G. Biroli, and M. Tarzia, Level statistics and localization transitions of Lévy matrices. *Phys. Rev. Lett.* **116** (2016), no. 1, Paper No. 010601 Zbl 1356.15018 MR 3555687

[64] K. Tikhomirov, Singularity of random Bernoulli matrices. *Ann. of Math. (2)* **191** (2020), no. 2, 593–634 Zbl 1458.15023 MR 4076632

[65] K. Tikhomirov and P. Youssef, Outliers in spectrum of sparse Wigner matrices. *Random Structures Algorithms* **58** (2021), no. 3, 517–605 MR 4234995

[66] C. A. Tracy and H. Widom, Introduction to random matrices. In *Geometric and Quantum Aspects of Integrable Systems (Scheveningen, 1992)*, pp. 103–130, Lecture Notes in Phys. 424, Springer, Berlin, 1993 Zbl 0791.15017 MR 1253763

[67] C. A. Tracy and H. Widom, On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** (1996), no. 3, 727–754 Zbl 0851.60101 MR 1385083

[68] V. H. Vu, Recent progress in combinatorial random matrix theory. *Probab. Surv.* **18** (2021), 179–200 Zbl 07367835 MR 4260513

[69] E. P. Wigner, On the distribution of the roots of certain symmetric matrices. *Ann. of Math. (2)* **67** (1958), 325–327 Zbl 0085.13203 MR 95527

[70] I. Zakharevich, A generalization of Wigner's law. *Comm. Math. Phys.* **268** (2006), no. 2, 403–414 Zbl 1147.82334 MR 2259200

Alice Guionnet
Ecole Normale Supérieure de Lyon, Université de Lyon, CNRS, 69342 Lyon, France;
alice.guionnet@ens-lyon.fr

DOI 10.4171/8ECM/30

An introduction to the mathematics of deep learning

Gitta Kutyniok

Abstract. Despite the outstanding success of deep neural networks in real-world applications, ranging from science to public life, most of the related research is empirically driven and a comprehensive mathematical foundation is still missing. At the same time, these methods have already shown their impressive potential in mathematical research areas such as imaging sciences, inverse problems, or numerical analysis of partial differential equations, sometimes by far outperforming classical mathematical approaches for particular problem classes.

The goal of this paper, which is based on a plenary lecture at the 8th European Congress of Mathematics in 2021, is to first provide an introduction into this new vibrant research area. We will then showcase some recent advances in two directions, namely the development of a mathematical foundation of deep learning and the introduction of novel deep learning-based approaches to solve inverse problems and partial differential equations.

1. Introduction

During the last years, deep neural networks have been key to spectacular successes in diverse applications such as autonomous driving, medical diagnosis, speech recognition, and telecommunication. It is by now evident that deep learning and, in general, artificial intelligence, will change in the future both public life and science in an unprecedented way; and this future has already begun. As an example in the sciences, Google's DeepMind's AlphaFold 2 has recently led to a breakthrough in highly accurate prediction of protein structures [20].

A strongly increasing impact on mathematics itself can also be witnessed. The field of inverse problems, predominantly in imaging science, was one of the first areas in mathematics, which embraced these novel methodologies. This area, which focusses on problems such as denoising, inpainting, super-resolution or computed tomography, is particularly accessible to learning methods, since there does not exist a precise model for what an image is. Almost all novel contributions, which improved

2020 Mathematics Subject Classification. Primary 68T07; Secondary 41A25, 42C15, 35C20, 65D18.
Keywords. Approximation theory, deep learning, inverse problems, parametric partial different equations.

the state of the art, employ such techniques. This, by now, already led to a change in paradigm in this field. We will discuss further details in Section 4.1.

Besides inverse problems, another large area of mathematical problem settings are partial differential equations. One can, in general, imagine using learning methods in solvers. It is, however, not immediately evident what the advantage of such an approach would be. The ability of deep neural networks to beat the curse of dimensionality then led to a change of paradigm in this area as well, and research at the intersection of numerical analysis of partial differential equations and deep learning accelerated since about 2017. Several milestones could already be celebrated as will be presented in Section 4.2.

As bright as the deep learning future appears to be, one has to also be aware of various major obstacles still waiting to be overcome. This was very prominently stated during the plenary talk at the main conference in artificial intelligence and machine learning, namely NIPS (today called NeurIPS) in 2017 on behalf of the Test-of-Time Award, in which Ali Rahimi from Google claimed that "Machine learning has become a form of alchemy". And, indeed, as we will discuss later, a fundamental understanding of deep learning algorithms is still missing, posing a great—and exciting—challenge to, in particular, mathematics.

This problem becomes even more severe when observing that in addition to a lack of theoretical foundation, causing, for instance, a very time-consuming and delicate training process, deep learning approaches also sometimes fail dramatically. One example of such failures are so-called adversarial examples, when small changes in the data lead to a radically different decision; a well-known problem in this regime is the sensitivity of self-driving cars to minor adaptions of traffic signs such as the placement of stickers. Another example is fairness, when biased training data causes deep learning approaches to, for instance, reach racist decisions.

Summarizing, there is a tremendous need for mathematics in the area of deep learning. One can identify two different research directions:

- *Mathematics for deep learning*. This direction aims for deriving a deep mathematical understanding of deep learning and asks questions such as "How can we make deep learning more robust?"
- *Deep learning for mathematics*. This direction focusses on mathematical problem settings such as inverse problems and numerical analysis of partial differential equations with the goal to employ deep learning techniques for superior solvers.

In this article, we will touch upon both research directions, showcasing some novel results and pointing out key future challenges for mathematics. In Section 2, we will first provide an introduction into deep learning from a mathematics viewpoint. We will then delve deeper into the first direction, namely *mathematics for deep learning*, and discuss the subarea of expressivity in more detail (Section 3). This will be followed in Section 4 by highlighting examples of the second direction, namely *deep learning for mathematics*. Finally, Section 5 is devoted to future perspectives for mathematics.

2. Deep neural networks

In 1943, McCulloch and Pitts had the vision to introduce artificial intelligence to the world [28]. At that time, their idea was to develop an algorithmic approach to learning by mimicking the functionality of the human brain. Due to the structure of the brain being composed of neurons with numerous interconnections, they introduced so-called artificial neurons as building blocks. The structure of a neuron in the human brain, in its most simple form, consists of dendrites through which signals are transmitted to its soma, while being scaled/amplified due to the structural properties of the respective dendrites. In the soma of the neuron, those incoming signals are accumulated, and a decision is reached whether to fire to other neurons or not, and also with which strength.

A mathematical definition of an artificial neuron is consequently defined as follows. In the sequel, we will build a neural network from such components with the weights and biases being the free parameters, which need to be trained.

Definition 2.1. An *artificial neuron* with *weights* $w_1, \dots, w_n \in \mathbb{R}$, *bias* $b \in \mathbb{R}$, and *activation function* $\rho : \mathbb{R} \to \mathbb{R}$ is defined as the function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x_1, \dots, x_n) = \rho\left(\sum_{i=1}^{n} x_i w_i + b\right) = \rho\big(\langle x, w \rangle + b\big),$$

where $w = (w_1, \dots, w_n)$ and $x = (x_1, \dots, x_n)$.

Let us now take a look at some examples of activation functions.

Example 2.2. (1) Heaviside function

$$\rho(x) = \begin{cases} 1, & x > 0, \\ 0, & x \le 0. \end{cases}$$

(2) Sigmoid function $\rho(x) = \frac{1}{1+e^{-x}}$.

(3) Rectifiable Linear Unit (ReLU) $\rho(x) = \max\{0, x\}$.

The most basic activation function is certainly the Heaviside function, leading to a yes/no decision. The sigmoid function is a smooth alternative. But the by far most extensively used activation function in basically all applications is the ReLU due to its simple piecewise linear structure, which is advantageous in the training process, and still allows superior performance.

2.1. The mathematical definition

An (artificial feed-forward) neural network is then built by concatenating artificial neurons to compositions of affine linear maps and activation functions. This leads to the following definition.

Definition 2.3. Let $d \in \mathbb{N}$ be the dimension of the input layer, let L be the number of layers, let $N_0 := d$, N_ℓ, $\ell = 1, \dots, L$, be the dimensions of the hidden and last layer, let $\rho : \mathbb{R} \to \mathbb{R}$ be a (non-linear) activation function, and, for $\ell = 1, \dots, L$, let T_ℓ be the affine-linear functions

$$T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, \quad T_\ell x = W^{(\ell)} x + b^{(\ell)},$$

with $W^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ being the weight matrices and $b^{(\ell)} \in \mathbb{R}^{N_\ell}$ the bias vectors of the ℓth layer. Then $\Phi : \mathbb{R}^d \to \mathbb{R}^{N_L}$ given by

$$\Phi(x) = T_L \rho\big(T_{L-1} \rho\big(\dots \rho\big(T_1(x)\big) \dots\big)\big), \quad x \in \mathbb{R}^d,$$

is called *(deep) neural network.*

We would like to stress that in many papers a distinction is made between a neural network and its realization, namely the function it realizes. The reason for this is that different architectures can lead to the same function. For this article, we will, however, avoid such technical delicacies.

2.2. Key research directions

Aiming to identify the key mathematical research directions in deep learning, let us take a high-level view of the typical application of a deep neural network; exemplarily we choose classification. One proceeds in the four—very coarsely explained—following steps.

(1) We assume that we are given samples $(x_i, f(x_i))_{i=1}^m$ of a function such as $f : \mathcal{M} \to \{1, 2, \dots, K\}$, where $\mathcal{M}$ might be a lower-dimensional manifold of $\mathbb{R}^d$. This is a customarily assumed setting in image classification. We then split this set into a training data set $(x_i, f(x_i))_{i=1}^{\tilde{m}}$, say, and a test data set $(x_i, f(x_i))_{i=\tilde{m}+1}^m$, say. The training data set is—as the name indicates—used for training, and the test data set for testing the performance of the trained network. Notice that the test data set stays hidden during the training process.

(2) Then an architecture of a deep neural network needs to be selected, i.e., a choice of L, $(N_\ell)_{\ell=1}^L$, and ρ. Sometimes selected entries of the weight matrices $(W^{(\ell)})_{\ell=1}^L$ are already set to zero at this point if one does not intend to train a fully connected neural network.

(3) Next, the affine-linear functions $(T_\ell)_{\ell=1}^L = (W^{(\ell)} \cdot + b^{(\ell)})_{\ell=1}^L$ are learnt by solving the optimization problem given by

$$\min_{(W^{(\ell)}, b^{(\ell)})_\ell} \sum_{i=1}^{\tilde{m}} \mathcal{L}\big(\Phi_{(W^{(\ell)}, b^{(\ell)})_\ell}(x_i), f(x_i)\big) + \lambda \mathcal{R}\big((W^{(\ell)}, b^{(\ell)})_\ell\big),$$

where $\mathcal{L}$ is a loss function to determine a measure of closeness between the network evaluated in the training samples and the (known) function values $f(x_i)$ and $\mathcal{R}$ is a regularization term to impose additional constraints on the weight matrices and bias vectors. The optimization problem is typically solved by stochastic gradient descent, yielding a network

$$\Phi_{(W^{(\ell)}, b^{(\ell)})_\ell} : \mathbb{R}^d \to \mathbb{R}^{N_L},$$

where

$$\Phi_{(W^{(\ell)}, b^{(\ell)})_\ell}(x) = T_L \rho\big(T_{L-1} \rho\big(\dots \rho\big(T_1(x)\big) \dots\big)\big).$$

(4) Finally, one employs the test data set to analyze whether

$$\Phi_{(W^{(\ell)}, b^{(\ell)})_\ell} \approx f,$$

i.e., whether and to which extent the training process was successful.

It is in fact very surprising that this procedure works this well these days, which has two main reasons: first, the drastic improvement of computing power allows the training of networks with hundreds of layers in the sense of *deep* neural networks. And, second, we are living in the age of data, hence vast amounts of training data is available. This being the empirical explanation, a profound mathematical explanation why, for instance, *deep* networks are superior to shallow ones or why the complex training data does not lead to the phenomenon of *overfitting* is to a large extent still missing.

2.2.1. Mathematics for deep learning. Based on these considerations, we can now formulate the four key mathematical research directions, first for *mathematics for deep learning*. We will each time also mention the main mathematical fields involved, thereby showing that almost each area of mathematics is touched and required.

- *Expressivity*. This direction aims to understand whether and to which extent aspects of a neural network architecture affect the performance of deep learning. Typically methods from applied harmonic analysis and approximation theory are used.
- *Learning*. The goal here is to analyze the training procedure with a key question being why the typically applied algorithm of stochastic gradient descent does

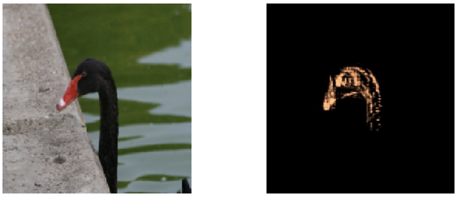


Figure 1. Illustration of an explanation for the classification as a "black swan" using RDE.

converge to suitable local minima even though the problem itself is highly nonconvex. This direction relies on techniques from areas such as algebraic/differential geometry, optimal control, and optimization.

- *Generalization.* This research direction is the least explored and maybe also the most difficult, sometimes called the "holy grail" of deep learning. It targets the out-of-sample error and asks questions such as "Why is depth beneficial" or "Why does high overparametrization not lead to overfitting?". Required methods belong in particular to the following areas: learning theory, probability theory, and statistics.

Notice that these three research directions are precisely related to the three components of the error of a statistical learning problem (cf. [4, (1.4) and Figure 1.2]), namely the approximation error from the hypothesis class, the optimization error from the algorithm itself, and the out-of-sample error.

Besides these more classical problem complexes, new directions have evolved. One of the most exciting directions might be the following, which until now lacks almost entirely a mathematical foundation.

- *Explainability.* Given a trained neural network, this area aims to analyze why certain decisions were reached, and which components of the input data were crucial for those. The range of required approaches is quite broad, including areas such as information theory or uncertainty quantification.

In practice, this direction is invaluable, since one often encounters the situation that a neural network is given and decisions have to be explained, for instance, to a customer. In the imaging situation, typical explanations are relevance maps assigning each pixel a relevance score for the decision such as layerwise-relevance propagation (LRP) [3] or rate-distortion explanation (RDE) [15]. For an example of such an explanation, we refer to Figure 1.

However, from a mathematical standpoint, one truly aims for a mathematical definition of the term "relevance" and an according theory of optimal relevance maps. Ideally, one would also like to have explanations beyond the pixel-based setting and for more challenging modalities. For a survey of some recent work in this direction, we refer to [21].

2.2.2. Deep learning for mathematics. As said before, the second main research thread is *deep learning for mathematics* in the sense of deep learning for mathematical problem settings. The two key research subfields are as follows.

- *Inverse problems.* The main goal is to improve classical model-based approaches by deep learning techniques. Since it is often highly beneficial to not entirely neglect domain knowledge such as the physics of the problem, one crucial question is how to optimally combine deep learning with model-based approaches. This direction relies on tools from imaging science, inverse problems, and microlocal analysis, to name a few.
- *Partial differential equations.* Research in this area targets foremost the question of how and to which extent deep neural networks are able to beat the curse of dimensionality. This direction obviously requires methods from areas such as numerical mathematics and partial differential equations.

3. Mathematics for deep learning

Deep learning-based methodologies for inverse problems and partial differential equations exploit deep neural networks as approximators. Thus, the first question to ask is whether deep neural networks are at least as good as all previous mathematical methods. This question belongs in the realm of the previously introduced area of expressivity, which will be the focus of this section, aiming to provide a (partial) answer.

3.1. Revisiting classical approximation theory

We start by revisiting classical approximation theory, and, in the sequel, analyze whether deep neural networks have at least similar approximation properties as classical methods.

In a nutshell, function approximation has the following goal. Given a class $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$ of interest—for later use it is sufficient for us to consider $L^2(\mathbb{R}^d)$—and a representation system $(\varphi_i)_{i \in I} \subseteq L^2(\mathbb{R}^d)$, which can be an orthonormal basis or, more generally, a frame, one aims to measure the suitability of $(\varphi_i)_{i \in I}$ for uniformly approximating functions from $\mathcal{C}$. For a budget N, the approximating function has then typically the form of a linear combination of N terms of the representation system. This leads to the following definition.

Definition 3.1. The *error of best N-term approximation* of some $f \in \mathcal{C}$ is given by

$$\sigma_N(f) := \inf_{I_N \subset I,\ \#I_N = N,\ (c_i)_{i \in I_N}} \Big\| f - \sum_{i \in I_N} c_i \varphi_i \Big\|_2 .$$

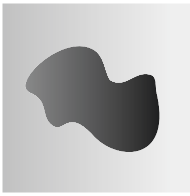


Figure 2. Illustration of a cartoon-like function.

The largest $\gamma > 0$ such that

$$\sup_{f \in \mathcal{C}} \sigma_N(f) = O(N^{-\gamma}), \quad \text{as } N \to \infty,$$

determines the *optimal (sparse) approximation rate* of $\mathcal{C}$ by $(\varphi_i)_{i \in I}$.

A closer look reveals that this viewpoint relates approximation accuracy to the complexity of the approximating system in terms of sparsity.

Also for later use, we will now introduce one example of a class $\mathcal{C}$ and a representation system $(\varphi_i)_{i \in I}$ along with an analysis of its optimal (sparse) approximation rate. The model class we will consider, called *cartoon-like functions* (see Figure 2), was first introduced in imaging science [10], since the predominant features of images are edge structures. Such anisotropic features also occur in other settings such as the solution of transport dominated equations, leading to a model class with much larger applicability.

Definition 3.2. The set of *cartoon-like functions* $\mathcal{E}^2(\mathbb{R}^2)$ is defined by

$$\mathcal{E}^2(\mathbb{R}^2) := \{f \in L^2(\mathbb{R}^2) : f = f_0 + f_1 \cdot \chi_B\},$$

where $\emptyset \neq B \subset [0,1]^2$ is simply connected with a C^2-curve with bounded curvature as its boundary, and $f_i \in C^2(\mathbb{R}^2)$ with supp $f_i \subseteq [0,1]^2$ and $\|f_i\|_{C^2} \leq 1$, $i = 0, 1$.

A lower bound for any optimal (sparse) approximation rate was derived in the same article (i.e., [10]). We would like to remark that the purpose of the technical requirement of "polynomial depth search" in the following theorem is to avoid degenerate cases of representation systems.

Theorem 3.3. *Allowing only polynomial depth search, we have the following optimal behavior for $f \in \mathcal{E}^2(\mathbb{R}^2)$:*

$$\sigma_N(f) \asymp N^{-1}, \quad \textit{as } N \to \infty.$$

The well-known wavelet systems [9] do only provide a suboptimal rate of $N^{-\frac{1}{2}}$ due to the fact that they are isotropic multiscale systems in the sense of scaling in both directions at a similar rate (cf. Figure 3).

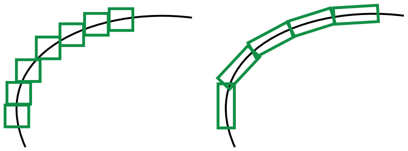


Figure 3. Schematic illustration of wavelet and shearlet approximation.

Various systems were suggested to provide optimal (sparse) approximations for cartoon-like functions. The first successful systems were curvelets [7], which, however, did not allow faithful implementations. This could be achieved by so-called shearlets, which were introduced in [26], see also the survey article [23]. For an illustration of the benefit of anisotropic scaling, we refer to Figure 3.

Shearlet systems are associated with three parameters: scale j, position m, and orientation k. For the precise definition, let A_{2^j} and $\widetilde{A}_{2^j}$, $j \in \mathbb{Z}$, denote the parabolic scaling matrices given by

$$A_{2^j} := \begin{pmatrix} 2^j & 0 \\ 0 & 2^{j/2} \end{pmatrix}$$

and $\widetilde{A}_{2^j} := \operatorname{diag}(2^{j/2}, 2^j)$, and let S_k, $k \in \mathbb{Z}$, be the shearing matrix given by

$$S_k := \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix}.$$

(Cone-adapted) discrete shearlet systems can then be defined as follows (cf. [24]).

Definition 3.4. The *(cone-adapted) discrete shearlet system* $\mathcal{SH}(\phi, \psi, \widetilde{\psi})$ generated by $\phi \in L^2(\mathbb{R}^2)$ and $\psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$ is the union of

$$\begin{gathered} \{\phi(\cdot - m) : m \in \mathbb{Z}^2\}, \\ \{2^{3j/4}\psi(S_k A_{2^j} \cdot - m) : j \geq 0,\ |k| \leq \lceil 2^{j/2} \rceil,\ m \in \mathbb{Z}^2\}, \\ \{2^{3j/4}\widetilde{\psi}(S_k^T \widetilde{A}_{2^j} \cdot - m) : j \geq 0,\ |k| \leq \lceil 2^{j/2} \rceil,\ m \in \mathbb{Z}^2\}. \end{gathered}$$

We denote the associated *shearlet transform* by

$$\mathrm{SH}(f) := \big(\langle f, g \rangle\big)_{g \in \mathcal{SH}(\phi,\psi,\widetilde{\psi})}, \quad f \in L^2(\mathbb{R}^2).$$

This system indeed satisfies the optimal (sparse) approximation rate for cartoon-like functions up to a log-factor, which is often regarded as negligible. The following statement is taken from [24], where also the precise hypotheses can be found.

Theorem 3.5. *Let $\phi, \psi, \widetilde{\psi} \in L^2(\mathbb{R}^2)$ be compactly supported, and let $\widehat{\psi}$, $\widehat{\widetilde{\psi}}$ satisfy certain decay condition. Then $\mathcal{SH}(\phi, \psi, \widetilde{\psi})$ provides an* optimally sparse approximation *of* $f \in \mathcal{E}^2(\mathbb{R}^2)$, *i.e.,*

$$\sigma_N(f) \lesssim N^{-1}(\log N)^{\frac{3}{2}}, \quad \text{as } N \to \infty.$$

Concluding our example for Definition 3.1, shearlet systems provide an (almost) optimal (sparse) approximation rate of N^{-1} for the class $\mathcal{C}$ of cartoon-like functions. For the interested reader, a faithful implementation of the shearlet transform as a 2D&3D (parallelized) fast shearlet transform can be found in www.ShearLab.org.

3.2. Universality of deep neural networks

Analyzing approximation problems for deep neural networks immediately bears the question of how to replace the notion of complexity of the approximating term, which was before measured in terms of sparsity. A typical approach for networks is a complexity measure in terms of memory requirements. Recall that the $\|\cdot\|_0$-“norm” counts the number of non-zero entries.

Definition 3.6. Retaining the same notation for deep neural networks as in Definition 2.3, the *complexity* $C(\Phi)$ of a deep neural network Φ is defined by

$$C(\Phi) := \sum_{\ell=1}^{L} \big(\|W^{(\ell)}\|_0 + \|b^{(\ell)}\|_0\big).$$

We will also in the sequel use the notion $\mathcal{NN}_{L,C,d,\rho}$ for the class of deep neural networks with no more than L layers, complexity of at most C, input dimension d, and activation function ρ. If no bound is given, we indicate this by writing ∞.

Thus, the key challenge is now to relate approximation accuracy to the complexity of the approximating network in terms of memory efficiency. A very classical result—and maybe the first main expressivity result from the time of the “first wave” of neural networks—is the universal approximation theorem [8, 17], which states that each continuous function on a compact domain can be approximated up to an arbitrary accuracy by a shallow neural network.

Theorem 3.7. *Let $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compact, $f : K \to \mathbb{R}$ continuous, $\rho : \mathbb{R} \to \mathbb{R}$ continuous and not a polynomial. Then, for each $\varepsilon > 0$, there exist $N \in \mathbb{N}$, $a_k, b_k \in \mathbb{R}$, and $w_k \in \mathbb{R}^d$, $1 \leq k \leq N$, such that*

$$\Big\| f - \sum_{k=1}^{N} a_k \rho\big(\langle w_k, \cdot\rangle - b_k\big) \Big\|_\infty \leq \varepsilon.$$

While this is certainly an interesting result, it is not satisfactory in terms of complexity, since this can be arbitrary large.

Aiming to derive an optimality result, we require a lower bound as a benchmark. One example of such a statement was proven in [5] in terms of a so-called optimal exponent $\gamma^*(\mathcal{C})$ from information theory to measure the complexity of $\mathcal{C} \subset L^2(\mathbb{R}^d)$. We should stress that only the essence of this result is stated without all details.

Theorem 3.8. *Let $d \in \mathbb{N}$, $\rho : \mathbb{R} \to \mathbb{R}$, and $\mathcal{C} \subset L^2(\mathbb{R}^d)$. Further, let*

$$\mathrm{Learn} : (0,1) \times \mathcal{C} \to \mathcal{N}\mathcal{N}_{\infty,\infty,d,\rho}$$

satisfy that, for each $f \in \mathcal{C}$ and $0 < \varepsilon < 1$,

$$\sup_{f \in \mathcal{C}} \big\| f - \mathrm{Learn}(\varepsilon, f) \big\|_2 \le \varepsilon.$$

Then, for all $\gamma < \gamma^*(\mathcal{C})$,

$$\varepsilon^\gamma \sup_{f \in \mathcal{C}} C\big(\mathrm{Learn}(\varepsilon, f)\big) \to \infty, \quad \textit{as } \varepsilon \to 0.$$

This now provides a conceptual lower bound independent of the learning algorithm. It in fact allows not only to construct deep neural networks, which are memory-optimal, but also to answer the question with which we started, namely whether deep neural networks are at least as good as all previous mathematical methods. We will affirm this for approximations by affine systems such as wavelets and shearlets.

One can now proceed as follows. Assume that we are given a specific function class such as cartoon-like images, and an associated representation system with an optimal approximation rate such as shearlets. Mimicking classical approximation theory—more specifically best N-term approximations—by neural networks leads to such memory-optimal neural networks, which at the same time perform at least as good as the associated representation system from an approximation standpoint.

One example of a resulting theorem is taken from [5]. Notice that this is in fact the optimal approximation rate (up to some ε), implying that the bound in Theorem 3.8 is sharp.

Theorem 3.9. *Let ρ be a suitably chosen activation function, and let $\varepsilon > 0$. Then, for all $f \in \mathcal{E}^2(\mathbb{R}^2)$ and $N \in \mathbb{N}$, there exists $\Phi \in \mathcal{N}\mathcal{N}_{3,O(N),2,\rho}$ with*

$$\|f - \Phi\|_2 \lesssim N^{-1+\varepsilon} \to 0, \quad \textit{as } N \to \infty.$$

Thus, one can conclude that deep neural networks achieve optimal approximation properties of all affine systems combined. Intriguingly, training the network architecture of the proof, the neural network does even learn approximations of classical affine systems such as shearlets; for more details see [5].

4. Deep learning for mathematics

Having established that deep neural networks are at least as good as various classical approximation methods, we will continue our journey in the deep learning world and next ask whether deep learning methods are even better than classical approaches. For this, we will now enter the area of *deep learning for mathematics* and turn towards the setting of inverse problems.

4.1. Inverse problems meet deep learning

We start by recalling a general classical approach to solve inverse problems. We will later discuss how to best combine it with deep learning in specific problem settings in the sense of taking the best out of the model- and data-world.

Assume that we are given an (ill-posed) inverse problem

$$Kf = g, \quad \text{where } K : X \to Y,$$

where X and Y are Hilbert spaces, say. In its most classical form in imaging science, K could be an operator which adds noise to an image, leading to a denoising problem. *Sparse regularization* is a conceptually general approach for recovering f from knowledge of g and K, see also [18]. It computes an approximate solution $f^\alpha \in X$, $\alpha > 0$, by solving

$$f^\alpha := \operatorname*{argmin}_f \Big[\underbrace{\|Kf - g\|^2}_{\text{data fidelity term}} + \alpha \cdot \underbrace{\big\| \big(\langle f, \varphi_i \rangle\big)_{i \in I} \big\|_1}_{\text{penalty term}} \Big],$$

where $(\varphi_i)_{i \in I}$ is a suitably selected—in the sense of providing sparse approximations of f—orthonormal basis or frame for X.

One class of approaches for combining deep learning with solvers such as sparse regularization are supervised approaches, which in their most direct form first apply the solver followed by the neural network [19]. A bit more sophisticated are approaches which replace certain procedures in the solver—such as a denoising part—by a deep neural network in the sense of plug-and-play [31] or using a specifically trained network [1]. Semi-supervised approaches aim to encode the regularization as a neural network (see, e.g., [27]), whereas deep image prior [32] are one example of what one might coin unsupervised approaches.

We will now focus on one specific inverse problem from imaging science and discuss one exemplary approach in more detail. This approach follows the philosophy to apply the model-based solver as far as it is reliable and only complement it by a deep neural network where necessary. The problem we aim to study is the inverse problem of (limited angle-) computed tomography.

A CT scanner samples the *Radon transform*, which is defined by

$$\mathcal{R}f(s,\theta) = \int_{-\infty}^{\infty} f\big(s\omega(\theta) + t\omega(\theta)^{\perp}\big)\,dt, \quad \text{for } (s,\theta) \in \mathbb{R} \times (0,\pi).$$

Here $\omega(\theta) := (\cos\theta, \sin\theta)$ is the unitary vector with orientation described by the angle θ with respect to the x_1-axis and $\omega(\theta)^{\perp} := (-\sin\theta, \cos\theta)$.

The problem of inverting the Radon transform becomes even harder if $\mathcal{R}f(s,\cdot)$ is only sampled on a proper subset $[-\phi,\phi]$ of $(0,\pi)$, which is the case in, for instance, electron tomography. In the sequel, we will refer to the respective Radon transform by $\mathcal{R}_{\phi}$. Classical solvers fail in this case due to the fact that a large connected region of the measurements is missing, while also being too complex for accurate modeling.

The key problem can in fact be regarded as recovering parts of the wavefront set of the original image, where—coarsely speaking—a wavefront set is the set of singularities of a distribution together with their directions; for a precise definition we refer to [16]. Since shearlets resolve the wavefront set [22], the following approach was suggested in [6], following the previously described philosophy:

- Step 1: *Reconstruct the visible.*
  Compute
  $$f^* := \operatorname*{argmin}_{f\geq 0} \|\mathcal{R}_{\phi}f - g\|_2^2 + \big\|\, \mathrm{SH}(f)\big\|_{1,w}.$$
  We then split the set of parameters (j,m,k) of shearlets into a visible set $\mathcal{I}_{\mathtt{vis}}$ and an invisible set $\mathcal{I}_{\mathtt{inv}}$ related to whether they are associated with shearlets within a range of acquired data or not, leading to the following cases:
  - ◇ for $(j,m,k) \in \mathcal{I}_{\mathtt{inv}}$, $\mathrm{SH}(f^*)_{(j,m,k)} \approx 0$;
  - ◇ for $(j,m,k) \in \mathcal{I}_{\mathtt{vis}}$, $\mathrm{SH}(f^*)_{(j,m,k)}$ is reliable and near perfect.
- Step 2: *Learn the invisible.*
  Train a neural network (U-net) Φ to compute
  $$\Phi : \mathrm{SH}(f^*)_{\mathcal{I}_{\mathtt{vis}}} \mapsto F,$$
  where F is an approximation of $\mathrm{SH}(f_{\mathtt{gt}})_{\mathcal{I}_{\mathtt{inv}}}$ and $f_{\mathtt{gt}}$ the ground truth image.
- Step 3: *Combine.*
  Finally, compute
  $$f_{\mathtt{LtI}} = \mathrm{SH}^T\big(\mathrm{SH}(f^*)_{\mathcal{I}_{\mathtt{vis}}} + F\big).$$

The numerical experiments in Figure 4 indicate the superiority of deep learning approaches in general and even more a careful combination of classical solvers with deep neural networks to pure model-based approaches.

This answers the question whether deep neural networks can perform even better than classical methods to the affirmative. We include with Figure 5 one additional

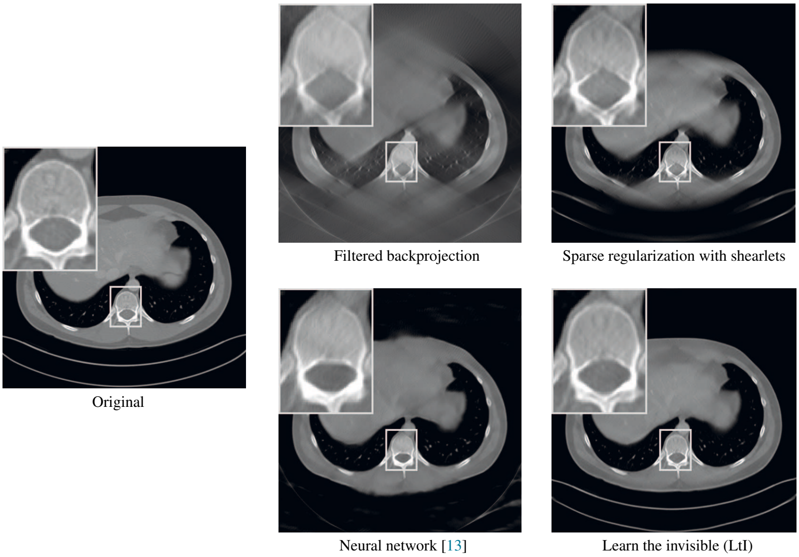

Original
Filtered backprojection
Sparse regularization with shearlets
Neural network [13]
Learn the invisible (LtI)

Figure 4. Illustration of the superiority of combined model-deep learning approaches.

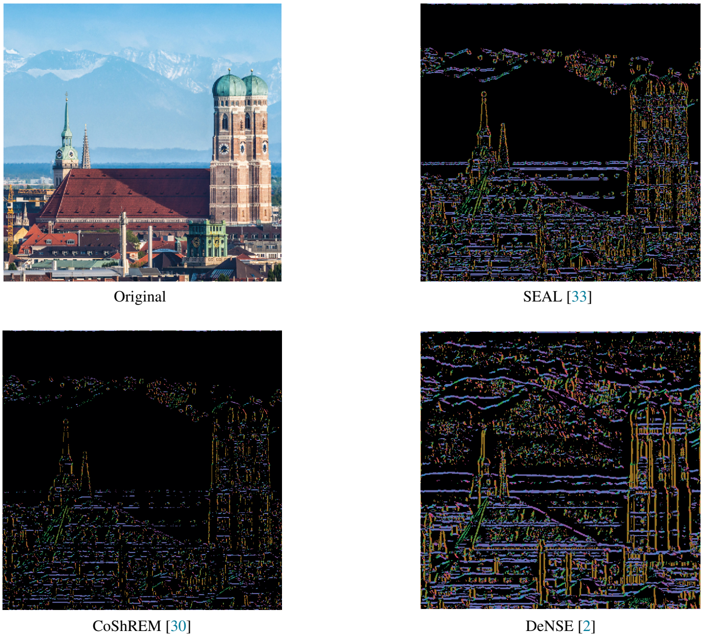


Figure 5. Illustration of another combined model-deep learning approach [2] in relation to pure model-based methods [30, 33].

example, which follows the same philosophy for the edge detection problem [2]. Without going into the details, this approach first uses shearlets as a coarse edge detector, followed by a deep neural network applied in shearlet domain.

4.2. Deep learning-based solvers for partial differential equations

Finally, we will provide a glimpse into the effectiveness of deep neural networks for solving partial differential equations, and provide an answer to the question of why one should use deep learning for solving partial differential equations at all.

Given a partial differential equation $\mathcal{L}(u) = f$, a common approach to solve this equation using a neural network Φ is to approximate the solution u by Φ, i.e., to train Φ such that

$$\mathcal{L}(\Phi) \approx f.$$

This requires to incorporate the partial differential equation into the loss function. Some of the key approaches in this realm are the Deep Ritz Method [11], the so-called physics-informed neural networks [29], or using a backwards stochastic partial differential equation reformulation [14].

We will now focus on a more general setting, namely parametric partial differential equations, which in fact arise in basically any branch of science and engineering such as in complex design problems or uncertainty quantification tasks. Let us now assume that we are given a parametric partial differential equation, $\mathcal{L}(u_y, y) = f_y$ with y being a parameter from a parameter space $\mathcal{Y} \subseteq \mathbb{R}^p$ and u_y the associated solution in a Hilbert space $\mathcal{H}$. Since in applications one typically faces a multi-query situation, the so-called *parametric map*, given by

$$\mathcal{Y} \ni y \mapsto u_y \in \mathcal{H} \quad \text{such that} \quad \mathcal{L}(u_y, y) = f_y,$$

needs to be solved several times. If p is very large, the curse of dimensionality could lead to an exponential computational cost.

It seems natural to ask whether deep neural networks can be of benefit in this situation in the sense of whether a network can approximate the parametric map leading to a flexible, universal approach which is hopefully not affected by the curse of dimensionality. For this, we first need to bring the problem into a finite-dimensional domain, which is done by a high-fidelity discretization, leading to the problem

$$\mathbb{R}^p \supseteq \mathcal{Y} \ni y \mapsto \mathbf{u}_y^{\mathrm{h}} \in \mathbb{R}^D \quad \text{such that} \quad b_y(u_y^h, v) = f_y(v) \text{ for all } v$$

with $b_y(u_y^h, v) = f_y(v)$ being the associated variational form and $\mathbf{u}_y^{\mathrm{h}}$ being the coefficient vector of u_y^h with respect to a suitable basis. We can now ask the following questions.

- Given $\varepsilon > 0$, does there exist a neural network Φ such that

$$\left\| \Phi(y) - \mathbf{u}_y^{\mathrm{h}} \right\| \leq \varepsilon, \quad \text{for all } y \in \mathcal{Y},$$

  and how does the complexity of Φ depend on p and D?
- How do neural networks perform numerically on this task?

The first question falls in the category of expressivity and would need to be complemented by an analysis of the learning procedure as well as the generalization error, as discussed in Section 2.2.1. Mathematical answers to those two questions are however at this point still out of reach, leaving only numerical experiments as an alternative.

The first question was indeed solved by explicitly constructing an associated deep neural network, while carefully monitoring its complexity. We state the result from [25] in a high level form.

Theorem 4.1. *There exists a neural network Φ which approximates the parametric map, i.e.,*

$$\left\| \Phi(y) - \mathbf{u}_y^{\mathrm{h}} \right\| \leq \varepsilon, \quad \textit{for all } y \in \mathcal{Y},$$

and the dependence of $C(\Phi)$ on p and D can be (polynomially) controlled.

With an extensive set-up of numerical experiments such as fixing a specific neural network architecture and the training procedure, it could then be shown in [12] that the numerical performance of deep neural networks for this task does also not suffer from the curse of dimensionality.

5. Conclusions

Deep learning shows impressive performance in real-world applications. However, a theoretical foundation is largely missing. Developing such a foundation requires various areas of mathematics as well as the development of new mathematics. The two main research areas are *mathematics for deep learning* with its subfields expressivity, learning, generalization, and explainability, and *deep learning for mathematics* aiming to apply deep learning to solve inverse problems and partial differential equations.

Let us conclude with seven mathematical key problems of deep learning as they were stated in [4]:

(1) What is the role of depth?

(2) Which aspects of a neural network architecture affect the performance of deep learning?

(3) Why does stochastic gradient descent converge to good local minima despite the non-convexity of the problem?

(4) Why do large neural networks not overfit?

(5) Why do neural networks perform well in very high-dimensional environments?

(6) Which features of data are learned by deep architectures?

(7) Are neural networks capable of replacing highly specialized numerical algorithms in natural sciences?

It is thus fair to say that there are exciting future perspectives for mathematics.

Acknowledgments. The author would like to thank Hector Andrade-Loarca, Ron Levie, and Philipp Scholl for their helpful feedback on an early version of this article.

Funding. This research was partly supported by the Bavarian High-Tech Agenda, DFG-SFB/TR 109 Grant C09, DFG-SPP 1798 Grant KU 1446/21-2, DFG-SPP 2298 Grant KU 1446/32-1, and NSF-Simons Research Collaboration on the Mathematical and Scientific Foundations of Deep Learning (MoDL) (NSF DMS 2031985).

References

[1] J. Adler and O. Öktem, Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems* **33** (2017), no. 12, Article No. 124007 Zbl 1394.92070 MR 3729789

[2] H. Andrade-Loarca, G. Kutyniok, O. Öktem, and P. C. Petersen, Extraction of digital wavefront sets using applied harmonic analysis and deep neural networks. *SIAM J. Imaging Sci.* **12** (2019), no. 4, 1936–1966 Zbl 1439.35009 MR 4031466

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10** (2015), e0130140

[4] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, The modern mathematics of deep learning. In *Mathematical Aspects of Deep Learning*, edited by P. Grohs and G. Kutyniok, Cambridge University Press, Cambridge, 2022

[5] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.* **1** (2019), no. 1, 8–45 Zbl 07468819 MR 3949699

[6] T. A. Bubba et al., Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography. *Inverse Problems* **35** (2019), no. 6, Article No. 064002 Zbl 1416.92099 MR 3975365

[7] E. J. Candès and D. L. Donoho, New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.* **57** (2004), no. 2, 219–266 Zbl 1038.94502 MR 2012649

[8] G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** (1989), no. 4, 303–314 Zbl 0679.94019 MR 1015670

[9] I. Daubechies, *Ten Lectures on Wavelets*. CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992 Zbl 0776.42018 MR 1162107

[10] D. L. Donoho, Sparse components of images and optimal atomic decompositions. *Constr. Approx.* **17** (2001), no. 3, 353–382 Zbl 0995.65150 MR 1828917

[11] W. E and B. Yu, The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. *Commun. Math. Stat.* **6** (2018), no. 1, 1–12 Zbl 1392.35306 MR 3767958

[12] M. Geist, P. Petersen, M. Raslan, R. Schneider, and G. Kutyniok, Numerical solution of the parametric diffusion equation by deep neural networks. *J. Sci. Comput.* **88** (2021), no. 1, Article No. 22 Zbl 1473.35331 MR 4268857

[13] J. Gu and J. C. Ye, Multi-scale wavelet domain residual learning for limited-angle CT reconstruction. In *Procs Fully3D*, pp. 443–447, 2017

[14] J. Han, A. Jentzen, and W. E, Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci. USA* **115** (2018), no. 34, 8505–8510 Zbl 1416.35137 MR 3847747

[15] C. Heiß, R. Levie, C. Resnick, G. Kutyniok, and J. Bruna, In-distribution interpretability for challenging modalities. In *ICML Workshop on ML Interpretability for Scientific Discovery*, 2020

[16] L. Hörmander, *The Analysis of Linear Partial Differential Operators. I. Distribution Theory and Fourier Analysis*. Classics in Mathematics, Springer, Berlin, 2003 Zbl 1028.35001 MR 1996773

[17] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** (1989), no. 5, 359–366 Zbl 1383.92015

[18] B. Jin, P. Maaß, and O. Scherzer, Sparsity regularization in inverse problems [preface]. *Inverse Problems* **33** (2017), no. 6, Article No. 060301 Zbl 1369.00102 MR 3988255

[19] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26** (2017), no. 9, 4509–4522 Zbl 1409.94275 MR 3670561

[20] J. Jumper et al., Highly accurate protein structure prediction with AlphaFold. *Nature* **596** (2021), 583–589

[21] S. Kolek, D. A. Nguyen, R. Levie, J. Bruna, and G. Kutyniok, A rate-distortion framework for explaining black-box model decisions. In *xxAI - Beyond Explainable AI: International Workshop*, edited by A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, pp. 91–115, Springer, Cham, 2022

[22] G. Kutyniok and D. Labate, Resolution of the wavefront set using continuous shearlets. *Trans. Amer. Math. Soc.* **361** (2009), no. 5, 2719–2754 Zbl 1169.42012 MR 2471937

[23] G. Kutyniok and D. Labate, Introduction to shearlets. In *Shearlets*, pp. 1–38, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, New York, 2012 Zbl 1251.42010 MR 2896274

[24] G. Kutyniok and W.-Q. Lim, Compactly supported shearlets are optimally sparse. *J. Approx. Theory* **163** (2011), no. 11, 1564–1589 Zbl 1226.42031 MR 2832720

[25] G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider, A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.* **55** (2022), no. 1, 73–125 Zbl 07493717 MR 4376560

[26] D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss, Sparse multidimensional representation using shearlets. In *Wavelets XI (San Diego, CA, 2005)*, edited by M. Papadakis, A. F. Laine, and M. A. Unser, pp. 254–262, SPIE Proc. 5914, SPIE, Bellingham, WA, 2005

[27] S. Lunz, O. Öktem, and C.-B. Schönlieb, Adversarial regularizers in inverse problems. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, pp. 8507–8516, 2018

[28] W. S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5** (1943), 115–133 Zbl 0063.03860 MR 10388

[29] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378** (2019), 686–707 Zbl 1415.68175 MR 3881695

[30] R. Reisenhofer, J. Kiefer, and E. J. King, Shearlet-based detection of flame fronts. *Exp. Fluids* **57** (2016), Article No. 41

[31] Y. Romano, M. Elad, and P. Milanfar, The little engine that could: regularization by denoising (RED). *SIAM J. Imaging Sci.* **10** (2017), no. 4, 1804–1844 Zbl 1401.62101 MR 3714346

[32] D. Ulyanov, A. Vedaldi, and V. Lempitsky, Deep image prior. In *The Ieee Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9446–9454, 2018

[33] Z. Yu et al., Simultaneous edge alignment and learning. In *Computer Vision – ECCV 2018*, edited by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, pp. 400–417, Springer, Cham, 2018

Gitta Kutyniok
Department of Mathematics, Ludwig-Maximilians-Universität München, Theresienstr. 39, 80333 München; Munich Center for Machine Learning (MCML), Geschwister-Scholl-Platz 1, 80539 München, Germany; and Department of Physics and Technology, University of Tromsø, Klokkargårdsbakken 35, 9019 Tromsø, Norway; kutyniok@math.lmu.de

DOI 10.4171/8ECM/25

Geometric valuation theory

Monika Ludwig

Abstract. A brief introduction to geometric valuation theory is given. The focus is on classification results for valuations on convex bodies and on function spaces.

1. Introduction

Measurement is part of the literal meaning of geometry and geometric valuation theory deals with measurement in the following sense. We want to associate to a geometric object a real number (or, more generally, an element of an abelian semigroup $\mathbb{A}$). For example, we can associate to a sufficiently regular subset of $\mathbb{R}^n$ its n-dimensional volume or the $(n-1)$-dimensional measure of its boundary. Let $\mathcal{S}$ be a class of subsets of $\mathbb{R}^n$. We call a function $\mathrm{Z} : \mathcal{S} \to \mathbb{A}$ a *valuation* if

$$\mathrm{Z}(K) + \mathrm{Z}(L) = \mathrm{Z}(K \cup L) + \mathrm{Z}(K \cap L)$$

for all $K, L \in \mathcal{S}$ with $K \cap L$, $K \cup L \in \mathcal{S}$ (and we set $\mathrm{Z}(\emptyset) := 0$). Thus, the valuation property is just the inclusion-exclusion principle applied to two sets. In particular, measures on $\mathbb{R}^n$ when restricted to elements of $\mathcal{S}$ are valuations but there are many additional interesting valuations.

In his Third Problem, Hilbert asked whether an elementary definition of volume on polytopes is possible. In 1900, it was known that it is possible on $\mathbb{R}^2$ but the question was open in higher dimensions. Let $\mathcal{P}^n$ be the set of convex polytopes in $\mathbb{R}^n$ and call $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}$ *simple* if $\mathrm{Z}(P) = 0$ for all lower dimensional polytopes. Using our terminology, Hilbert's Third Problem turns out to be equivalent to the question whether every simple, rigid motion invariant valuation $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}$ is a multiple of n-dimensional volume for $n \geq 3$. Dehn [46] solved Hilbert's Third Problem by constructing a simple, rigid motion invariant valuation that is not a multiple of volume and thereby showed that an elementary definition of volume is not possible for $n \geq 3$.

Blaschke [30] took the important next step by asking for classification results for invariant valuations on $\mathcal{P}^n$ and on the space of convex bodies, $\mathcal{K}^n$, that is, of

2020 Mathematics Subject Classification. Primary 52B45; Secondary 46E35, 52A20, 26B25.
Keywords. Valuation, convex body, convex function, Sobolev space.

non-empty, compact, convex sets in $\mathbb{R}^n$. For a class $\mathcal{S}$ of subsets of $\mathbb{R}^n$, we say that a function $\mathrm{Z}: \mathcal{S} \to \mathbb{A}$ is G *invariant* for a group G acting on $\mathbb{R}^n$ if $\mathrm{Z}(\phi K) = \mathrm{Z}(K)$ for all $\phi \in G$ and $K \in \mathcal{S}$. Blaschke's question is motivated by Klein's Erlangen Program. We will describe some of the results that were obtained in this direction, in particular, focusing on the special linear group, $\mathrm{SL}(n)$, and the group of (orientation preserving) rotations, $\mathrm{SO}(n)$. Often additional regularity assumptions are required and for $\mathbb{A}$, a topological semigroup, we consider continuous and upper semicontinuous valuations, where the topology on $\mathcal{K}^n$ and its subspaces is induced by the Hausdorff metric.

In addition to classification results and their applications, structural results for spaces of valuations have attracted a lot of attention in recent years. We refer to the books and surveys [14, 17, 21]. Valuations were also considered on various additional spaces, in particular, on manifolds (see [12]). We will restrict our attention to subspaces of $\mathcal{K}^n$ and to recent results on valuations on spaces of real valued functions. On a space X of (extended) real valued functions, a function $\mathrm{Z}: X \to \mathbb{A}$ is called a *valuation* if

$$\mathrm{Z}(u) + \mathrm{Z}(v) = \mathrm{Z}(u \vee v) + \mathrm{Z}(u \wedge v)$$

for all $u, v \in X$ such that also their pointwise maximum $u \vee v$ and pointwise minimum $u \wedge v$ belong to X. Since spaces of convex bodies can be embedded in various function spaces in such a way that union and intersection of convex bodies correspond to pointwise minimum and maximum of functions, this notion generalizes the classical notion.

2. Affine valuations on convex bodies

The first classification result in geometric valuation theory is due to Blaschke. He worked on polytopes and aimed at a complete classification of rigid motion invariant valuations. However, at a certain step, he had to assume also $\mathrm{SL}(n)$ invariance and established the following result (and the corresponding result on polytopes).

Theorem 2.1 (Blaschke [30]). *A functional* $\mathrm{Z}: \mathcal{K}^n \to \mathbb{R}$ *is a continuous, translation and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0, c_n \in \mathbb{R}$ *such that*

$$\mathrm{Z}(K) = c_0 V_0(K) + c_n V_n(K)$$

for every $K \in \mathcal{K}^n$.

Here, $V_0(K) := 1$ is the Euler characteristic of K and $V_n(K)$ is its n-dimensional volume. It has become customary to refer to results that involve invariance (or covariance) with respect to $\mathrm{SL}(n)$ as affine results and the title of this section is to be understood in this sense.

We will first describe results for affine valuations on polytopes and then on general convex bodies. While on $\mathcal{P}^n$ a complete classification of $\mathrm{SL}(n)$ invariant valuations has been established, we require additional regularity assumptions on $\mathcal{K}^n$. Such assumptions are also used on important subspaces of $\mathcal{P}^n$ and $\mathcal{K}^n$. We will also describe results for affine valuations with values in tensor spaces, spaces of convex bodies, and related spaces.

2.1. SL(n) invariant valuations on convex polytopes

We call a function $\zeta : [0,\infty) \to \mathbb{R}$ a *Cauchy function* if

$$\zeta(x+y) = \zeta(x) + \zeta(y)$$

for every $x, y \in [0,\infty)$. Cauchy functions are well understood and can be completely described (if we assume the axiom of choice) by their values on a Hamel basis.

The following result gives a complete classification of translation and $\mathrm{SL}(n)$ invariant valuations on polytopes and is closely related to Theorem 2.1.

Theorem 2.2 ([94]). *A functional* $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}$ *is a translation and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0 \in \mathbb{R}$ *and a Cauchy function* $\zeta : [0,\infty) \to \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c_0 V_0(P) + \zeta\big(V_n(P)\big)$$

for every $P \in \mathcal{P}^n$.

Even without translation invariance, a complete classification can be obtained (see [94]). We state the case when the valuation is in addition continuous. We write $[0, P]$ for the convex hull of the origin and $P \in \mathcal{P}^n$.

Theorem 2.3 ([94]). *A functional* $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}$ *is a continuous and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0, c_n, d_n \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c_0 V_0(P) + c_n V_n(P) + d_n V_n\big([0, P]\big)$$

for every $P \in \mathcal{P}^n$.

Corresponding results are known on the space, $\mathcal{P}^n_0$, of polytopes containing the origin (see [94]).

Let $\mathcal{P}^n_{(0)}$ be the space of convex polytopes in $\mathbb{R}^n$ that contain the origin in their interiors. Here, we have additional interesting valuations connected to polarity. For $K \in \mathcal{K}^n$, define its polar by

$$K^* := \big\{y \in \mathbb{R}^n : \langle x, y\rangle \le 1 \text{ for all } x \in K\big\},$$

where $\langle x, y\rangle$ is the inner product of $x, y \in \mathbb{R}^n$. If $P \in \mathcal{P}^n_{(0)}$, then $P^* \in \mathcal{P}^n_{(0)}$. Hence, setting

$$V_n^*(P) := V_n(P^*),$$

we obtain a finite valued functional on $\mathcal{P}^n_{(0)}$ and it follows easily from properties of polarity that it is a valuation.

Valuations on $\mathcal{P}^n_{(0)}$ were first considered in [84], where a classification of Borel measurable, $\mathrm{SL}(n)$ invariant, and homogeneous valuations was established. Here, we say that $\mathrm{Z} : \mathcal{P}^n_{(0)} \to \mathbb{R}$ is *homogeneous* if there is $q \in \mathbb{R}$ such that

$$\mathrm{Z}(tP) = t^q\,\mathrm{Z}(P)$$

for every $P \in \mathcal{P}^n_{(0)}$ and $t > 0$. We say that Z is *Borel measurable* if the pre-image of every open set is a Borel set. We use corresponding notions on $\mathcal{K}^n$ and related spaces.

The results from [84] were strengthened by Haberl and Parapatits.

Theorem 2.4 (Haberl and Parapatits [55, 57]). *A functional* $\mathrm{Z} : \mathcal{P}^n_{(0)} \to \mathbb{R}$ *is a Borel measurable and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0, c_n, c_{-n} \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c_0 V_0(P) + c_n V_n(P) + c_{-n} V_n^*(P)$$

for every $P \in \mathcal{P}^n_{(0)}$.

The regularity assumption is again required to exclude discontinuous solutions of the Cauchy functional equation. It is an open problem to establish a complete classification without such assumption.

We remark that lattice polytopes, that is, convex polytopes with vertices in the integer lattice $\mathbb{Z}^n$, are important in many fields and subjects. The Betke–Kneser theorem [28] gives a complete classification of valuations on this class that are invariant with respect to translations by integer vectors and by so-called unimodular transformations (which can be described by matrices with integer coefficients and determinant ±1). For more information on valuations on lattice polytopes, see [32].

2.2. Affine surface areas

For $K \in \mathcal{K}^n$, the *affine surface area* of K is defined by

$$\Omega(K) := \int_{\partial K} \kappa(K, x)^{\frac{1}{n+1}}\,\mathrm{d}x, \tag{2.1}$$

where $\kappa(K, x)$ is the generalized Gaussian curvature of ∂K at x and integration is with respect to the $(n-1)$-dimensional Hausdorff measure. For smooth convex surfaces, this definition is classical (see [29]). It is also classical that Ω is translation and

$\mathrm{SL}(n)$ invariant for smooth surfaces. The extension of the definition of affine surface area to general convex bodies was obtained more recently in a series of papers by Leichtweiß [73], Lutwak [98], and Schütt and Werner [126]. There it is also proved that Ω is translation and $\mathrm{SL}(n)$ invariant on $\mathcal{K}^n$. The notion of affine surface area is fundamental in affine differential geometry. Moreover, since many basic problems in discrete and stochastic geometry are translation and $\mathrm{SL}(n)$ invariant, affine surface area has found numerous applications in these fields (see [47, 50]). It follows easily from (2.1) that Ω vanishes on polytopes and therefore is not continuous. The long conjectured upper semicontinuity of affine surface area (for smooth surfaces as well as for general convex surfaces) was proved by Lutwak [98]. For a proof that Ω is a valuation, see [125].

The following result gives a classification of upper semicontinuous, translation and $\mathrm{SL}(n)$ invariant valuations and represents a strengthening of Theorem 2.1. It provides a characterization of affine surface area.

Theorem 2.5 ([92]). *A functional* $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}$ *is an upper semicontinuous, translation and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0, c_n \in \mathbb{R}$ *and* $c \geq 0$ *such that*

$$\mathrm{Z}(K) = c_0 V_0(K) + c_n V_n(K) + c\,\Omega(K)$$

for every $K \in \mathcal{K}^n$.

For $n = 2$, this result was proved in [80], where also applications to asymptotic approximation by polytopes were obtained.

A complete classification of translation and $\mathrm{SL}(n)$ invariant valuations on $\mathcal{K}^n$ appears to be out of reach. Already a weakening of upper semicontinuity to, say, Baire-one (that is, a pointwise limit of continuous functionals) would be interesting and would have applications in discrete and stochastic geometry.

Let $\mathcal{K}^n_{(0)}$ be the space of convex bodies in $\mathbb{R}^n$ containing the origin in their interiors. For such a convex body with smooth boundary, the *centro-affine surface area* is a classical notion that can be defined by

$$\Omega_n(K) := \int_{\partial K} \kappa_0(K, x)^{\frac{1}{2}} \,\mathrm{d}V_K(x),$$

where $\mathrm{d}V_K(x) := \langle x, u_K(x)\rangle \,\mathrm{d}x$ with $u_K(x)$ the outer unit normal vector to K at x is (up to a constant) the cone measure on ∂K and

$$\kappa_0(K, x) := \frac{\kappa(K, x)}{\big\langle x, u_K(x)\big\rangle^{n+1}}.$$

It is classical that Ω_n is $\mathrm{GL}(n)$ invariant. Lutwak [100] extended this notion to general convex bodies in $\mathcal{K}^n_{(0)}$ and showed that Ω_n is upper semicontinuous.

The following result gives a complete classification of upper semicontinuous and $\mathrm{GL}(n)$ invariant valuations on $\mathcal{K}^n_{(0)}$ and provides a characterization of centro-affine surface area.

Theorem 2.6 ([93]). *A functional* $\mathrm{Z} : \mathcal{K}^n_{(0)} \to \mathbb{R}$ *is an upper semicontinuous and* $\mathrm{GL}(n)$ *invariant valuation if and only if there are* $c_0 \in \mathbb{R}$ *and* $c \geq 0$ *such that*

$$\mathrm{Z}(K) = c_0 V_0(K) + c\,\Omega_n(K)$$

for every $K \in \mathcal{K}^n_{(0)}$.

Lutwak [100] defined the so-called L^p-affine surface areas which were characterized in [93] as upper semicontinuous, $\mathrm{SL}(n)$ invariant, homogeneous valuations.

A more general notion, now called *Orlicz affine surface area*, was introduced in [93]. Let

$$\mathrm{Conc}[0,\infty) := \Big\{\zeta : [0,\infty) \to [0,\infty) : \zeta \text{ concave}, \lim_{t\to 0} \zeta(t) = \lim_{t\to\infty} \frac{\zeta(t)}{t} = 0\Big\}.$$

The following result gives a classification of upper semicontinuous, $\mathrm{SL}(n)$ invariant valuations on $\mathcal{K}^n_{(0)}$ and provides a characterization of Orlicz affine surface areas.

Theorem 2.7 ([55, 93]). *A functional* $\mathrm{Z}\colon \mathcal{K}^n_{(0)} \to \mathbb{R}$ *is an upper semicontinuous and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are* $c_0, c_n, c_{-n} \in \mathbb{R}$ *and* $\zeta \in \mathrm{Conc}[0,\infty)$ *such that*

$$\mathrm{Z}(K) = c_0 V_0(K) + c_n V_n(K) + c_{-n} V_n^*(K) + \int_{\partial K} \zeta\big(\kappa_0(K,x)\big)\,\mathrm{d}V_K(x)$$

for every $K \in \mathcal{K}^n_{(0)}$.

Here, the classification of upper semicontinuous, $\mathrm{SL}(n)$ invariant valuations vanishing on polytopes from [93] is combined with Theorem 2.4 by Haberl and Parapatits.

2.3. Vector and tensor valuations

We say that $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}^n$ is $\mathrm{SL}(n)$ *equivariant* if

$$\mathrm{Z}(\phi P) = \phi\,\mathrm{Z}(P)$$

for all $\phi \in \mathrm{SL}(n)$ and $P \in \mathcal{P}^n$. We use corresponding definitions for subspaces of $\mathcal{P}^n$.

The study of $\mathrm{SL}(n)$ equivariant vector valuations on convex polytopes containing the origin in their interiors was started in [82], where a classification of Borel measurable, $\mathrm{SL}(n)$ equivariant, homogeneous valuations was established. Haberl and Parapatits strengthened this result and obtained the following complete classification, which we state for $n \geq 3$.

Theorem 2.8 (Haberl and Parapatits [57, 58]). *A function* $\mathrm{Z} : \mathcal{P}^n_{(0)} \to \mathbb{R}^n$ *is a Borel measurable and* $\mathrm{SL}(n)$ *equivariant valuation if and only if there is* $c \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c m(P)$$

for every $P \in \mathcal{P}^n_{(0)}$.

Here, for $P \in \mathcal{P}^n$, the moment vector $m(P)$ is defined by $m(P) := \int_P x \, \mathrm{d}x$.

Zeng and Ma showed that it is possible to obtain a complete classification of vector valuations on convex polytopes without any regularity assumptions. We state their result for $n \geq 3$.

Theorem 2.9 (Zeng and Ma [137]). *A function* $\mathrm{Z} : \mathcal{P}^n \to \mathbb{R}^n$ *is an* $\mathrm{SL}(n)$ *equivariant valuation if and only if there are* $c, d \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c m(P) + d m\big([0, P]\big)$$

for every $P \in \mathcal{P}^n$.

In the same paper, a complete classification result is also established for $n = 2$. The obtained valuations depend on Cauchy functions.

Also higher rank tensor valuations are important in the geometry of convex bodies. In particular, the moment matrix $M^{2,0}(K)$ of a convex body K is a most valuable tool through its connection to the Legendre ellipsoid and the notion of isotropic position. In a certain way dual is the so-called LYZ ellipsoid, which was introduced by Lutwak, Yang, and Zhang [102, 103]. Associated to this ellipsoid is the LYZ matrix, which was characterized as a matrix valuation on convex polytopes containing the origin in [85]. The LYZ matrix corresponds to the Fisher information matrix [89, 102, 103] important in statistics and information theory.

Haberl and Parapatits [58] extended the result from [85] to general symmetric tensor valuations. For $p \geq 1$, let $\mathbb{T}^p(\mathbb{R}^n)$ denote the space of symmetric p-tensors on $\mathbb{R}^n$. We identify $\mathbb{R}^n$ with its dual space and regard each symmetric p-tensor as a symmetric p-linear functional on $(\mathbb{R}^n)^p$. We say that $\mathrm{Z} : \mathcal{P}^n_{(0)} \to \mathbb{T}^p(\mathbb{R}^n)$ is $\mathrm{SL}(n)$ *equivariant* if

$$\mathrm{Z}(\phi P)(y_1, \dots, y_p) = \mathrm{Z}(P)(\phi^{-1} y_1, \dots, \phi^{-1} y_p)$$

for all $y_1, \dots, y_p \in \mathbb{R}^n$, all $\phi \in \mathrm{SL}(n)$, and all $P \in \mathcal{P}^n_{(0)}$. We state the result by Haberl and Parapatits for $n \geq 3$ and $p \geq 2$.

Theorem 2.10 (Haberl and Parapatits [58]). *A function* $\mathrm{Z} : \mathcal{P}^n_{(0)} \to \mathbb{T}^p(\mathbb{R}^n)$ *is a Borel measurable,* $\mathrm{SL}(n)$ *equivariant valuation if and only if there are* $c, d \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c M^{p,0}(P) + d M^{0,p}(P^*)$$

for every $P \in \mathcal{P}^n_{(0)}$.

Here, the pth moment tensor of a convex polytope $P \in \mathcal{P}^n_{(0)}$ is defined by

$$M^{p,0}(P) := \frac{1}{p!} \int_P x^p \,\mathrm{d}x, \tag{2.2}$$

where x^p is the p-fold symmetric tensor product of $x \in \mathbb{R}^n$ and the pth LYZ tensor is

$$M^{0,p}(P) := \int_{\mathbb{S}^{n-1}} y^p \,\mathrm{d}S_{n-1,p}(P, y),$$

where $S_{n-1,p}(P, \cdot)$ is the L^p surface area measure of P, which is a central notion in the L^p Brunn–Minkowski theory (see [99, 100]).

For classifications of matrix valuation on $\mathcal{P}^n$ without regularity assumptions, see [108, 109], and for tensor valuations on lattice polytopes, see [95]. Continuous tensor valuations on complex vector spaces are classified in [4].

2.4. Convex body valued valuations and related notions

Affinely associated convex bodies play an important role in convex geometry (see [122, Chapter 10]). We have already mentioned the Legendre and the LYZ ellipsoid and describe here results on valuations $\mathrm{Z} : \mathcal{K}^n \to \mathcal{K}^n$, where we choose suitable additions on $\mathcal{K}^n$. The most classical choice is the *Minkowski addition*, where for $K, L \in \mathcal{K}^n$,

$$K + L := \{x + y : x \in K,\ y \in L\},$$

and such valuations are called *Minkowski valuations*.

The first classification result for Minkowski valuations was obtained in [83] and strengthened in [86]. It provides a characterization of projection bodies, a notion that was introduced by Minkowski.

Theorem 2.11 ([86]). *An operator* $\mathrm{Z} : \mathcal{P}^n \to \mathcal{K}^n$ *is a translation invariant,* $\mathrm{SL}(n)$ *contravariant Minkowski valuation if and only if there is* $c \geq 0$ *such that*

$$\mathrm{Z}\, P = c\,\Pi P$$

for every $P \in \mathcal{P}^n$.

Here, we describe convex bodies by their support functions, where for $K \in \mathcal{K}^n$, the *support function* $h(K, \cdot) : \mathbb{R}^n \to \mathbb{R}$ is given by

$$h(K, y) := \max\big\{\langle x, y\rangle : x \in K\big\}.$$

The support function is homogeneous of degree 1 and convex and any such function is the support function of a convex body. For $K \in \mathcal{K}^n$, the *projection body* of K is defined by

$$h(\Pi K, y) := V_{n-1}(K|y^\perp)$$

for $y \in \mathbb{S}^{n-1}$, where $y^\perp$ is the hyperplane orthogonal to y and $K|y^\perp$ denotes the image of the orthogonal projection of K onto $y^\perp$. We say that $\mathrm{Z} : \mathcal{P}^n \to \mathcal{K}^n$ is $\mathrm{SL}(n)$ *contravariant* if

$$\mathrm{Z}(\phi P) = \phi^{-t}\,\mathrm{Z}\,P$$

for all $\phi \in \mathrm{SL}(n)$ and $P \in \mathcal{P}^n$, where ϕ^{-t} is the inverse of the transpose of ϕ. For more information on projection bodies and their many applications, see [48, 122].

We say that $\mathrm{Z} : \mathcal{P}^n \to \mathcal{K}^n$ is $\mathrm{SL}(n)$ *equivariant* if

$$\mathrm{Z}(\phi P) = \phi\,\mathrm{Z}\,P$$

for all $\phi \in \mathrm{SL}(n)$ and $P \in \mathcal{P}^n$. The following result establishes a classification $\mathrm{SL}(n)$ equivariant valuations.

Theorem 2.12 ([86]). *An operator* $\mathrm{Z} : \mathcal{P}^n \to \mathcal{K}^n$ *is a translation invariant,* $\mathrm{SL}(n)$ *equivariant Minkowski valuation if and only if there is* $c \geq 0$ *such that*

$$\mathrm{Z}\,P = c\,\mathrm{D}\,P$$

for every $P \in \mathcal{P}^n$.

Here, the operator $P \mapsto \mathrm{D}\,P := \{x - y : x, y \in P\}$ assigns to P its *difference body* (see [48, 122]).

A classification of $\mathrm{SL}(n)$ equivariant, homogeneous Minkowski valuations on the space, $\mathcal{K}_0^n$, of convex bodies containing the origin was obtained in [86]. The result was strengthened by Haberl [53], who was able to drop the assumption of homogeneity. Let $n \geq 3$.

Theorem 2.13 (Haberl [53]). *An operator* $\mathrm{Z} : \mathcal{K}_0^n \to \mathcal{K}^n$ *is a continuous,* $\mathrm{SL}(n)$ *equivariant Minkowski valuation if and only if there are* $c_0 \in \mathbb{R}$ *and* $c_1, c_2, c_3 \geq 0$ *such that*

$$\mathrm{Z}\,K = c_0 m(K) + c_1 K + c_2(-K) + c_3\,\mathrm{M}\,K$$

for every $K \in \mathcal{K}_0^n$.

Here, the *moment body*, $\mathrm{M}\,K$, of K is defined by

$$h(\mathrm{M}\,K, y) := \int_K \big|\langle x, y\rangle\big|\,\mathrm{d}x$$

for $y \in \mathbb{R}^n$. When divided by the volume of K, the moment body of K is called its *centroid body* and is a classical and important notion going back to at least Dupin (see [48,122]). Results corresponding to Theorem 2.13 for $\mathrm{SL}(n)$ contravariant Minkowski valuations were obtained in [53, 86]. On the space, $\mathcal{P}_0^n$, of convex polytopes containing the origin, classification results for $\mathrm{SL}(n)$ contravariant Minkowski valuations were established in [53, 86] without assuming continuity and additional operators appear. For the $\mathrm{SL}(n)$ equivariant case, such results were established in [76].

We remark that the results from Theorem 2.13 and the corresponding results in the $\mathrm{SL}(n)$ equivariant case were complemented in [124, 135] by classification results for continuous, homogeneous Minkowski valuations on $\mathcal{K}^n$. A complete classification for $\mathrm{SL}(n)$ equivariant Minkowski valuations on $\mathcal{P}_0^n$ was established in [53]. On the space of convex bodies that contain the origin in their interiors, moment bodies allow to define $\mathrm{SL}(n)$ equivariant Minkowski valuations using polarity. For continuous, $\mathrm{SL}(n)$ equivariant, homogeneous valuations, a complete classification on this space was established in [88]. For Minkowski valuations on lattice polytopes, see [33].

Classification results for Minkowski valuations on complex vector spaces were established by Abardia and Bernig [1–3]. They introduce and characterize complex projection and difference bodies.

An important extension of the classical Brunn–Minkowski theory is the more recent L^p Brunn–Minkowski theory (see [99,100]). For $p > 1$, the L^p sum of convex bodies $K, L \in \mathcal{K}_0^n$ is defined by

$$h^p(K +_p L, y) := h^p(K, y) + h^p(L, y)$$

for $y \in \mathbb{R}^n$. An L^p Minkowski valuation $\mathrm{Z} : \mathcal{K}^n \to \mathcal{K}_0^n$ is a valuation where on $\mathcal{K}_0^n$ this addition is chosen. Classification results were obtained in [76,86,117,118] and led to the definition of asymmetric L^p projection and moment bodies (see [86]). Inequalities for these new classes of operators were established by Haberl and Schuster [59]. They generalize the L^p Petty projection and the L^p Busemann–Petty moment inequalities, which were established by Lutwak, Yang, and Zhang [101], and were, in turn, generalized within the Orlicz–Brunn–Minkowski inequality by Lutwak, Yang, and Zhang [105, 106]. For information on valuations in this setting, see [77].

A classical notion of addition on full dimensional convex bodies in $\mathbb{R}^n$ is Blaschke addition, which is defined using the sum of surface area measures of convex bodies and the solution of the classical Minkowski problem. The so-called *Blaschke valuations* were classified in [52]. For information on the corresponding question within the L^p Brunn–Minkowski theory, see [79].

The dual Brunn–Minkowski theory, established by Lutwak [96], is, in a certain way, dual to the classical theory. Star bodies replace convex bodies and radial addition (defined by the addition of radial functions) corresponds to Minkowski addition. Intersection bodies in the dual Brunn–Minkowski theory correspond to projection bodies in the classical theory. Intersection bodies were critical in the solution of the Busemann–Petty problem [97,139]. A classification of radial valuations and a characterization of the intersection body operator was established in [87]. Replacing radial addition by L^p radial addition leads to L^p radial valuations (see [51,54] for classification results).

Since convex bodies can be described by support functions and star bodies by radial functions, a natural extension of the results described above is a classification of valuations $\mathrm{Z} : \mathcal{K}^n \to F(\mathbb{R}^n)$, where $F(\mathbb{R}^n)$ is a suitable space of functions on $\mathbb{R}^n$. Such results were obtained by Li [74, 75] and by Li and Ma [78], where a characterization of the Laplace transform on convex bodies is established. Another way to describe convex bodies is by suitable measures and a classification of measure valued valuations was obtained by Haberl and Parapatits [56], where characterization results of surface area measures and of L^p surface area measures were established.

3. The Hadwiger theorem on convex bodies

The classical Steiner formula states that the volume of the outer parallel set of a convex body at distance $r > 0$ can be expressed as a polynomial in r of degree at most n. Using that the outer parallel set of $K \in \mathcal{K}^n$ at distance $r > 0$ is just the Minkowski sum of K and rB^n (the ball of radius r), we get

$$V_n(K + rB^n) = \sum_{j=0}^{n} r^{n-j} \kappa_{n-j} V_j(K)$$

for every $r > 0$, where κ_j is the j-dimensional volume of the unit ball in $\mathbb{R}^j$ (with the convention that $\kappa_0 := 1$). The coefficients $V_j(K)$ are known as the *intrinsic volumes* of K. Up to normalization and numbering, they coincide with the classical quermassintegrals. In particular, $V_{n-1}(K)$ is proportional to the surface area of K and $V_1(K)$ to its mean width (cf. [122]).

The celebrated Hadwiger theorem gives a characterization of intrinsic volumes and a complete classification of continuous, translation and rotation invariant valuations. For $n = 2$, it follows from the positive solution to Hilbert's Third Problem in this case. It was proved for $n = 3$ in [60] and then for general $n \geq 3$ in [61].

Theorem 3.1 (Hadwiger [61]). *A functional $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}$ is a continuous, translation and rotation invariant valuation if and only if there are $c_0, \dots, c_n \in \mathbb{R}$ such that*

$$\mathrm{Z}(K) = c_0 V_0(K) + \cdots + c_n V_n(K)$$

for every $K \in \mathcal{K}^n$.

The Hadwiger theorem leads to effortless proofs of numerous results in integral geometry and geometric probability (see [63,69]). An alternate proof of the Hadwiger theorem is due to Klain [67].

We will describe results on translation invariant and rotation equivariant valuations with values in tensor spaces and spaces of convex bodies. We remark that upper semicontinuous, translation and rotation invariant valuations were only classified in the planar case (see [81]).

3.1. Vector and tensor valuation

The first classification of vector valuations was established by Hadwiger and Schneider [64] using rotation equivariant valuations $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}^n$, that is, valuations such that

$$\mathrm{Z}(\phi K) = \phi\, \mathrm{Z}(K)$$

for all $\phi \in \mathrm{SO}(n)$ and $K \in \mathcal{K}^n$.

Theorem 3.2 (Hadwiger and Schneider [64]). *A function $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}^n$ is a continuous, translation covariant, rotation equivariant valuation if and only if there are $c_1, \dots, c_{n+1} \in \mathbb{R}$ such that*

$$\mathrm{Z}(K) = c_1\, \mathrm{M}_1^{1,0}(K) + \dots + c_{n+1}\, \mathrm{M}_{n+1}^{1,0}(K)$$

for every $K \in \mathcal{K}^n$.

Here $\mathrm{M}_i^{1,0}(K) := \Phi_i^{1,0}(K)$ are the *intrinsic vectors* of K (see (3.1) below) and see (3.2) for the definition of translation covariance.

The theorem by Hadwiger and Schneider was extended by Alesker [5,7] (based on [6]) to a classification of continuous, translation covariant, rotation equivariant tensor valuations on $\mathcal{K}^n$. Just as the intrinsic volumes can be obtained from the Steiner polynomial, the moment tensor (defined in (2.2)) satisfies the Steiner formula

$$\mathrm{M}^{p,0}(K + rB^n) = \sum_{j=0}^{n+p} r^{n+p-j} \kappa_{n+p-j} \sum_{k \geq 0} \Phi_{j-p+k}^{p-k,k}(K) \tag{3.1}$$

for $K \in \mathcal{K}^n$ and $r \geq 0$. The coefficients $\Phi_k^{p,s}(K)$ are called the *Minkowski tensors* of K (see [122, Section 5.4]). Recall that $\mathbb{T}^p(\mathbb{R}^n)$ is the space of symmetric p-tensors on $\mathbb{R}^n$ and let $Q \in \mathbb{T}^2(\mathbb{R}^n)$ be the metric tensor, that is, $Q(x, y) := \langle x, y \rangle$ for $x, y \in \mathbb{R}^n$.

Theorem 3.3 (Alesker [5]). *A function $\mathrm{Z} : \mathcal{K}^n \to \mathbb{T}^p(\mathbb{R}^n)$ is a continuous, translation covariant, rotation equivariant valuation if and only if Z can be written as linear combination of the symmetric tensor products $Q^l\, \Phi_k^{m,s}$ with $2l + m + s = p$.*

Here, a valuation $\mathrm{Z} : \mathcal{K}^n \to \mathbb{T}^p(\mathbb{R}^n)$ is called *translation covariant* if there exist associated functions $\mathrm{Z}^j : \mathcal{K}^n \to \mathbb{T}^j(\mathbb{R}^n)$ for $j = 0, \dots, p$ such that

$$\mathrm{Z}(K + y) = \sum_{j=0}^{p} \mathrm{Z}^{r-j}(K) \frac{y^j}{j!} \tag{3.2}$$

for all $y \in \mathbb{R}^n$ and $K \in \mathcal{K}^n$, where on the right side we sum over symmetric tensor products. We say that Z is G *equivariant* for a group G acting on $\mathbb{R}^n$ if

$$\mathrm{Z}(\phi K)(y_1, \dots, y_p) = \mathrm{Z}(K)(\phi^t y_1, \dots, \phi^t y_p)$$

for all $y_1, \dots, y_p \in \mathbb{R}^n$, all transformation $\phi \in G$, and all $K \in \mathcal{K}^n$, where ϕ^t is the transpose of ϕ.

For a classification of local tensor valuations, see [65], and for applications in various fields, including astronomy and material sciences, see [66].

3.2. Convex body valued valuations

An operator $\mathrm{Z} : \mathcal{K}^n \to \mathcal{K}^n$ is called *Minkowski additive* if

$$\mathrm{Z}(K + L) = \mathrm{Z}(K) + \mathrm{Z}(L)$$

for all $K, L \in \mathcal{K}^n$. Since $K + L = K \cup L + K \cap L$ for $K, L \in \mathcal{K}^n$ with $K \cup L \in \mathcal{K}^n$, it is easy to see that every Minkowski additive operator is a Minkowski valuation. While the first classification results for Minkowski valuations were established in [83], Schneider [120] earlier obtained the first classification results for rotation equivariant Minkowski additive operators. For continuous, translation invariant, rotation equivariant Minkowski valuations, so far no complete classification has been established. But the following representation is known to hold. Let $\mathcal{M}_{\mathrm{cen}}(\mathbb{S}^{n-1})$ and $C_{\mathrm{cen}}(\mathrm{S}^{n-1})$ denote the spaces of signed Borel measures and continuous functions on S^{n-1}, respectively, having their center of mass at the origin.

Theorem 3.4 (Schuster and Wannerer [123]). *If $\mathrm{Z} : \mathcal{K}^n \to \mathcal{K}^n$ is a continuous, translation invariant, rotation equivariant Minkowski valuation, then there are uniquely determined constants $c_0, c_n \geq 0$ and $\mathrm{SO}(n-1)$ invariant measures $\mu_i \in \mathcal{M}_{\mathrm{cen}}(\mathbb{S}^{n-1})$ for $1 \leq i \leq n-2$, as well as an $\mathrm{SO}(n-1)$ invariant function $\zeta_{n-1} \in C_{\mathrm{cen}}(\mathrm{S}^{n-1})$ such that*

$$h(\mathrm{Z}\, K, \cdot) = c_0 + \sum_{i=1}^{n-2} S_i(K, \cdot) * \mu_i + S_{n-1}(K, \cdot) * \zeta_{n-1} + c_n V_n(K)$$

for every $K \in \mathcal{K}^n$.

The Borel measures $S_i(K, \cdot)$ on $\mathbb{S}^{n-1}$ are Aleksandrov's area measures (see [122]) of $K \in \mathcal{K}^n$. The convolution of functions and measures on $\mathbb{S}^{n-1}$ is induced from the group $\mathrm{SO}(n)$ by identifying $\mathbb{S}^{n-1}$ with the homogeneous space $\mathrm{SO}(n)/\,\mathrm{SO}(n-1)$ (see [123]). The above representation formula has to be read in the sense of equality of measures and $h(\mathrm{Z}\, K, \cdot)$ is identified with the measure with this density.

4. More on invariant valuations on convex bodies

Translation invariant valuations on polytopes were classified using simplicity or mild regularity assumptions. Hadwiger [62] established a complete classification of simple, weakly continuous, translation invariant valuations on convex polytopes. Here, informally, a valuation is *weakly continuous* if it is continuous under parallel displacements of the facets of a polytope. Hadwiger's result was extended by McMullen [112] to the following result.

Theorem 4.1 (McMullen [112]). *A functional* $\mathrm{Z}: \mathcal{P}^n \to \mathbb{R}$ *is a weakly continuous, translation invariant valuation if and only if*

$$\mathrm{Z}(P) = \sum_{j=0}^{n} \sum_{F \in \mathcal{F}_j(P)} \mathrm{Y}_j\big(N(P,F)\big) V_j(F)$$

for every $P \in \mathcal{P}^n$ *where* $\mathrm{Y}_j : \mathcal{Q}^{n-j} \to \mathbb{R}$ *is a simple valuation.*

Here, $\mathcal{F}_j(P)$ is the set of j-dimensional faces of P and $N(P,F)$ is the normal cone to P at F while $\mathcal{Q}^k$ is the system of all closed polyhedral convex cones of dimension at most k. We remark that valuations on convex polyhedral cones (or, equivalently, on spherical polytopes) are not yet well understood and the problems to classify simple, rotation invariant valuations on spherical polytopes and on spherical convex bodies are open on spheres of dimension ≥ 3 (even if continuity is assumed). Kusejko and Parapatits [72] extended Hadwiger's result and established a complete classification of simple, translation invariant valuations on polytopes using Cauchy functions.

Hadwiger [63] proved that simple, continuous, translation invariant valuations on $\mathcal{K}^n$ have a *homogeneous decomposition*. His result was extended by McMullen [110].

Theorem 4.2 (McMullen [110]). *If* $\mathrm{Z}: \mathcal{K}^n \to \mathbb{R}$ *is a continuous and translation invariant valuation, then*

$$\mathrm{Z} = \mathrm{Z}_0 + \cdots + \mathrm{Z}_n,$$

where $\mathrm{Z}_j : \mathcal{K}^n \to \mathbb{R}$ *is a continuous, translation invariant valuation that is homogeneous of degree* j*.*

It is easy to see that every continuous, translation invariant valuation that is homogeneous of degree 0 is a multiple of the Euler characteristic. For the degrees of homogeneity $j = n$ and $j = n - 1$, the following results hold.

Theorem 4.3 (Hadwiger [63]). *A functional* $\mathrm{Z}: \mathcal{P}^n \to \mathbb{R}$ *is a translation invariant valuation that is homogeneous of degree n if and only if there is* $c \in \mathbb{R}$ *such that*

$$\mathrm{Z}(P) = c\, V_n(P)$$

for every $P \in \mathcal{P}^n$*.*

Theorem 4.4 (McMullen [111]). *A functional* $\mathrm{Z}\colon \mathcal{K}^n \to \mathbb{R}$ *is a continuous and translation invariant valuation which is homogeneous of degree* $(n-1)$ *if and only if there is* $\zeta \in C(\mathbb{S}^{n-1})$ *such that*

$$\mathrm{Z}(K) = \int_{\mathbb{S}^{n-1}} \zeta(y)\,\mathrm{d}S_{n-1}(K,y)$$

for every $K \in \mathcal{K}^n$*. The function* ζ *is uniquely determined up to addition of the restriction of a linear function.*

Continuous, translation invariant valuations that are homogeneous of degree 1 were classified by Goodey and Weil [49].

While a complete classification of continuous, translation invariant valuations on $\mathcal{K}^n$ is out of reach, Alesker [9] proved the following result.

Theorem 4.5 (Alesker [9]). *For* $0 \le j \le n$*, linear combinations of the valuations*

$$\left\{K \mapsto V\big(K[j], K_1, \dots, K_{n-j}\big) : K_1, \dots, K_{n-j} \in \mathcal{K}^n\right\}$$

are dense in the space of continuous and translation invariant valuations that are homogeneous of degree j*.*

Here, $V(K[j], K_1, \dots, K_{n-j})$ is the mixed volume of K taken j times and $K_1, \dots, K_{n-j}$ while the topology on the space of continuous, translation invariant valuations is induced by the norm $\|\,\mathrm{Z}\,\| := \sup\{|\,\mathrm{Z}(K)| : K \in \mathcal{K}^n,\ K \subseteq B^n\}$. Alesker's result confirms a conjecture by McMullen [111] and is based on Alesker's so-called irreducibility theorem, which was proved in [9] and which has far-reaching consequences.

For simple valuations, the following complete classification was established by Klain and Schneider.

Theorem 4.6 (Klain [67], Schneider [121]). *A functional* $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}$ *is a simple, continuous, translation invariant valuation if and only if there are* $c \in \mathbb{R}$ *and an odd function* $\zeta \in C(\mathbb{S}^{n-1})$ *such that*

$$\mathrm{Z}(K) = \int_{\mathbb{S}^{n-1}} \zeta(y)\,\mathrm{d}S_{n-1}(K,y) + c\,V_n(K)$$

for every $K \in \mathcal{K}^n$*. The function* ζ *is uniquely determined up to addition of the restriction of a linear function.*

Klain [67] used his classification of simple valuations in his proof of the Hadwiger theorem. For an alternate proof of Theorem 4.6, see [72].

A valuation $\mathrm{Z} : \mathcal{K}^n \to \mathbb{R}$ is called *translatively polynomial* if $x \mapsto \mathrm{Z}(P + x)$ is a polynomial in the coordinates of $x \in \mathbb{R}^n$ for all $K \in \mathcal{K}^n$. Alesker [6] established a complete classification of continuous, translatively polynomial, rotation invariant valuations on $\mathcal{K}^n$. Theorem 3.3 is the version of this result for tensor valuations.

Classification results for continuous, translation invariant valuations that are invariant under indefinite orthogonal groups were established by Alesker and Faifman [16] and Bernig and Faifman [23]. For subgroups of the orthogonal group $\mathrm{O}(n)$, the following result holds.

Theorem 4.7 (Alesker [8,12]). *For a compact subgroup G of $\mathrm{O}(n)$, the linear space of continuous, translation and G invariant valuations on $\mathcal{K}^n$ is finite dimensional if and only if G acts transitively on $\mathbb{S}^{n-1}$.*

As the classification of the such subgroups G is known, it was a natural task (which was already proposed in [8]) to find bases for spaces of G invariant valuations (see [9–11,13,19,20,22,24–27] for results on real valued valuations and [31,136] for results on tensor and measure valued valuations).

5. Affine valuations on function spaces

We describe classification results for valuations on function spaces that correspond to the results in Section 2. Let $F(\mathbb{R}^n)$ be a space of functions $f : \mathbb{R}^n \to [-\infty, \infty]$ and let G be a subgroup of $\mathrm{GL}(n)$. An operator $\mathrm{Z} : F(\mathbb{R}^n) \to \mathbb{A}$ is G *invariant* if

$$\mathrm{Z}(f \circ \phi^{-1}) = \mathrm{Z}(f)$$

for all $\phi \in G$ and $f \in F(\mathbb{R}^n)$. If G acts on $\mathbb{A}$, we say that an operator $\mathrm{Z} : F(\mathbb{R}^n) \to \mathbb{A}$ is G *contravariant* if for some $q \in \mathbb{R}$,

$$\mathrm{Z}(f \circ \phi^{-1}) = |\det \phi|^q \phi^{-t}\, \mathrm{Z}(f)$$

for all $\phi \in G$ and $f \in F(\mathbb{R}^n)$. It is G *equivariant* if for some $q \in \mathbb{R}$,

$$\mathrm{Z}(f \circ \phi^{-1}) = |\det \phi|^q \phi\, \mathrm{Z}(f)$$

for all $\phi \in G$ and $f \in F(\mathbb{R}^n)$. It is called *homogeneous* if for some $q \in \mathbb{R}$,

$$\mathrm{Z}(sf) = |s|^q\, \mathrm{Z}(f)$$

for all $s \in \mathbb{R}$ and $f \in F(\mathbb{R}^n)$ such that $sf \in F(\mathbb{R}^n)$. An operator is called *affinely contravariant* if it is translation invariant, $\mathrm{GL}(n)$ contravariant, and homogeneous.

5.1. Valuations on Sobolev spaces

For $p \geq 1$, let $W^{1,p}(\mathbb{R}^n)$ be the Sobolev space of functions belonging to $L^p(\mathbb{R}^n)$ whose distributional first-order derivatives belong to $L^p(\mathbb{R}^n)$.

The following result corresponds to Theorem 2.11. Let $\mathcal{K}^n_c$ be the set of origin-symmetric convex bodies in $\mathbb{R}^n$. Let $n \geq 3$.

Theorem 5.1 ([90]). *An operator* $\mathrm{Z} : W^{1,1}(\mathbb{R}^n) \to \mathcal{K}^n_c$ *is a continuous, affinely contravariant Minkowski valuation if and only if there is* $c \geq 0$ *such that*

$$\mathrm{Z}(f) = c\,\Pi\,\langle f\rangle$$

for every $f \in W^{1,1}(\mathbb{R}^n)$.

Here, for $f \in W^{1,1}(\mathbb{R}^n)$, the *LYZ body* $\langle f\rangle$ is defined by Lutwak, Yang, and Zhang [104] as the unique origin-symmetric convex body in $\mathbb{R}^n$ such that

$$\int_{\mathbb{S}^{n-1}} \zeta(y)\,\mathrm{d}S_{n-1}\big(\langle f\rangle, y\big) = \int_{\mathbb{R}^n} \zeta\big(\nabla f(x)\big)\,\mathrm{d}x \tag{5.1}$$

for every even continuous function $\zeta : \mathbb{R}^n \to \mathbb{R}$ that is homogeneous of degree 1. Equation (5.1) is a functional version of the classical even Minkowski problem.

Combined with (5.1), it follows from the definition of projection bodies and surface area measures that for $f \in W^{1,1}(\mathbb{R}^n)$ and $y \in \mathbb{S}^{n-1}$,

$$h\big(\Pi\,\langle f\rangle, y\big) = \frac{1}{2}\int_{\mathbb{R}^n} \big|\langle \nabla f(x), y\rangle\big|\,\mathrm{d}x.$$

We remark that the convex body $\langle f\rangle$ has proved to be critical in geometric analysis: the affine Sobolev–Zhang inequality [138] is a volume inequality for the polar body of $\Pi\,\langle f\rangle$, which strengthens and implies the Euclidean case of the classical Sobolev inequality, and it was proved in [104] that $\langle f\rangle$ describes the optimal Sobolev norm of $f \in W^{1,1}(\mathbb{R}^n)$. Tuo Wang [133] studied the LYZ operator $f \mapsto \langle f\rangle$ on the space of functions of bounded variation. Here, the LYZ operator is not a valuation anymore but Wang [134] established a characterization as an affinely covariant Blaschke semi-valuation.

The following classification of tensor valuation corresponds to Theorem 2.10 for $p = 2$. Let $n \geq 3$.

Theorem 5.2 ([89]). *An operator* $\mathrm{Z} : W^{1,2}(\mathbb{R}^n) \to \mathbb{T}^2(\mathbb{R}^n)$ *is a continuous, affinely contravariant valuation if and only if there is* $c \in \mathbb{R}$ *such that*

$$\mathrm{Z}(f) = c\,\mathrm{J}(f^2)$$

for every $f \in W^{1,2}(\mathbb{R}^n)$.

Here, we write $\mathrm{J}(h)$ for the *Fisher information matrix* of the weakly differentiable function $h : \mathbb{R}^n \to [0, \infty)$, that is, the $n \times n$ matrix with entries

$$\mathrm{J}_{ij}(h) := \int_{\mathbb{R}^n} \frac{\partial \log h(x)}{\partial x_i}\frac{\partial \log h(x)}{\partial x_j} h(x)\,\mathrm{d}x. \tag{5.2}$$

We remark that the Fisher information matrix plays an important role in information theory and statistics (see [45]). In general, Fisher information is a measure of the minimum error in the maximum likelihood estimate of a parameter in a distribution. The Fisher information matrix (5.2) describes such an error for a random vector of density h with respect to a location parameter.

For results on real valued valuations on Sobolev spaces, see [107].

5.2. Valuations on convex functions

We write $\operatorname{Conv}(\mathbb{R}^n)$ for the space of convex functions $u : \mathbb{R}^n \to (-\infty, \infty]$ that are lower semicontinuous and proper, that is, $u \not\equiv \infty$. We equip $\operatorname{Conv}(\mathbb{R}^n)$ and its subspaces with the topology induced by epi-convergence (see [119]). Let

$$\operatorname{Conv}_{\mathrm{coe}}(\mathbb{R}^n) := \big\{u \in \operatorname{Conv}(\mathbb{R}^n) : \lim_{|x|\to\infty} u(x) = \infty\big\}$$

be the space of *coercive*, convex functions, where $|x|$ is the Euclidean norm of $x \in \mathbb{R}^n$. The following result corresponds to Theorem 2.1.

Theorem 5.3 ([38]). *A functional* $\mathrm{Z} : \operatorname{Conv}_{\mathrm{coe}}(\mathbb{R}^n) \to [0, \infty)$ *is a continuous, translation and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are a continuous function* $\zeta_0 : \mathbb{R} \to [0, \infty)$ *and a continuous function* $\zeta_n : \mathbb{R} \to [0, \infty)$ *with finite* $(n-1)$*th moment such that*

$$\mathrm{Z}(u) = \zeta_0\big(\min_{x\in\mathbb{R}^n} u(x)\big) + \int_{\operatorname{dom} u} \zeta_n\big(u(x)\big)\,\mathrm{d}x$$

for every $u \in \operatorname{Conv}_{\mathrm{coe}}(\mathbb{R}^n)$.

Here, a function $\zeta : \mathbb{R} \to [0, \infty)$ has finite kth moment if $\int_0^\infty t^k \zeta(t)\,\mathrm{d}t < \infty$ and $\operatorname{dom} u$ is the *domain* of u, that is, $\operatorname{dom} u := \{x \in \mathbb{R}^n : u(x) < \infty\}$.

Let $\operatorname{Conv}(\mathbb{R}^n;\mathbb{R})$ be the space of finite valued convex functions, that is, of convex functions $u : \mathbb{R}^n \to \mathbb{R}$. We say that $u \in \operatorname{Conv}(\mathbb{R}^n)$ is *super-coercive* if

$$\lim_{|x|\to\infty} \frac{u(x)}{|x|} = \infty.$$

Let $\operatorname{Conv}_{\mathrm{sc}}(\mathbb{R}^n;\mathbb{R})$ be the space of *super-coercive*, finite valued, convex functions. The following result corresponds to Theorem 2.4.

Theorem 5.4 (Mussnig [114]). *A functional* $\mathrm{Z} : \operatorname{Conv}_{\mathrm{sc}}(\mathbb{R}^n;\mathbb{R}) \to [0, \infty)$ *is a continuous, translation and* $\mathrm{SL}(n)$ *invariant valuation if and only if there are a continuous* $\zeta_0 : \mathbb{R} \to [0, \infty)$, *a continuous* $\zeta_n : \mathbb{R} \to [0, \infty)$ *with finite* $(n-1)$*th moment, and a continuous* $\zeta_{-n} : \mathbb{R} \to [0, \infty)$ *whose support is bounded from above such that*

$$\mathrm{Z}(u) = \zeta_0\big(\min_{x\in\mathbb{R}^n} u(x)\big) + \int_{\mathbb{R}^n} \zeta_n\big(u(x)\big)\,\mathrm{d}x + \int_{\mathbb{R}^n} \zeta_{-n}\big(u(x)\big)\,\mathrm{d}\operatorname{MA}(u, x)$$

for every $u \in \operatorname{Conv}_{\mathrm{sc}}(\mathbb{R}^n;\mathbb{R})$.

Here, $\mathrm{MA}(u,\cdot)$ denotes the Monge–Ampère measure of u, which is also called the nth Hessian measure. See [113] for a result on coercive functions in $\mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$.

The following results correspond to Theorems 2.11 and 2.12. Let $n \geq 3$.

Theorem 5.5 ([37]). *An operator* $\mathrm{Z}:\mathrm{Conv}_{\mathrm{coe}}(\mathbb{R}^n)\to\mathcal{K}^n$ *is a continuous, monotone, translation invariant,* $\mathrm{SL}(n)$ *contravariant Minkowski valuation if and only if there is a continuous, decreasing* $\zeta:\mathbb{R}\to[0,\infty)$ *with finite* $(n-2)$*th moment such that*

$$\mathrm{Z}(u)=\Pi\,\langle\zeta\circ u\rangle$$

for every $u\in\mathrm{Conv}_{\mathrm{coe}}(\mathbb{R}^n)$.

For $u\in\mathrm{Conv}_{\mathrm{coe}}(\mathbb{R}^n)$ and suitable $\zeta\in C(\mathbb{R})$, define the *level set body* $[\zeta\circ u]\in\mathcal{K}^n$ by

$$h\big([\zeta\circ u],y\big):=\int_0^\infty h\big(\{\zeta\circ u\geq t\},y\big)\,\mathrm{d}t$$

for $y\in\mathbb{R}^n$. Hence the level set body is a Minkowski average of the level sets.

Theorem 5.6 ([37]). *An operator* $\mathrm{Z}:\mathrm{Conv}_{\mathrm{coe}}(\mathbb{R}^n)\to\mathcal{K}^n$ *is a continuous, monotone, translation invariant,* $\mathrm{SL}(n)$ *equivariant Minkowski valuation if and only if there is a continuous, decreasing* $\zeta:\mathbb{R}\to[0,\infty)$ *with finite integral over* $[0,\infty)$ *such that*

$$\mathrm{Z}(u)=\mathrm{D}\,[\zeta\circ u]$$

for every $u\in\mathrm{Conv}_{\mathrm{coe}}(\mathbb{R}^n)$.

We remark that the results in this section can be easily translated to classification results for valuations on log-concave functions. In this setting, the results on convex body valued valuations were strengthened by Mussnig [115].

6. The Hadwiger theorem on convex functions

We call a functional $\mathrm{Z}:\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)\to\mathbb{R}$ *epi-translation invariant* if

$$\mathrm{Z}(u\circ\tau^{-1}+c)=\mathrm{Z}(u)$$

for all translations $\tau:\mathbb{R}^n\to\mathbb{R}^n$ and $c\in\mathbb{R}$. Hence $\mathrm{Z}(u)$ is not changed by translations of the epi-graph of u. To state the Hadwiger theorem on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$, we need to define functional versions of the intrinsic volumes. Let $C_b((0,\infty))$ be the set of continuous functions on $(0,\infty)$ with bounded support. For $0\leq j\leq n-1$, let

$$D_j^n:=\Big\{\zeta\in C_b\big((0,\infty)\big):\lim_{s\to 0+}s^{n-j}\zeta(s)=0,\\ \lim_{s\to 0+}\int_s^\infty t^{n-j-1}\zeta(t)\,\mathrm{d}t\text{ exists and is finite}\Big\}.$$

In addition, let D^n_n be the set of functions $\zeta \in C_b((0,\infty))$ where $\lim_{s\to 0^+} \zeta(s)$ exists and is finite, and set $\zeta(0) := \lim_{s\to 0^+} \zeta(s)$.

Theorem 6.1 ([39]). *For $0 \le j \le n$ and $\zeta \in D^n_j$, there exists a unique, continuous, epi-translation and rotation invariant valuation $\mathrm{V}_{j,\zeta}\colon \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \to \mathbb{R}$ such that*

$$\mathrm{V}_{j,\zeta}(u) = \int_{\mathbb{R}^n} \zeta\big(|\nabla u(x)|\big)\big[\mathrm{D}^2 u(x)\big]_{n-j}\,\mathrm{d}x$$

for every $u \in \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \cap C^2_+(\mathbb{R}^n)$.

Here, $\mathrm{D}^2 u$ is the Hessian matrix of u and $[\mathrm{D}^2 u(x)]_k$ the kth elementary symmetric functions of the eigenvalues of $\mathrm{D}^2 u(x)$ (with the convention that $[\mathrm{D}^2 u(x)]_0 :\equiv 1$) while $C^2_+(\mathbb{R}^n)$ is the space of twice continuously differentiable functions with positive definite Hessian. We remark that $\mathrm{V}_{0,\zeta}$ is constant on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$.

The following result is the Hadwiger theorem on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$. Here, a functional $\mathrm{Z} : \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \to \mathbb{R}$ is said to be *rotation invariant* if $\mathrm{Z}(u \circ \vartheta^{-1}) = \mathrm{Z}(u)$ for every $\vartheta \in \mathrm{SO}(n)$. Let $n \ge 2$.

Theorem 6.2 ([39]). *A functional $\mathrm{Z} : \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \to \mathbb{R}$ is a continuous, epi-translation and rotation invariant valuation if and only if there are functions $\zeta_0 \in D^n_0, \dots, \zeta_n \in D^n_n$ such that*

$$\mathrm{Z}(u) = \mathrm{V}_{0,\zeta_0}(u) + \cdots + \mathrm{V}_{n,\zeta_n}(u)$$

for every $u \in \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$.

A comparison of Theorems 3.1 and 6.2 shows that for $0 \le j \le n$ and $\zeta \in D^n_j$, the functional $\mathrm{V}_{j,\zeta}$ plays a role corresponding to that of the jth intrinsic volume V_j. Hence, we call $\mathrm{V}_{j,\zeta}$ a jth *functional intrinsic volume* on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$. It is connected to the classical intrinsic volume by

$$\mathrm{V}_{j,\zeta}(\mathrm{I}_K) = c\,V_j(K)$$

for $K \in \mathcal{K}^n$ where I_K is the convex indicator function (that is, $\mathrm{I}_K(x) = 0$ for $x \in K$ and $\mathrm{I}_K(x) = \infty$ otherwise) and c depends only on j, n, and ζ (see [42]).

We call a functional $\mathrm{Z} : \mathrm{Conv}(\mathbb{R}^n; \mathbb{R}) \to \mathbb{R}$ *dually epi-translation invariant* if

$$\mathrm{Z}(v + \ell + c) = \mathrm{Z}(v)$$

for all linear functions $\ell : \mathbb{R}^n \to \mathbb{R}$ and $c \in \mathbb{R}$. Using the convex conjugate or Legendre transform of $u \in \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$, given by

$$u^*(y) := \sup_{x\in\mathbb{R}^n} \big(\langle x, y\rangle - u(x)\big)$$

for $y \in \mathbb{R}^n$, we see that $v \mapsto \mathrm{Z}(v)$ is dually epi-translation invariant on $\mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$ if and only if $u \mapsto \mathrm{Z}(u^*)$ is epi-translation invariant on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$. It was proved in [40] that Z is a continuous valuation on $\mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$ if and only if $\mathrm{Z}^* \colon \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \to \mathbb{R}$, defined by

$$\mathrm{Z}^*(u) := \mathrm{Z}(u^*),$$

is a continuous valuation on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$. This fact permits us to transfer results valid for valuations on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$ to results for valuations on $\mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$ and vice versa.

The following result is obtained from Theorem 6.1 by using convex conjugation.

Theorem 6.3 ([39]). *For $0 \le j \le n$ and $\zeta \in D_j^n$, the functional $\mathrm{V}_{j,\zeta}^* \colon \mathrm{Conv}(\mathbb{R}^n;\mathbb{R}) \to \mathbb{R}$ is a continuous, dually epi-translation and rotation invariant valuation such that*

$$\mathrm{V}_{j,\zeta}^*(v) = \int_{\mathbb{R}^n} \zeta\big(|x|\big)\big[\mathrm{D}^2 v(x)\big]_j \,\mathrm{d}x \tag{6.1}$$

for every $v \in \mathrm{Conv}(\mathbb{R}^n;\mathbb{R}) \cap C_+^2(\mathbb{R}^n)$.

Here, $\mathrm{V}_{j,\zeta}^*(v) := \mathrm{V}_{j,\zeta}(v^*)$ for $0 \le j \le n$ and $\zeta \in D_j^n$. Theorem 6.2 has the following dual version. Let $n \ge 2$.

Theorem 6.4 ([39]). *A functional $\mathrm{Z} : \mathrm{Conv}(\mathbb{R}^n;\mathbb{R}) \to \mathbb{R}$ is a continuous, dually epi-translation and rotation invariant valuation if and only if there are functions $\zeta_0 \in D_0^n$, $\dots, \zeta_n \in D_n^n$ such that*

$$\mathrm{Z}(v) = \mathrm{V}_{0,\zeta_0}^*(v) + \cdots + \mathrm{V}_{n,\zeta_n}^*(v)$$

for every $v \in \mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$.

For $\zeta \in D_j^n$, the functional $\mathrm{V}_{j,\zeta}^*$ is connected to the classical intrinsic volume by

$$\mathrm{V}_{j,\zeta}^*(h_K) = c V_j(K)$$

for $K \in \mathcal{K}^n$, where c depends only on j, n, and ζ (see [42]).

Applications of the Hadwiger theorem on convex functions including integral geometric formulas and additional representations of functional intrinsic volumes can be found in [42].

7. More on invariant valuations on function spaces

For continuous, epi-translation invariant valuations on $\mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$, the existence of a homogeneous decomposition corresponding to Theorem 4.2 was established in [41], that is, every such valuation is a linear combination of continuous, epi-translation invariant valuations that are epi-homogeneous of degree j and $0 \le j \le n$. Here Z is called *epi-homogeneous* of degree j if $\mathrm{Z}(u)$ is multiplied by t^j when the epi-graph of u is multiplied by $t > 0$. It is not difficult to see that every continuous, epi-translation invariant valuation that is epi-homogeneous of degree 0 is constant.

The following classification corresponding to Theorem 4.3 was established in [41].

Theorem 7.1 ([41]). *A functional* $\mathrm{Z}\colon \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n) \to \mathbb{R}$ *is an epi-translation invariant valuation that is epi-homogeneous of degree n if and only if there is* $\zeta \in C_c(\mathbb{R}^n)$ *such that*

$$\mathrm{Z}(u) = \int_{\operatorname{dom} u} \zeta\big(\nabla u(x)\big)\,\mathrm{d}x$$

for every $u \in \mathrm{Conv}_{\mathrm{sc}}(\mathbb{R}^n)$.

Here, $C_c(\mathbb{R}^n)$ is the space of continuous functions with compact support. The result corresponding to Theorem 7.1 on $\mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$ is stated next.

Theorem 7.2 ([41]). *A functional* $\mathrm{Z} : \mathrm{Conv}(\mathbb{R}^n;\mathbb{R}) \to \mathbb{R}$ *is a dually epi-translation invariant valuation that is homogeneous of degree n if and only if there is* $\zeta \in C_c(\mathbb{R}^n)$ *such that*

$$\mathrm{Z}(v) = \int_{\mathbb{R}^n} \zeta(x)\,\mathrm{d}\,\mathrm{MA}(v, x)$$

for every $v \in \mathrm{Conv}(\mathbb{R}^n;\mathbb{R})$.

See [41], for more information on homogeneous decompositions and why such results do not hold for many spaces of convex functions. For more results on valuations on convex functions, see [15, 34, 70, 71], and for results on valuations on quasi-concave functions, see [35, 36].

While formally not results for valuations on function spaces, classification results for valuations on star shaped sets in $\mathbb{R}^n$ were the motivation for some of the results on function spaces. Let $\mathcal{S}^n(\mathbb{R}^n)$ be the space of sets $S \subset \mathbb{R}^n$ which are star shaped with respect to the origin and whose radial functions $\rho(S,\cdot) : \mathbb{S}^{n-1} \to [0,\infty]$, given by

$$\rho(S, x) := \sup\{r \ge 0 : rx \in S\},$$

are in $L^n(\mathbb{S}^{n-1})$. Let $\mathcal{S}_0$ be the space of star bodies, that is, of star shaped sets with continuous radial functions. We remark that $\mathcal{S}_0^n$ is the space used in the dual Brunn–Minkowski theory (see [48, 96]). Note that union and intersection on $\mathcal{S}^n(\mathbb{R}^n)$ and on $\mathcal{S}_0^n$ correspond to the pointwise maximum and minimum for radial functions. We equip $\mathcal{S}^n(\mathbb{R}^n)$ with the topology induced by the L^n norm on $\mathbb{S}^{n-1}$ and $\mathcal{S}_0^n$ with the topology induced by the maximum norm.

Klain [68] established the following classification results on star shaped sets.

Theorem 7.3 (Klain [68]). *A functional* $\mathrm{Z} : \mathcal{S}^n(\mathbb{R}^n) \to \mathbb{R}$ *is a continuous, rotation invariant valuation with* $\mathrm{Z}(\{0\}) = 0$ *if and only if there is* $\zeta \in C([0,\infty))$ *with the properties that* $\zeta(0) = 0$ *and* $|\zeta(t)| \le c + d|t|^n$ *for all* $t \in \mathbb{R}$ *for some* $c, d \ge 0$ *such that*

$$\mathrm{Z}(S) = \int_{\mathbb{S}^{n-1}} \zeta\big(\rho(S,y)\big)\,\mathrm{d}y$$

for every $S \in \mathcal{S}^n(\mathbb{R}^n)$.

If the valuation Z in Theorem 7.3 is in addition positively homogeneous of degree p, then $\zeta(t) = ct^p$ with $c \in \mathbb{R}$ and $0 \le p \le n$ and hence Z is a dual mixed volume (as defined by Lutwak [96]).

Tsang [130] obtained classification results for valuations on $L^p(X,\mu)$, when X is a non-atomic measure space. Here we state a special case of his results that complements Theorem 7.3. Let $p \ge 1$.

Theorem 7.4 (Tsang [130]). *A functional* $\mathrm{Z} : L^p(\mathbb{R}^n) \to \mathbb{R}$ *is a continuous, translation invariant valuation that vanishes on the null function if and only if there is* $\zeta \in C(\mathbb{R})$ *with the property that* $|\zeta(t)| \le c|t|^p$ *for all* $t \in \mathbb{R}$ *for some* $c \ge 0$ *such that*

$$\mathrm{Z}(f) = \int_{\mathbb{R}^n} \zeta\big(f(x)\big)\,\mathrm{d}x$$

for every $f \in L^p(\mathbb{R}^n)$.

We remark that also Theorem 7.3 can be written as a classification result on the space of non-negative functions in $L^n(\mathbb{S}^{n-1})$ (also see [130]). For results on tensor and Minkowski valuations on L^p space, see [91, 116, 131].

Villanueva [132] obtained classification results for non-negative valuations on star bodies. In [127], Tradacete and Villanueva showed that a result corresponding to the classification from Theorem 7.3 is valid on $\mathcal{S}_0^n$. A complete classification on $\mathcal{S}_0^n$ is given in the following result.

Theorem 7.5 (Tradacete and Villanueva [128]). *A functional* $\mathrm{Z} : \mathcal{S}_0^n \to \mathbb{R}$ *is a continuous valuation if and only if there are a finite Borel measure* μ *on* $\mathbb{S}^{n-1}$ *and a function* $\zeta : [0,\infty) \times \mathbb{S}^{n-1} \to \mathbb{R}$ *that fulfills the strong Carathéodory condition with respect to* μ *such that*

$$\mathrm{Z}(S) = \int_{\mathbb{S}^{n-1}} \zeta\big(\rho(S,y),y\big)\,\mathrm{d}\mu(y)$$

for every $u \in \mathcal{S}_0^n$.

Here, we say that $\zeta : [0,\infty) \times \mathbb{S}^{n-1} \to \mathbb{R}$ fulfills the *strong Carathéodory condition* with respect to μ if $\zeta(s,\cdot)$ is Borel measurable for all $s \geq 0$ and $\zeta(\cdot, y)$ is continuous for μ almost every $y \in \mathbb{S}^{n-1}$, while for every $t > 0$ there is $\xi_t \in L^1(\mathbb{S}^{n-1}, \mu)$ such that $\zeta(s, y) \leq \xi_t(y)$ for $s < t$ and μ almost every $y \in \mathbb{S}^{n-1}$. We remark that Theorem 7.5 can be rewritten as a result on valuations on non-negative functions in $C(\mathbb{S}^{n-1})$.

Classification results for valuations on Lipschitz functions on $\mathbb{S}^{n-1}$ were obtained in [43, 44] and on Banach lattices in [129]. A Hadwiger theorem for valuations on definable functions was established in [18].

Funding. M. Ludwig was supported, in part, by the Austrian Science Fund (FWF): P 34446.

References

[1] J. Abardia, Difference bodies in complex vector spaces. *J. Funct. Anal.* **263** (2012), no. 11, 3588–3603 Zbl 1262.52012 MR 2984076

[2] J. Abardia, Minkowski valuations in a 2-dimensional complex vector space. *Int. Math. Res. Not. IMRN* **2015** (2015), no. 5, 1247–1262 Zbl 1320.52019 MR 3340354

[3] J. Abardia and A. Bernig, Projection bodies in complex vector spaces. *Adv. Math.* **227** (2011), no. 2, 830–846 Zbl 1217.52009 MR 2793024

[4] J. Abardia-Evéquoz, K. J. Böröczky, M. Domokos, and D. Kertész, SL$(m, \mathbb{C})$-equivariant and translation covariant continuous tensor valuations. *J. Funct. Anal.* **276** (2019), no. 11, 3325–3362 Zbl 1431.52019 MR 3944297

[5] S. Alesker, Continuous valuations on convex sets. *Geom. Funct. Anal.* **8** (1998), no. 2, 402–409 Zbl 0919.52010 MR 1616167

[6] S. Alesker, Continuous rotation invariant valuations on convex sets. *Ann. of Math. (2)* **149** (1999), no. 3, 977–1005 Zbl 0941.52002 MR 1709308

[7] S. Alesker, Description of continuous isometry covariant valuations on convex sets. *Geom. Dedicata* **74** (1999), no. 3, 241–248 Zbl 0935.52006 MR 1669363

[8] S. Alesker, On P. McMullen's conjecture on translation invariant valuations. *Adv. Math.* **155** (2000), no. 2, 239–263 Zbl 0971.52004 MR 1794712

[9] S. Alesker, Description of translation invariant valuations on convex sets with solution of P. McMullen's conjecture. *Geom. Funct. Anal.* **11** (2001), no. 2, 244–272 Zbl 0995.52001 MR 1837364

[10] S. Alesker, SU(2)-invariant valuations. In *Geometric Aspects of Functional Analysis*, pp. 21–29, Lecture Notes in Math. 1850, Springer, Berlin, 2004 Zbl 1064.52010 MR 2087147

[11] S. Alesker, Valuations on convex sets, non-commutative determinants, and pluripotential theory. *Adv. Math.* **195** (2005), no. 2, 561–595 Zbl 1078.52011 MR 2146354

[12] S. Alesker, Theory of valuations on manifolds: a survey. *Geom. Funct. Anal.* **17** (2007), no. 4, 1321–1341 Zbl 1132.52018 MR 2373020

[13] S. Alesker, Plurisubharmonic functions on the octonionic plane and Spin(9)-invariant valuations on convex sets. *J. Geom. Anal.* **18** (2008), no. 3, 651–686 Zbl 1165.32015 MR 2420758

[14] S. Alesker, *Introduction to the Theory of Valuations*. CBMS Reg. Conf. Ser. Math. 126, American Mathematical Society, Providence, RI, 2018 Zbl 1398.52001 MR 3820854

[15] S. Alesker, Valuations on convex functions and convex sets and Monge–Ampère operators. *Adv. Geom.* **19** (2019), no. 3, 313–322 Zbl 1445.52011 MR 3982569

[16] S. Alesker and D. Faifman, Convex valuations invariant under the Lorentz group. *J. Differential Geom.* **98** (2014), no. 2, 183–236 Zbl 1312.52007 MR 3238311

[17] S. Alesker and J. H. G. Fu, *Integral Geometry and Valuations*. Adv. Courses Math. CRM Barcelona, Birkhäuser/Springer, Basel, 2014 Zbl 1302.53004 MR 3380549

[18] Y. Baryshnikov, R. Ghrist, and M. Wright, Hadwiger's Theorem for definable functions. *Adv. Math.* **245** (2013), 573–586 Zbl 1286.52006 MR 3084438

[19] A. Bernig, A Hadwiger-type theorem for the special unitary group. *Geom. Funct. Anal.* **19** (2009), no. 2, 356–372 Zbl 1180.53076 MR 2545241

[20] A. Bernig, Integral geometry under G_2 and $\mathrm{Spin}(7)$. *Israel J. Math.* **184** (2011), 301–316 Zbl 1262.53067 MR 2823979

[21] A. Bernig, Algebraic integral geometry. In *Global Differential Geometry*, pp. 107–145, Springer Proc. Math. 17, Springer, Heidelberg, 2012 Zbl 1252.52002 MR 3289841

[22] A. Bernig, Invariant valuations on quaternionic vector spaces. *J. Inst. Math. Jussieu* **11** (2012), no. 3, 467–499 Zbl 1248.53059 MR 2931316

[23] A. Bernig and D. Faifman, Valuation theory of indefinite orthogonal groups. *J. Funct. Anal.* **273** (2017), no. 6, 2167–2247 Zbl 1373.52017 MR 3669033

[24] A. Bernig and J. H. G. Fu, Hermitian integral geometry. *Ann. of Math. (2)* **173** (2011), no. 2, 907–945 Zbl 1230.52014 MR 2776365

[25] A. Bernig and G. Solanes, Classification of invariant valuations on the quaternionic plane. *J. Funct. Anal.* **267** (2014), no. 8, 2933–2961 Zbl 1305.53076 MR 3255479

[26] A. Bernig and G. Solanes, Kinematic formulas on the quaternionic plane. *Proc. Lond. Math. Soc. (3)* **115** (2017), no. 4, 725–762 Zbl 1430.52017 MR 3716941

[27] A. Bernig and F. Voide, Spin-invariant valuations on the octonionic plane. *Israel J. Math.* **214** (2016), no. 2, 831–855 Zbl 1347.53065 MR 3544703

[28] U. Betke and M. Kneser, Zerlegungen und Bewertungen von Gitterpolytopen. *J. Reine Angew. Math.* **358** (1985), 202–208 Zbl 0567.52002 MR 797683

[29] W. Blaschke, *Differentialgeometrie II*. Springer, Berlin, 1923

[30] W. Blaschke, *Vorlesungen über Integralgeometrie. H. 2.* Teubner, Berlin, 1937 Zbl 0016.27703

[31] K. J. Böröczky, M. Domokos, and G. Solanes, Dimension of the space of unitary equivariant translation invariant tensor valuations. *J. Funct. Anal.* **280** (2021), no. 4, Paper No. 108862 Zbl 1472.52004 MR 4181164

[32] K. J. Böröczky and M. Ludwig, Valuations on lattice polytopes. In *Tensor Valuations and their Applications in Stochastic Geometry and Imaging*, pp. 213–234, Lecture Notes in Math. 2177, Springer, Cham, 2017 Zbl 1376.52022 MR 3702374

[33] K. J. Böröczky and M. Ludwig, Minkowski valuations on lattice polytopes. *J. Eur. Math. Soc. (JEMS)* **21** (2019), no. 1, 163–197 Zbl 06997332 MR 3880207

[34] L. Cavallina and A. Colesanti, Monotone valuations on the space of convex functions. *Anal. Geom. Metr. Spaces* **3** (2015), no. 1, 167–211 Zbl 1321.26027 MR 3377373

[35] A. Colesanti and N. Lombardi, Valuations on the space of quasi-concave functions. In *Geometric Aspects of Functional Analysis*, pp. 71–105, Lecture Notes in Math. 2169, Springer, Cham, 2017 Zbl 1375.52010 MR 3645116

[36] A. Colesanti, N. Lombardi, and L. Parapatits, Translation invariant valuations on quasi-concave functions. *Studia Math.* **243** (2018), no. 1, 79–99 Zbl 1471.26005 MR 3803252

[37] A. Colesanti, M. Ludwig, and F. Mussnig, Minkowski valuations on convex functions. *Calc. Var. Partial Differential Equations* **56** (2017), no. 6, Paper No. 162 Zbl 1400.52014 MR 3715395

[38] A. Colesanti, M. Ludwig, and F. Mussnig, Valuations on convex functions. *Int. Math. Res. Not. IMRN* **2019** (2019), no. 8, 2384–2410 Zbl 1436.52014 MR 3942165

[39] A. Colesanti, M. Ludwig, and F. Mussnig, A Hadwiger theorem on convex functions. I. 2020, arXiv:2009.03702

[40] A. Colesanti, M. Ludwig, and F. Mussnig, Hessian valuations. *Indiana Univ. Math. J.* **69** (2020), no. 4, 1275–1315 Zbl 1445.26011 MR 4124129

[41] A. Colesanti, M. Ludwig, and F. Mussnig, A homogeneous decomposition theorem for valuations on convex functions. *J. Funct. Anal.* **279** (2020), no. 5, 108573, 25 Zbl 1446.26010 MR 4097279

[42] A. Colesanti, M. Ludwig, and F. Mussnig, A Hadwiger theorem on convex functions. II. 2021, arXiv:2109.09434

[43] A. Colesanti, D. Pagnini, P. Tradacete, and I. Villanueva, A class of invariant valuations on Lip(S^{n-1}). *Adv. Math.* **366** (2020), 107069, 37 Zbl 1441.52013 MR 4070303

[44] A. Colesanti, D. Pagnini, P. Tradacete, and I. Villanueva, Continuous valuations on the space of Lipschitz functions on the sphere. *J. Funct. Anal.* **280** (2021), no. 4, Paper No. 108873 Zbl 1462.26004 MR 4181166

[45] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. 2nd edn., Wiley-Interscience, Hoboken, NJ, 2006 Zbl 1140.94001 MR 2239987

[46] M. Dehn, Ueber den Rauminhalt. *Math. Ann.* **55** (1901), no. 3, 465–478 Zbl 32.0486.01 MR 1511157

[47] L. Fejes Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*. Die Grundlehren der mathematischen Wissenschaften 65, Springer, Berlin, 1972 Zbl 0229.52009 MR 0353117

[48] R. J. Gardner, *Geometric Tomography*. 2nd edn., Encyclopedia Math. Appl. 58, Cambridge University Press, New York, 2006 Zbl 1102.52002 MR 2251886

[49] P. Goodey and W. Weil, Distributions and valuations. *Proc. London Math. Soc. (3)* **49** (1984), no. 3, 504–516 Zbl 0526.52004 MR 759301

[50] P. M. Gruber, *Convex and Discrete Geometry*. Grundlehren Math. Wiss. 336, Springer, Berlin, 2007 Zbl 1139.52001 MR 2335496

[51] C. Haberl, Star body valued valuations. *Indiana Univ. Math. J.* **58** (2009), no. 5, 2253–2276 Zbl 1183.52003 MR 2583498

[52] C. Haberl, Blaschke valuations. *Amer. J. Math.* **133** (2011), no. 3, 717–751 Zbl 1229.52003 MR 2808330

[53] C. Haberl, Minkowski valuations intertwining with the special linear group. *J. Eur. Math. Soc. (JEMS)* **14** (2012), no. 5, 1565–1597 Zbl 1270.52018 MR 2966660

[54] C. Haberl and M. Ludwig, A characterization of L_p intersection bodies. *Int. Math. Res. Not. IMRN* **2006** (2006), Article No. 10548 Zbl 1115.52006 MR 2250020

[55] C. Haberl and L. Parapatits, The centro-affine Hadwiger theorem. *J. Amer. Math. Soc.* **27** (2014), no. 3, 685–705 Zbl 1319.52006 MR 3194492

[56] C. Haberl and L. Parapatits, Valuations and surface area measures. *J. Reine Angew. Math.* **687** (2014), 225–245 Zbl 1295.52018 MR 3176613

[57] C. Haberl and L. Parapatits, Moments and valuations. *Amer. J. Math.* **138** (2016), no. 6, 1575–1603 Zbl 1368.52010 MR 3595495

[58] C. Haberl and L. Parapatits, Centro-affine tensor valuations. *Adv. Math.* **316** (2017), 806–865 Zbl 1373.52018 MR 3672921

[59] C. Haberl and F. E. Schuster, General L_p affine isoperimetric inequalities. *J. Differential Geom.* **83** (2009), no. 1, 1–26 Zbl 1185.52005 MR 2545028

[60] H. Hadwiger, Beweis eines Funktionalsatzes für konvexe Körper. *Abh. Math. Sem. Univ. Hamburg* **17** (1951), 69–76 Zbl 0042.16402 MR 41468

[61] H. Hadwiger, Additive Funktionale k-dimensionaler Eikörper. I. *Arch. Math.* **3** (1952), 470–478 Zbl 0049.12202 MR 55707

[62] H. Hadwiger, Translationsinvariante, additive und schwachstetige Polyederfunktionale. *Arch. Math. (Basel)* **3** (1952), 387–394 Zbl 0048.28801 MR 54699

[63] H. Hadwiger, *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*. Springer, Berlin, 1957 Zbl 0078.35703 MR 0102775

[64] H. Hadwiger and R. Schneider, Vektorielle Integralgeometrie. *Elem. Math.* **26** (1971), 49–57 Zbl 0216.44003 MR 283737

[65] D. Hug and R. Schneider, Local tensor valuations. *Geom. Funct. Anal.* **24** (2014), no. 5, 1516–1564 Zbl 1366.52004 MR 3261633

[66] E. B. V. Jensen and M. Kiderlen, Rotation invariant valuations. In *Tensor Valuations and Their Applications in Stochastic Geometry and Imaging*, pp. 185–212, Lecture Notes in Math. 2177, Springer, Cham, 2017 Zbl 1369.52021 MR 3702373

[67] D. A. Klain, A short proof of Hadwiger's characterization theorem. *Mathematika* **42** (1995), no. 2, 329–339 Zbl 0835.52010 MR 1376731

[68] D. A. Klain, Invariant valuations on star-shaped sets. *Adv. Math.* **125** (1997), no. 1, 95–113 Zbl 0889.52007 MR 1427802

[69] D. A. Klain and G.-C. Rota, *Introduction to Geometric Probability*. Lezioni Lincee, Cambridge University Press, Cambridge, 1997 Zbl 0896.60004 MR 1608265

[70] J. Knoerr, Smooth valuations on convex functions. 2021, arXiv:2006.12933v3

[71] J. Knoerr, The support of dually epi-translation invariant valuations on convex functions. *J. Funct. Anal.* **281** (2021), no. 5, Paper No. 109059 Zbl 07456688 MR 4252807

[72] K. Kusejko and L. Parapatits, A valuation-theoretic approach to translative-equidecomposability. *Adv. Math.* **297** (2016), 174–195 Zbl 1346.52005 MR 3498797

[73] K. Leichtweiß, Über einige Eigenschaften der Affinoberfläche beliebiger konvexer Körper. *Results Math.* **13** (1988), no. 3-4, 255–282 Zbl 0645.53004 MR 941335

[74] J. Li, Affine function-valued valuations. *Int. Math. Res. Not. IMRN* **2020** (2020), no. 22, 8197–8233 Zbl 1459.52011 MR 4216687

[75] J. Li, SL(n) covariant function-valued valuations. *Adv. Math.* **377** (2021), Paper No. 107462 Zbl 1458.52005 MR 4186005

[76] J. Li and G. Leng, L_p Minkowski valuations on polytopes. *Adv. Math.* **299** (2016), 139–173 Zbl 1352.52018 MR 3519466

[77] J. Li and G. Leng, Orlicz valuations. *Indiana Univ. Math. J.* **66** (2017), no. 3, 791–819 Zbl 1386.52014 MR 3663326

[78] J. Li and D. Ma, Laplace transforms and valuations. *J. Funct. Anal.* **272** (2017), no. 2, 738–758 Zbl 1353.44001 MR 3571907

[79] J. Li, S. Yuan, and G. Leng, L_p-Blaschke valuations. *Trans. Amer. Math. Soc.* **367** (2015), no. 5, 3161–3187 Zbl 1327.52025 MR 3314805

[80] M. Ludwig, A characterization of affine length and asymptotic approximation of convex discs. *Abh. Math. Sem. Univ. Hamburg* **69** (1999), 75–88 Zbl 0954.52002 MR 1722923

[81] M. Ludwig, Upper semicontinuous valuations on the space of convex discs. *Geom. Dedicata* **80** (2000), no. 1-3, 263–279 Zbl 0960.52002 MR 1762513

[82] M. Ludwig, Moment vectors of polytopes. *Rend. Circ. Mat. Palermo (2) Suppl.* (2002), no. 70, part II, 123–138 Zbl 1113.52031 MR 1962589

[83] M. Ludwig, Projection bodies and valuations. *Adv. Math.* **172** (2002), no. 2, 158–168 Zbl 1019.52003 MR 1942402

[84] M. Ludwig, Valuations of polytopes containing the origin in their interiors. *Adv. Math.* **170** (2002), no. 2, 239–256 Zbl 1015.52012 MR 1932331

[85] M. Ludwig, Ellipsoids and matrix-valued valuations. *Duke Math. J.* **119** (2003), no. 1, 159–188 Zbl 1033.52012 MR 1991649

[86] M. Ludwig, Minkowski valuations. *Trans. Amer. Math. Soc.* **357** (2005), no. 10, 4191–4213 Zbl 1077.52005 MR 2159706

[87] M. Ludwig, Intersection bodies and valuations. *Amer. J. Math.* **128** (2006), no. 6, 1409–1428 Zbl 1115.52007 MR 2275906

[88] M. Ludwig, Minkowski areas and valuations. *J. Differential Geom.* **86** (2010), no. 1, 133–161 Zbl 1215.52004 MR 2772547

[89] M. Ludwig, Fisher information and matrix-valued valuations. *Adv. Math.* **226** (2011), no. 3, 2700–2711 Zbl 1274.62064 MR 2739790

[90] M. Ludwig, Valuations on Sobolev spaces. *Amer. J. Math.* **134** (2012), no. 3, 827–842 Zbl 1255.52013 MR 2931225

[91] M. Ludwig, Covariance matrices and valuations. *Adv. in Appl. Math.* **51** (2013), no. 3, 359–366 Zbl 1303.62016 MR 3084504

[92] M. Ludwig and M. Reitzner, A characterization of affine surface area. *Adv. Math.* **147** (1999), no. 1, 138–172 Zbl 0947.52003 MR 1725817

[93] M. Ludwig and M. Reitzner, A classification of SL(n) invariant valuations. *Ann. of Math. (2)* **172** (2010), no. 2, 1219–1267 Zbl 1223.52007 MR 2680490

[94] M. Ludwig and M. Reitzner, SL(n) invariant valuations on polytopes. *Discrete Comput. Geom.* **57** (2017), no. 3, 571–581 Zbl 1369.52019 MR 3614772

[95] M. Ludwig and L. Silverstein, Tensor valuations on lattice polytopes. *Adv. Math.* **319** (2017), 76–110 Zbl 1390.52023 MR 3695869

[96] E. Lutwak, Dual mixed volumes. *Pacific J. Math.* **58** (1975), no. 2, 531–538 Zbl 0273.52007 MR 380631

[97] E. Lutwak, Intersection bodies and dual mixed volumes. *Adv. in Math.* **71** (1988), no. 2, 232–261 Zbl 0657.52002 MR 963487

[98] E. Lutwak, Extended affine surface area. *Adv. Math.* **85** (1991), no. 1, 39–68 Zbl 0727.53016 MR 1087796

[99] E. Lutwak, The Brunn–Minkowski–Firey theory. I. Mixed volumes and the Minkowski problem. *J. Differential Geom.* **38** (1993), no. 1, 131–150 Zbl 0788.52007 MR 1231704

[100] E. Lutwak, The Brunn–Minkowski–Firey theory. II. Affine and geominimal surface areas. *Adv. Math.* **118** (1996), no. 2, 244–294 Zbl 0853.52005 MR 1378681

[101] E. Lutwak, D. Yang, and G. Zhang, L_p affine isoperimetric inequalities. *J. Differential Geom.* **56** (2000), no. 1, 111–132 Zbl 1034.52009 MR 1863023

[102] E. Lutwak, D. Yang, and G. Zhang, A new ellipsoid associated with convex bodies. *Duke Math. J.* **104** (2000), no. 3, 375–390 Zbl 0974.52008 MR 1781476

[103] E. Lutwak, D. Yang, and G. Zhang, The Cramer–Rao inequality for star bodies. *Duke Math. J.* **112** (2002), no. 1, 59–81 Zbl 1021.52008 MR 1890647

[104] E. Lutwak, D. Yang, and G. Zhang, Optimal Sobolev norms and the L^p Minkowski problem. *Int. Math. Res. Not. IMRN* **2006** (2006), Article No. 62987 Zbl 1110.46023 MR 2211138

[105] E. Lutwak, D. Yang, and G. Zhang, Orlicz centroid bodies. *J. Differential Geom.* **84** (2010), no. 2, 365–387 Zbl 1206.49050 MR 2652465

[106] E. Lutwak, D. Yang, and G. Zhang, Orlicz projection bodies. *Adv. Math.* **223** (2010), no. 1, 220–242 Zbl 1437.52006 MR 2563216

[107] D. Ma, Real-valued valuations on Sobolev spaces. *Sci. China Math.* **59** (2016), no. 5, 921–934 Zbl 1338.46019 MR 3484491

[108] D. Ma, Moment matrices and $\mathrm{SL}(n)$ equivariant valuations on polytopes. *Int. Math. Res. Not. IMRN* **2021** (2021), no. 14, 10469–10489 Zbl 07456026 MR 4285727

[109] D. Ma and W. Wang, LYZ matrices and $\mathrm{SL}(n)$ contravariant valuations on polytopes. *Canad. J. Math.* **73** (2021), no. 2, 383–398 Zbl 1471.52013 MR 4230379

[110] P. McMullen, Valuations and Euler-type relations on certain classes of convex polytopes. *Proc. London Math. Soc. (3)* **35** (1977), no. 1, 113–135 Zbl 0353.52001 MR 448239

[111] P. McMullen, Continuous translation-invariant valuations on the space of compact convex sets. *Arch. Math. (Basel)* **34** (1980), no. 4, 377–384 Zbl 0424.52003 MR 593954

[112] P. McMullen, Weakly continuous valuations on convex polytopes. *Arch. Math. (Basel)* **41** (1983), no. 6, 555–564 Zbl 0526.52003 MR 731639

[113] F. Mussnig, Volume, polar volume and Euler characteristic for convex functions. *Adv. Math.* **344** (2019), 340–373 Zbl 1429.52019 MR 3897436

[114] F. Mussnig, $\mathrm{SL}(n)$ invariant valuations on super-coercive convex functions. *Canad. J. Math.* **73** (2021), no. 1, 108–130 Zbl 1470.26018 MR 4201535

[115] F. Mussnig, Valuations on log-concave functions. *J. Geom. Anal.* **31** (2021), no. 6, 6427–6451 Zbl 1475.26011 MR 4267652

[116] M. Ober, L_p-Minkowski valuations on L^q-spaces. *J. Math. Anal. Appl.* **414** (2014), no. 1, 68–87 Zbl 1331.52022 MR 3165294

[117] L. Parapatits, $\mathrm{SL}(n)$-contravariant L_p-Minkowski valuations. *Trans. Amer. Math. Soc.* **366** (2014), no. 3, 1195–1211 Zbl 1286.52007 MR 3145728

[118] L. Parapatits, $\mathrm{SL}(n)$-covariant L_p-Minkowski valuations. *J. Lond. Math. Soc. (2)* **89** (2014), no. 2, 397–414 Zbl 1296.52010 MR 3188625

[119] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Grundlehren Math. Wiss. 317, Springer, Berlin, 1998 Zbl 0888.49001 MR 1491362

[120] R. Schneider, Equivariant endomorphisms of the space of convex bodies. *Trans. Amer. Math. Soc.* **194** (1974), 53–78 Zbl 0287.52004 MR 353147

[121] R. Schneider, Simple valuations on convex bodies. *Mathematika* **43** (1996), no. 1, 32–39 Zbl 0864.52009 MR 1401706

[122] R. Schneider, *Convex Bodies: the Brunn–Minkowski Theory*. expanded edn., Encyclopedia Math. Appl. 151, Cambridge University Press, Cambridge, 2014 Zbl 1287.52001 MR 3155183

[123] F. Schuster and T. Wannerer, Minkowski valuations and generalized valuations. *J. Eur. Math. Soc. (JEMS)* **20** (2018), no. 8, 1851–1884 Zbl 1398.52018 MR 3854893

[124] F. E. Schuster and T. Wannerer, GL(n) contravariant Minkowski valuations. *Trans. Amer. Math. Soc.* **364** (2012), no. 2, 815–826 Zbl 1246.52009 MR 2846354

[125] C. Schütt, On the affine surface area. *Proc. Amer. Math. Soc.* **118** (1993), no. 4, 1213–1218 Zbl 0784.52004 MR 1181173

[126] C. Schütt and E. Werner, The convex floating body. *Math. Scand.* **66** (1990), no. 2, 275–290 Zbl 0739.52008 MR 1075144

[127] P. Tradacete and I. Villanueva, Radial continuous valuations on star bodies. *J. Math. Anal. Appl.* **454** (2017), no. 2, 995–1018 Zbl 1368.52011 MR 3658809

[128] P. Tradacete and I. Villanueva, Continuity and representation of valuations on star bodies. *Adv. Math.* **329** (2018), 361–391 Zbl 1400.52015 MR 3783417

[129] P. Tradacete and I. Villanueva, Valuations on Banach lattices. *Int. Math. Res. Not. IMRN* **2020** (2020), no. 1, 287–319 Zbl 1476.46032 MR 4050568

[130] A. Tsang, Valuations on L^p-spaces. *Int. Math. Res. Not. IMRN* **2010** (2010), no. 20, 3993–4023 Zbl 1211.52013 MR 2738348

[131] A. Tsang, Minkowski valuations on L^p-spaces. *Trans. Amer. Math. Soc.* **364** (2012), no. 12, 6159–6186 Zbl 1279.52008 MR 2965739

[132] I. Villanueva, Radial continuous rotation invariant valuations on star bodies. *Adv. Math.* **291** (2016), 961–981 Zbl 1338.52015 MR 3459034

[133] T. Wang, The affine Sobolev–Zhang inequality on BV($\mathbb{R}^n$). *Adv. Math.* **230** (2012), no. 4-6, 2457–2473 Zbl 1257.46016 MR 2927377

[134] T. Wang, Semi-valuations on BV($\mathbb{R}^n$). *Indiana Univ. Math. J.* **63** (2014), no. 5, 1447–1465 Zbl 1320.46035 MR 3283557

[135] T. Wannerer, GL(n) equivariant Minkowski valuations. *Indiana Univ. Math. J.* **60** (2011), no. 5, 1655–1672 Zbl 1270.52019 MR 2997003

[136] T. Wannerer, The module of unitarily invariant area measures. *J. Differential Geom.* **96** (2014), no. 1, 141–182 Zbl 1296.53149 MR 3161388

[137] C. Zeng and D. Ma, SL(n) covariant vector valuations on polytopes. *Trans. Amer. Math. Soc.* **370** (2018), no. 12, 8999–9023 Zbl 1406.52031 MR 3864403

[138] G. Zhang, The affine Sobolev inequality. *J. Differential Geom.* **53** (1999), no. 1, 183–202 Zbl 1040.53089 MR 1776095

[139] G. Zhang, A positive solution to the Busemann–Petty problem in $\mathbb{R}^4$. *Ann. of Math. (2)* **149** (1999), no. 2, 535–543 Zbl 0937.52004 MR 1689339

Monika Ludwig
Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien,
Wiedner Hauptstraße 8-10/1046, 1040 Wien, Austria; monika.ludwig@tuwien.ac.at

DOI 10.4171/8ECM/35

Metric measure spaces and synthetic Ricci bounds: Fundamental concepts and recent developments

Karl-Theodor Sturm

Abstract. Metric measure spaces with synthetic Ricci bounds have attracted great interest in recent years, accompanied by spectacular breakthroughs and deep new insights. In this survey, I will provide a brief introduction to the concept of lower Ricci bounds as introduced by Lott–Villani and myself, and illustrate some of its geometric, analytic, and probabilistic consequences, among them Li–Yau estimates, coupling properties for Brownian motions, sharp functional and isoperimetric inequalities, rigidity results, and structural properties like rectifiability and rectifiability of the boundary. In particular, I will explain its crucial interplay with the heat flow and its link to the curvature-dimension condition formulated in functional-analytic terms by Bakry–Émery. This equivalence between the Lagrangian and the Eulerian approach then will be further explored in various recent research directions: (i) time-dependent Ricci bounds which provide a link to (super-) Ricci flows for singular spaces, (ii) second-order calculus, upper Ricci bounds, and transformation formulas, (iii) distribution-valued Ricci bounds which, e.g., allow singular effects of non-convex boundaries to be taken into account.

1. Synthetic Ricci bounds for metric measure spaces

1.1. Metric spaces

The class of *metric spaces* (X,d) is a far-reaching generalization of the class of *Riemannian manifolds* (M,g). It allows for rich geometric structures including singularities, branching, change of dimension as well as fractional and infinite dimensions.

Already in the middle of the last century, A. D. Aleksandrov [1, 2] has proposed his fundamental concepts of lower and upper bounds for generalized sectional curvature for metric spaces. Especially these lower bounds are particularly well behaved with respect to the so-called Gromov–Hausdorff metric on the class of compact metric spaces as observed by Gromov [77, 78]:

2020 Mathematics Subject Classification. Primary 35K05; Secondary 31C25, 47D08, 30L99.
Keywords. Ricci curvature, metric measure space, curvature-dimension condition, optimal transport, Bakry–Émery, synthetic Ricci bound.

- for each $K \in \mathbb{R}$, the class

$$\big\{(\mathsf{X}, \mathsf{d}) \text{ with sect. curv. } \geq K\big\}$$

  is closed under GH-convergence;
- for each $K, L, N \in \mathbb{R}$, the class

$$\big\{(\mathsf{X}, \mathsf{d}) \text{ with sect. curv.} \geq K, \text{ dimension} \leq N, \text{ diameter} \leq L\big\}$$

  is compact.

In the sequel, many properties of Riemannian manifolds and geometric estimates which only depend on one-sided curvature bounds could be proven for such metric spaces (X, d) with synthetic (upper or lower) curvature bounds. For spaces with synthetic lower bounds on the sectional curvature, also a far-reaching analytic calculus was developed with foundational contributions by Burago–Gromov–Perel'man [24], Kuwae–Machigashira–Shioya [101], Zhang–Zhu [149].

However, for most properties and estimates in geometric analysis, spectral theory and stochastic analysis on manifolds, no quantitative assumptions on the sectional curvature are needed but—as observed in the seminal works of Yau, Cheeger, Colding, Elworthy, Malliavin, Bismut, Perel'man and many others—merely a lower bound on the *Ricci curvature*

$$\mathrm{Ric} \geq K\mathsf{g}.$$

Since the Ricci tensor is the trace of the sectional curvature, i.e.,

$$\mathrm{Ric}_x(v_i, v_i) := \sum_{j \neq i} \mathrm{Sec}_x(v_i, v_j) \quad \text{if } \{v_i\}_{i=1,\ldots,n} \text{ ONB of } T_x N,$$

assumptions on lower bounded Ricci curvature are less restrictive than assumptions on lower bounded sectional curvature. Replacing (synthetic) sectional curvature bounds by Ricci bounds, the previously mentioned Gromov's compactness theorem turns into a precompactness theorem:

- For any choice of $K, L, N \in \mathbb{R}$, the class of Riemannian manifolds (M, g) with Ricci curvature $\geq K$, dimension $\leq N$, and diameter $\leq L$ is relatively compact with respect to mGH-convergence.

Properties of mGH-limits of Cauchy sequences in such classes (so-called *Ricci limit spaces*) have been studied in great detail by Cheeger–Colding [32–34]; see also [35, 36, 39].

As already pointed out by Gromov, the right setting to deal with the completions of these classes is the class of *metric measure spaces*. However, what was missing for decades was a synthetic formulation of lower Ricci bounds, applicable not only to Riemannian manifolds (and their limits) but also to metric measure spaces.

1.2. Metric measure spaces

Here and in the sequel, a *metric measure space* (briefly *mm-space*) will always mean a triple $(\mathsf{X},\mathsf{d},\mathsf{m})$ consisting of

- a space X,
- a complete separable metric d on X,
- a locally finite Borel measure m on it.

It is called normalized (or mm_1-space) iff in addition $\mathsf{m}(\mathsf{X}) = 1$.

A primary goal since many years has been to find a formulation of generalized Ricci curvature bounds $\mathrm{Ric}(\mathsf{X},\mathsf{d},\mathsf{m}) \geq K$ which is

- equivalent to $\mathrm{Ric}_x(v,v) \geq K\|v\|^2$ if X is a Riemannian manifold,
- stable under convergence,
- intrinsic, synthetic (like curvature bounds in Aleksandrov geometry),
- sufficient for many geometric, analytic, and spectral theoretic conclusions.

In independent works, such a formulation has been proposed by the author [136, 137] and by Lott–Villani [107], based on the concept of optimal transport and relying on previous works by Brenier [21], Gangbo [60], McCann [112, 113], Otto [128], Otto–Villani [129], Cordero-Erausquin–McCann–Schmuckenschläger [40], and von Renesse–Sturm [145].

The synthetic lower Ricci bound for an mm-space $(\mathsf{X},\mathsf{d},\mathsf{m})$ will be defined through the interplay of two quantities on X:

- the *Kantorovich–Wasserstein distance*

$$W_2(\mu_1,\mu_2) := \inf\left\{\left(\int_{\mathsf{X}\times\mathsf{X}} \mathsf{d}^2(x,y)\,dq(x,y)\right)^{1/2} : q \in \mathsf{Cpl}(\mu_1,\mu_2)\right\} \quad (1.1)$$

  on the space $\mathcal{P}(\mathsf{X})$ of Borel probability measures on X where

$$\mathsf{Cpl}(\mu_1,\mu_2) := \left\{q \in \mathcal{P}(\mathsf{X}\times\mathsf{X}),\ (\pi_1)_*q = \mu_1,\ (\pi_2)_*q = \mu_2\right\}$$

  denotes the set of *couplings* of two probability measures μ_1, μ_2,

- the *Boltzmann entropy*

$$S(\mu) = \mathrm{Ent}(\mu|m) = \begin{cases} \int_{\mathsf{X}} \rho\log\rho\,d\mathsf{m}, & \text{if } \mu = \rho\cdot\mathsf{m}, \\ +\infty, & \text{if } \mu \not\ll \mathsf{m}, \end{cases} \quad (1.2)$$

  regarded as a functional on $\mathcal{P}(\mathsf{X})$.

The first of these quantities is defined merely using the metric d on X, the second one merely using the measure m on X.

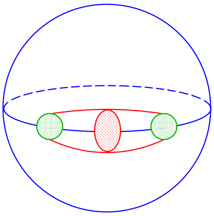


Figure 1.

Remark 1.1. En passant, we record some nice properties of the underlying metric d on X which carry over to the Kantorovich–Wasserstein metric on the *Wasserstein space* $\mathcal{P}_2(X) = \{\mu \in \mathcal{P}(\mathsf{X}) : \int_{\mathsf{X}} \mathsf{d}^2(x, x_0)\mu(dx) < \infty\}$:

- $(\mathcal{P}_2(\mathsf{X}), W_2)$ is a complete separable metric space,
- $(\mathcal{P}_2(\mathsf{X}), W_2)$ is a *compact* space or a *length* space or an *Aleksandrov* space with curvature ≥ 0 if and only if (X, d) is so.

1.3. Synthetic Ricci bounds for metric measure spaces

Following [107, 136, 137], we now present the so-called *curvature-dimension condition* $\mathsf{CD}(K, N)$ to be considered as a synthetic formulation for "Ricci curvature $\geq K$ and dimension $\leq N$". For convenience, we first treat the case $N = \infty$, where no constraint on the dimension is imposed.

Definition 1.2. We say that a metric measure space $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ has *Ricci curvature* $\geq K$ or that it satisfies the *curvature-dimension condition* $\mathsf{CD}(K, \infty)$ iff $\forall \mu_0, \mu_1 \in \mathcal{P}_2(\mathsf{X})$, there exists W_2-geodesic $(\mu_t)_{t\in[0,1]}$ connecting them such that

$$S(\mu_t) \leq (1-t)S(\mu_0) + tS(\mu_1) - \frac{K}{2}t(1-t)W_2^2(\mu_0, \mu_1). \tag{1.3}$$

Remark 1.3. In other words, the $\mathsf{CD}(K, \infty)$-condition holds true if and only if the Boltzmann entropy is *weakly K-convex* on $\mathcal{P}_2(\mathsf{X})$, see Figure 1. Recall that S is called *K-convex* on $\mathcal{P}_2(\mathsf{X})$ iff (1.3) holds true for *all* W_2-geodesics $(\mu_t)_{t\in[0,1]}$ in $\mathcal{P}_2(\mathsf{X})$. The reason for requiring the weaker version is the stability under convergence of the latter (see below).

The second case which allows for an easy formulation is $K = 0$. Here for finite $N \in \mathbb{R}_+$, the formulation is based on the Renyi-type entropy

$$S_N(\nu|\mathsf{m}) := -\int_{\mathsf{X}} \rho^{1-1/N}\, d\mathsf{m} \quad \text{for } \nu = \rho \cdot \mathsf{m} + \nu_s.$$

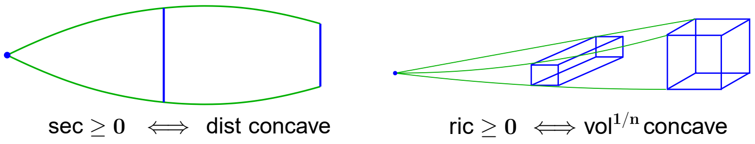


Figure 2.

Definition 1.4. We say that $(\mathsf{X},\mathsf{d},\mathsf{m})$ satisfies the *curvature-dimension condition* $\mathsf{CD}(0,N)$ iff $\forall \mu_0,\mu_1\in\mathcal{P}_2(\mathsf{X})$, there exists W_2-geodesic $(\mu_t)_{t\in[0,1]}$ connecting them such that

$$S_N(\mu_t|\mathsf{m})\le(1-t)S_N(\mu_0|\mathsf{m})+tS_N(\mu_1|\mathsf{m}). \tag{1.4}$$

Remark 1.5. It is quite instructive to observe that

$$S_N(\nu|\mathsf{m})=-\mathsf{m}(A)^{1/N}\quad\text{if } \nu \text{ is unif. distrib. on } A\subset X.$$

Thus the curvature-dimension condition $\mathsf{CD}(0,N)$ can be vaguely interpreted as a kind of concavity property for the N-th root of the volume, see Figure 2. This should be seen in context with the facts that (i) on N-dimensional spaces, the N-th root of the volume has the dimension of a length, (ii) nonnegative sectional curvature in the sense of Aleksandrov can be regarded as a concavity property of distances, and (iii) Ricci curvature should be regarded as the average of the sectional curvatures.

1.4. The curvature-dimension condition $\mathsf{CD}(K,N)$

The curvature-dimension condition $\mathsf{CD}(K,N)$ for general pairs of K,N is more involved. It was introduced in [137]. (Based on that, later on Lott–Villani [106] also introduced a slight modification of it—the difference, however, will be irrelevant for the sequel. In their original paper [107], they consider only the case $K/N=0$, where the effects of dimension and curvature are decoupled.)

Definition 1.6. Given that $K,N\in\mathbb{R}$ (with $N\ge 1$), we say that an mm-space $(\mathsf{X},\mathsf{d},\mathsf{m})$ satisfies the *curvature-dimension condition* $\mathsf{CD}(K,N)$ iff $\forall\rho_0\mathsf{m},\rho_1\mathsf{m}\in\mathcal{P}_2(\mathsf{X})$, there exists W_2-geodesic $(\rho_t\mathsf{m})_{t\in[0,1]}$ connecting them and a W_2-optimal coupling q of them such that

$$\begin{aligned}\int_{\mathsf{X}}\rho_t^{1-1/N}(z)\,d\mathsf{m}(z)\ge\int_{\mathsf{X}\times\mathsf{X}}&\big[\tau_{K,N}^{(1-t)}(\gamma_0,\gamma_1)\cdot\rho_0^{-1/N}(\gamma_0)\\&+\tau_{K,N}^{(t)}(\gamma_0,\gamma_1)\cdot\rho_1^{-1/N}(\gamma_1)\big]\,dq(\gamma_0,\gamma_1).\end{aligned}\tag{1.5}$$

Here the *distortion coefficients* are given by

$$\tau_{K,N}^{(t)}(x,y):=t^{\frac{1}{N}}\left(\frac{\sin\left(\sqrt{\frac{K}{N-1}}t\,\mathrm{d}(x,y)\right)}{\sin\left(\sqrt{\frac{K}{N-1}}\,\mathrm{d}(x,y)\right)}\right)^{\frac{N-1}{N}}$$

in case $K > 0$, analogous formula with $\sin\sqrt{\cdots}$ replaced by $\sinh\sqrt{-\cdots}$ in case $K < 0$, and $\tau^{(t)}_{K,N}(x, y) := t$ in case $K = 0$.

The interpretation of $\mathsf{CD}(K, N)$ as a synthetic formulation for “Ricci curvature $\geq K$, dimension $\leq N$” is justified by the Riemannian case.

Theorem 1.7 ([137] extending [40, 135, 145]). *For Riemannian manifolds* (M, g)*,*

$$\mathsf{CD}(K, N) \Leftrightarrow \mathrm{Ric}_{\mathsf{M}} \geq K \quad \textit{and} \quad \dim_{\mathsf{M}} \leq N.$$

Further examples of metric measure spaces satisfying a $\mathsf{CD}(K,N)$-condition include weighted Riemannian spaces, Ricci limit spaces, Aleksandrov spaces, and Finsler spaces. If one slightly extends the concept of “metric” towards “pseudo metric”, it also includes path spaces (e.g. the Wiener space with $K = 1$, $N = \infty$) and configuration spaces.

Moreover, many further examples are obtained by *constructions* as limits, products, cones, suspensions, or warped products.

2. Geometric aspects

The broad interest in—and the great success of—the concept of the curvature-dimension condition $\mathsf{CD}(K, N)$ is due to

- its equivalence to classical lower Ricci bounds in the Riemannian setting,
- its stability under convergence and under various constructions, and
- the fact that it implies almost all of the geometric and functional analytic estimates (with sharp constants!) from Riemannian geometry which depend only on (the dimension and on) lower bounds on the Ricci curvature.

2.1. Volume growth

Let us summarize some of the most fundamental geometric estimates.

Theorem 2.1 (Bonnet–Myers diameter bound [137]). *The* $\mathsf{CD}(K, N)$*-condition with finite N and positive K implies compactness of* X *and*

$$\mathrm{diam}(\mathsf{X}) \leq \sqrt{\frac{N-1}{K}} \cdot \pi. \tag{2.1}$$

Theorem 2.2 (Bishop–Gromov volume growth estimate [137]). *Under* $\mathsf{CD}(K, N)$ *with finite N, for every $x_0 \in \mathsf{X}$, the volume growth function $r \mapsto \mathsf{m}(B_r(x_0))$ is absolutely continuous and its weak derivative $s(r) := \frac{\partial}{\partial r}\mathsf{m}(B_r(x_0))$ satisfies*

$$s(r)/s(R) \quad \geq \quad \sin\left(\sqrt{\frac{K}{N-1}}r\right)^{N-1} \Big/ \sin\left(\sqrt{\frac{K}{N-1}}R\right)^{N-1} \tag{2.2}$$

for all $0 < r < R$ *with the usual re-interpretation of the RHS if* $K \leq 0$ *(i.e., replacing all* $\sin(\sqrt{K}\cdots)$ *by* $\sinh(\sqrt{-K}\cdots)$ *in the case* $K < 0$*).*

As in the smooth Riemannian setting, this differential inequality immediately implies the integrated version:

$$\frac{\mathsf{m}\big(B_r(x_0)\big)}{\mathsf{m}\big(B_R(x_0)\big)} \geq \frac{\int_0^r \sin\Big(\sqrt{\frac{K}{N-1}}t\Big)^{N-1} dt}{\int_0^R \sin\Big(\sqrt{\frac{K}{N-1}}t\Big)^{N-1} dt}$$

for all $0 < r < R$, and thus in particular

$$\mathsf{m}\big(B_R(x_0)\big) \leq C r^N \exp\big(\sqrt{(N-1)K^- R}\big).$$

The results so far assumed that N is finite. In the case $N = \infty$, the $\mathsf{CD}(K, N)$-condition implies a *novel volume growth estimate* [136], not known before in the Riemannian setting,

$$\mathsf{m}\big(B_R(x_0)\big) \leq \exp\left(\frac{K^-}{2}R^2 + c_1 R + c_0\right). \tag{2.3}$$

It can be seen as complementary to the concentration of measure phenomenon. The sharpness is illustrated by the following example.

Example 2.3. Consider $\mathsf{X} = \mathbb{R}$, $\mathsf{d} = |\cdot|$, and $d\mathsf{m}(x) = \exp(\frac{\kappa}{2}|x|^2)$ for $\kappa > 0$. Then $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ satisfies $\mathsf{CD}(-\kappa, \infty)$, and $\mathsf{m}(B_R(x)) \geq \exp(\frac{\kappa}{2}(R - \frac{1}{2})^2)$ for all x and $R \geq \frac{1}{2}$.

The curvature-dimension condition $\mathsf{CD}(K, N)$ also implies numerous further geometric estimates, among them the *Brunn–Minkowski inequality* [137] and the *Borell–Brascamp–Lieb inequality* [11]. What remained an open problem for many years was the Lévy–Gromov isoperimetric inequality which only recently was proven by Cavalletti–Mondino.

Theorem 2.4 (Lévy–Gromov isoperimetric inequality [30]). *Let* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *be an essentially non-branching mm-space which satisfies* $\mathsf{CD}(K, N)$ *and let* $\widehat{\mathsf{X}}$ *be a* $\mathsf{CD}(K, N)$*-model space. Then for every subset* $E \subset \mathsf{X}$ *and every spherical cap* $B \subset \widehat{\mathsf{X}}$,

$$\frac{|\partial E|}{|\mathsf{X}|} \geq \frac{|\partial B|}{|\widehat{\mathsf{X}}|} \quad \text{if } \frac{|E|}{|\mathsf{X}|} = \frac{|B|}{|\widehat{\mathsf{X}}|}. \tag{2.4}$$

Here $|\cdot|$ *denotes the respective volume or surface measure.*

2.2. The space of spaces

Two mm_1-spaces will be called *isomorphic*—and henceforth identified—iff there exists a measure preserving isometry between the supports of the respective measures. It is a quite remarkable observation that the space Ξ of isomorphism classes of normalized mm$_1$-spaces itself is a geodesic space.

The *L^p-transportation distance* between mm$_1$-spaces $(\mathsf{X}_0, \mathsf{d}_0, \mathsf{m}_0)$ and $(\mathsf{X}_1, \mathsf{d}_1, \mathsf{m}_1)$ is defined for $p \in [1, \infty)$ as

$$\mathbb{D}_p\big((\mathsf{X}_0, \mathsf{d}_0, \mathsf{m}_0), (\mathsf{X}_1, \mathsf{d}_1, \mathsf{m}_1)\big) = \inf_{\mathsf{d},\mathsf{m}} \left(\int_{\mathsf{X}_0 \times \mathsf{X}_1} \mathsf{d}(x_0, x_1)^p d\mathsf{m}(x_0, x_1) \right)^{1/p},$$

where the infimum is taken over all *couplings* m of m_0 and m_1 and over all *couplings* d of d_0 and d_1 (i.e., metrics on $\mathsf{X}_0 \sqcup \mathsf{X}_1$ which coincide with d_0 on X_0 and with d_1 on X_1), [136]. With slight modifications, this definition also extends to $p = \infty$ and $p \in (0, 1)$. Furthermore, for $p = 0$ we define in the spirit of the Ky Fan metric

$$\mathbb{D}_0\big((\mathsf{X}_0, \mathsf{d}_0, \mathsf{m}_0), (\mathsf{X}_1, \mathsf{d}_1, \mathsf{m}_1)\big) = \inf_{\mathsf{d},\mathsf{m}} \inf \big\{\varepsilon > 0 : \mathsf{m}\big\{\mathsf{d}(x_0, x_1) > \varepsilon\big\} \leq \varepsilon\big\}.$$

A closely related concept is the *L^p-distortion distance* between mm$_1$-spaces defined for $p \in [1, \infty)$ as

$$\begin{aligned}&\mathbf{\Delta}_p\big((\mathsf{X}_0, \mathsf{d}_0, \mathsf{m}_0), (\mathsf{X}_1, \mathsf{d}_1, \mathsf{m}_1)\big)\\ &= \inf_{\mathsf{m}} \left(\int_{\mathsf{X}_0 \times \mathsf{X}_1} \int_{\mathsf{X}_0 \times \mathsf{X}_1} \big|\mathsf{d}_0(x_0, y_0) - \mathsf{d}_1(x_1, y_1)\big|^p d\mathsf{m}(x_0, x_1) d\mathsf{m}(y_0, y_1) \right)^{1/p},\end{aligned}$$

where the infimum is taken over all *couplings* m of m_0 and m_1, and again with slight modifications also extended to $p = \infty$, $p \in (0, 1)$, and $p = 0$. Under uniform control of the moments of the involved metric measure spaces, the topologies induced by all these metrics are the same and coincide with that of *Gromov's box distance* $\square_\lambda$ and with that of measured Gromov–Hausdorff convergence.

Lemma 2.5 ([76, 116, 138]). (a) $\forall p \in [0, \infty)$*:* $\mathbb{D}_p$ *is complete whereas* $\mathbf{\Delta}_p$ *is not complete,*

(b) $\mathbb{D}_p$*-convergence* $\Leftrightarrow$ $\mathbb{D}_0$*-convergence and convergence of* p*-th moments,*

(c) $\mathbf{\Delta}_p$*-convergence* $\Leftrightarrow$ $\mathbf{\Delta}_0$*-convergence and convergence of* p*-th moments,*

(d) $\mathbb{D}_0$*-convergence* $\Leftrightarrow$ $\mathbf{\Delta}_0$*-convergence* $\Leftrightarrow$ $\square_\lambda$*-convergence.*

The main result here is that the *space of spaces* is an Aleksandrov space.

Theorem 2.6 ([138]). *The metric space* $(\Xi_2, \mathbf{\Delta}_2)$ *of isomorphism classes of mm$_1$-spaces is a* geodesic space *with* nonnegative curvature.

The tangent space (for the space of spaces) at a given mm$_1$-space admits an explicit representation and so does the *symmetry group*, with the latter e.g. in terms of *optimal self-couplings*. Of particular interest are finite dimensional subspaces of the space of spaces.

Proposition 2.7. *For each $n \in \mathbb{N}$, the subspace of n-point spaces (i.e., mm$_1$-spaces with equal mass on n-points) is a* Riemannian orbifold *with nonnegative curvature.*

2.3. Stability, compactness

Converging sequences of mm$_1$-spaces can always be embedded into common metric spaces. The stability of the $\mathsf{CD}(K, N)$-condition then simply amounts to the lower semicontinuity of the Renyi-type entropy for weakly convergent sequences of probability measures.

Theorem 2.8. *The curvature-dimension condition is* stable *under $\mathbb{D}_0$-convergence of mm$_1$-spaces.*

The volume growth estimates entailed by the $\mathsf{CD}(K, N)$-condition, together with the stability of the latter under convergence, allow us to turn Gromov's pre-compactness theorem under Ricci bounds into a compactness theorem.

Theorem 2.9. *For every triple $K, N, L \in \mathbb{R}$, the space of all mm$_1$-spaces $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ that satisfy $\mathsf{CD}(K, N)$ and have diameter $\leq L$ is* compact.

2.4. Local to global

A crucial property of curvature bounds both in Riemannian geometry and in the geometry of Aleksandrov spaces is the *local-to-global property*: sharp global estimates follow from uniform local curvature assumptions. For the synthetic Ricci bounds for mm-spaces, this is a highly non-trivial claim. To deal with it, we restrict ourselves to non-branching geodesic spaces.

The first *globalization theorem* was obtained in the case $K/N = 0$, where curvature and dimension effects are de-coupled.

Proposition 2.10 ([107,136,137]). *If $K=0$ or $N=\infty$, then every mm-space $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ satisfies*

$$\mathsf{CD}(K, N) \textit{ locally } \Leftrightarrow \mathsf{CD}(K, N) \textit{ globally.}$$

Further progress then was based on the *reduced curvature-dimension condition* $\mathsf{CD}^*(K, N)$ defined similarly as $\mathsf{CD}(K, N)$ but now with the distortion coefficient $\tau^{(t)}_{K,N}(x, y)$ in (1.5) replaced by the reduced coefficients

$$\sigma^{(t)}_{K,N}(x, y) := \sin\left(\sqrt{\frac{K}{N}}t\, \mathrm{d}(x, y)\right) \Big/ \sin\left(\sqrt{\frac{K}{N}}\mathrm{d}(x, y)\right).$$

Proposition 2.11 ([12]). *For all $K, N \in \mathbb{R}$ and all mm-spaces,*

$$\mathsf{CD}(K, N) \textit{ locally } \Leftrightarrow \mathsf{CD}^*(K, N) \textit{ locally } \Leftrightarrow \mathsf{CD}^*(K, N) \textit{ globally.}$$

Only recently, the globalization theorem could be proven in full generality by Cavalletti–Milman (with a minor extension by Zhenhao Li removing the finiteness assumption for the underlying measure). Their approach is based on Klartag's [95] *needle decomposition* and the *localization technique* developed by Cavalletti–Mondino [30].

Theorem 2.12 ([29, 103]).

$$\mathsf{CD}(K,N) \textit{ locally} \Leftrightarrow \mathsf{CD}(K,N) \textit{ globally.}$$

3. Analytic aspects

A deeper understanding of the role of synthetic lower Ricci bounds on singular spaces will be obtained through links with spectral properties of the Laplacians and estimates for heat kernels on such spaces.

3.1. Heat flow on metric measure spaces

There are two different (seemingly unrelated) approaches to define the *heat equation* on an mm-space $(\mathsf{X},\mathsf{d},\mathsf{m})$:

- either as a gradient flow in $L^2(\mathsf{X},\mathsf{m})$ for the *energy*

$$\mathcal{E}(u) = \frac{1}{2}\int_X |\nabla u|^2\,dm = \liminf_{v\to u \text{ in } L^2} \frac{1}{2}\int_X (\mathrm{lip}_x v)^2\,dm(x)$$

  with $\mathrm{lip}_x v(x) = \limsup_{y\to x} \frac{|v(x)-v(y)|}{\mathsf{d}(x,y)}$ and $|\nabla u| =$ minimal weak upper gradient,
- or as a gradient flow in $\mathcal{P}_2(\mathsf{X})$ for the *Boltzmann entropy*

$$\mathrm{Ent}(u) = \int_X u \log u\,dm.$$

The former approach (the traditional point of view) has the advantage that the energy—if it exists—is always convex and thus guarantees the existence of the gradient flow. Its disadvantage is that it relies on the concept of weakly differentiable functions. However, all analytic problems related to the notion of energy have been fully resolved in the trilogy [3–5] by Ambrosio–Gigli–Savaré.

The latter approach (the novel perspective of Otto) has the advantage that the entropy is always obviously well defined. However, for its gradient flow to exist, additional assumptions are required, e.g. that the entropy is semi-convex. Up to minor technicalities, this simply says that the underlying mm-space has lower bounded synthetic Ricci curvature. Under this minimal assumption, indeed, both approaches coincide.

Theorem 3.1 ([3]). *For every mm-space* $(\mathsf{X},\mathsf{d},\mathsf{m})$ *that satisfies* $\mathsf{CD}(K,\infty)$ *for some* $K \in \mathbb{R}$*, the energy approach and the entropy approach coincide.*

Example 3.2. There are plenty of examples to which this result applies. The most prominent among them (and the authors who first proved it) are

(i) *Euclidean space* $\mathbb{R}^n$: Jordan–Kinderlehrer–Otto [85],

(ii) *Riemann manifolds* (M, g): Ohta [124], Savaré [133], Villani [144], Erbar [45],

(iii) *Finsler spaces* $(\mathsf{M}, \mathsf{F}, \mathsf{m})$: Ohta–Sturm [126],

(iv) *Aleksandrov spaces*: Gigli–Kuwada–Ohta [67].

Example 3.3. In many other cases not covered by any CD-condition, we know that the energy approach and the entropy approach coincide:

(a) *Heisenberg group* (unbounded curvature): Juillet [86],

(b) *Wiener space* (degenerate distance): Fang–Shao–Sturm [58],

(c) *Configuration space* (degenerate distance): Erbar–Huesmann [50],

(d) *Neumann Laplacian* (unbounded curvature if nonconvex): Lierl–Sturm [104],

(e) *Dirichlet Laplacian* (no mass conservation): Profeta–Sturm [131],

(f) *Discrete spaces* (no W_2-geodesics): Maas [109], Mielke [117],

(g) *Lévy semigroups* (no W_2-geodesics): Erbar [46],

(h) *Metric graphs* (unbounded curvature): Erbar–Forkert–Maas–Mugnolo [49].

In the latter examples (e), (f), and (g), the concept of “gradient flow for the Boltzmann entropy” has to be slightly adapted.

3.2. Curvature-dimension condition: Eulerian vs. Lagrangian

Besides the Lagrangian formulation of synthetic Ricci bounds in terms of semiconvexity properties of the entropy, there is also a Eulerian formulation in terms of the energy: the celebrated *curvature-dimension (or* Γ_2*) condition of Bakry–Émery*. It is a groundbreaking observation that both formulations are equivalent in great generality.

For this equivalence to hold, we now make the standing assumption that $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ is *infinitesimally Hilbertian*, i.e., the energy $\mathcal{E}$ is quadratic or, in other words, Laplacian and heat flow are linear. For convenience, we will also assume that the mm-space under consideration has the *Sobolev-to-Lipschitz property* and volume growth bounded by e^{Cr^2}. Note that both of these latter properties follow from the validity of the Lagrangian $\mathsf{CD}(K, N)$-condition.

Theorem 3.4 ([4, 5, 52]). *Under the above assumptions, the following properties are equivalent:*

(i) *the synthetic Ricci bound* $\mathsf{CD}(K, N)$*, briefly reformulated as*

$$\text{Hess}\, S - \frac{1}{N}(\nabla S)^{\otimes 2} \geq K \quad \textit{on } \big(\mathcal{P}_2(\mathsf{X}), W_2\big),$$

(ii) *the transport estimate*

$$W_2^2(P_s\mu, P_t\nu) \le e^{-K\tau} W_2^2(\mu,\nu) + 2N \frac{1-e^{-K\tau}}{K\tau}(\sqrt{s}-\sqrt{t})^2$$

with $\tau := \frac{2}{3}(s + \sqrt{st} + t)$*,*

(iii) *the gradient estimate*

$$|\nabla P_t u|^2 + \frac{4Kt^2}{N(e^{2Kt}-1)}|\Delta P_t u|^2 \le e^{-2Kt} P_t|\nabla u|^2,$$

(iv) *the Bochner inequality*

$$\frac{1}{2}\Delta|\nabla u|^2 - \langle \nabla u, \nabla \Delta u\rangle \ge K \cdot |\nabla u|^2 + \frac{1}{N}(\Delta u)^2,$$

also known as Bakry–Émery criterion and written in comprehensive form as

$$\Gamma_2(u) \ge K \cdot \Gamma(u) + \frac{1}{N}(\Delta u)^2.$$

These equivalences allow for easy explanations and/or intuitive interpretations. The equivalence (iii)⇔(iv), indeed, is known since decades as a basic result of the so-called Γ-calculus of Markov semigroups [13,14], and easily follows by differentiating $s \mapsto P_{t-s}(|\nabla P_s u|^2)$. The equivalence (i)⇔(ii), from a heuristic point of view, is a consequence of the fact that the heat flow is the gradient flow for the entropy with respect to the metric W_2. Finally, the equivalence (ii)⇔(iii) is the important *Kuwada duality* which extends the celebrated *Kantorovich–Rubinstein duality* towards $p \ne 1$, $q \ne \infty$. The rigorous proofs of the above equivalences by Ambrosio–Gigli–Savaré [4, 5] (for the case $N = \infty$) and Erbar–Kuwada–Sturm [52] (for the general case) are rather sophisticated and mark milestones in the development of the theory. For an alternative approach in the general case, see also [7].

Remark 3.5. The Bakry–Émery estimate

$$\Gamma_2(u) - K \cdot |\nabla u|^2 \ge \frac{1}{N}(\Delta u)^2 \quad (\forall u)$$

has a remarkable *self-improvement property* [13–15, 57, 134] asserting that it implies the seemingly stronger estimate

$$\begin{aligned}\Gamma_2(u) - K \cdot |\nabla u|^2 &\ge \frac{1}{N}(\Delta u)^2 + \frac{N}{N-1}\Big|\big|\nabla|\nabla u|\big| - \frac{1}{N}|\Delta u|\Big|^2 \\ &= \big|\nabla|\nabla u|\big| + \frac{1}{N-1}\Big|\big|\nabla|\nabla u|\big| - |\Delta u|\Big|^2 \quad (\forall u).\end{aligned}$$

This leads to improved gradient estimates and improved transport estimates which e.g. in the case $N = \infty$ read as

$$|\nabla P_t u| \leq e^{-Kt} P_t |\nabla u|, \quad W_\infty(P_t\mu, P_t\nu) \leq e^{-Kt} W_\infty(\mu, \nu).$$

3.3. $\mathsf{RCD}(K, N)$-spaces—functional inequalities

We will say that an mm-space satisfies *the* $\mathsf{RCD}(K, N)$*-condition* iff it satisfies the $\mathsf{CD}(K, N)$-condition and iff it is infinitesimally Hilbertian. For these mm-spaces, the full machinery of geometric analysis and Riemannian calculus can be developed and far-reaching structural assertions can be derived.

Here we have to restrict ourselves to present only a selection of the many results proven so far. And we will not formulate detailed estimates (except for the first result), we will just mention the respective results.

Theorem 3.6. *The following estimates hold true (each of them with sharp constants) on any mm-space which satisfies an* $\mathsf{RCD}(K, N)$*-condition for some* $K \in \mathbb{R}$ *and for* $N \leq \infty$*:*

- *Poincaré/Lichnerowicz inequality* [106]*:* $\lambda_1 \geq \frac{N}{N-1} K$,

moreover, for $N < \infty$*:*

- *Laplace comparison* [64],
- *Bochner's inequality* [7, 52],
- *Li-Yau differential Harnack inequality, Gaussian heat kernel estimates* [61],
- *Sobolev, Cheeger, and Buser inequalities* [44, 130],

whereas for $N = \infty$*:*

- *Talagrand- and logarithmic Sobolev inequalities* [106],
- *Wang's Harnack inequality* [102]*, upper Gaussian heat kernel estimate* [143]*, and Ledoux's inequality* [44].

In all the previous results, the dimensional parameter has always been a number $N \geq 1$ (which in turn then even implies that $N \geq \dim_{\mathcal{H}}(\mathsf{X})$). Quite remarkably, various of these results also admit versions where the *dimensional parameter* N *is a negative number*; see e.g. [110, 111, 119, 125, 127].

3.4. $\mathsf{RCD}(K, N)$-spaces—splitting and rigidity

In the smooth Riemannian setting, an important consequence of nonnegative Ricci curvature is the Cheeger–Gromoll splitting theorem. In order to extend this to metric measure spaces, it is essential to assume that the underlying spaces are infinitesimally Hilbertian.

Theorem 3.7 (Splitting theorem [63]). *If an mm-space* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *satisfies* $\mathsf{RCD}(0, N)$ *and* contains a line*, then* $\mathsf{X} = \mathbb{R} \times \mathsf{X}'$ *for some* $\mathsf{RCD}(0, N-1)$*-space* $(\mathsf{X}', \mathsf{d}', \mathsf{m}')$.

The counterpart to the splitting theorem for positive lower Ricci bound is Cheng's maximal diameter theorem.

Theorem 3.8 (Maximal diameter theorem [91]). *If an mm-space* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *satisfies* $\mathsf{RCD}(N-1, N)$ *and has* diameter π*, then* X *is the spherical suspension of some* $\mathsf{RCD}(N-2, N-1)$*-space* $(\mathsf{X}', \mathsf{d}', \mathsf{m}')$.

In the smooth Riemannian setting, the maximal diameter theorem provides a more far-reaching conclusion, namely, that X is the round N-sphere. In the singular setting, however, this conclusion is false [91].

On the other hand, such a far-reaching conclusion can be drawn from the maximality of the spherical size.

Theorem 3.9 (Maximal spherical size theorem [56]). *If an mm-space* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *satisfies* $\mathsf{RCD}(N-1, N)$ *and*

$$-\int_{\mathsf{X}}\int_{\mathsf{X}} \cos\big(\mathsf{d}(x,y)\big)\, d\mathsf{m}(x)\, d\mathsf{m}(y) \geq 0, \tag{3.1}$$

then $N \in \mathbb{N}$ *and* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *is isomorphic to the* N*-dimensional round sphere* $\mathbb{S}^N$.

Closely related to the maximal diameter theorem is Obata's theorem on the minimality of the spectral gap.

Theorem 3.10 (Obata's theorem [92]). *If an* $\mathsf{RCD}(N-1, N)$*-space* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *has* spectral gap N*, then it is the spherical suspension of some* $\mathsf{RCD}(N-2, N-1)$*-space* $(\mathsf{X}', \mathsf{d}', \mathsf{m}')$.

This splitting theorem indeed also admits an extension to $N = \infty$ which states that an mm-space $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ that satisfies $\mathsf{RCD}(1, \infty)$ and has spectral gap 1 splits off a Gaussian factor [66].

3.5. $\mathsf{RCD}(K, N)$-spaces—structure theory

Since blow-ups of $\mathsf{RCD}(K, N)$-spaces are $\mathsf{RCD}(0, N)$-spaces which contain lines, a sophisticated iterated application of the splitting theorem will lead to deep insights into tangent spaces and local structure of RCD-spaces.

Theorem 3.11 (Rectifiability and constancy of dimension [23, 120]). *If* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *satisfies* $\mathsf{RCD}(K, N)$*, then*

(a) $\mathsf{X} = \bigcup_{k=1}^{\lfloor N \rfloor} \mathcal{R}_k \cup \mathcal{N}$, $m(\mathcal{N}) = 0$,

(b) *each* $\mathcal{R}_k$ *is covered by countably many measurable sets which are* $(1+\varepsilon)$*-biLipschitz equivalent to subsets of* $\mathbb{R}^k$,

(c) m *and* $\mathcal{H}^k$ *are mutually abs. cont. on* $\mathcal{R}_k$,

and even more,

(d) *there exists* $n \in \mathbb{N}$ *such that* $m(\mathcal{R}_k) = 0$ *for all* $k \neq n$.

Besides the two landmark contributions to this structure theory mentioned above, numerous important results were obtained [6, 43, 69, 90]. Particularly nice insights could be obtained in the case $N = 2$.

Corollary 3.12 ([108]). $\mathsf{RCD}(K, 2)$-*spaces with* $\mathsf{m} = \mathcal{H}^2$ *are Aleksandrov spaces.*

Further challenges then concern the boundaries of mm-spaces. Various concepts how to define them and related results were presented in [42,88,89]. Important contributions to the analysis of tangent cones and to the regularity theory for non-collapsed RCD-spaces were provided in [8,82,94]. Based on these results, a precise description could be derived.

Theorem 3.13 ([22]). *Let* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *be a non-collapsed* $\mathsf{RCD}(K, N)$-*space (with* $\mathsf{m} = \mathcal{H}^N$, $N \in \mathbb{N}$*). Then*

(a) *there exists a stratification* $\mathcal{S}^0 \subset \mathcal{S}^0 \subset \cdots \subset \mathcal{S}^{N-1} = \mathcal{S} = \mathsf{X} \setminus \mathcal{R}_N$,

(b) *the boundary* $\partial \mathsf{X} := \overline{\mathcal{S}^{N-1} \setminus \mathcal{S}^{N-2}}$ *is* $(N-1)$-*rectifiable,*

(c) $T_x\mathsf{X} \simeq \mathbb{R}^{N-1} \times \mathbb{R}_+$ *for* $x \in \mathcal{S}^{N-1} \setminus \mathcal{S}^{N-2}$,

(d) $\mathsf{X} \setminus \mathcal{S}^{N-2}$ *is a topological manifold with boundary.*

4. Recent developments

The concept of synthetic Ricci bounds for singular spaces turned out to be extremely fruitful, both for theory and applications. A rich theory of mm-spaces satisfying such uniform lower Ricci bounds has been established. The last 15 years have seen a wave of impressive results—many of them going far beyond the previously described scope.

In the following, we will first present in detail recent developments concerning

- heat flow on time-dependent mm-spaces and super-Ricci flows,
- second-order calculus, upper Ricci bounds, and transformation formulas,
- distribution-valued lower Ricci bounds,

and then briefly summarize several further developments.

4.1. Heat flow on time-dependent mm-spaces and super-Ricci flows

Whereas construction and properties of the heat flow on "static" metric measure space $(\mathsf{X}, \mathsf{d}, \mathsf{m})$—in particular, its relation to synthetic lower bounds on the Ricci curvature—by now are well understood in great generality, analogous questions for time-dependent families of mm-spaces $(\mathsf{X}_t, \mathsf{d}_t, \mathsf{m}_t)$, $t \in I = (0, T)$, until recently remained widely open:

- How do we define a heat propagator $(P_{t,s})_{t \geq s}$ acting on functions in $L^2(\mathsf{X}_s, \mathsf{m}_s)$ and/or its dual $(\widehat{P}_{t,s})_{s \leq t}$ acting on measures on X_t?
  Can they be regarded as gradient flows of (time-dependent) energy or entropy functionals in function/measure spaces with time-dependent norms or metrics?
- Is there a parabolic analogue to synthetic lower Ricci bounds? Can one formulate it as "dynamic convexity" of a time-dependent entropy functional? How is this related to the notion of super-Ricci flows for families of Riemannian manifolds?
- Are there "parabolic versions" of the functional inequalities that characterize synthetic lower Ricci bounds?

Within recent years, for families of mm-spaces $(\mathsf{X}, \mathsf{d}_t, \mathsf{m}_t)$, $t \in (0, T)$, such that

- for every $t \in I$ the mm-space $(\mathsf{X}, \mathsf{d}_t, \mathsf{m}_t)$ satisfies an $\mathsf{RCD}(K, N)$-condition,
- there exists some regular t-dependence of d_t and m_t,

these questions found affirmative answers.

Definition 4.1 ([140]). A family of mm-spaces $(\mathsf{X}, \mathsf{d}_t, \mathsf{m}_t)_{t \in (0,T)}$ is called *super-Ricci flow* iff the function

$$\mathrm{Ent}\colon (0, T) \times \mathcal{P}(X) \to (-\infty, \infty], \quad (t, \mu) \mapsto \mathrm{Ent}_t(\mu) := \mathrm{Ent}(\mu | m_t)$$

is *dynamically convex* on $\mathcal{P}(X)$—equipped with the 1-parameter family of metrics W_t ($= L^2$-Kantorovich–Wasserstein metrics with respect to d_t)—in the following sense: for all μ^0, μ^1 and a.e. t there exists a W_t-geodesic $(\mu^a)_{a \in [0,1]}$ such that

$$\partial_a \mathrm{Ent}_t(\mu^0) - \partial_a \mathrm{Ent}_t(\mu^1) \leq \frac{1}{2} \partial_t W_t^2(\mu^0, \mu^1). \tag{4.1}$$

Example 4.2. A family of Riemannian manifolds $(\mathsf{M}, \mathsf{g}_t)$, $t \in (0, T)$ is a super-Ricci flow in the previous sense iff

$$\mathrm{Ric}_t + \frac{1}{2} \partial_t \mathsf{g}_t \geq 0.$$

Recall that $(\mathsf{M}, \mathsf{g}_t)_{t \in (0,T)}$ is called *Ricci flow* if $\mathrm{Ric}_t + \frac{1}{2} \partial_t \mathsf{g}_t = 0$. These properties can be regarded as the parabolic analogue to nonnegative (or vanishing, resp.) Ricci curvature for static manifolds.

Whereas in the static setting the gradient flow for the energy and the gradient flow for the entropy characterize the same evolution (either in terms of densities or in terms of measures), this is no longer the case in the dynamic setting: here one is characterizing the forward evolution whereas the other one is characterizing the backward evolution.

Theorem 4.3 ([98]). *In the previous setting, there exists a well-defined heat propagator $(P_{t,s})_{t\geq s}$ acting on functions in $L^2(\mathsf{X},\mathsf{m}_s)$ and its dual $(\hat{P}_{t,s})_{s\leq t}$ acting on measures on* X. *Moreover,*

(1) *$\forall u \in \mathcal{D}om(\mathcal{E})$, $\forall s \in I$, the heat flow $t \mapsto u_t = P_{t,s}u$ is the unique* forward gradient flow *for the Cheeger energy $\frac{1}{2}\mathcal{E}_s$ in $L^2(\mathsf{X},\mathsf{m}_s)$.*

(2) *$\forall \mu \in \mathcal{D}om(\mathrm{Ent})$, $\forall t \in I$, the dual heat flow $s \mapsto \mu_s = \hat{P}_{t,s}\mu$ is the unique* backward gradient flow *for the Boltzmann entropy* Ent_t *in $(\mathcal{P}(\mathsf{X}), W_t)$ provided that $(\mathsf{X},\mathsf{d}_t,\mathsf{m}_t)$ is a super-Ricci flow.*

Both gradient flows can be obtained as limits of corresponding steepest-descend schemes (aka JKO-schemes) adapted to the time-dependent setting [97].

In analogy to Theorem 3.4, the Lagrangian characterization of super-Ricci flows (in terms of dynamic convexity of the entropy) turns out to be equivalent to a Eulerian characterization (in terms of a dynamic Γ_2-inequality), to a gradient estimate for the forward evolution, and to a transport estimate (as well as to a pathwise Brownian coupling property) for the backward evolution.

Theorem 4.4 ([98]). *The following are equivalent:*

(a) $\partial_a \mathrm{Ent}_t(\mu^a)|_{a=0} - \partial_a \mathrm{Ent}_t(\mu^a)|_{a=1} \leq \frac{1}{2}\partial_t W_t^2(\mu^0,\mu^1)$,

(b) $W_s(\hat{P}_{t,s}\mu, \hat{P}_{t,s}\nu) \leq W_t(\mu,\nu)$,

(c) *$\forall x, y$, $\forall t$, there exist coupled backward Brownian motions $(X_s, Y_s)_{s\leq t}$ starting at t in (x,y) such that $\mathsf{d}_s(X_s,Y_s) \leq \mathsf{d}_t(x,y)$ a.s. for all $s \leq t$,*

(d) $|\nabla_t(P_{t,s}u)|^2 \leq P_{t,s}(|\nabla_s u|^2)$,

(e) $\Gamma_{2,t} \geq \frac{1}{2}\partial_t\Gamma_t$, *where* $\Gamma_{2,t}(u) = \frac{1}{2}\Delta_t|\nabla_t u|^2 - \langle \nabla_t u, \nabla_t \Delta_t u\rangle$.

This result in particular extends a previous characterization of super-Ricci flows of *smooth* families of Riemannian manifolds in terms of the previous assertion (b) by McCann–Topping [115] and in terms of the previous assertion (c) by Arnaudon–Coulibaly–Thalmaier [10].

There is a whole zoo of further functional inequalities which characterize super-Ricci flows. Several implications for the subsequent assertions were new even in the static case.

Theorem 4.5 ([99]). *Each of the following assertions is equivalent to any of the above or, in other words, to $(\mathsf{X},\mathsf{d}_t,\mathsf{m}_t)_{t\in I}$ being a super-Ricci flow:*

(f) *local Poincaré inequalities:*

$$2(t-s)\Gamma_t(P_{t,s}u) \leq P_{t,s}(u^2) - (P_{t,s}u)^2 \leq 2(t-s)P_{t,s}(\Gamma_s u),$$

(g) *local logarithmic Sobolev inequalities:*

$$(t-s)\frac{\Gamma_t(P_{t,s}u)}{P_{t,s}u} \le P_{t,s}(u\log u) - (P_{t,s}u)\log(P_{t,s}u) \le (t-s)P_{t,s}\left(\frac{\Gamma_s u}{u}\right),$$

(h) *dimension-free Harnack inequality:* $\forall\alpha > 1$

$$(P_{t,s}u)^\alpha(y) \le P_{t,s}u^\alpha(x)\cdot\exp\left(\frac{\alpha\mathsf{d}_t^2(x,y)}{4(\alpha-1)(t-s)}\right),$$

(i) *log Harnack inequality:*

$$P_{t,s}(\log u)(x) \le \log P_{t,s}u(y) + \frac{\mathsf{d}_t^2(x,y)}{4(t-s)}.$$

With these concepts and results, a robust theory of super-Ricci flows is established—being regarded as a parabolic analogue to singular spaces with lower Ricci bounds. In the smooth case, deeper insights and more powerful estimates require to restrict oneself to *Ricci flows* rather than super-Ricci flows; see e.g. [16, 81, 96, 100]. To deal with similar questions in the singular case, first of all we need a synthetic notion of upper Ricci bounds; see the next subsection.

For related current research on lower Ricci bounds in time-like directions on Lorentzian manifolds and on Einstein equation in general relativity, see [31, 114, 122].

4.2. Second-order calculus, upper Ricci bounds, and transformation formulas

So far, on RCD-space we only dealt with the canonical first-order calculus for (real-valued) functions on these spaces. The setting, however, allows us to go far beyond this.

Theorem 4.6 ([18, 62, 65, 70–72, 121]). *Given an* $\mathsf{RCD}(K,\infty)$*-space* $(\mathsf{X},\mathsf{d},\mathsf{m})$*, there exist well established concepts of*

- *a powerful second-order order calculus on* X *including a consistent notion of Ricci tensor (the lower bound of which coincides with the synthetic lower Ricci bound in terms of semiconvexity of the entropy),*
- *the heat flow on* 1*-forms on* X *which among others leads to the celebrated Hess–Schrader–Uhlenbrock inequality*

$$|P_t\,\mathsf{d}f| \le e^{-Kt}P_t|\mathsf{d}f|,$$

- *harmonic maps from* X *into metric spaces* $(\mathsf{Y},\mathsf{d}_\mathsf{Y})$*, typically of nonpositive curvature, based on Sobolev calculus and approximation of energy densities for maps between metric spaces, providing Lipschitz continuity of these maps.*

In a different direction, a challenging goal is to provide synthetic characterizations of upper Ricci bounds $\mathrm{Ric} \le L$. Indeed, various of the (equivalent) synthetic characterizations of lower Ricci bounds admit partial converses. However, these converse characterizations are not necessarily equivalent to each other. Moreover, any such characterizations will certainly be not as powerful as the corresponding lower bound. Typically, the upper Ricci bounds are asymptotic estimates whereas the lower Ricci bounds are uniform estimates.

Theorem 4.7 ([142]). *Weak synthetic characterizations of upper Ricci bounds for an* $\mathsf{RCD}(K,N)$*-space* $(\mathsf{X},\mathsf{d},\mathsf{m})$

- *in terms of partial L-concavity of the Boltzmann entropy and*
- *in terms of the heat kernel asymptotics*

are equivalent to each other.

More precisely, a weak upper bound L for the Ricci curvature is given by

$$L := \sup_z \limsup_{x,y\to z} \eta(x,y),$$

where for all $x, y \in X$,

$$\begin{aligned}
&\eta(x,y) := \\
&= \lim_{\varepsilon\to 0}\, \inf \Big\{ \frac{1}{W_2^2(\rho^0,\rho^1)} \cdot \big[\partial_a^- S(\rho^a)\big|_{a=1} - \partial_a^+ S(\rho^a)\big|_{a=0}\big] : \big(\rho^a\big)_{a\in[0,1]} \text{ geodesic}, \\
&\qquad\qquad S(\rho^0)<\infty,\ S(\rho^1)<\infty,\ \mathrm{supp}[\rho^0]\subset B_\varepsilon(x),\ \mathrm{supp}[\rho^1]\subset B_\varepsilon(y)\Big\} \\
&= \lim_{\varepsilon\to 0}\, \inf \big\{ -\partial_t^+ \log W_2(P_t\mu, P_t\nu)\big|_{t=0} : \mathrm{supp}[\mu]\subset B_\varepsilon(x),\ \mathrm{supp}[\nu]\subset B_\varepsilon(y)\big\}.
\end{aligned}$$

Remark 4.8. For weighted Riemannian manifolds $(\mathsf{M},\mathsf{g},e^{-f}d\mathrm{vol}_{\mathsf{g}})$,

$$\mathsf{Ric}_f(x,y) \le \eta(x,y) \le \mathsf{Ric}_f(x,y) + \sigma(x,y)\cdot \tan^2\big(\sqrt{\sigma(x,y)}\,\mathsf{d}(x,y)/2\big)$$

provided x and y are not conjugate. Here $\mathsf{Ric}_f(x,y) = \int_0^1 \mathrm{Ric}_f(\dot\gamma^a,\dot\gamma^a)/|\dot\gamma^a|^2\, da$ denotes the average Bakry–Émery–Ricci curvature along the (unique) geodesic $\gamma = (\gamma^a)_{a\in[0,1]}$ from x to y, and $\sigma(x,y)$ denotes the maximal modulus of the Riemannian curvature along this geodesic.

Similar as other approaches (e.g. [123]), these weak upper Ricci bounds will not be able to detect the positive Ricci curvature sitting in the tip of a cone over a circle of length $< 2\pi$. A slightly stronger notion will detect it.

Theorem 4.9 ([56]). *If a metric cone has both sided ("strong") Ricci bounds K and L in the sense of* $\mathsf{RCD}(K,\infty)$ *and*

$$-\liminf_{x,y\to z}\liminf_{t\to 0} \frac{1}{t}\log\frac{W_2\big(P_t\delta_x, P_t\delta_y\big)}{\mathsf{d}(x,y)} \le L \quad (\forall z\in \mathsf{X}),$$

then it is the flat Euclidean space (of some integer dimension).

A crucial property of the class of RCD-spaces is that it is preserved under transformations of measure and metric of the underlying spaces, and that there exist explicit formulas for the transformation of the parameters K and N in the curvature-dimension condition $\mathsf{CD}(K,N)$.

To be more specific, let an mm-space $(\mathsf{X},\mathsf{d},\mathsf{m})$ be given as well as continuous ("weight") functions V, W on X. In terms of them, define the transformed mm-space $(\mathsf{X},\mathsf{d}',\mathsf{m}')$ with $\mathsf{m}' := e^V \mathsf{m}$ and

$$\mathsf{d}'(x,y) := \inf\left\{ \int_0^1 |\dot\gamma_t| \cdot e^{W(\gamma_t)}\, dt : \gamma : [0,1] \to X \text{ rectifiable}, \gamma_0 = x,\ \gamma_1 = y \right\}.$$

If $\int |\nabla u|^2\, d\mathsf{m}$ on $L^2(\mathsf{X},\mathsf{m})$ denotes the Dirichlet form ("Cheeger energy") associated with $(\mathsf{X},\mathsf{d},\mathsf{m})$, then the Dirichlet form associated with the transformed mm-space is given by

$$\int |\nabla u|^2 e^{V-2W}\, d\mathsf{m} \quad \text{on } L^2(\mathsf{X}, e^V \mathsf{m}).$$

Theorem 4.10 ([80, 139]). *If* $(\mathsf{X},\mathsf{d},\mathsf{m})$ *satisfies* $\mathsf{RCD}(K,N)$ *for finite* $K, N \in \mathbb{R}$ *and if* $V, W \in W^{2,\infty}(X)$*, then for each* $N' > N$ *there exists an explicitly given* K' *such that* $(\mathsf{X},\mathsf{d}',\mathsf{m}')$ *satisfies* $\mathsf{RCD}(K',N')$.

(If $W = 0$*, then also* $N = N' = \infty$ *is admissible; if* $V = NW$*, then also* $N' = N$ *is admissible.)*

Let us illustrate this result in three special cases of particular importance:

- $W = 0$ ("drift transformation"):

$$K' = K - \sup_{f,x} \frac{1}{|\nabla f|^2}\left[\operatorname{Hess} V(\nabla f, \nabla f) + \frac{1}{N'-N} \langle \nabla V, \nabla f\rangle^2 \right](x);$$

- $V = 2W$ ("time change"):

$$K' = \inf_x e^{-2W}\left[K - \Delta W - \frac{\big[(N-2)(N'-2)\big]_+}{N'-N} |\nabla W|^2 \right](x);$$

- $V = NW$ ("conformal transformation"): $N' = N$ and

$$K' = \inf_x e^{-2W}\Big[K - \big[\Delta W + (N-2)|\nabla W|^2\big] \\ - \sup_f \frac{N-2}{|\nabla f|^2}\big[\operatorname{Hess} W(\nabla f, \nabla f) - \langle \nabla W, \nabla f\rangle^2\big] \Big](x).$$

The first of these cases is well studied in the setting of Bakry–Émery calculus (and also in the setting of synthetic Ricci bounds for mm-spaces). It is the only case where also $N = \infty$ is admitted. The last of these cases is well known in Riemannian geometry but has not been considered before in singular settings. A particular feature of the second case is that the transformation formula for the Ricci bound only depends on bounds for $|\nabla W|$ and ΔW (and thus extends to distribution-valued Ricci bounds in case of $W \in \mathrm{Lip}(X)$; see the next subsection).

4.3. Distribution-valued Ricci bounds

Uniform lower Ricci bounds of the form $\mathsf{CD}(K, \infty)$ on mm-spaces

- are preserved for Neumann Laplacian on convex subsets, but
- never hold for Neumann Laplacian on non-convex subsets.

The goal thus is

- to find appropriate modification for non-convex subsets,
- to replace constant K, by function k, measure κ, distribution, etc.

Theorem 4.11 ([20]). *Given an infinitesimally Hilbertian mm-space* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *and a lower bounded, lower semicontinuous function* $k : \mathsf{X} \to \mathbb{R}$*, the following are equivalent:*

(i) *curvature-dimension condition* $\mathsf{CD}(k, \infty)$ *with variable* k*:* $\forall \mu_0, \mu_1 \in \mathcal{P}(X)$*, there exists* W_2*-geodesic* $(\mu_t)_t = (e_{t*}\boldsymbol{\nu})_t$ *such that* $\forall t \in [0,1]$ *with* $g_{s,t} := (1-s)t \wedge s(1-t)$*,*

$$\mathrm{Ent}(\mu_t) \le (1-t)\,\mathrm{Ent}(\mu_0) + t\,\mathrm{Ent}(\mu_1) - \int \int_0^1 k(\gamma_s) g_{s,t}\, ds |\dot\gamma|^2 \boldsymbol{\nu}(d\gamma),$$

(ii) *gradient estimate:*

$$|\nabla P_t u|(x) \le \mathbb{E}_x\big[e^{-\int_0^t k(B_s)ds} \cdot |\nabla u|(B_t)\big],$$

(iii) *Bochner's inequality* $\mathsf{BE}_2(k, \infty)$*:*

$$\frac{1}{2}\Delta|\nabla u|^2 - \langle \nabla u, \nabla \Delta u\rangle \ge k \cdot |\nabla u|^2,$$

(iv) $\forall \mu_1,\ \mu_2$*, there exists a coupled pair of Brownian motions* $(B^1_{t/2})_{t\ge 0}$*,* $(B^2_{t/2})_{t \ge 0}$ *with given initial distributions such that a.s. for all* $s < t$

$$\mathsf{d}(B^1_t, B^2_t) \le e^{-\int_s^t \bar k(B^1_r, B^2_r)dr} \cdot \mathsf{d}(B^1_s, B^2_s)$$

with $\bar k(x_0, x_1) := \sup\{\int_0^1 k(\gamma_u)du : \gamma_0 = x_0,\ \gamma_1 = x_1,\ \gamma$ *geodesic*$\}$.

For extensions to (k, N)-versions, see [52, 93, 141].

To proceed towards distribution-valued Ricci bounds, define the spaces $W^{1,p}(\mathsf{X})$ for $p \in [1,\infty]$, put $W^{1,\infty}_*(\mathsf{X}) := \{f \in W^{1,2}_{loc}(\mathsf{X}) : \||\nabla f|\|_{L^\infty} < \infty\}$, and denote by $W^{-1,\infty}(\mathsf{X})$ the topological dual of

$$W^{1,1+}(\mathsf{X}) := \Big\{f \in L^1(\mathsf{X}) : f_n := f \wedge n \vee (-n) \in W^{1,2}(\mathsf{X}),\ \sup_n \big\||\nabla f_n|\big\|_{L^1} < \infty\Big\}.$$

Definition 4.12. Given $\kappa \in W^{-1,\infty}(\mathsf{X})$, we say that the Bochner inequality $\mathsf{BE}_1(\kappa,\infty)$ holds iff $|\nabla f| \in W^{1,2}$ for all $f \in \mathrm{D}(\Delta)$, and

$$-\int_X \big\langle \nabla|\nabla f|, \nabla\phi\big\rangle + \frac{1}{|\nabla f|}\langle \nabla f, \nabla \Delta f\rangle \phi \, dm \geq \big\langle |\nabla f|\phi, \kappa\big\rangle_{W^{1,1},W^{-1,\infty}}$$

for all $f \in \mathrm{D}(\Delta)$ with $\Delta f \in W^{1,2}$ and all nonnegative $\phi \in W^{1,2}$.

Given $\kappa \in W^{-1,\infty}(X)$, we define a closed, lower bounded bilinear form $\mathcal{E}^\kappa$ on $L^2(X)$ by

$$\mathcal{E}^\kappa(f,g) := \mathcal{E}(f,g) + \langle fg, \kappa\rangle_{W^{1,1+},W^{-1,\infty}}$$

for $f, g \in \mathrm{Dom}(\mathcal{E}^\kappa) := W^{1,2}(\mathsf{X})$. Associated to it, there is a strongly continuous, positivity preserving semigroup $(P^\kappa_t)_{t\geq 0}$ on $L^2(\mathsf{X})$.

Theorem 4.13 ([141]). *The Bochner inequality* $\mathsf{BE}_1(\kappa,\infty)$ *is equivalent to the gradient estimate*

$$|\nabla P_t f| \leq P^\kappa_t\big(|\nabla f|\big). \tag{4.2}$$

To gain a better understanding of the semigroup $(P^\kappa_t)_{t\geq 0}$, assume that $\kappa = -\underline{\Delta}\psi$ for some $\psi \in W^{1,\infty}$.

Theorem 4.14 ([37, 141]). *Then*

$$\mathcal{E}^\kappa(f,g) = \mathcal{E}(f,g) + \mathcal{E}(fg,\psi) \tag{4.3}$$

and

$$P^\kappa_{t/2} f(x) = \mathbb{E}_x\big[e^{N^\psi_t} f(B_t)\big], \tag{4.4}$$

where $(\mathbb{P}_x, (B_t)_{t\geq 0})$ *denotes Brownian motion starting in* $x \in X$, *and* N^ψ *is the zero energy part in the Fukushima decomposition; i.e.,* $N^\psi_t = \psi(B_t) - \psi(B_0) - M^\psi_t$.

If $\psi \in \mathrm{Dom}(\Delta)$, then $N^\psi_t = \frac{1}{2}\int_0^t \Delta\psi(B_s)ds$—in consistency with the previous theorem (Theorem 4.11).

Remark 4.15. The concept of *tamed spaces* proposed by Erbar–Rigoni–Sturm–Tamanini [55] generalizes the previous approach to distribution-valued lower Ricci bounds in various respects:

- the objects under consideration are strongly local, quasi-regular Dirichlet spaces $(\mathsf{X}, \mathcal{E}, \mathsf{m})$ (rather than infinitesimally Hilbertian mm-spaces $(\mathsf{X}, \mathsf{d}, \mathsf{m})$);

- the Ricci bounds are formulated in terms of distributions $\kappa \in W^{-1,2}_{qloc}(\mathsf{X})$ (rather than $\kappa \in W^{-1,\infty}(\mathsf{X})$); for such distributions κ which lie quasi locally in the dual of $W^{1,2}(\mathsf{X})$, the previous ansatz for defining the semigroup $(P^\kappa_t)_{t>0}$ still works with appropriate sequences of localizing stopping times;
- in addition, the distributions κ are assumed to be moderate in the sense that

$$\sup_{t\le 1, x\in \mathsf{X}} P^\kappa_t 1(x) < \infty.$$

This reminds of the Kato condition but is significantly more general since it does not require any decomposition of κ into positive and negative parts. It always holds if $\kappa = -\underline{\Delta}\psi$ for some $\psi \in \mathrm{Lip}_b(X)$.

Example 4.16. The prime examples of *tamed spaces* are provided by the following:

(a) ground state transformation of Hamiltonian for molecules [19, 79]; it yields curvature bounds in terms of unbounded functions in the Kato class;

(b) Riemannian Lipschitz manifolds with lower Ricci bound in the Kato class [27, 28, 132];

(c) time change of $\mathsf{RCD}(K,N)$-spaces with $W \in \mathrm{Lip}_b(\mathsf{X})$ (cf. Theorem 4.10); it typically yields curvature bounds κ which are not signed measures;

(d) restriction of $\mathsf{RCD}(K,N)$-spaces to (convex or non-convex) subsets $\mathsf{Y} \subset \mathsf{X}$ or, in other words, Laplacian with Neumann boundary conditions; it yields curvature bounds in terms of signed measures $\kappa = km + \ell\sigma$; see below.

Assume that $(\mathsf{X},\mathsf{d},\mathsf{m})$ satisfies an $\mathsf{RCD}(k,N)$-condition with variable $k : \mathsf{X} \to \mathbb{R}$ and finite N. Let a closed subset $\mathsf{Y} \subset \mathsf{X}$ be given which can be represented as sublevel set $\mathsf{Y} = \{V \le 0\}$ for some semiconvex function $V : \mathsf{X} \to \mathbb{R}$ with $|\nabla V| = 1$ on $\partial\mathsf{Y}$. Typically, V is the signed distance functions $V = \mathsf{d}(\,\cdot\,,\mathsf{Y}) - \mathsf{d}(\,\cdot\,,\mathsf{X}\setminus\mathsf{Y})$.

A function $\ell : \mathsf{X} \to \mathbb{R}$ is regarded as "generalized lower bound for the curvature (or second fundamental) form of $\partial\mathsf{Y}$" iff it is a synthetic lower bound for the Hessian of V.

Example 4.17. Assume that X is an Aleksandrov space with sectional curvature ≥ 0 and that $\mathsf{Y} \subset \mathsf{X}$ satisfies an exterior ball condition: $\forall z \in \partial\mathsf{Y}$, there exists a ball $B_r(x) \subset \complement\mathsf{Y}$ with $z \in \partial B_r(x)$. Then $\ell(z) := -\frac{1}{r(z)}$ is a lower bound for the curvature of $\partial\mathsf{Y}$.

Under weak regularity assumptions, the distributional Laplacian $\sigma_{\mathsf{Y}} := \underline{\Delta} V^+$ is a (nonnegative) measure which then will be regarded as "the surface measure of $\partial\mathsf{Y}$".

Theorem 4.18 ([141]). *Under weak regularity assumptions on V and ℓ, the restricted space $(\mathsf{Y},\mathsf{d}_{\mathsf{Y}},\mathsf{m}_{\mathsf{Y}})$ satisfies a Bakry–Émery condition $\mathsf{BE}_1(\kappa,\infty)$ with a signed measure valued Ricci bound*

$$\kappa = k\cdot m_{\mathsf{Y}} + \ell\cdot\sigma_{\mathsf{Y}}. \tag{4.5}$$

Thus the Neumann heat semigroup on Y *satisfies*

$$|\nabla P_t^{\mathsf{Y}} u|(x) \leq \mathbb{E}_x\big[|\nabla u|(B_t)\cdot e^{-\int_0^t k(B_s)ds}\cdot e^{-\int_0^t \ell(B_s)dL_s}\big], \tag{4.6}$$

where $(B_{s/2})_{s\geq 0}$ *denotes the Brownian motion in* Y *and* $(L_s)_{s\geq 0}$ *the continuous additive functional associated with* σ_{Y}*.*

For smooth subsets in Riemannian manifolds, this kind of gradient estimate—with $(L_s)_{s\geq 0}$ being the *local time* of the boundary—has been firstly derived by Hsu [84]; cf. also [38, 146].

Let us illustrate the power of the above estimates with two simple examples: the ball and its complement.

Corollary 4.19. *Let* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *be an* N*-dimensional Aleksandrov space (*$N \geq 3$*) with* $\mathrm{Ric} \geq -1$ *and* $\mathsf{sec} \leq 0$*. Then for* $\mathsf{Y} := \mathsf{X} \setminus B_r(z)$*,*

$$\big|\nabla P^{\mathsf{Y}}_{t/2} f\big|(x) \leq \mathbb{E}^{\mathsf{Y}}_x\Big[e^{t/2+\frac{1}{2r}L_t^{\partial\mathsf{Y}}}\cdot\big|\nabla f(B_t^{\mathsf{Y}})\big|\Big].$$

In particular, $\mathrm{Lip}(P^{\mathsf{Y}}_{t/2} f) \leq \sup_x \mathbb{E}^{\mathsf{Y}}_x[e^{t/2+\frac{1}{2r}L_t^{\partial\mathsf{Y}}}]\cdot\mathrm{Lip}(f)$ *and*

$$|\nabla P^{Y}_{t/2} f|^2(x) \leq e^{Ct+C'\sqrt{t}}\cdot P^{\mathsf{Y}}_{t/2}|\nabla f|^2(x). \tag{4.7}$$

Upper and lower bounds of curvature (here 0 and -1, resp.) can be chosen to be any numbers. Note that *no estimate* of the form

$$\big|\nabla P^{\mathsf{Y}}_{t/2} f\big|^2(x) \leq e^{Ct}\cdot P^{\mathsf{Y}}_{t/2}\big|\nabla f\big|^2(x)$$

can hold true due to the non-convexity of Y. Thus it is *necessary* to take into account the singular contribution arising from the negative curvature of the boundary.

In the next example, the singular contribution arising from the positive curvature of the boundary can be ignored. However, taking it into account will significantly *improve* the gradient estimate.

Corollary 4.20. *Let* $(\mathsf{X}, \mathsf{d}, \mathsf{m})$ *be an* N*-dimensional Aleksandrov space with* $\mathrm{Ric} \geq 0$ *and* $\mathsf{sec} \leq 1$*. Then for* $\mathsf{Y} := \overline{B}_r(z)$ *for some* $z \in \mathsf{X}$ *and* $r \in (0, \pi/4)$*,*

$$\big|\nabla P^{Y}_{t/2} f\big|(x) \leq \mathbb{E}^{\mathsf{Y}}_x\big[e^{-\frac{\cot r}{2}L_t^{\partial\mathsf{Y}}}\cdot\big|\nabla f(B_t^{Y})\big|\big].$$

In particular, $\mathrm{Lip}(P^{Y}_{t/2} f) \leq \sup_x \mathbb{E}^{\mathsf{Y}}_x[e^{-\frac{\cot r}{2}L_t^{\partial\mathsf{Y}}}\cdot\mathrm{Lip}(f)]$ *and*

$$\big|\nabla P^{\mathsf{Y}}_{t/2} f\big|^2(x) \leq e^{-t\frac{N-1}{2}\cot^2 r+1}\cdot P^{\mathsf{Y}}_{t/2}|\nabla f|^2(x). \tag{4.8}$$

Taking into account the curvature of the boundary allows us to derive a positive lower bound for the spectral gap (without involving any diameter bound and despite possibly vanishing Ricci curvature in the interior).

Corollary 4.21. *In the previous setting,* $\lambda_1 \geq \frac{N-1}{2} \cot^2 r$*.*

4.4. Synthetic Ricci bounds—extended settings

In order to summarize recent developments concerning synthetic Ricci bounds for singular spaces, let us recall the previously presented

(1) *heat flow on time-dependent mm-spaces and super-Ricci flows,*

(2) *second-order calculus, upper Ricci bounds, and transformation formulas,*

(3) *distribution-valued lower Ricci bounds,*

and then move on to further developments in extended settings

(4) *discrete mm-spaces:* for discrete mm-spaces $(\mathsf{X}, \mathsf{d}, \mathsf{m})$, the synthetic Ricci bounds as introduced above will be meaningless since there will be no non-constant geodesics with respect to the Kantorovich–Wasserstein metric W_2 as defined in (1.1). This disadvantage can be overcome by resorting to a modified Kantorovich–Wasserstein metric based on a subtle discrete version of the Benamou–Brenier formula. This way, the heat flow can again be characterized as the gradient flow of the entropy [109, 117].
And synthetic Ricci bounds defined in terms of semiconvexity of the entropy with respect to this modified metric are intimately linked to equilibration properties of the heat flow; see e.g. [47, 48, 53, 54, 75]. Challenging questions address homogenization [68,73,74] and evolution under curvature flows [51]. Related—but in general different—concepts of synthetic Ricci bounds are based on discrete versions of the Bakry–Émery condition; see e.g. [17,41, 59, 105, 147].

(5) *non-commutative spaces:* inspired by the synthetic Ricci bounds for discrete spaces, an analogous concept also has been proposed for non-commutative spaces, with remarkable insights e.g. for (ergodic) quantum Markov semigroups on tracial or finite-dimensional unital C^*-algebras, in particular, equilibration rate estimates for the fermionic Ornstein–Uhlenbeck semigroup and for Bose Ornstein–Uhlenbeck semigroups [9, 25, 26, 83, 118, 148].

(6) *Dirichlet boundary conditions:* for a long time, it seemed that OT techniques could not be used to analyze the heat flow with Dirichlet boundary conditions. Only recently, Profeta–Sturm [131] overcame the problem of mass absorption by considering *charged particles* (which are either particles or anti-particles), and this way succeeded in finding a characterization for the heat flow as a gradient flow for the entropy. Passing from particles to charged particles technically corresponds to passing from a space X to its *doubling*. Functional inequalities for the Dirichlet heat flow thus are closely linked to those for the doubled space. For recent progress concerning the challenging *problem of gluing convex subsets in* RCD*-spaces*, see [87].

Funding. The author gratefully acknowledges financial support from the European Research Council through the ERC AdG "RicciBounds" (grant agreement 694405) as well as funding by the Deutsche Forschungsgemeinschaft through the excellence cluster "Hausdorff Center for Mathematics" and the collaborative research center CRC 1060.

References

[1] A. D. Aleksandrov, *Intrinsic Geometry of Convex Surfaces*. OGIZ, Moscow-Leningrad, 1948 Zbl 0038.35201 MR 0029518

[2] A. D. Alexandrow, *Die innere Geometrie der konvexen Flächen*. Akademie, Berlin, 1955 Zbl 0065.15102 MR 0071041

[3] L. Ambrosio, N. Gigli, and G. Savaré, Calculus and heat flow in metric measure spaces and applications to spaces with Ricci bounds from below. *Invent. Math.* **195** (2014), no. 2, 289–391 Zbl 1312.53056 MR 3152751

[4] L. Ambrosio, N. Gigli, and G. Savaré, Metric measure spaces with Riemannian Ricci curvature bounded from below. *Duke Math. J.* **163** (2014), no. 7, 1405–1490 Zbl 1304.35310 MR 3205729

[5] L. Ambrosio, N. Gigli, and G. Savaré, Bakry–Émery curvature-dimension condition and Riemannian Ricci curvature bounds. *Ann. Probab.* **43** (2015), no. 1, 339–404 Zbl 1307.49044 MR 3298475

[6] L. Ambrosio, S. Honda, and D. Tewodrose, Short-time behavior of the heat kernel and Weyl's law on $\mathrm{RCD}^*(K,N)$ spaces. *Ann. Global Anal. Geom.* **53** (2018), no. 1, 97–119 Zbl 1390.58015 MR 3746517

[7] L. Ambrosio, A. Mondino, and G. Savaré, Nonlinear diffusion equations and curvature conditions in metric measure spaces. *Mem. Amer. Math. Soc.* **262** (2019), no. 1270, v+121 Zbl 1477.49003 MR 4044464

[8] G. Antonelli, E. Brué, and D. Semola, Volume bounds for the quantitative singular strata of non collapsed RCD metric measure spaces. *Anal. Geom. Metr. Spaces* **7** (2019), no. 1, 158–178 Zbl 1428.53051 MR 4015195

[9] P. Antonini and F. Cavalletti, Geometry of Grassmannians and optimal transport of quantum states. 2021, arXiv:2104.02616

[10] M. Arnaudon, K. A. Coulibaly, and A. Thalmaier, Brownian motion with respect to a metric depending on time: definition, existence and applications to Ricci flow. *C. R. Math. Acad. Sci. Paris* **346** (2008), no. 13-14, 773–778 Zbl 1144.58019 MR 2427080

[11] K. Bacher, On Borell–Brascamp–Lieb inequalities on metric measure spaces. *Potential Anal.* **33** (2010), no. 1, 1–15 Zbl 1190.53035 MR 2644212

[12] K. Bacher and K.-T. Sturm, Localization and tensorization properties of the curvature-dimension condition for metric measure spaces. *J. Funct. Anal.* **259** (2010), no. 1, 28–56 Zbl 1196.53027 MR 2610378

[13] D. Bakry, Transformations de Riesz pour les semi-groupes symétriques. II. Étude sous la condition $\Gamma_2 \geq 0$. In *Séminaire de probabilités, XIX, 1983/84*, pp. 145–174, Lecture Notes in Math. 1123, Springer, Berlin, 1985 Zbl 0561.42011 MR 889473

[14] D. Bakry, L'hypercontractivité et son utilisation en théorie des semigroupes. In *Lectures on Probability Theory (Saint-Flour, 1992)*, pp. 1–114, Lecture Notes in Math. 1581, Springer, Berlin, 1994 Zbl 0856.47026 MR 1307413

[15] D. Bakry and Z. Qian, Some new results on eigenvectors via dimension, diameter, and Ricci curvature. *Adv. Math.* **155** (2000), no. 1, 98–153 Zbl 0980.58020 MR 1789850

[16] R. H. Bamler and B. Kleiner, Uniqueness and stability of Ricci flow through singularities. *Acta Math.* **228** (2022), no. 1, 1–215 MR 4448680

[17] F. Bauer, J. Jost, and S. Liu, Ollivier–Ricci curvature and the spectrum of the normalized graph Laplace operator. *Math. Res. Lett.* **19** (2012), no. 6, 1185–1205 Zbl 1297.05143 MR 3091602

[18] M. Braun, Heat flow on 1-forms under lower Ricci bounds. Functional inequalities, spectral theory, and heat kernel. *J. Funct. Anal.* **283** (2022), no. 7, Paper No. 109599 MR 4444737

[19] M. Braun and B. Güneysu, Heat flow regularity, Bismut–Elworthy–Li's derivative formula, and pathwise couplings on Riemannian manifolds with Kato bounded Ricci curvature. *Electron. J. Probab.* **26** (2021), Paper No. 129 Zbl 07478658 MR 4343567

[20] M. Braun, K. Habermann, and K.-T. Sturm, Optimal transport, gradient estimates, and pathwise Brownian coupling on spaces with variable Ricci bounds. *J. Math. Pures Appl. (9)* **147** (2021), 60–97 Zbl 1459.58013 MR 4213679

[21] Y. Brenier, Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** (1991), no. 4, 375–417 Zbl 0738.46011 MR 1100809

[22] E. Bruè, A. Naber, and D. Semola, Boundary regularity and stability for spaces with Ricci bounded below. *Invent. Math.* **228** (2022), no. 2, 777–891 Zbl 07514027 MR 4411732

[23] E. Brué and D. Semola, Constancy of the dimension for $\mathrm{RCD}(K,N)$ spaces via regularity of Lagrangian flows. *Comm. Pure Appl. Math.* **73** (2020), no. 6, 1141–1204 Zbl 1442.35054 MR 4156601

[24] Y. Burago, M. Gromov, and G. Perel'man, A. D. Aleksandrov spaces with curvatures bounded below. *Uspekhi Mat. Nauk* **47** (1992), no. 2(284), 3–51, 222 MR 1185284

[25] E. A. Carlen and J. Maas, An analog of the 2-Wasserstein metric in non-commutative probability under which the fermionic Fokker–Planck equation is gradient flow for the entropy. *Comm. Math. Phys.* **331** (2014), no. 3, 887–926 Zbl 1297.35241 MR 3248053

[26] E. A. Carlen and J. Maas, Non-commutative calculus, optimal transport and functional inequalities in dissipative quantum systems. *J. Stat. Phys.* **178** (2020), no. 2, 319–378 Zbl 1445.46049 MR 4055244

[27] G. Carron, Geometric inequalities for manifolds with Ricci curvature in the Kato class. *Ann. Inst. Fourier (Grenoble)* **69** (2019), 3095–3167 Zbl 1455.53065 MR 4286831

[28] G. Carron, I. Mondello, and D. Tewodrose, Limits of manifolds with a Kato bound on the Ricci curvature. 2021, arXiv:2102.05940

[29] F. Cavalletti and E. Milman, The globalization theorem for the curvature-dimension condition. *Invent. Math.* **226** (2021), no. 1, 1–137 Zbl 1479.53049 MR 4309491

[30] F. Cavalletti and A. Mondino, Sharp and rigid isoperimetric inequalities in metric-measure spaces with lower Ricci curvature bounds. *Invent. Math.* **208** (2017), no. 3, 803–849 Zbl 1375.53053 MR 3648975

[31] F. Cavalletti and A. Mondino, Optimal transport in Lorentzian synthetic spaces, synthetic timelike Ricci curvature lower bounds and applications. 2020, arXiv:2004.08934

[32] J. Cheeger and T. H. Colding, On the structure of spaces with Ricci curvature bounded below. I. *J. Differential Geom.* **46** (1997), no. 3, 406–480 Zbl 0902.53034 MR 1484888

[33] J. Cheeger and T. H. Colding, On the structure of spaces with Ricci curvature bounded below. II. *J. Differential Geom.* **54** (2000), no. 1, 13–35 Zbl 1027.53042 MR 1815410

[34] J. Cheeger and T. H. Colding, On the structure of spaces with Ricci curvature bounded below. III. *J. Differential Geom.* **54** (2000), no. 1, 37–74 Zbl 1027.53043 MR 1815411

[35] J. Cheeger, W. Jiang, and A. Naber, Rectifiability of singular sets of noncollapsed limit spaces with Ricci curvature bounded below. *Ann. of Math. (2)* **193** (2021), no. 2, 407–538 Zbl 1469.53083 MR 4226910

[36] J. Cheeger and A. Naber, Lower bounds on Ricci curvature and quantitative behavior of singular sets. *Invent. Math.* **191** (2013), no. 2, 321–339 Zbl 1268.53053 MR 3010378

[37] Z.-Q. Chen and T.-S. Zhang, Girsanov and Feynman–Kac type transformations for symmetric Markov processes. *Ann. Inst. H. Poincaré Probab. Statist.* **38** (2002), no. 4, 475–505 Zbl 1004.60077 MR 1914937

[38] L. Cheng, A. Thalmaier, and J. Thompson, Functional inequalities on manifolds with non-convex boundary. *Sci. China Math.* **61** (2018), no. 8, 1421–1436 Zbl 1407.58012 MR 3833744

[39] T. H. Colding and A. Naber, Sharp Hölder continuity of tangent cones for spaces with a lower Ricci curvature bound and applications. *Ann. of Math. (2)* **176** (2012), no. 2, 1173–1229 Zbl 1260.53067 MR 2950772

[40] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, A Riemannian interpolation inequality à la Borell, Brascamp and Lieb. *Invent. Math.* **146** (2001), no. 2, 219–257 Zbl 1026.58018 MR 1865396

[41] D. Cushing, S. Kamtue, J. Koolen, S. Liu, F. Münch, and N. Peyerimhoff, Rigidity of the Bonnet–Myers inequality for graphs with respect to Ollivier Ricci curvature. *Adv. Math.* **369** (2020), 107188, 53 Zbl 1440.05069 MR 4096132

[42] G. De Philippis and N. Gigli, Non-collapsed spaces with Ricci curvature bounded from below. *J. Éc. polytech. Math.* **5** (2018), 613–650 Zbl 1409.53038 MR 3852263

[43] G. De Philippis, A. Marchese, and F. Rindler, On a conjecture of Cheeger. In *Measure Theory in Non-Smooth Spaces*, pp. 145–155, Partial Differ. Equ. Meas. Theory, De Gruyter Open, Warsaw, 2017 Zbl 1485.53052 MR 3701738

[44] N. De Ponti and A. Mondino, Sharp Cheeger–Buser type inequalities in $\mathsf{RCD}(K,\infty)$ spaces. *J. Geom. Anal.* **31** (2021), no. 3, 2416–2438 Zbl 1475.53040 MR 4225812

[45] M. Erbar, The heat equation on manifolds as a gradient flow in the Wasserstein space. *Ann. Inst. Henri Poincaré Probab. Stat.* **46** (2010), no. 1, 1–23 Zbl 1215.35016 MR 2641767

[46] M. Erbar, Gradient flows of the entropy for jump processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** (2014), no. 3, 920–945 Zbl 1311.60091 MR 3224294

[47] M. Erbar and M. Fathi, Poincaré, modified logarithmic Sobolev and isoperimetric inequalities for Markov chains with non-negative Ricci curvature. *J. Funct. Anal.* **274** (2018), no. 11, 3056–3089 Zbl 1386.53020 MR 3782987

[48] M. Erbar, M. Fathi, and A. Schlichting, Entropic curvature and convergence to equilibrium for mean-field dynamics on discrete spaces. *ALEA Lat. Am. J. Probab. Math. Stat.* **17** (2020), no. 1, 445–471 Zbl 1441.82017 MR 4105926

[49] M. Erbar, D. Forkert, J. Maas, and D. Mugnolo, Gradient flow formulation of diffusion equations in the Wasserstein space over a metric graph. *Netw. Heterog. Media* **17** (2022), no. 5, 687–717 MR 4459624

[50] M. Erbar and M. Huesmann, Curvature bounds for configuration spaces. *Calc. Var. Partial Differential Equations* **54** (2015), no. 1, 397–430 Zbl 1335.53047 MR 3385165

[51] M. Erbar and E. Kopfer, Super Ricci flows for weighted graphs. *J. Funct. Anal.* **279** (2020), no. 6, 108607, 51 Zbl 1444.53061 MR 4099474

[52] M. Erbar, K. Kuwada, and K.-T. Sturm, On the equivalence of the entropic curvature-dimension condition and Bochner's inequality on metric measure spaces. *Invent. Math.* **201** (2015), no. 3, 993–1071 Zbl 1329.53059 MR 3385639

[53] M. Erbar and J. Maas, Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ration. Mech. Anal.* **206** (2012), no. 3, 997–1038 Zbl 1256.53028 MR 2989449

[54] M. Erbar, J. Maas, and P. Tetali, Discrete Ricci curvature bounds for Bernoulli–Laplace and random transposition models. *Ann. Fac. Sci. Toulouse Math. (6)* **24** (2015), no. 4, 781–800 Zbl 1333.60088 MR 3434256

[55] M. Erbar, C. Rigoni, K.-T. Sturm, and L. Tamanini, Tamed spaces—Dirichlet spaces with distribution-valued Ricci bounds. *J. Math. Pures Appl. (9)* **161** (2022), 1–69 Zbl 07503507 MR 4403622

[56] M. Erbar and K.-T. Sturm, Rigidity of cones with bounded Ricci curvature. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 1, 219–235 Zbl 1478.53078 MR 4186467

[57] M. Erbar and K.-T. Sturm, On the self-improvement of the curvature-dimension condition with finite dimension. 2022, to appear

[58] S. Fang, J. Shao, and K.-T. Sturm, Wasserstein space over the Wiener space. *Probab. Theory Related Fields* **146** (2010), no. 3-4, 535–565 Zbl 1201.37095 MR 2574738

[59] M. Fathi and Y. Shu, Curvature and transport inequalities for Markov chains in discrete spaces. *Bernoulli* **24** (2018), no. 1, 672–698 Zbl 1396.60084 MR 3706773

[60] W. Gangbo, An elementary proof of the polar factorization of vector-valued functions. *Arch. Rational Mech. Anal.* **128** (1994), no. 4, 381–399 Zbl 0828.57021 MR 1308860

[61] N. Garofalo and A. Mondino, Li–Yau and Harnack type inequalities in $\mathsf{RCD}^*(K,N)$ metric measure spaces. *Nonlinear Anal.* **95** (2014), 721–734 Zbl 1286.58016 MR 3130557

[62] N. Gigli, Second order analysis on $(P_2(M), W_2)$. *Mem. Amer. Math. Soc.* **216** (2012), no. 1018, xii+154 Zbl 1253.58008 MR 2920736

[63] N. Gigli, An overview of the proof of the splitting theorem in spaces with non-negative Ricci curvature. *Anal. Geom. Metr. Spaces* **2** (2014), no. 1, 169–213 Zbl 1310.53031 MR 3210895

[64] N. Gigli, On the differential structure of metric measure spaces and applications. *Mem. Amer. Math. Soc.* **236** (2015), no. 1113, vi+91 Zbl 1325.53054 MR 3381131

[65] N. Gigli, On the regularity of harmonic maps from $\mathsf{RCD}(K,N)$ to $\mathsf{CAT}(0)$ spaces and related results. 2022, arXiv:2204.04317

[66] N. Gigli, C. Ketterer, K. Kuwada, and S.-i. Ohta, Rigidity for the spectral gap on Rcd(K,∞)-spaces. *Amer. J. Math.* **142** (2020), no. 5, 1559–1594 Zbl 1462.58014 MR 4150652

[67] N. Gigli, K. Kuwada, and S.-i. Ohta, Heat flow on Alexandrov spaces. *Comm. Pure Appl. Math.* **66** (2013), no. 3, 307–331 Zbl 1267.58014 MR 3008226

[68] N. Gigli and J. Maas, Gromov–Hausdorff convergence of discrete transportation metrics. *SIAM J. Math. Anal.* **45** (2013), no. 2, 879–899 Zbl 1268.49054 MR 3045651

[69] N. Gigli and E. Pasqualetto, Behaviour of the reference measure on RCD spaces under charts. *Comm. Anal. Geom.* **29** (2021), no. 6, 1391–1414 Zbl 07473916 MR 4367429

[70] N. Gigli, E. Pasqualetto, and E. Soultanis, Differential of metric valued Sobolev maps. *J. Funct. Anal.* **278** (2020), no. 6, 108403, 24 Zbl 1433.53067 MR 4054105

[71] N. Gigli and L. Tamanini, Second order differentiation formula on RCD(K,N) spaces. *Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl.* **29** (2018), no. 2, 377–386 Zbl 1394.53050 MR 3797990

[72] N. Gigli and A. Tyulenev, Korevaar–Schoen's directional energy and Ambrosio's regular Lagrangian flows. *Math. Z.* **298** (2021), no. 3-4, 1221–1261 Zbl 1471.53077 MR 4282128

[73] P. Gladbach, E. Kopfer, and J. Maas, Scaling limits of discrete optimal transport. *SIAM J. Math. Anal.* **52** (2020), no. 3, 2759–2802 Zbl 1447.49062 MR 4110822

[74] P. Gladbach, E. Kopfer, J. Maas, and L. Portinale, Homogenisation of dynamical optimal transport on periodic graphs. 2021, arXiv:2110.15321

[75] N. Gozlan, C. Roberto, P.-M. Samson, and P. Tetali, Displacement convexity of entropy and related inequalities on graphs. *Probab. Theory Related Fields* **160** (2014), no. 1-2, 47–94 Zbl 1332.60037 MR 3256809

[76] A. Greven, P. Pfaffelhuber, and A. Winter, Convergence in distribution of random metric measure spaces (Λ-coalescent measure trees). *Probab. Theory Related Fields* **145** (2009), no. 1-2, 285–322 Zbl 1215.05161 MR 2520129

[77] M. Gromov, *Structures métriques pour les variétés riemanniennes*. Textes Math. 1, CEDIC, Paris, 1981 Zbl 0509.53034 MR 682063

[78] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces*. Mod. Birkhäuser Class., Birkhäuser, Boston, MA, 2007 Zbl 1113.53001 MR 2307192

[79] B. Güneysu and M. von Renesse, Molecules as metric measure spaces with Kato-bounded Ricci curvature. *C. R. Math. Acad. Sci. Paris* **358** (2020), no. 5, 595–602 Zbl 1475.53045 MR 4149858

[80] B.-X. Han and K.-T. Sturm, Curvature-dimension conditions under time change. *Ann. Mat. Pura Appl. (4)* **201** (2022), no. 2, 801–822 Zbl 07495357 MR 4386845

[81] R. Haslhofer and A. Naber, Weak solutions for the Ricci flow. I. 2015, arXiv:1504.00911

[82] S. Honda, New differential operator and noncollapsed RCD spaces. *Geom. Topol.* **24** (2020), no. 4, 2127–2148 Zbl 1452.53041 MR 4173928

[83] D. F. Hornshaw, Quantum optimal transport for approximately finite-dimensional C^*-algebras. 2019, arXiv:1910.03312v1

[84] E. P. Hsu, Multiplicative functional for the heat equation on manifolds with boundary. *Michigan Math. J.* **50** (2002), no. 2, 351–367 Zbl 1037.58024 MR 1914069

[85] R. Jordan, D. Kinderlehrer, and F. Otto, The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29** (1998), no. 1, 1–17 Zbl 0915.35120 MR 1617171

[86] N. Juillet, Diffusion by optimal transport in Heisenberg groups. *Calc. Var. Partial Differential Equations* **50** (2014), no. 3-4, 693–721 Zbl 1378.53044 MR 3216830

[87] V. Kapovitch, C. Ketterer, and K.-T. Sturm, On gluing Alexandrov spaces with lower Ricci curvature bounds. 2020, *Comm. Anal. Geom.* (to appear); arXiv:2003.06242

[88] V. Kapovitch, A. Lytchak, and A. Petrunin, Metric-measure boundary and geodesic flow on Alexandrov spaces. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 1, 29–62 Zbl 07328107 MR 4186463

[89] V. Kapovitch and A. Mondino, On the topology and the boundary of N-dimensional $\mathsf{RCD}(K, N)$ spaces. *Geom. Topol.* **25** (2021), 445–495 Zbl 1466.53050 MR 4226234

[90] M. Kell and A. Mondino, On the volume measure of non-smooth spaces with Ricci curvature bounded below. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **18** (2018), no. 2, 593–610 Zbl 1393.53034 MR 3801291

[91] C. Ketterer, Cones over metric measure spaces and the maximal diameter theorem. *J. Math. Pures Appl. (9)* **103** (2015), no. 5, 1228–1275 Zbl 1317.53064 MR 3333056

[92] C. Ketterer, Obata's rigidity theorem for metric measure spaces. *Anal. Geom. Metr. Spaces* **3** (2015), no. 1, 278–295 Zbl 1327.53051 MR 3403434

[93] C. Ketterer, On the geometry of metric measure spaces with variable curvature bounds. *J. Geom. Anal.* **27** (2017), no. 3, 1951–1994 Zbl 1375.53057 MR 3667417

[94] Y. Kitabeppu, A sufficient condition to a regular set being of positive measure on RCD spaces. *Potential Anal.* **51** (2019), no. 2, 179–196 Zbl 1420.53050 MR 3983504

[95] B. Klartag, Needle decompositions in Riemannian geometry. *Mem. Amer. Math. Soc.* **249** (2017), no. 1180, v+77 Zbl 1457.53028 MR 3709716

[96] B. Kleiner and J. Lott, Singular Ricci flows I. *Acta Math.* **219** (2017), no. 1, 65–134 Zbl 1396.53090 MR 3765659

[97] E. Kopfer, Gradient flow for the Boltzmann entropy and Cheeger's energy on time-dependent metric measure spaces. *Calc. Var. Partial Differential Equations* **57** (2018), no. 1, Paper No. 20 Zbl 1398.35098 MR 3740400

[98] E. Kopfer and K.-T. Sturm, Heat flow on time-dependent metric measure spaces and super-Ricci flows. *Comm. Pure Appl. Math.* **71** (2018), no. 12, 2500–2608 Zbl 1408.58020 MR 3869036

[99] E. Kopfer and K.-T. Sturm, Functional inequalities for the heat flow on time-dependent metric measure spaces. *J. Lond. Math. Soc. (2)* **104** (2021), no. 2, 926–955 Zbl 1478.35008 MR 4311115

[100] K. Kuwada and X.-D. Li, Monotonicity and rigidity of the $\mathcal{W}$-entropy on $\mathsf{RCD}(0, N)$ spaces. *Manuscripta Math.* **164** (2021), no. 1-2, 119–149 Zbl 1461.53032 MR 4203686

[101] K. Kuwae, Y. Machigashira, and T. Shioya, Sobolev spaces, Laplacian, and heat kernel on Alexandrov spaces. *Math. Z.* **238** (2001), no. 2, 269–316 Zbl 1001.53017 MR 1865418

[102] H. Li, Dimension-free Harnack inequalities on $\mathrm{RCD}(K, \infty)$ spaces. *J. Theoret. Probab.* **29** (2016), no. 4, 1280–1297 Zbl 1443.58021 MR 3571246

[103] Z. Li, Remark on Cavalletti–Milman's globalization theorem. To appear

[104] J. Lierl and K.-T. Sturm, Neumann heat flow and gradient flow for the entropy on non-convex domains. *Calc. Var. Partial Differential Equations* **57** (2018), no. 1, Paper No. 25 Zbl 06845942 MR 3742815

[105] S. Liu, F. Münch, and N. Peyerimhoff, Bakry–Émery curvature and diameter bounds on graphs. *Calc. Var. Partial Differential Equations* **57** (2018), no. 2, Paper No. 67 Zbl 1394.53047 MR 3776357

[106] J. Lott and C. Villani, Weak curvature conditions and functional inequalities. *J. Funct. Anal.* **245** (2007), no. 1, 311–333 Zbl 1119.53028 MR 2311627

[107] J. Lott and C. Villani, Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)* **169** (2009), no. 3, 903–991 Zbl 1178.53038 MR 2480619

[108] A. Lytchak and S. Stadler, Ricci curvature in dimension 2. *J. Eur. Math. Soc. (JEMS)* **25** (2023), no. 3, 845–867 Zbl 07683501 MR 4577954

[109] J. Maas, Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.* **261** (2011), no. 8, 2250–2292 Zbl 1237.60058 MR 2824578

[110] M. Magnabosco and C. Rigoni, Optimal maps and local-to-global property in negative dimensional spaces with Ricci curvature bounded from below. 2021, arXiv:2105.12017

[111] M. Magnabosco, C. Rigoni, and G. Sosa, Convergence of metric measure spaces satisfying the CD condition for negative values of the dimension parameter. 2021, arXiv:2104.03588

[112] R. J. McCann, Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80** (1995), no. 2, 309–323 Zbl 0873.28009 MR 1369395

[113] R. J. McCann, Polar factorization of maps on Riemannian manifolds. *Geom. Funct. Anal.* **11** (2001), no. 3, 589–608 Zbl 1011.58009 MR 1844080

[114] R. J. McCann, Displacement convexity of Boltzmann's entropy characterizes the strong energy condition from general relativity. *Camb. J. Math.* **8** (2020), no. 3, 609–681 Zbl 1454.53058 MR 4192570

[115] R. J. McCann and P. M. Topping, Ricci flow, entropy and optimal transportation. *Amer. J. Math.* **132** (2010), no. 3, 711–730 Zbl 1203.53065 MR 2666905

[116] F. Mémoli, Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* **11** (2011), no. 4, 417–487 Zbl 1244.68078 MR 2811584

[117] A. Mielke, A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity* **24** (2011), no. 4, 1329–1346 Zbl 1227.35161 MR 2776123

[118] A. Mielke, Dissipative quantum mechanics using GENERIC. In *Recent Trends in Dynamical Systems*, pp. 555–585, Springer Proc. Math. Stat. 35, Springer, Basel, 2013 Zbl 1315.81061 MR 3110143

[119] E. Milman, Beyond traditional curvature-dimension I: new model spaces for isoperimetric and concentration inequalities in negative dimension. *Trans. Amer. Math. Soc.* **369** (2017), no. 5, 3605–3637 Zbl 1362.53047 MR 3605981

[120] A. Mondino and A. Naber, Structure theory of metric measure spaces with lower Ricci curvature bounds. *J. Eur. Math. Soc. (JEMS)* **21** (2019), no. 6, 1809–1854 Zbl 1468.53039 MR 3945743

[121] A. Mondino and D. Semola, Lipschitz continuity and Bochner–Eells–Sampson inequality for harmonic maps from $\mathsf{RCD}(K,N)$ spaces to CAT(0) spaces. 2022, arXiv:2202.01590

[122] A. Mondino and S. Suhr, An optimal transport formulation of the Einstein equations of general relativity. *J. Eur. Math. Soc. (JEMS)* **25** (2023), no. 3, 933–994 Zbl 07683504 MR 4577957

[123] A. Naber, Characterizations of bounded Ricci curvature on smooth and nonsmooth spaces. 2013, arXiv:1306.6512

[124] S.-i. Ohta, Gradient flows on Wasserstein spaces over compact Alexandrov spaces. *Amer. J. Math.* **131** (2009), no. 2, 475–516 Zbl 1169.53053 MR 2503990

[125] S.-i. Ohta, (K,N)-convexity and the curvature-dimension condition for negative N. *J. Geom. Anal.* **26** (2016), no. 3, 2067–2096 Zbl 1347.53034 MR 3511469

[126] S.-i. Ohta and K.-T. Sturm, Heat flow on Finsler manifolds. *Comm. Pure Appl. Math.* **62** (2009), no. 10, 1386–1433 Zbl 1176.58012 MR 2547978

[127] S.-i. Ohta and A. Takatsu, Displacement convexity of generalized relative entropies. *Adv. Math.* **228** (2011), no. 3, 1742–1787 Zbl 1223.53032 MR 2824568

[128] F. Otto, The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations* **26** (2001), no. 1-2, 101–174 Zbl 0984.35089 MR 1842429

[129] F. Otto and C. Villani, Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173** (2000), no. 2, 361–400 Zbl 0985.58019 MR 1760620

[130] A. Profeta, The sharp Sobolev inequality on metric measure spaces with lower Ricci curvature bounds. *Potential Anal.* **43** (2015), no. 3, 513–529 Zbl 1333.53050 MR 3430465

[131] A. Profeta and K.-T. Sturm, Heat flow with Dirichlet boundary conditions via optimal transport and gluing of metric measure spaces. *Calc. Var. Partial Differential Equations* **59** (2020), no. 4, Paper No. 117 Zbl 1439.35186 MR 4117516

[132] C. Rose, Li–Yau gradient estimate for compact manifolds with negative part of Ricci curvature in the Kato class. *Ann. Global Anal. Geom.* **55** (2019), no. 3, 443–449 Zbl 1411.35134 MR 3936228

[133] G. Savaré, Gradient flows and diffusion semigroups in metric spaces under lower curvature bounds. *C. R. Math. Acad. Sci. Paris* **345** (2007), no. 3, 151–154 Zbl 1125.53064 MR 2344814

[134] G. Savaré, Self-improvement of the Bakry–Émery condition and Wasserstein contraction of the heat flow in $\mathrm{RCD}(K,\infty)$ metric measure spaces. *Discrete Contin. Dyn. Syst.* **34** (2014), no. 4, 1641–1661 Zbl 1275.49087 MR 3121635

[135] K.-T. Sturm, Convex functionals of probability measures and nonlinear diffusions on manifolds. *J. Math. Pures Appl. (9)* **84** (2005), no. 2, 149–168 Zbl 1259.49074 MR 2118836

[136] K.-T. Sturm, On the geometry of metric measure spaces. I. *Acta Math.* **196** (2006), no. 1, 65–131 Zbl 1105.53035 MR 2237206

[137] K.-T. Sturm, On the geometry of metric measure spaces. II. *Acta Math.* **196** (2006), no. 1, 133–177 Zbl 1106.53032 MR 2237207

[138] K.-T. Sturm, The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. 2012, *Mem. Amer. Math. Soc.* (to appear); arXiv:1208.0434

[139] K.-T. Sturm, Ricci tensor for diffusion operators and curvature-dimension inequalities under conformal transformations and time changes. *J. Funct. Anal.* **275** (2018), no. 4, 793–829 Zbl 1419.58020 MR 3807777

[140] K.-T. Sturm, Super-Ricci flows for metric measure spaces. *J. Funct. Anal.* **275** (2018), no. 12, 3504–3569 Zbl 1401.37040 MR 3864508

[141] K.-T. Sturm, Distribution-valued Ricci bounds for metric measure spaces, singular time changes, and gradient estimates for Neumann heat flows. *Geom. Funct. Anal.* **30** (2020), no. 6, 1648–1711 Zbl 1475.53047 MR 4182834

[142] K.-T. Sturm, Remarks about synthetic upper Ricci bounds for metric measure spaces. *Tohoku Math. J. (2)* **73** (2021), no. 4, 539–564 Zbl 07473223 MR 4355059

[143] L. Tamanini, From Harnack inequality to heat kernel estimates on metric measure spaces and applications. 2019, arXiv:1907.07163

[144] C. Villani, *Optimal Transport. Old and New*. Grundlehren Math. Wiss. 338, Springer, Berlin, 2009 Zbl 1156.53003 MR 2459454

[145] M.-K. von Renesse and K.-T. Sturm, Transport inequalities, gradient estimates, entropy, and Ricci curvature. *Comm. Pure Appl. Math.* **58** (2005), no. 7, 923–940 Zbl 1078.53028 MR 2142879

[146] F.-Y. Wang, Second fundamental form and gradient of Neumann semigroups. *J. Funct. Anal.* **256** (2009), no. 10, 3461–3469 Zbl 1165.58014 MR 2504531

[147] M. Weber, E. Saucan, and J. Jost, Characterizing complex networks with Forman–Ricci curvature and associated geometric flows. *J. Complex Netw.* **5** (2017), no. 4, 527–550 MR 3801701

[148] M. Wirth, A dual formula for the noncommutative transport distance. *J. Stat. Phys.* **187** (2022), no. 2, Paper No. 19 Zbl 07506512 MR 4406600

[149] H.-C. Zhang and X.-P. Zhu, Yau's gradient estimates on Alexandrov spaces. *J. Differential Geom.* **91** (2012), no. 3, 445–522 Zbl 1258.53075 MR 2981845

Karl-Theodor Sturm
Hausdorff Center for Mathematics, University of Bonn, 53115 Bonn, Germany;
sturm@uni-bonn.de

DOI 10.4171/8ECM/05

Torsion in algebraic groups and problems which arise

Umberto Zannier

Abstract. This article is based on the lecture that I had the honor and pleasure to deliver at the 8th European Congress of Mathematics in Portorož, Slovenia (originally planned for June 2020, then shifted to June 2021 for public health reasons). In the talk I tried to give an overview of some issues linked to torsion in algebraic groups, focusing on some recent research. Taking into account the purposes of reaching a large audience of mathematicians, from all subjects, I started with elementary general concepts, recalling some historical steps, before shifting to more specific themes which I was more familiar with. In these notes, I maintained the same principles, and only slightly expanded the contents of the lecture; indeed, I have not gone into any detailed argument.

1. Torsion in commutative algebraic groups

Torsion (etymology): the word *torsion* (in mathematics) often denotes a quantity, suitably defined in differential terms, which measures local "twisting" of a curve in Euclidean space (roughly speaking, it expresses "how far" the curve locally is from being a plane curve). However, in this exposition we shall adopt the usual *algebraic* meaning, namely according to the following definition.

Definition. An element g in a group G is *torsion* if $g^m = 1 =$ identity of G, for some integer $m > 0$. (Such an m is called an *exponent* for g, whereas the minimal such m is called *the* – exact – *order* of g.)

This terminology apparently is not unrelated to the former one, as it seemingly originated from the structure of homology groups of spaces obtained by *twisting*. For instance, the real projective plane $\mathbb{P}_2(\mathbb{R})$, defined by gluing antipodal points in a closed half sphere, has the torsion group $\mathbb{Z}/2$ as its first homology group (over $\mathbb{Z}$).

A torsion element g as above generates a so-called *finite cyclic* group; now the etymology comes from the *circle*, because the powers g^n repeat *cyclically*: $\ldots, g, g^2, \ldots, g^{m+1} = g, g^{m+2} = g^2, \ldots$ and generally $g^{n+m} = g^n$, $n = 0, 1, \ldots$.

2020 Mathematics Subject Classification. Primary 11-02; Secondary 11Gxx, 14Gxx, 14Kxx, 14Lxx.
Keywords. Number theory, Diophantine geometry, algebraic groups.

Indeed the *circle* comes into the picture beyond this simple intuition, through its topology (especially the fundamental group).

1.1. Algebraic groups

We shall consider some examples of torsion elements, and their structure, in *algebraic groups*: roughly speaking an algebraic group is defined first as an algebraic variety, i.e., a set of points satisfying a given system of algebraic equations in an affine or projective space, and then one has a group law expressed by polynomials in the coordinates.

An algebraic group is an irreducible variety if and only if it is connected, and, in general, it is anyway a finite union of translates of the connected component of the identity element (which is a normal subgroup).

In this article we shall meet only *commutative algebraic groups*, a property which entails that torsion elements form a subgroup.

For simplicity we shall consider only algebraic groups and points defined over the field $\mathbb{C}$ of complex numbers and tacitly identify such a group with the set of its complex points. (However, this does not mean that we shall disregard the actual *minimal* field of definition of the points of interest for us, a field which may be small and is highly important for arithmetical information.)

Examples.

Additive algebraic group. The *additive algebraic group*, denoted by $\mathbb{G}_\mathrm{a}$, is simply the affine line $\mathbb{A}^1$ as an algebraic variety. The group law is expressed additively by $(x, y) \mapsto x + y$. The set of complex points $\mathbb{G}_\mathrm{a}(\mathbb{C})$ of $\mathbb{G}_\mathrm{a}$ is simply $\mathbb{C}$.

A torsion element $g \in \mathbb{C}$ of exponent m now satisfies $mg = 0$, hence $g = 0$, which means that there is no torsion other than 0 (as over any field of characteristic zero, whereas every element is torsion of exponent p in positive characteristic p).

Multiplicative algebraic group. The *multiplicative algebraic group* $\mathbb{G}_\mathrm{m}$ is the affine line deprived of the origin $\mathbb{A}^1 - \{0\}$ as an algebraic variety, with the algebraic group law $(x, y) \mapsto xy$. The set $\mathbb{G}_\mathrm{m}(\mathbb{C})$ of its complex points is the multiplicative group of nonzero complex numbers $\mathbb{C}^* := \mathbb{C} \setminus \{0\}$.

A torsion element $g \in \mathbb{C}^*$ of exponent m satisfies $g^m = 1$, so the torsion elements are precisely the (complex) roots of unity. There are m having exponent m; these lie on the *unit circle* $S_1 := \{z \in \mathbb{C} : |z| = 1\}$, and they form the vertices of a regular m-gon in the complex plane $\mathbb{C}$.

Note the (analytic) exponential map $z \mapsto e^{2\pi i z}$, which sends homomorphically $\mathbb{C}$ onto $\mathbb{C}^*$ and has a kernel $\mathbb{Z} \cong \pi_1(\mathbb{C}^*)$; this is a non-divisible group, which explains torsion elements in the image (whereas there are no nontrivial ones in the domain).

Through this map we have the *analytic* isomorphism $\mathbb{G}_{\mathrm{m}}(\mathbb{C}) = \mathbb{C}^* \cong \mathbb{C}/\mathbb{Z}$, the quotient of $\mathbb{C}$ by a *discrete* subgroup (of rank 1).

Elliptic curves. The additive and multiplicative algebraic groups are curves (i.e., have dimension 1). However, they do not exhaust the possibilities for a curve to be a connected algebraic group. Indeed, (the) other fundamental examples are given by (complex) *elliptic curves*. They can be defined in the *projective* plane $\mathbb{P}_2$ by equations of the shape

$$E\colon y^2 = 4x^3 - g_2x - g_3 \text{ in } \mathbb{A}^2 + \text{point at infinity } O := (0:1:0) \text{ in } \mathbb{P}_2,$$

where $g_2, g_3 \in \mathbb{C}$ are such that $4x^3 - g_2x - g_3$ has no multiple roots, i.e.,

$$g_2^3 - 27g_3^2 \neq 0.$$

What is particularly remarkable is the existence of an algebraic-group (commutative) law among the points of each such curve. Namely, if we prescribe that the origin is the point at infinity O, then to add points $P, Q \in E$, we first draw the line through P, Q (or the tangent to E if $P = Q$) which (taking into account multiplicities) will intersect E in a third point R. The group law is such that $P + Q + R = O$, whereas $P + Q$ is the point opposite to R with respect to the x-axis.

This outstanding law (called classically the *chord and tangent process*, apparently observed first by Newton) indeed may be expressed by polynomials in the homogeneous coordinates and satisfies the group axioms (the associative law being not entirely trivial to check). It is fundamental in several respects, e.g., in the theory of Diophantine equations, since, when the curve has rational coefficients, it produces rational points out of rational ones.

Somewhat similarly to the case of $\mathbb{G}_{\mathrm{m}}$, each of these curves is found to be *analytically* isomorphic to a (compact) *complex torus* $\mathbb{C}/L$, where L is again a suitable discrete subgroup, however now of maximal rank 2, i.e., a lattice in $\mathbb{C}$. The isomorphism occurs through the *Weierstrass exponential map*: $z \mapsto (\wp_L(z), \wp_L'(z))$, where $\wp_L$ is the Weierstrass function associated to L:

$$\wp_L(z) = z^{-2} + \sum_{l \in L - \{0\}} \big((z-l)^{-2} - l^{-2}\big).$$

This function is meromorphic on $\mathbb{C}$ and admits L as its group of periods. (It sends L to O.)

The addition $+$ on $\mathbb{C}$ (and on $\mathbb{C}/L$) then explains the group law on E in the sense that the former is transported to the latter by the said exponential.

In particular, it appears that now there are m^2 torsion elements of exponent m.

The complex elliptic curves, up to complex isomorphism, form a family of dimension 1 (parameterized by the so-called j-invariant $j(E) = 1728g_2^3/(g_2^3 - 27g_3^2)$, which can assume any complex value). Together with $\mathbb{G}_{\mathrm{a}}$ and $\mathbb{G}_{\mathrm{m}}$, they exhaust the isomorphism classes of complex (connected) algebraic groups of dimension 1.

Abelian varieties. *Abelian varieties* are the irreducible (or, equivalently, connected) projective algebraic groups. They are automatically commutative (the terminology "abelian" arising for a different reason).

Elliptic curves represent precisely the abelian varieties of dimension 1. But abelian varieties exist in any dimension: a simple example is a power E^r of an elliptic curve E, though this is extremely special. Other very important (though still special) abelian varieties arise as *Jacobians of (smooth) algebraic curves* of genus $g > 0$; such a Jacobian has dimension g.

Like for elliptic curves, every complex abelian variety, say of dimension g, is *analytically* isomorphic to a complex torus, i.e., a quotient $\mathbb{C}^g/L$ where L is a (full) lattice; however for $g > 1$ not every complex torus is an abelian variety, a certain subtle "bilinear" condition on the lattice (existence of a Riemann *form*), in heavy part arithmetical, being necessary and sufficient.

Products. We may obtain other algebraic groups by taking products, e.g., of the form $\mathbb{G}_{\mathrm{a}}^r \times \mathbb{G}_{\mathrm{m}}^s$; the complex points are now vectors in $\mathbb{C}^{r+s}$, where the last s coordinates are nonzero and where the operations are coordinatewise (additive on the first r coordinates, multiplicative on the last s ones). For topological reasons the powers $\mathbb{G}_{\mathrm{m}}^s$ are sometimes called *(complex multiplicative) tori.*

Similarly, we may take products among the other algebraic groups we have seen. However, one should take into account that there exist *extensions* of algebraic groups, i.e., exact sequences $0 \to G_1 \to G \to G_2 \to 0$, where G is not (necessarily isomorphic to) the product $G_1 \times G_2$ (examples occur already in dimension 2, on taking $G_1 = \mathbb{G}_{\mathrm{a}}$ or $\mathbb{G}_{\mathrm{m}}$ and $G_2 =$ an elliptic curve). When $G_1 = \mathbb{G}_{\mathrm{m}}^s$ and G_2 is an abelian variety, any G in such an exact sequence is called a *semiabelian variety.*

1.2. Some results about torsion in algebraic groups

Additive case. We have already noted that $\mathbb{G}_{\mathrm{a}}$ has no nontrivial torsion in characteristic zero, thus in particular over $\mathbb{C}$.

Multiplicative case. In this case, we have recalled that the torsion elements of $\mathbb{G}_{\mathrm{m}}(\mathbb{C}) = \mathbb{C}^*$ are the complex roots of unity.

Through the exponential map $z \to e^{2\pi i z}$, the roots of unity correspond to $z =$ rational number, which raises a link with Number Theory.

Roots of unity naturally appear in describing *discrete periodical phenomena.* For instance one finds here *finite Fourier series*, i.e., linear combinations of exponential functions (on $\mathbb{Z}$) having roots of unity as bases; they are a discrete counterpart of the famous series expansions introduced systematically by Fourier, at the very heart of Harmonic Analysis. The finite Fourier series represent all periodic functions on $\mathbb{Z}$

and are of the utmost relevance in myriads of topics and applications, including Coding Theory, Combinatorics, Fast Multiplication, Group Theory, (Analytic) Number Theory, Numerical Analysis, and so on.

An effectivity issue. Let us pause to note that, already for roots of unity, even to *decide* whether a specific number (or, more generally, specific point of a given algebraic group) is torsion seems not completely obvious. For instance, consider the following:

$$\text{Small challenge:} \quad \textit{Is } \alpha := \frac{3+4\sqrt{-1}}{5} \textit{ a root of unity?}$$

Note that α indeed lies in the unit circle S_1, as $3^2 + 4^2 = 5^2$. So the natural test of computing $|\alpha|$ does not disprove the sought eventuality for this particular number.

Now, we can check whether $\alpha^2, \alpha^3, \ldots$, or any *given* power α^m, is or is not equal to 1, but a possible torsion-exponent is not bounded *a priori*; so, unless we find 1 at some stage, we are left with an open possibility for the next check.

In conclusion, a little reflection may be needed to (see how to) answer such (type of) question(s) algorithmically in the general case. (Now a negative answer may be obtained using the Euler function value $\phi(q)$ for the degree over $\mathbb{Q}$ of a root of unity of exact order q, which bounds the possible torsion order in terms of the degree of the given number. This works generally, but for the actual question maybe the simplest way is to observe that α is not an algebraic integer.[1] We invite the interested reader to seek several different arguments for answering the question.)

From a number theoretical viewpoint, Gauss (*Disquisitiones Arithmeticae 1801*) was the first to study in depth the arithmetical properties of roots of unity. In particular, this led him to criteria for *constructing a regular n-gon with ruler and compass* (ancient problem of Greek mathematics). For instance this is possible for

$$n = \mathbf{3}, 4, \mathbf{5}, 6, 8, 10, 12, 15, 16, \mathbf{17}, \ldots, \quad \text{but not for } n = 7, 9, 11, 13, 14, 18, 19, \ldots.$$

As is well known, Fermat primes $2^{2^k} + 1$ play a heavy role here....

In fact, already a few years before the publication of the *Disquisitiones*, Gauss had succeeded to construct the regular polygon of 17 sides, obtaining in practice the remarkable equality

$$16\cos\frac{2\pi}{17} = -1 + \sqrt{17} + \sqrt{34 - 2\sqrt{17}} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{170 + 38\sqrt{17}}}.$$

We may say that Gauss anticipated the *Galois theory of the cyclotomic fields*; in fact, in particular he defined the so-called *Gaussian periods*, which *a posteriori* turn

[1] This fits with a well-known theorem of Kronecker: "Roots of unity are those algebraic *integers* having all conjugates of complex absolute value 1", which may be rephrased as: "An algebraic number is a root of unity if and only if *all* its absolute values are 1".

out to be suitable invariants for subgroups of automorphisms. (They may be also conceived as values of certain *finite Fourier series* alluded to above.) For instance, Gauss obtained from them "explicit" generators for all the subfields of a *principal* cyclotomic field $\mathbb{Q}(e^{2\pi i/p})$, where p is a prime number. (They are highly important for other reasons as well.) In particular, Gauss expressed, through the famous *Gauss sums*, any number $\sqrt{n}$, $n \in \mathbb{Z}$, as a sum of roots of unity, which is not at all obvious. This also started the theory of abelian extensions of $\mathbb{Q}$ and of number fields (so-called "Class-field theory").

In this case of the roots of unity (through viewpoints introduced by Deuring, ..., Tate, ..., Grothendieck, ...) the Galois groups which arise may be seen as an algebraic manifestation and realization of the *monodromy (group) of the circle* $S_1 = \{z \in \mathbb{C} : |z| = 1\}$. For instance, we have

$$\pi_1(S_1) = \pi_1(\mathbb{C}^*) = \mathbb{Z},$$

and its finite quotients are the $\mathbb{Z}/(m)$ which correspond to the finite covers of S_1, and so to the homomorphisms

$$S_1 \to S_1\colon \quad z \mapsto z^m,$$

with kernel the group $U_m = e^{2\pi i \mathbb{Z}/m} \cong \mathbb{Z}/(m)$ of mth roots of unity. So U_m is the topological covering group and the Galois group over $\mathbb{Q}$ acts on it, and we have

$$\mathrm{Aut}(U_m) \cong \big(\mathbb{Z}/(m)\big)^* \cong \mathrm{Gal}\,\big(\mathbb{Q}(U_m)/\mathbb{Q}\big) = \mathrm{Gal}\,\big(\mathbb{Q}(e^{2\pi i/m})/\mathbb{Q}\big),$$

as proved essentially by Gauss. So the algebraic Galois group of the corresponding field extension equals the (abstract) automorphism group of the topological covering group.

Elliptic case. Now the theory of torsion elements is again highly interesting, rich, and actually (much) more difficult than in the cyclotomic case. We have already recalled that there are m^2 elements of exponent m. The coordinates of these points generate (over the ground field $\mathbb{Q}(g_2, g_3)$) a field which is found to contain the cyclotomic field $\mathbb{Q}(e^{2\pi i/m})$, so we may say that the cyclotomic case recalled above falls just as a special piece of the elliptic theory.

The torsion points now lie on a space which may be identified with the product $S_1 \times S_1$ of two circles (a torus), and the topological covering group corresponding to torsion points of order m is now $(\mathbb{Z}/(m))^2$. The elements of the Galois group again correspond to automorphisms of the covering group and thus may be viewed inside the finite matrix group $\mathrm{GL}_2(\mathbb{Z}/(m))$. A fundamental issue is to understand the image of the Galois group (as m varies). This Galois theory somewhat depends on the coefficients g_2, g_3.

The "generic" case of transcendental j-invariant had been dealt with by Fricke & Weber between the XIX and XX centuries: they proved that the image is essentially the "largest possible one" (i.e., $\mathrm{SL}_2(\mathbb{Z}/(m))$ if we work over $\mathbb{C}$).

The algebraic case lies much deeper, and suppose to fix ideas that $g_2, g_3 \in \mathbb{Q}$. There are two essentially different subcases, according to whether the ring of endomorphisms of the elliptic curve is "trivial" (i.e., $\mathbb{Z}$) or not; the latter case, called *Complex Multiplication*, is "exceptional" in various ways (now the endomorphism ring is an *order* in some imaginary quadratic field).

Already Gauss (beginning of Chapter VII of *Disquisitiones*) foresaw the interest and depth of this issue in some of these situations, which he interpreted as analogous to cyclotomy, i.e., as the (arithmetical) theory of the dissection of a *lemniscate* (in place of a circle) into equal parts. It is also interesting that the general case of prime m had been considered by Galois (in a letter to Chevalier, 29 May 1832), especially from the viewpoint of *solvability by radicals* of the corresponding algebraic equations.

We skip any other detail and only recall a few basic more modern achievements.

Some elliptic results

- A deep landmark result on the above-mentioned Galois image is Serre's *Open Image Theorem* (70s): in a sense it extends Gauss' achievements (and more) to the most general elliptic case, proving that the Galois image is as large as possible (compatibly with the endomorphism ring) up to bounded index. (We omit a precise statement, which would lead us outside the scope of these notes.)
- Another very important and deep theorem is due to Mazur (70s), who proved that for $g_2, g_3 \in \mathbb{Q}$ the possible torsion orders of *rational* torsion points never exceed 12. This result corresponds to finding all rational points on suitable *modular curves*, providing a link of the present topic with major questions in the theory of Diophantine equations.
- Merel 1994, with some new ingredient, extended this kind of result to number fields other than $\mathbb{Q}$ (some independent work being due to Kamienny & Parent, and previously to Demianenko & Manin in the case of prime-power torsion order).

Case of abelian varieties. The arithmetic and Galois structure of torsion elements is even more difficult than the special elliptic case. But nowadays there has been great progress, thanks to the work of Deligne, Bogomolov, Faltings, Serre, . . . , Masser & Wüstholz, . . . , Mazur, Ribet, Pink, Tamagawa, Cadoret, ...

2. Algebraic relations among torsion points

We have recalled some results on individual torsion points. Let us now see some problems on *relations* among torsion points.

An old significant example comes from Gordan 1877 who studied the equation

$$\cos 2\pi x + \cos 2\pi y + \cos 2\pi z + 1 = 0, \quad x, y, z \in \mathbb{Q},$$

with the purpose of classifying the finite subgroups of $\mathrm{PGL}_2(\mathbb{C})$. On writing

$$2\cos 2\pi z = e^{2\pi i z} + e^{-2\pi i z},$$

we see that this amounts to a certain algebraic equation among the three roots of unity $e^{2\pi i x}$, $e^{2\pi i y}$, $e^{2\pi i z}$, or else a certain (inhomogeneous) linear relation among three roots of unity and their reciprocals.

Later on, general *linear* equations in roots of unity were studied systematically, in particular by Mann 1965 and Conway & Jones 1976, in the setting of what the latter authors called *trigonometric diophantine equations*. These results in particular bounded the maximal torsion order in a linear equation with nonzero constant term and no vanishing subsums (with coefficients in $\mathbb{Q}$). As a very special instance, their conclusions very easily imply that

> *the only triangles with rational sides and angles rational multiples of π are equilateral,*

and similar results follow for polygons with a given number of sides.

More recent applications appear, for example, in the work of Gross, Hironaka, and McMullen (to cyclotomic factors of *E_n-Coxeter polynomials*, 2009), of Bourgain, Gamburd, and Sarnak (to *Markov surfaces* 2016), of Kedlaya, Kolpakov, Poonen, and Rubinstein (to rational angles among vectors in $\mathbb{R}^3$, 2020), and in a joint work of the author with Dvornicich & Veneziano (to rational angles in plane lattices, 2020).

Uniform *quantitative* results (regarding the number of solutions of a given linear equation) were proved, e.g., by Schlickewei, Evertse, Beukers & Smyth, and in a joint work of the author with Bombieri, also towards the conjecture of Lang to be discussed in the next section. (Subsequently these results have been quantitatively refined by several authors, including Amoroso & Viada and Martinez.)

2.1. The conjecture of Lang

Independently of the above authors, Lang had raised in the 60s the related problem of studying polynomial equations:

$$F(\theta, \zeta) = 0, \quad \theta, \zeta \text{ roots of unity, of } \textit{unrestricted} \text{ exponent.}$$

Note that such a pair (θ, ζ) is a *torsion point* on the plane curve $F(x, y) = 0$, viewed inside $\mathbb{G}_{\mathrm{m}}^2$.

Let F be given. As expected by Lang, there can be infinitely many solutions of the said shape only if F has a binomial factor of the shape $ax^m + by^n$ or $ax^m y^n + b$; this was quickly proved by Ihara, Serre, and Tate (and proofs may be got also through the previously mentioned results on linear equations, on considering monomial terms).

This was later extended to arbitrary dimensions by M. Laurent and Sarnak & Adams, proving (among other things) the *conjecture of Lang*:

> *An algebraic subvariety of* $\mathbb{G}_{\mathrm{m}}^r$ *can have infinitely many torsion points only if it contains a positive dimensional* special *subvariety, i.e., a translate of an algebraic subgroup by a torsion point.*

More precisely, their results yield the following theorem.

Theorem 2.1. *Let* Σ *be any set of torsion points inside* $\mathbb{G}_{\mathrm{m}}^r$. *The Zariski closure of* Σ *is a finite union of translates (by torsion points) of algebraic subgroups.*

It is moreover not difficult to see that any connected algebraic subgroup of $\mathbb{G}_{\mathrm{m}}^r$ can be defined by finitely many equations of the shape $x_1^{a_1} \cdots x_r^{a_r} = 1$ and is (algebraically) isomorphic to some $\mathbb{G}_{\mathrm{m}}^h$ $(h \leq r)$. Hence in practice the principle is that

> *every prescribed algebraic relation within varying torsion elements can be explained in finite terms by a multiplicative structure of algebraic group.*

Methods. The Galois theory of Gauss is a precious tool in all these achievements, though also other ingredients are relevant.

As for the previously mentioned work, later this had several applications.

For instance we quote the work by Sarnak (on *Betti numbers of congruence groups*, 1994), by Ailon & Rudnick (on $\gcd(f(t)^n - 1, g(t)^n - 1)$, 2004), by Kurasov & Sarnak (on *crystalline measures*, 2020).

2.2. Multiplicative relations on curves – unlikely intersections

The mentioned issues on torsion points may be extended to deal with more general *multiplicative relations* among coordinates of points on a curve X inside a torus $\mathbb{G}_{\mathrm{m}}^r$. That is, we weaken the condition that all the coordinates are torsion and only impose that a certain number of independent multiplicative relations hold among the coordinates.

It is easy to see that if we prescribe on the irreducible curve X a single multiplicative relation, i.e., of the shape $x_1^{a_1} \cdots x_r^{a_r} = 1$ with $(x_1, \dots, x_r) \in X$, then we obtain infinitely many points as $(a_1, \dots, a_r)$ varies through all nonzero integer vectors; this corresponds to intersect X with the union of all proper algebraic subgroups of $\mathbb{G}_{\mathrm{m}}^r$. However, it turns out that imposing another such relation, independent with the former but otherwise arbitrary (which corresponds to intersect X with the union of algebraic subgroups of codimension ≥ 2), yields only finitely many points, unless the curve X is *special* in the sense that it is contained in a *proper* algebraic subgroup of $\mathbb{G}_{\mathrm{m}}^{r}$. (For $r = 2$ we obtain nothing new since imposing two independent relations yields torsion coordinates, but for $r > 2$ this is a weaker restriction, making the theorem stronger.)

With the more demanding assumption that X is not contained in any proper *translate* of algebraic subgroup, this was dealt with in a joint work of the author with Bombieri & Masser 1999, and later proved in the sharper form by Maurin 2008, relying partly on methods by Rémond. (This case is more difficult; for instance it contains implicitly the so-called Mordell–Lang context for tori.) A different approach for this stronger theorem was found later by the former authors with Habegger 2010 (this time *using* the results of Mordell–Lang type). A still further approach with the stronger assumption appeared in a joint work of the author with Capuano, Masser, and Pila 2016, based on the counting method alluded to below; this argument has the advantage of extending to the abelian context (but not containing the sharper form).

Several results followed by other authors as well, also for some higher-dimensional analogues, and further in the abelian case.

The topic, sometimes called *Unlikely Intersections*, was independently raised also by Zilber 2002 (with entirely independent motivations from Logic) and again independently by Pink 2005.[2] They put forward certain general conjectures still widely open (those of Pink embracing still further realms). These conjectures dealt also with abelian varieties in place of tori, where exact analogues may be stated. We shall briefly discuss this context in the next subsection.

2.3. The conjecture of Manin–Mumford

A motivation for the above-mentioned problems stated by Lang had been a conjecture formulated independently by Manin & Mumford in the 60s.

Manin–Mumford conjecture. *A curve of genus ≥ 2 embedded in its Jacobian variety has only finitely many torsion points.*

This may be indeed seen as an analogue (of more difficult nature) for abelian varieties of some of the above problems for multiplicative tori. It became a theorem due to Raynaud in the 80s; he was able to prove, more generally, the analogue of Lang's conjecture above, and for arbitrary abelian varieties (not merely Jacobians) and arbitrary subvarieties. Several other proofs then followed, due, e.g., to Serre, Coleman, Hindry, Buium, Hrushovski, Pink & Roessler, M. Baker & Ribet.

Still other proofs (by Bilu 1997 for $\mathbb{G}_{\mathrm{m}}^{r}$ and Szpiro, Ullmo, and S. Zhang 1997 for the abelian case) gave stronger results of *Galois equidistribution* of the conjugates of torsion points when the degree of the field of definition of the points grows. Moreover,

[2] Certain rather special cases had been raised earlier by Schinzel, with still different language and motivations, coming mainly from his theory of reducibility of lacunary polynomials.

these proofs worked also for points of "small height".[3] Remarkable *uniform* estimates here (e.g., for the number of torsion points on the curve) have been given very recently by Kuehne 2021.

A further proof was found in a joint work of Pila and the author (2009): this relied on the analytic isomorphism of a complex abelian variety with a complex torus, in which torsion points correspond to rational points (as in the case of roots of unity). Then one reduces to *counting rational points on suitable analytic subvarieties* of the torus and comparing bounds from below (coming from the large degree of torsion points – work by Masser) and from above. This last step is done through estimates by Bombieri & Pila 1989, Pila, and finally Pila–Wilkie 2006. In turn, this last work involves the (model)-theory of the so-called *o-minimal structures* (developed by van der Dries et al.).

2.4. "Special points" and the André–Oort conjecture

As far-reaching analogues of torsion points, one may consider the so-called *special points* in *Shimura varieties*. An important kind of such varieties arises as moduli spaces of abelian varieties with certain properties (i.e., parametrizing abelian varieties of given dimension with supplementary symmetries). The *special points*, playing the role of torsion points, are those corresponding to Complex-Multiplication abelian varieties. Moreover, one may also define *special subvarieties* of positive dimension, analogues of the translates (by torsion points) of algebraic subgroups (in the conjecture of Lang) or of abelian subvarieties (in the theorem of Raynaud, in turn analogue of Lang's for abelian varieties).

We skip any formal definition, since the context is quite technical, but we note that one may formulate statements analogue to the above ones. A very relevant instance is the *André–Oort conjecture*, raised by André 1989 and Oort 1990s independently. Once that the above terminology has been introduced in precise terms, a possible phrasing of it is as follows:

> *The Zariski closure of a set of special points is a finite union of special subvarieties.*

This formulation reminds of what we have seen in the multiplicative and abelian cases.

After the proof of a special case by André (i.e., the significant case of CM-points on a curve in the plane $\mathbb{A}^2$, viewed as representing pairs of elliptic curves), the above-

[3] The *height* of a point with algebraic coordinates is a real non-negative number which measures its arithmetical complexity; one may define a *canonical* height on the algebraic points of a commutative algebraic group, which vanishes precisely at torsion points; we cannot pause here on this concept, introduced first by Weil, despite its great relevance.

mentioned *counting method* was applied by Pila to this context, proving substantially more general instances.

A final step for the full conjecture (for the moduli space $\mathcal{A}_g$) was finally devised by Tsimerman 2015 after many important intermediate results and steps, in particular by Colmez, Edixhoven, Gao, Klingler, Pila, Pila & Tsimerman, Ullmo & Yafaev, Yuan & S. Zhang, ... (A still more general form of the conjecture has been obtained very recently by Pila, Shankar, and Tsimerman, relying also on further ingredients provided by Binyamini and by Esnault & Groechening.)

Several other results in a similar spirit have been obtained and much work in the context is still in progress.

Dynamical analogues. Still further analogues of special points occur in *dynamics*, which we describe roughly as the study of iterates $f, f^{\circ 2} := f \circ f, \dots, f^{\circ n}, \dots$ of a map $f : X \to X$ from a space X to itself. The simplest examples (already leading often to very difficult problems) occur with rational maps $f : \mathbb{P}_1 \to \mathbb{P}_1$. As possible analogues of torsion points one can consider *preperiodic points* for f, i.e., the points $x \in X$ such that $f^{\circ n}(x) = f^{\circ m}(x)$ for some integers $n > m$ (so that the sequence $(f^{\circ r}(x))_{r \in \mathbb{N}}$ is finite). For instance, if $X = \mathbb{G}_{\mathrm{m}}$, $f(x) = x^d$ (any $d \geq 2$), the preperiodic points are precisely the torsion points. One may formulate analogues of the above statements, and some quite nontrivial remarkable results have been proved, mainly due to the work of M. Baker, Bell, DeMarco, Ghioca, Hsia, Mavraki, Scanlon, Silverman, Szpiro, Tucker, Yuan, S. Zhang, ..., among others. However, only a partial picture has been obtained to date in this direction compared to the original context of torsion points, and even a satisfactory formulation of suitable *complete* conjectures seems not to have been reached so far.

3. Torsion in families of algebraic groups

We have briefly discussed torsion in individual algebraic groups, and algebraic relations among them. To go one step further, we can consider torsion conditions in algebraic groups (and points) varying in families. The multiplicative group $\mathbb{G}_{\mathrm{m}}$ does not admit genuine "variation", but already for elliptic curves we have truly *nonconstant families*. A typical and historically relevant instance of this is the *Legendre family* of elliptic curves, defined by

$$\mathcal{L}_\lambda \colon y^2 = x(x-1)(x-\lambda) + \text{point at infinity } O,$$

where λ is a complex parameter in $\mathbb{C} - \{0, 1\}$. For each $b \in \mathbb{C} - \{0, 1\}$ up to two exceptions, there are only six values of λ producing a curve isomorphic to $\mathcal{L}_b$, and each complex elliptic curve is isomorphic to some $\mathcal{L}_b$, so the family indeed is intrinsically not "constant".

Regarding families of points (also called *sections*), we may consider, just as a simple instance, the points

$$P_\lambda = \big(2, \sqrt{2(2-\lambda)}\big) \in \mathcal{L}_\lambda,$$

where the choice of the sign is immaterial for us.

It may be shown that

(1) P_λ is not identically torsion on $\mathcal{L}_\lambda$ (i.e., there is no integer $m > 0$ such that $mP_\lambda = O$ for all λ), but

(2) P_b becomes torsion on $\mathcal{L}_b$ (of unrestricted exponent) for an infinite, even dense, set of $b \in \mathbb{C}$. This set consists of algebraic numbers, and the corresponding minimal torsion exponents tend to ∞;

(3) these numbers b have *bounded height*. So for instance *there are only finitely many rational or even quadratic irrational ones*, and in fact the degree over $\mathbb{Q}$ of these numbers tends to ∞. (Néron had previously shown that they form a so-called *thin* set in any given number field.)

Property (1) follows from the general principle that torsion points are *unramified* except above the locus of *bad reduction*. Property (2) may be proved through the Betti map, mentioned below. Property (3) follows from results by Silverman & Tate 1980s. Properties (2) and (3) actually hold for all sections (defined over $\overline{\mathbb{Q}}$) satisfying (1).

Further Galois-equidistribution results for these numbers b are due to DeMarco & Mavraki 2019. Note that the equidistribution here does not concern the (conjugates of the) hypothetical torsion points, but regards the (conjugates of the) values b for which P_b is a torsion point. Hence this result has a quite different meaning with respect to the previously mentioned equidistribution theorems of Bilu and Szpiro, Ullmo, and Zhang. This equidistribution implies in particular the above-mentioned complex density.

For the actual choice of family (using the Betti map appearing below) one can also prove density of the relevant b in the real half-line $(-\infty, 2)$, so that $P_b \in \mathcal{L}_b(\mathbb{R})$. On the other hand, a joint work of the author with Lawrence observes that we never have p-adic density for this set.

3.1. Masser's problem and the Pink conjectures

Masser considered a second family of points, for instance

$$Q_\lambda = \big(\lambda + 1, \sqrt{\lambda(\lambda+1)}\big) \in \mathcal{L}_\lambda.$$

The same remarks (1), (2), and (3) hold for this family, and moreover P_λ, Q_λ may be shown to be *generically linearly independent* on $\mathcal{L}_\lambda$, i.e., if $rP_\lambda + sQ_\lambda = O$ for certain integers r, s and all λ, then $r = s = 0$.

From (3) we see that the values b of λ for which each point becomes separately torsion form a *sparse* set, so Masser asked the following.

Masser's question. Is the "*doubly sparse*" set

$$\{b \in \mathbb{C} : P_b, Q_b \text{ are } \textit{both} \text{ torsion on } \mathcal{L}_b\}$$

even a finite set?

Here Galois groups of torsion points do not give enough information, essentially because the degree of the relevant numbers "b" is unbounded (and actually tends to ∞).

Using the above-mentioned counting method (and other tools), Masser & the author (2008) gave an affirmative answer to the question, actually to its natural generalization to other pairs of families and sections.

Later this was further extended to arbitrary algebraic pencils of abelian varieties and in other directions (e.g., of *Unlikely Intersections* type), also by M. Baker, Barroero, Bertrand, Capuano, Daw, DeMarco, Dill, Habegger, G. Jones, Orr, Pila, Pillay, H. Schmidt, Stoll, Tsimerman, ...

Some of these results may be seen as *relative* analogues of the Manin–Mumford conjecture (i.e., where the ambient abelian variety moves in a family), and some other ones as *dynamical* analogues (i.e., when the torsion points are replaced by *preperiodic points* with respect to suitable rational maps).

The problem of Masser was recognized as a special case of conjectures by Pink (and also of Zilber in other cases). As alluded above, these conjectures deal with much more general contexts (including the André–Oort one) and are still widely open.

3.2. The Betti map

The counting method alluded to above worked for families and points defined over $\bar{\mathbb{Q}}$, but some of the tools failed over $\mathbb{C}$. This obstacle was overcome in a joint work of the author with Corvaja & Masser 2017 by *specialization*, to reduce to the algebraic case.

This gave as a byproduct somewhat analogous conclusions for families parameterized by spaces of dimension > 1.

Specialization appeared delicate because of certain possible degeneracies, difficult to exclude *a priori*. To get control on this, a relevant tool came from the so-called (real analytic) *Betti map*: it gives the real coordinates of the point, in terms of a lattice basis for the torus representing the abelian variety, the basis varying locally holomorphically in the family.

Example 3.1. In the case of the Legendre family, consider a lattice $L_\lambda \subset \mathbb{C}$ such that $\mathbb{C}/L_\lambda \cong \mathcal{L}_\lambda$ (for instance through a Weierstrass exponential giving the Legendre equation). Then, e.g., in the region $\mathcal{R} \subset \mathbb{C}$ defined by $\max(|\lambda|, |1-\lambda|) < 1$, by formulae going back to the XIX century, one can express a $\mathbb{Z}$-basis of L_λ in terms of hypergeometric functions, in fact as $L_\lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$, where $\omega_1 = i\pi F(1-\lambda)$, $\omega_2 = \pi F(\lambda)$ and where $F(\lambda) = \sum \binom{-1/2}{n}^2 \lambda^n$. For a given $\lambda \in \mathcal{R}$ and a point $Q \in \mathcal{L}_\lambda$ we may take a representative for Q in $\mathbb{C}/L_\lambda$ of the shape $\beta_1\omega_1 + \beta_2\omega_2$ with $\beta_1, \beta_2 \in \mathbb{R}/\mathbb{Z}$. Then by definition the β_i are the Betti coordinates of Q and the Betti map takes the value (β_1, β_2) at Q.[4]

The Betti map is highly relevant in our context because its *rational values correspond precisely to torsion points*. We have already mentioned some proofs where essential use is made of this map.

The Betti map appeared implicitly already in a work by Manin 1960s and was recently studied (for higher dimensions) in a work of Voisin 2019 and of André, Corvaja & the author 2020, with further contributions by Gao and applications by Voisin and by Dimitrov–Gao–Habegger and Kuehne.

4. Some applications

4.1. Pell equations in polynomials

The *Pell equation*

$$x^2 - y^2 D = 1, \quad D \text{ non-square positive integer},$$

to be solved in *integers* $x, y \neq 0$, is a celebrated Diophantine equation, proposed in fact by Fermat in the XVII century but actually having roots in ancient mathematics. It is linked with many important issues in Number Theory, such as integral points on curves (especially general affine conics), class-numbers and units of quadratic rings, orthogonal groups over $\mathbb{Z}$, Diophantine approximation and continued fractions, and so on.

There is also a *polynomial* analogue, more recent and apparently less known, but in fact also old, studied for instance already by Abel 1826, where $D = D(t)$ is a (complex, for instance) polynomial of even degree $2d$ and not a square, and one seeks *polynomial* solutions $x(t), y(t) \neq 0$. Following a suggestion of Serre, this equation may then be called *Pell–Abel equation*.

As in the classical case, a possible *nontrivial* solution generates infinitely many ones through the formulae $x_n \pm y_n\sqrt{D} = (x \pm y\sqrt{D})^n$, $n \in \mathbb{Z}$ (and all solutions are generated in this way, up to sign, from a “minimal” one).

[4] This map may be defined in any given open simply connected region, like the above $\mathcal{R}$, and we can cover the domain $\mathbb{C} \setminus \{0, 1\}$ with such regions. Then the map depends on a choice of basis for a given region and is subject to monodromy as we travel through loops meeting several regions.

For the Pell–Abel equation, when deg $D = 2$ there are always solutions (over $\mathbb{C}$), and the polynomials $x(t)$, $y(t)$ which arise are related to Chebyshev polynomials. But if deg $D \geq 4$, contrary to the classical case, it is generally unexpected to have solutions (unless we work over a finite field). This assertion can be put on a rigorous ground for instance using the Betti map. Indeed, the issue is linked with *torsion* in the Jacobian of the smooth complete hyperelliptic curve $H = H_D$ defined affinely by $u^2 = D(t)$. In fact, denoting $\infty_\pm$ the poles of t on H, it is not difficult to prove that *(nontrivial) solutions exist if and only if the class of the divisor* $\infty_+ - \infty_- \in \mathrm{div}(H)$ *has finite order in the divisor class group – i.e., in the Jacobian – of H.*[5] Abel gave a translation of such condition in terms of the continued fraction for $\sqrt{D(t)}$ being periodic (as happens in the classical case).

The polynomials $D(t)$ for which nontrivial solutions of the Pell–Abel equation exist are sometimes called *Pellian.*

In a joint work of the author with Masser we studied some 1-parameter families for fixed d, like $D_\lambda(t) = D_{d,\lambda}(t) := t^{2d} + t + \lambda$. As said, for $d = 1$, D_b is always Pellian. For $d = 2$ we easily realized that $D_b(t)$ is Pellian for infinitely many $b \in \mathbb{C}$ (satisfying (3) of Section 3), whereas we proved that for $d = 3$ there are only finitely many such values. We then extended the analysis to arbitrary 1-dimensional families of polynomials $D(t)$ of higher degree and those results would lead, for example, to the following theorem.

Theorem 4.1. *For any $d \geq 3$, there are only finitely many $b \in \mathbb{C}$ for which the Pell–Abel equation for $D_{d,b}(t)$ is solvable.*

We note that 0 lies in all these sets $\mathcal{P}_d := \{b \in \mathbb{C} : D_{d,b}$ is Pellian$\}$, for we have

$$(2t^{2d-1} + 1)^2 - (2t^{d-1})^2 D_{d,0}(t) = 1.$$

Open question. Is the union $\bigcup_{d\geq 3} \mathcal{P}_d$ of these finite sets itself finite?

If the answer is at all affirmative, it appears to require new tools to be proved, since the method that we used to deal with each single degree $d \geq 3$ is not completely uniform as d varies.

The Pell–Abel equation, similarly to the original version, appears in many mathematical topics; just to mention a recent instance, it has been studied by Kollar in connection with decidability issues and the Hilbert X problem over function fields. (We recall that the usual Pell equation had been used by Matijasevic in his final step solving the original Hilbert X problem.)

[5] A *generalized Jacobian* has to be considered if $D(t)$ has multiple roots. This link with Jacobians may be viewed as somewhat analogue of Dirichlet class number formula for real quadratic fields, the analogy being closer if we work over finite fields.

4.2. Integration in finite terms

The problem of expressing indefinite integrals in terms of "simple" functions goes back to long ago and appeared among the first examples and motivations for *differential algebra*. In this direction, we recall for instance the following (more or less classical) definition.

Definition. We call *Integrable in Finite Terms* (abbr. IFT) a differential whose (indefinite) integral can be expressed by a finite tower of operations either of algebraic type, or by taking exponentials or by taking logarithms (starting from rational functions). We also call *elementary* an integral which can be likewise expressed.

Even recently, much attention has been given to the study of possible algebraic relations among (definite) integrals of algebraic functions, special cases of *periods* (after Grothendieck, ..., Kontsevich & Zagier, ...), a topic not entirely unrelated with this theme.

We have already mentioned Abel in connection with Pell's equations in polynomials, and indeed his research involved also elementary integration. Subsequently the matter was studied by authors like Chebyshev, Liouville, Ritt, Kolchin, ..., giving rise for instance to Differential Galois theory.

More recently, J. Davenport investigated *pencils* of algebraic differentials, to be integrated in finite terms; he sought to understand *whether*,

> *if the general member of the family cannot be likewise integrated, the same happens for the special members, up to finitely many exceptions.*

Together with Masser we found how to establish when this type of assertion is correct, and we also found some counterexamples.

By means of a criterion of Risch and other considerations, it turns out that the analysis for such results in fact involves *torsion*, now in *generalized Jacobians*, which are algebraic groups obtained as extensions of usual Jacobians by products of groups of type $\mathbb{G}_{\mathrm{a}}$ or $\mathbb{G}_{\mathrm{m}}$.

Jointly with Masser, we carried out this, applying in particular some of the above results, and here are special cases of the output (all results joint with Masser 2018–2020).

Theorem 4.2. *There are only finitely many $b \in \mathbb{C}$ such that the integral $\int \frac{(2z+b)\,\mathrm{d}z}{\sqrt{z^4+z+b}}$ is elementary.*

Example 4.3. The special value $b = 1/2$ is in the said finite set:

$$\int \frac{(2z + 1/2)\, dz}{\sqrt{z^4 + z + 1/2}}$$
$$= \frac{1}{2} \log \left(4z^4 - 4z^3 + 2z^2 + 2z - 1 + (4z^2 - 4z + 2)\sqrt{z^4 + z + 1/2}\right).$$

This corresponds to a torsion point of order 4 in an extension by $\mathbb{G}_{\mathrm{a}}$ of the elliptic curve $w^2 = z^4 + z + (1/2)$.

The next example yields a negative answer towards Davenport's issue.

Example 4.4 (Counterexample). The differential $\frac{z\,\mathrm{d}z}{(z^2-t^2)\sqrt{z^3-z}}$ (over $\mathbb{C}(t)$) is not identically IFT but it becomes IFT for infinitely many specializations $t \to b$.

In this example, note the underlying elliptic curve $w^2 = z^3 - z$ with CM: this is no coincidence, since it can be shown that if the (usual) Jacobian of the underlying curve (corresponding to the differential) does not contain CM elliptic curves, then Davenport's expectation is correct.

4.3. Elliptical billiards

Further applications of some of the results are to *elliptical billiards*, namely billiard tables whose border is an ellipse and such that consecutive segments of billiard trajectory obey the usual law of reflection at the border.

Work going back to Poncelet and Jacobi shows that to such a billiard one can associate an elliptic family. In fact, it may be shown by nice arguments of Geometry, of type almost going back to Euclid, that all segments in a *given* billiard trajectory are tangent to a same conic, confocal with the ellipse, the so-called *caustic*. This caustic varies in a family of dimension 1. If the caustic is given, then the set of pairs (P, l), where P lies on the ellipse and l is a line through P tangent to the caustic, describes a curve of genus 1 embedded in $\mathbb{P}_1^2$. This curve becomes an elliptic curve after choice of an origin, whence, as the caustic varies, we obtain the said elliptic family.

A choice of a slope for a billiard shot from a given point yields a section of this family (depending on the point and parameterized by the slope). The *torsion* values of such a section correspond to the trajectories which are *periodic*, whose analysis is a main issue in the study of billiards.[6]

In this frame, on applying some of the above-mentioned results, in a recent joint work with Corvaja (2021) we deduced certain finiteness theorems for periodic trajectories in such billiards. For instance, we have the following conclusion.

Theorem 4.5. *For each $\alpha \in (0, \pi)$ there are only finitely many periodic pairs of billiard shots from a given point in an elliptical billiard such that the initial directions form an angle α.*

This may be shown to be not generally true for rectangular billiards.

[6] Part of this is a special case of a famous theorem of Poncelet, dealing with more general pairs of conics. The context has been generalized to higher dimensions by Griffiths & Harris 1977, which raises again questions related to the present realm.

Another finiteness conclusion (proved however with results of "Unlikely Intersections" type going beyond torsion – see above) concerns the set $T_{P,Q,R}$ of billiard trajectories which connect two given points P, Q and pass through another given point R: for instance, we have the following theorem.

Theorem 4.6. *If Q is a* hole *(i.e., lies on the boundary) and P, R are not both foci of the ellipse, then the set $T_{P,Q,R}$ is finite.*

It is somewhat curious that some of these results in the degenerate case of a circular billiard are related to the above discussion around Lang's conjecture.

Still further conclusions in the same spirit may be stated, e.g., concerning *boomerang* billiard shots. The link with the algebraic theory of elliptic families also shows how arithmetic information may affect chaotic behavior in an elliptical billiard. For instance, *shots from a given point, and having slope of large enough arithmetic height, cannot lead to periodic trajectories* (we tacitly deal here with ellipses and points defined over the algebraic numbers, which implies that the shot-slope is algebraic too if we have periodicity). This kind of implication seems not to have previously appeared in the theory of billiards.

5. Final remarks

Some open issues:

(1) To prove further cases of the conjectures of Pink and Zilber.

(2) To achieve effectivity in the counting of rational points appearing in some of the proofs.
This last issue is related to the theory of *o-minimality* in Model Theory. Some crucial recent work towards effectivity is due to Binyamini, and also to Daw, Jones, Schmidt, ...

(3) To prove finiteness in families where also the degrees vary.

Some of the methods from o-minimality have been developed (by Cluckers, Comte, Forey, and Loeser) in the p-adic context, and already applied by Chambert-Loir and Loeser 2017.

One expects here further applications.

6. References

I have realized that giving references for all the topics that we have touched would lead to a very long list, with some difficult choices and a heavy risk of leaving out something relevant. So, I have decided to quote just two of my own publications on these subjects, whose union contains a relevant quantity of references.

(1) The book [1] on Unlikely Intersections was written about 10 years ago: much work has appeared later, but the book contains an account of a substantial part of the contents of these notes, and many references.

(2) The more recent survey paper [2] contains further descriptions and more updated bibliography with respect to the former reference.

Acknowledgements. I thank the Scientific Committee of the 8th ECM for the appreciated invitation to deliver a plenary lecture, at the basis of these notes. I also thank all the Organizers for their excellent job, their hospitality and their precious help in several circumstances.

As to the present notes, I thank Yves André, Pietro Corvaja, David Masser, Barry Mazur, Zeev Rudnick, Peter Sarnak, Jean-Pierre Serre, and Don Zagier for their kind attention and several relevant comments.

References

[1] U. Zannier, *Some Problems of Unlikely Intersections in Arithmetic and Geometry*. Ann. of Math. Stud. 181, Princeton University Press, Princeton, NJ, 2012 Zbl 1246.14003 MR 2918151

[2] U. Zannier, Some specialization theorems for families of abelian varieties. *Münster J. Math.* **13** (2020), no. 2, 597–619 Zbl 1455.14087 MR 4130694

Umberto Zannier
Scuola Normale Superiore, 7 Piazza dei Cavalieri, 56126 Pisa, Italy; umberto.zannier@sns.it

Invited lectures

DOI 10.4171/8ECM/17

Positive harmonic functions on the Heisenberg group I

Yves Benoist

Abstract. We present the classification of positive harmonic functions on the Heisenberg group in the case of the southwest measure.

1. Introduction

In this self-contained paper, we present the classification of the positive harmonic functions on the Heisenberg group $H_3(\mathbb{Z})$ in the special case of the *southwest measure*. This example is striking because the famous partition functions occur as positive harmonic functions. In this case, our main result tells us that roughly all positive harmonic functions are combinations of characters and partition functions (Theorem 1.1).

We will also explain with no proof how this result can be extended to finite positive measures on $H_3(\mathbb{Z})$ (Theorem 3.8). The proof of this extension can be found in [2].

1.1. The partition function $p(x, y, z)$ as a potential

We first introduce the "partition function" $p(x, y, z)$ for any integers x, y, z in $\mathbb{Z}$.

1.1.1. The partition function. This function counts the "number of Young diagrams of area z," also called "partitions of z," included in a rectangle with side lengths x and y (see Figure 1). More precisely, when x, y, and z are non-negative, one has that

$$\begin{aligned} p(x, y, z) = \big|\big\{(n_1, \ldots, n_y) \in \mathbb{Z}^y \mid x \geq n_1 \geq \cdots \geq n_y \geq 0 \\ \text{and } n_1 + \cdots + n_y = z\big\}\big|, \end{aligned} \tag{1.1}$$

and $p(x, y, z) = 0$ otherwise. The integers n_i are the lengths of the rows of the partition. By convention, for $x \geq 0$, one has that $p(x, 0, z) = 0$ when $z \neq 0$, and

2020 Mathematics Subject Classification. Primary 31C05; Secondary 20F18.
Keywords. Nilpotent group, potential theory, partition function, random walk.

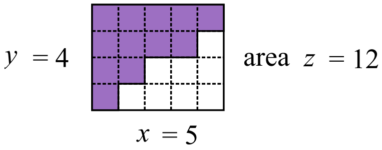


Figure 1. The partition $12 = 5 + 4 + 2 + 1$ is included in a 5×4 rectangle.

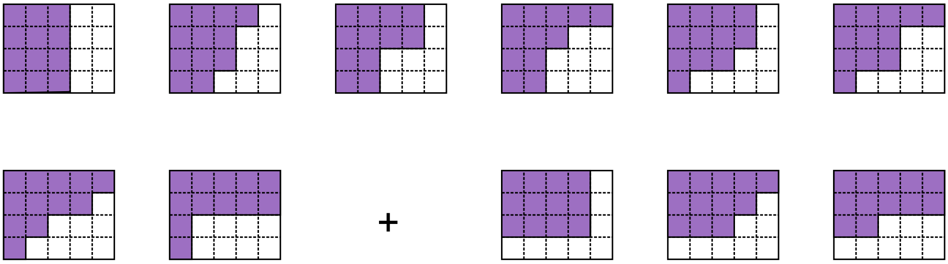


Figure 2. The 11 partitions in the equality $p(5, 4, 12) = p(4, 4, 8) + p(5, 3, 12)$.

that $p(x, 0, 0) = 1$. This partition function satisfies the functional equation, for all $g = (x, y, z)$ in $\mathbb{Z}^3$, $g \neq (0, 0, 0)$,

$$p(x, y, z) = p(x - 1, y, z - y) + p(x, y - 1, z). \tag{1.2}$$

One checks it by splitting this set of partitions according to the color of the lower-left case of the rectangle as in Figure 2.

1.1.2. The Heisenberg group. Recall that the Heisenberg group $G := H_3(\mathbb{Z})$ is the set $\mathbb{Z}^3$ of triples seen as matrices

$$(x, y, z) := \left(\begin{smallmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{smallmatrix}\right).$$

It is endowed with the product

$$(x_0, y_0, z_0)(x, y, z) = (x_0 + x, y_0 + y, z_0 + z + x_0 y). \tag{1.3}$$

Let μ_0 be the southwest measure on G. It is given by

$$\mu_0 = \delta_{a^{-1}} + \delta_{b^{-1}}, \quad \text{where } a := (1, 0, 0) \text{ and } b = (0, 1, 0). \tag{1.4}$$

Let $e := (0, 0, 0)$ be the unity of G and let $\mathbf{1}_{\{e\}}$ be the characteristic function of $\{e\}$. Equation (1.2) can be rewritten as, for all $g = (x, y, z)$ in G,

$$p(g) = p(a^{-1}g) + p(b^{-1}g) + \mathbf{1}_{\{e\}}(g). \tag{1.5}$$

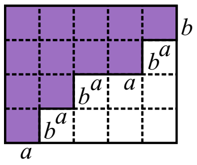


Figure 3. The partition $12 = 5 + 4 + 2 + 1$ associated to the word $w = ababaabab$ gives the element $g = g_w = ababaabab = (5, 4, 12) \in H_3(\mathbb{Z})$.

In particular, the function $f = p$ satisfies

$$f(g) \geq P_{\mu_0} f(g), \quad \text{where } P_{\mu_0} f(g) := f(a^{-1}g) + f(b^{-1}g). \qquad (1.6)$$

This inequality (1.6) tells us that the function f is a μ_0-superharmonic function on the Heisenberg group G.

1.1.3. The potential. More precisely, *the partition function $p(g)$ is the potential of μ_0 at e*. This means that one has the equality

$$p = \sum_{n \geq 0} P^n_{\mu_0} \mathbf{1}_{\{e\}}.$$

Indeed, as can be seen in Figure 3, for g in G,

$$p(g) \textit{ is the number of ways to write } g \textit{ as a word in } a \textit{ and } b. \qquad (1.7)$$

A function h on G is said to be μ_0-harmonic if it satisfies

$$h(g) = P_{\mu_0} h(g), \quad \text{for all } g \text{ in } G, \text{ or equivalently} \qquad (1.8)$$

$$h(x, y, z) = h(x - 1, y, z - y) + h(x, y - 1, z), \quad \text{for all } (x, y, z) \text{ in } \mathbb{Z}^3. \qquad (1.9)$$

1.2. Construction of positive harmonic functions

We want to classify all the positive[1] solutions of (1.6), i.e., all the positive μ_0-superharmonic functions h on G. We begin with five remarks.

1.2.1. Choquet theorem. By a theorem of Choquet in [5], every positive superharmonic function h is an average of extremal[2] positive superharmonic functions h_α. Moreover, when h is harmonic, the h_α are harmonic. By Riesz decomposition theorem [13, Thm. 2.1.4], every positive μ_0-superharmonic function can be written in a unique way as the sum of a potential[3] and a positive μ_0-harmonic function. Therefore, it is enough to describe the extremal positive μ_0-harmonic functions on G.

[1] A function f on G is said to be positive if $f(g) \geq 0$ for all g in G and $f \neq 0$.

[2] A positive (super)harmonic function is said to be extremal if it cannot be written as the sum of two non-proportional positive (super)harmonic functions.

[3] A potential is a function of the form $f = \sum_{n \geq 0} P^n_{\mu_0} F$ for a positive function F on G.

	↑ z									
0	**0**	1	5	10	15	18	20	21	22	
0	**0**	1	4	8	11	13	14	15	15	
0	**0**	1	4	7	9	10	11	11	11	
0	**0**	1	3	5	6	7	7	7	7	
0	**0**	1	3	4	5	5	5	5	5	
0	**0**	1	2	3	3	3	3	3	3	
0	**0**	1	2	2	2	2	2	2	2	
0	**0**	1	1	1	1	1	1	1	1	
0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	→ y
0	**0**	0	0	0	0	0	0	0	0	

Figure 4. The function $p_y(z)$ satisfies $p_y(z) = p_y(z-y) + p_{y-1}(z)$.

1.2.2. Choquet–Deny theorem. If we look for a μ_0-harmonic function h which does not depend on z, then (1.9) becomes

$$h(x,y) = h(x-1,y) + h(x,y-1), \quad \text{for all } (x,y) \text{ in } \mathbb{Z}^2. \tag{1.10}$$

This equation tells us that the function h is μ_0-harmonic on the abelian quotient $\mathbb{Z}^2$ of G. According to a theorem of Choquet and Deny in [6], since the support of the measure μ_0 spans the group $\mathbb{Z}^2$, every extremal positive μ_0-harmonic function on this abelian group is proportional to a character[4]:

$$\chi(x,y,z) = r^x s^y \quad \text{with } r,s > 0 \text{ and } \frac{1}{r} + \frac{1}{s} = 1. \tag{1.11}$$

1.2.3. The partition function as a harmonic function. If we look for a μ_0-harmonic function h which does not depend on x, then (1.9) becomes

$$h(y,z) = h(y,z-y) + h(y-1,z), \quad \text{for all } (y,z) \text{ in } \mathbb{Z}^2. \tag{1.12}$$

A nice example is given in Figure 4 by the partition function $(y,z) \mapsto p_y(z)$, where

$$\begin{aligned} p_y(z) &= \sup_{x\in\mathbb{Z}} p(x,y,z) = \lim_{x\to\infty} p(x,y,z) = p(z,y,z) \\ &= \text{the number of partitions of } z \text{ with at most } y \text{ rows.} \end{aligned} \tag{1.13}$$

Hence the function $h_0(x,y,z) := p_y(z)$ is a μ_0-harmonic function on G.

[4] The proof is very short. One notices that equality (1.10) gives a decomposition of h as a sum of two positive harmonic functions and hence both of them are proportional to h.

1.2.4. Margulis first theorem. According to the first theorem of Margulis, a theorem he proved in [10] when he was not yet twenty, the Choquet–Deny theorem is still true on a finitely generated nilpotent group G as soon as the support of the measure spans G as a semigroup (see Fact 3.7). This is why it might look surprising at first glance that there exists a positive μ_0-harmonic function h_0 on $H_3(\mathbb{Z})$ which is not invariant by the center. The reason it exists is that the support of μ_0 spans G as a group but not as a semigroup. What is more surprising is that this "new" positive harmonic function h_0 is given by the famous partition function $p_y(z)$.

1.2.5. Switching and translating harmonic functions. We denote by σ the automorphism of G exchanging a and b. It is given by

$$\sigma(x, y, z) = (y, x, xy - z).$$

Since the function h_0 is μ_0-harmonic, the function

$$h_1 := h_0 \circ \sigma : (x, y, z) \mapsto p_x(xy - z)$$

is also μ_0-harmonic. For g_0 in G, we denote by $\rho_{g_0} : g \mapsto gg_0$ the right translation by g_0 on G. The translated functions $h_0 \circ \rho_{g_0} : g \mapsto h_0(gg_0)$ and $h_1 \circ \rho_{g_0} : g \mapsto h_1(gg_0)$ are also μ_0-harmonic.

1.3. Classification of positive harmonic functions

We can now state our main result for the southwest measure μ_0 introduced in (1.4).

1.3.1. Main result and strategy.

Theorem 1.1. *Let h be an extremal positive μ_0-harmonic function on the Heisenberg group $G := H_3(\mathbb{Z})$. Then, up to a multiplicative scalar,*

- *either $h = \chi$ is a μ_0-harmonic character $\chi(x, y, z) = r^x s^y$ as in (1.11)*
- *or $h = h_0 \circ \rho_{g_0}$ is a translate of the function $h_0(x, y, z) := p_y(z)$*
- *or $h = h_1 \circ \rho_{g_0}$ is a translate of the function $h_1(x, y, z) := p_x(xy - z)$.*

This classification has been announced on May 28th 2019 in a short informal videotaped speech at the Cetraro conference "Dynamics of group actions." This video can be found on the author's web page.

As we will see, the partition function $p(x, y, z)$ will play a crucial role in the proof of Theorem 1.1. Indeed, in Chapter 2, we will prove a ratio limit theorem for the partition function $p(x, y, z)$. In Chapter 3, we will deduce from this ratio limit theorem the proof of Theorem 1.1.

Notice that the positive μ_0-harmonic function h_0 vanishes. In particular, it does not satisfy the Harnack inequality. This contrasts with the case studied in [10], where the support of μ spans G as a semigroup.

In the last section (Section 3.4), we will present the classification of the positive μ-harmonic functions, for all finitely supported measures μ on G.

1.3.2. Dealing with a probability measure. At first glance it might look a little bit weird to deal with a μ_0-harmonic function for a measure μ_0 which is not a probability measure. We could have worked instead with the probability measure

$$\widetilde{\mu}_0 = \frac{1}{2}(\delta_{a^{-1}} + \delta_{b^{-1}}), \quad \text{where } a := (1,0,0) \text{ and } b = (0,1,0)$$

which is the law for the *southwest random walk on* $\mathbb{H}_3(\mathbb{Z})$. The $\widetilde{\mu}_0$-harmonic functions $\tilde{h}$ on G are the functions satisfying

$$\tilde{h} = P_{\widetilde{\mu}_0} h, \quad \text{where } P_{\widetilde{\mu}_0} h(x,y,z) = \frac{1}{2}\big(\tilde{h}(x-1,y,z-y) + \tilde{h}(x,y-1,z)\big)$$

is the expected value of the function h after one step of the random walk.

It is easy to see that

$$h(x,y,z) \text{ is } \mu_0\text{-harmonic if and only if } 2^{-x-y} h(x,y,z) \text{ is } \widetilde{\mu}_0\text{-harmonic.}$$

Therefore, classifying positive μ_0-harmonic functions is equivalent to classifying positive $\widetilde{\mu}_0$-harmonic functions. The main reason we are using μ_0 instead of $\widetilde{\mu}_0$ is to get rid of all these factors 2^{-x-y}.

1.3.3. Extremal superharmonic functions. We have seen in (1.5) that the partition function p is μ_0-superharmonic and more precisely that it is the potential of μ_0 at e. For every g_0 in G, the function $p \circ \rho_{g_0}$ is also a potential of μ_0 at g_0^{-1}. By Riesz decomposition theorem, those potentials are exactly the extremal positive μ_0-superharmonic functions on G which are not harmonic. Therefore,

> *every extremal positive μ_0-superharmonic function f on G which is not harmonic is a translate $f = p \circ \rho_{g_0}$ of the function $p(x,y,z)$.*

We would like to end this introduction by pointing out other limit theorems for random walks on the Heisenberg group and other nilpotent groups as [3,4,7,8] even though we will not use them here.

2. The partition function

The aim of this chapter is to prove the ratio limit theorem (Proposition 2.2) for the partition function $p(x,y,z)$.

2.1. The unimodality of the partition functions

We recall that, for $x, y, z \geq 0$, the partition function $p(x, y, z)$ counts the number of partitions of z included in a rectangle with side lengths x and y. See definition (1.1) and Figure 1.

This function is non-zero for $0 \leq z \leq xy$ and satisfies the equalities

$$p(x, y, z) = p(y, x, z) = p(x, y, xy - z). \tag{2.1}$$

This function is well studied. For instance, one has the following fact.

Fact 2.1 (Cayley, Sylvester 1850). *The sequence $z \mapsto p(x, y, z)$ is unimodal; i.e., it is increasing for $z \leq xy/2$.*

The proof below relies on the theory of finite dimensional representations of the Lie algebra $\mathfrak{sl}(2, \mathbb{R})$. This proof is due to Hughes in [9]. See [12] for an elementary proof and [14, p. 522] for a survey of various generalizations.

Sketch of proof of Fact 2.1. Let $n := x + y$ and let (Y, H, X) be the principal $\mathfrak{sl}_2$-triple in the Lie algebra $\mathfrak{g} := \mathfrak{sl}(n, \mathbb{R})$ so that $H = \operatorname{diag}(n-1, n-3, \ldots, -n+1)$. This Lie algebra $\mathfrak{g}$ has a natural representation in the space $V := \Lambda^x(\mathbb{R}^n)$. One checks that $p(x, y, z) = \dim V_{xy-2z}$, where V_λ denotes the eigenspace of H in V for the eigenvalue λ. The theory of representations of $\mathfrak{sl}(2, \mathbb{R})$ tells us that for $\lambda > 0$, one always has that $\dim V_\lambda \leq \dim V_{\lambda-2}$. ■

2.2. The ratio limit theorem

Here is the *ratio limit theorem for* $p(x, y, z)$.

Proposition 2.2. *One has that*

$$\lim_{\substack{z \to \infty \\ xy - z \to \infty}} \frac{p(x, y, z-1)}{p(x, y, z)} = 1.$$

This limit is taken along sequences of positive triples (x, y, z) such that $z \to \infty$ and $xy - z \to \infty$.

With this generality this theorem seems to be new, even though there already exist very precise estimates of $p(x, y, z)$ in certain ranges. For instance, when $x, y \geq z$, the partition function $p(x, y, z) = p(z, z, z)$ depends only on z. It is the classical partition function $p(z)$ which admits a famous asymptotic expansion due to Hardy and Ramanujan in 1920 (see [1, Chap. 5]). These estimates have been extended to larger ranges of (x, y, z) as in [11, 15]. We will not use them.

The proof of Proposition 2.2 is tricky but elementary. The rough idea is to introduce a relation between the set of partitions w of z and the set of partitions w' of $z - 1$ such that "most of the time" when w and w' are related, they are related to approximately the same number of partitions (see Lemma 2.5).

Because of (2.1), we can assume that $y \leq x$ and $z \leq xy/2$.

2.3. When the height of the rectangles is bounded

In this section, we deal with the easy case when the height y remains bounded.

Lemma 2.3. *For all $y \geq 1$, one has that*

$$\lim_{\substack{x,z\to\infty \\ z\leq xy/2}} \frac{p(x,y,z-1)}{p(x,y,z)} = 1.$$

Note that in this limit y is fixed, and x, z go to ∞ with $z \leq xy/2$.

Proof of Lemma 2.3. This follows from Lemma 2.4 and the inequalities

$$0 \leq p(x,y,z) - p(x,y,z-1) \leq p(x,y-1,z).$$

The first inequality is the unimodality of the partition function.

For the second inequality, just notice that one can inject the set of partitions of z of height exactly y inside the set of partitions of $z-1$ of height at most y by removing the last square in the bottom row of each partition. ∎

We have used the following lemma.

Lemma 2.4. (a) *For all $x, y, z \geq 1$, one has that $p(x,y,z) \leq z^{y-1}$.*

(b) *For all $y \geq 1$, there exists $\alpha_y > 0$ such that, for all $x, z \geq 1$ with $z \leq xy/2$, one has that $p(x,y,z) \geq \alpha_y z^{y-1}$.*

Proof of Lemma 2.4. (a) The lengths of the last $y-1$ rows of the partition are bounded by $z-1$ and the first row is deduced from the others.

(b) Choose $y-1$ integers $m_1, \dots, m_{y-1}$ in the interval $[0, \frac{z}{y^2}]$ and keep only those for which the system

$$n_1 - n_2 = m_1, \dots,\ n_{y-1} - n_y = m_{y-1} \quad \text{and} \quad n_1 + \cdots + n_y = z$$

has a solution $(n_1, \dots, n_y)$ in $\mathbb{Z}^y$. But then one has that

$$\begin{aligned} n_y &= \frac{1}{y}\big(z - m_1 - 2m_2 - \cdots - (y-1)m_{y-1}\big) \geq 0, \\ n_1 &= n_y + m_1 + \cdots + m_{y-1} \leq \frac{z}{y} + \frac{z}{y} \leq x. \end{aligned}$$

This gives about $\frac{1}{y}(\frac{z}{y^2})^{y-1}$ partitions of z with $x \geq n_1 \geq \cdots \geq n_y \geq 0$. ∎

2.4. Inner and outer corner of a partition

We now introduce notations that will strengthen the connection between partitions and words in the Heisenberg group.

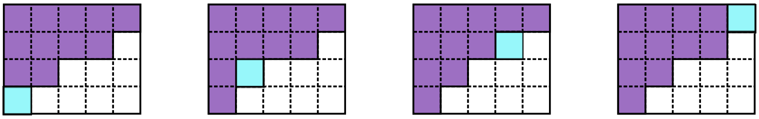


Figure 5. The fiber $\pi^{-1}(w)$ of the word $w = ababaabab$ has size $f_w = 4$.

We recall that $a = (1,0,0)$ and $b = (0,1,0)$ are the generators of the Heisenberg group $G = H_3(\mathbb{Z})$. Let

$$G^+ := \{g = (x,y,z) \in G \mid x,y \geq 0 \text{ and } 0 \leq z \leq xy\}$$

be the semigroup generated by a and b and let

$$c = aba^{-1}b^{-1} = (0,0,1) \tag{2.2}$$

be the generator of the center Z of G.

Let $B_n := \{a,b\}^n$ be the set of finite words w in a, b of length $\ell_w = n$ and let $B := \bigcup_{n\geq 0} B_n$. Using the product law in G, to each word $w \in B$, we can associate an element g_w in G^+. The partition function gives the size of the fibers of this map:

$$p(g) = |B_g|, \quad \text{where } B_g := \{w \in B \mid g_w = g\}. \tag{2.3}$$

Indeed, as explained in Figure 3, when $g = (x,y,z)$, each word w in B_g corresponds uniquely to a partition of z included in a rectangle with side lengths x and y. We introduce now the following relation $\mathcal{R}$ on B:

$$\mathcal{R} := \{(w,w') \in B \times B \mid w = w_0abw_1 \text{ and } w' = w_0baw_1 \\ \text{for some } w_0, w_1 \text{ in } B\}.$$

Let $\pi : \mathcal{R} \to B$ and $\pi' : \mathcal{R} \to B$ be the two projections

$$\pi(w,w') = w \quad \text{and} \quad \pi(w,w') = w'.$$

For w, w' in B, the cardinality of the fiber $f_w := |\pi^{-1}(w)|$ is the number of pairs ab occurring in the word w. The size f_w is also the number of *inner corners* of the partition associated to w (see Figure 5). Similarly the cardinality of the fiber $f'_{w'} := |\pi'^{-1}(w')|$ is the number of pairs ba occurring in the word w'. It is equal to the number of *outer corners* of the partition associated to w'.

The following lemma compares the size of these fibers.

Lemma 2.5. (a) *For all $(w,w') \in \mathcal{R}$, one has that $g_w = g_{w'}c$.*

(b) *For all $(w,w') \in \mathcal{R}$, one has that $|f_w - f'_{w'}| \leq 2$.*
In particular, one also has that $f_w \leq 3f'_{w'}$.

Proof of Lemma 2.5. (a) This follows from the equality $c = aba^{-1}b^{-1}$.

(b) Comparing the number of pairs ab and pairs ba occurring in w and in w', one gets $|f_w - f_{w'}| \le 1$ and $|f_{w'} - f'_{w'}| \le 1$. ■

2.5. Partitions with bounded number of corners

We will need to control the number $p_{\le i}(x, y, z)$ of partitions of z included in a rectangle with side length x, y that have at most i inner corner.

The following Lemma 2.6 tells us that $p_{\le i}(x, y, z)$ is negligible compared to the total number of partitions $p(x, y, z)$.

Lemma 2.6. *For all $i \ge 0$, one has that*

$$\lim_{\substack{x,y,z\to\infty \\ z\le xy/2}} \frac{p_{\le i}(x, y, z)}{p(x, y, z)} = 0.$$

The limit is taken along sequences where all coordinates x, y, z go to ∞ and $z \le xy/2$.

Proof of Lemma 2.6. Use the following slight upgrade of Lemma 2.4. ■

Lemma 2.7. (a) *For all $x, y, z, i \ge 1$, one has that $p_{\le i}(x, y, z) \le (2z)^{2i}$.*

(b) *For all $j > 1$, there exists $z_0 = z_0(j) \ge 1$ such that, for all $x, y, z \ge 1$ with $4j \le y \le x$ and $z_0 \le z \le xy/2$, one has that $p(x, y, z) \ge z^j$.*

Proof of Lemma 2.7. It is similar to Lemma 2.4.

(a) We can assume that $x = y = z$. We want to choose integers $a_1, \dots, a_i \ge 1$ and $m_1, \dots, m_i \ge 0$, bounded by z such that $a_1m_1 + \cdots + a_im_i = z$. There are at most $(2z)^{2i}$ possibilities.

(b) We give a rough count. Choose $L_y \le y$ as large as possible such that, setting $\ell_y = [L_y/2]$ and $\ell_x = [z/L_y]$, one has that $\ell_y \le \ell_x \le x/2$. There exists a partition w_0 of z with L_y rows and all of whose rows have length ℓ_x or $\ell_x + 1$. For every sequence $\ell_x > m_1 \ge \cdots \ge m_{\ell_y} \ge 0$, we can modify this partition w_0 by adding m_j spots to the jth highest row of w_0 and removing m_j spots to the jth lowest row of w_0, for all $j \le \ell_y$. This gives N different partitions of z, where $N := \binom{\ell_x+\ell_y-1}{\ell_y} \ge \max(2, \ell_x/\ell_y)^{\ell_y}$. Hence, one has that $p(x, y, z) \ge N$.

First case: when $z \le y^2/2$. In this case, we have that $L_y = [\sqrt{2z}]$.

One gets $N \ge 2^{\ell_y} \ge 2^{\sqrt{z}/2} \ge z^j$.

Second case: when $z \ge y^2/2$. In this case, we have that $L_y = y$.

If $z \le y^4$, one gets $N \ge 2^{\ell_y} \ge 2^{\sqrt[4]{z}/4} \ge z^j$.

If $z \ge y^4$, one gets $N \ge (\frac{\ell_x}{\ell_y})^{\ell_y} \ge (\frac{z}{y^2})^{\ell_y} \ge \sqrt{z}^{\ell_y} \ge z^{y/4} \ge z^j$. ■

2.6. When the height of the rectangles is unbounded

We can now explain the proof of the ratio limit theorem.

Proof of Proposition 2.2. By (2.1) and Lemma 2.3, we can assume that the three positive integers x, y, z are going to ∞ with $y \leq x$ and $z \leq xy/2$. For $g = (x, y, z)$ in G^+, one sets $\mathcal{R}_g := \{(w, w') \in \mathcal{R} \mid g_w = g\}$, and one computes

$$p(g) = |B_g| = \varepsilon_g + \sum_{(w,w')\in\mathcal{R}_g} \frac{1}{f_w}, \tag{2.4}$$

where $\varepsilon_g = 1$ if $\mathcal{R}_g = \emptyset$ and $\varepsilon_g = 0$ otherwise. Similarly, by Lemma 2.5 (a), one has that

$$p(gc^{-1}) = |B_{gc^{-1}}| = \varepsilon'_g + \sum_{(w,w')\in\mathcal{R}_g} \frac{1}{f'_{w'}}, \tag{2.5}$$

where $\varepsilon'_g = 0$ or 1. Combining (2.4), (2.5), and Lemma 2.5 (b), one gets

$$\left|p(g) - p(gc^{-1})\right| \leq 2 + \sum_{(w,w')\in\mathcal{R}_g} \frac{2}{f_w f'_{w'}} \leq 2 + \sum_{(w,w')\in\mathcal{R}_g} \frac{6}{f_w^2} \leq 2 + \sum_{\substack{w\in B_g \\ f_w\neq 0}} \frac{6}{f_w}.$$

We recall that $p_{\leq i}(g)$ is the number of w with $f_w \leq i$. Therefore, one has that

$$\left|p(g) - p(gc^{-1})\right| \leq 2 + 6p_{\leq i}(g) + \frac{6}{i}p(g) \quad \text{for all } i \geq 1.$$

We let x, y, z go to infinity with $z \leq xy/2$. According to Lemma 2.6, for all $i \geq 1$, the ratios $p_{\leq i}(g)/p(g)$ converge to 0. Therefore,

$$\limsup \left|\frac{p(gc^{-1})}{p(g)} - 1\right| \leq \frac{6}{i},$$

and therefore the sequence $\frac{p(gc^{-1})}{p(g)}$ converges to 1 as required. ■

3. Positive harmonic functions

We now start the classification of extremal positive μ_0-harmonic functions h. In Section 3.1, we deal with the case where h has a non-zero limit along an orbit of a^{-1} or b^{-1}. In Sections 3.2 and 3.3, we deal with the case where h goes to zero along all orbits of a^{-1} and b^{-1}. In Section 3.4, we present the generalization of this classification to any finitely supported measure μ on G.

3.1. The partition function as a harmonic function

In this section, we characterize the functions $h_0 \circ \rho_{g_0}$ and $h_1 \circ \rho_{g_0}$ among extremal positive μ_0-harmonic functions by their behavior along the orbits $a^{-\mathbb{N}} g_0$ and $b^{-\mathbb{N}} g_0$ of G.

We recall that $a = (1, 0, 0)$ and $b = (0, 1, 0)$ are the generators of the Heisenberg group $G = H_3(\mathbb{Z})$, that $\mu_0 = \delta_{a^{-1}} + \delta_{b^{-1}}$, and that h_0 and h_1 are the μ_0-harmonic functions $h_0(x, y, z) = p_y(z)$ and $h_1(x, y, z) = p_x(xy - z)$.

We first begin by an alternative construction of the function h_0. Let H_0 be the abelian subgroup of G generated by a and let $\psi_0 := \mathbf{1}_{H_0}$ be the characteristic function of H_0. One has that

$$\psi_0(x, y, z) = \begin{cases} 1 & \text{when } y = z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 3.1. *One has the equality* $h_0 = \lim_{n\to\infty} P^n_{\mu_0}\psi_0$.

Remark. Since the function ψ_0 is μ_0-subharmonic, i.e., $\psi_0 \le P_{\mu_0}\psi_0$, the sequence $n \mapsto P^n_{\mu_0}\psi_0$ is increasing.

Proof of Lemma 3.1. One can compute explicitly this function $P^n_{\mu_0}\psi_0$. It does not depend on x. Indeed, $P^n_{\mu_0}\psi_0(x, y, z)$ is the number of ways of writing the element $(n - y, y, z)$ as a word w of length n in a and b. This proves the equality, involving the partition function,

$$P^n_{\mu_0}\psi_0(x, y, z) = p(n - y, y, z).$$

Letting n go to ∞, we conclude using (1.13). ■

Lemma 3.2. *Let $g_0 \in G$ and let h be an extremal positive μ_0-harmonic function on G such that* $\limsup_{n\to\infty} h(a^{-n} g_0) > 0$. *Then one has that $h = \lambda h_0 \circ \rho_{g_0}$ with $\lambda > 0$.*

In particular, the positive μ_0-harmonic function $h_0 \circ \rho_{g_0}$ is extremal.

Proof of Lemma 3.2. We can assume that $g_0 = e$. Since the function h is positive and μ_0-harmonic, the sequence $n \mapsto h(a^{-n})$ is decreasing. Hence it has a limit λ. By assumption, this limit λ is positive. By construction, one has the equality $h \ge \lambda\psi_0$. Since h is μ_0-harmonic, one also has the inequality $h \ge \lambda P^n_{\mu_0}\psi_0$ for all $n \ge 0$. Therefore, by Lemma 3.1, one gets $h \ge \lambda h_0$. Since h is extremal, it has to be proportional to h_0 and therefore one has that $h = \lambda h_0$.

It remains to check that h_0 is extremal. If one can write $h_0 = h'_0 + h''_0$ with both h'_0 and h''_0 positive μ_0-harmonic, for at least one of them, say h'_0, the sequence $h'_0(a^{-n})$ does not converge to 0 for $n \to \infty$. Hence, by the previous discussion, h'_0 is proportional to h_0. This proves that h_0 is extremal. ■

Exchanging the role of a and b we get the following corollary.

Corollary 3.3. *Let h be an extremal positive μ_0-harmonic function on G such that* $\limsup_{n\to\infty} h(b^{-n}g_0) > 0$. *Then one has that $h = \lambda h_1 \circ \rho_{g_0}$ for some $\lambda > 0$.*

In particular, the positive μ_0-harmonic function $h_1 \circ \rho_{g_0}$ is extremal.

3.2. Harmonic functions that decay on cosets

We now discuss positive harmonic functions on G that decay to 0 along the orbits $a^{-\mathbb{N}}g_0$ and $b^{-\mathbb{N}}g_0$.

Let G_n be the subset of G consisting of elements of "degree" n,

$$G_n = \{g = (x, y, z) \in G \mid x + y = n\}.$$

By definition and by (1.7), a positive μ_0-harmonic function h on G satisfies the equality, for all $n \geq 1$,

$$h(g_0) = \sum_{w\in B_n} h(g_w^{-1}g_0) = \sum_{g\in G_n} p(g)h(g^{-1}g_0). \tag{3.1}$$

For an integer $A > 0$, we set

$$\begin{aligned} G_{n,A} &= \{g = (x, y, z) \in G_n \mid z \leq A\}, \\ G^\sigma_{n,A} &= \{g = (x, y, z) \in G_n \mid xy - z \leq A\}. \end{aligned} \tag{3.2}$$

The following lemma tells us when the contributions of $G_{n,A}$ and $G^\sigma_{n,A}$ in formula (3.1) are negligible.

Lemma 3.4. *Let h be a positive μ_0-harmonic function on G such that*

$$\lim_{n\to\infty} h(a^{-n}g_0) = 0 \quad \textit{and} \quad \lim_{n\to\infty} h(b^{-n}g_0) = 0 \quad \textit{for all } g_0 \textit{ in } G. \tag{3.3}$$

Then, for all $A > 0$ and g_0 in G, one has that

$$\lim_{n\to\infty} \sum_{g\in G_{n,A}\cup G^\sigma_{n,A}} p(g)h(g^{-1}g_0) = 0. \tag{3.4}$$

Proof of Lemma 3.4. It is enough to prove (3.4) with $g_0 = e$. Moreover, since $G^\sigma_{n,A}$ is the image of $G_{n,A}$ by the involution σ which exchanges a and b, it is enough to prove (3.4) with $g \in G_{n,A}$. Equivalently, it is enough to prove that

$$\lim_{n\to\infty} \sum_{w\in B_{n,A}} h(g_w^{-1}) = 0, \quad \text{where } B_{n,A} := \{w \in B_n \mid g_w^{-1} \in G_{n,A}\}. \tag{3.5}$$

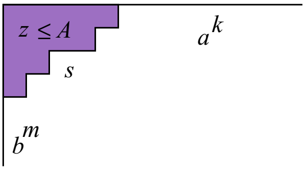


Figure 6. The decomposition $w = b^m s a^k$ for a word $w \in B_{n,A}$.

Note that, when $n > A$, every word $w \in B_{n,A}$ can be written as

$$w = b^m s a^k$$

with $s \in B_{A+1}$ a word of length $A + 1$ (see Figure 6). One splits the set $B_{n,A}$ according to $m \geq A$ or $m < A$. Therefore, for $n \geq 2A$, one has the inclusion

$$B_{n,A} \subset b^A B_{n-A} \cup B_{2A} a^{n-2A}.$$

Therefore, using (3.1), one gets the inequalities

$$\begin{aligned}\sum_{w \in B_{n,A}} h(g_w^{-1}) &\leq \sum_{w \in B_{n-A}} h(g_w^{-1} b^{-A}) + \sum_{w \in B_{2A}} h(a^{-(n-2A)} g_w^{-1}) \\ &= h(b^{-A}) + \sum_{w \in B_{2A}} h(a^{-(n-2A)} g_w^{-1}).\end{aligned}$$

For all $\varepsilon > 0$, we choose A large enough so that, by the second assumption (3.3), one has that $h(b^{-A}) \leq \varepsilon$. Then the last sum is a sum over the fixed finite set B_{2A}, and, by the first assumption (3.3), this last sum converges to 0 when n goes to infinity. This proves (3.5) as required. ■

3.3. Using the ratio limit theorem

Combining Lemma 3.4 with the ratio limit theorem, we can finish the last case of the proof of Theorem 1.1.

Lemma 3.5. *Let h be a positive μ_0-harmonic function on G such that, for all g_0 in G, $\lim_{n\to\infty} h(a^{-n} g_0) = \lim_{n\to\infty} h(b^{-n} g_0) = 0$. Then h is invariant by the center $Z = c^{\mathbb{Z}}$ of G.*

Proof of Lemma 3.5. Using (3.1) with g_0 and $g_0 c$, we compute

$$h(g_0) - h(g_0 c) = \sum_{g \in G_n} \big(p(g) - p(gc)\big) h(g^{-1} g_0). \tag{3.6}$$

We fix $\varepsilon > 0$. According to the ratio limit theorem (Proposition 2.2), there exists an integer $A > 0$ such that, for all $g = (x, y, z)$ in G^+ with $z \geq A$ and $xy - z \geq A$, one has that

$$\left|p(g) - p(gc)\right| \leq \varepsilon p(g). \tag{3.7}$$

Therefore, using (3.6), (3.7), and definition (3.2), one gets

$$\left|h(g_0) - h(g_0 c)\right| \leq \sum_{g \in G_n} \varepsilon p(g) h(g^{-1} g_0) + \sum_{g \in G_{n,A} \cup G^{\sigma}_{n,A}} p(g)\big(h(g^{-1} g_0) + h(g^{-1} g_0 c)\big).$$

By (3.1), the first term is equal to $\varepsilon h(g_0)$. Therefore, using twice Lemma 3.4 and letting n go to infinity, one gets $|h(g_0) - h(g_0 c)| \leq \varepsilon h(g_0)$. Since ε is arbitrary small, this proves that $h(g_0) = h(g_0 c)$ as required. ■

Corollary 3.6. *Let h be an extremal positive μ_0-harmonic function on G such that, for all g_0 in G, $\lim_{n\to\infty} h(a^{-n} g_0) = \lim_{n\to\infty} h(b^{-n} g_0) = 0$. Then h is a character of G.*

In particular, every μ_0-harmonic character of G is an extremal positive μ_0-harmonic function.

Proof of Corollary 3.6. By Lemma 3.5, the function h is μ_0-harmonic on the abelian group G/Z. By Choquet–Deny theorem, it is a character.

It remains to check that a μ_0-harmonic character χ is extremal. Assume that $\chi = h' + h''$ with both h' and h'' positive μ_0-harmonic. For all g_0 in G, the sequences $h'(a^{-n} g_0)$ and $h'(b^{-n} g_0)$ converge to 0 for $n \to \infty$. Hence, by the previous discussion and by Choquet's theorem, the function h' is an integral $h' = \int_C \chi' \, \mathrm{d}\sigma(\chi')$, where σ is a finite positive measure on the set C of (harmonic) character χ' of G. Since $h' \leq \chi$, the measure σ must be supported by χ. This proves that χ is extremal. ■

This ends the proof of Theorem 1.1.

3.4. Extension to finitely supported measures

In this section, we give the classification of the positive μ-harmonic functions on the Heisenberg group for all finitely supported measure μ.

Let $G = H_3(\mathbb{Z})$ be the Heisenberg group and let S be a finite subset of G. We denote by G_S the subgroup of G generated by S. Let $\mu = \sum_{s \in S} \mu_s \delta_s$ be a positive measure on G with support S.

We recall that a function h on G is said to be μ-harmonic if

$$h = P_\mu h, \quad \text{where } P_\mu h(g) := \sum_{s \in S} \mu_s h(sg). \tag{3.8}$$

We want to describe the cone $\mathcal{H}^+$ of positive μ-harmonic functions h on G. By Choquet's theorem, it is enough to describe the extremal rays of this cone $\mathcal{H}^+$.

There are two constructions of extremal positive μ-harmonic functions.

3.4.1. The harmonic characters χ. By definition, the μ-harmonic characters are the characters $\chi : G \to \mathbb{R}_{>0}$ of G such that $\sum_{s\in S} \mu_s \chi(s) = 1$. Such a function $h = \chi$ is an extremal positive μ-harmonic function on G which is invariant by the center Z of G.

We now recall Margulis's theorem which tells us that this first construction is the only possible when $G_\mu^+ = G$.

Fact 3.7 (Margulis). *Let μ be a finite positive measure on a finitely generated nilpotent group G. If the semigroup G_μ^+ generated by the support of μ is equal to G, then every extremal positive μ-harmonic function h on G is a character.*

Sketch of proof of Fact 3.7 for $G = H_3(\mathbb{Z})$. Because of the assumption $G_\mu^+ = G$, we can assume that $\mu_c > 0$ and $\mu_a > 0$. The first part of the argument is as in the abelian case: since $h(x, y, z) \geq \mu_c h(x, y, z + 1)$, these two μ-harmonic functions are proportional and we get that, for some $t > 0$, one has that $h(x, y, z) = h(x, y, 0)t^z$. We now want to prove that $t = 1$.

Let K_t be the set of positive harmonic functions $h_0(x, y, z) = \psi_0(x, y)t^z$ with $h_0(e) = 1$. Since $G_\mu^+ = G$, the convex set K_t is compact for the pointwise convergence. The element $a \in G$ acts continuously by "right-translation and renormalization" on K_t. By Schauder's fixed point theorem, this action has a fixed point h_0 in K_t. It can be written as $h_0(x, y, z) = r^x \varphi_0(y)t^z$ with $r > 0$. But then one writes $h_0(g) \geq \mu_a h_0(ag)$ for all g in G, or equivalently $\varphi_0(y) \geq \mu_a r \varphi_0(y)t^y$ for all $y \in \mathbb{Z}$. This proves that $t = 1$. ■

When $G_\mu^+ \neq G$, a second construction is possible.

3.4.2. The functions h_{S_0,χ_0} induced from a harmonic character. Let $S_0 \subset S$ be an abelian subset. Denote by $\mu_{S_0} := \sum_{s\in S_0} \mu_s \delta_s$ the measure restriction of μ to S_0. Let χ_0 be a μ_{S_0}-harmonic character of G_{S_0}. We extend χ_0 as a function

$$\psi_0 := \chi_0 \mathbf{1}_{G_{S_0}}$$

on G which is 0 outside G_{S_0}. This function ψ_0 is μ-subharmonic, so that the sequence $P_\mu^n \psi_0$ is increasing. We set

$$h_{S_0,\chi_0} = \lim_{n\to\infty} P_\mu^n \psi_0.$$

We can tell exactly for which pairs (S_0, χ_0) the function h_{S_0,χ_0} is finite (see [2]). In this case, the function h_{S_0,χ_0} is an extremal positive μ-harmonic function on G.

We can now state the extension of Theorem 1.1 to a more general finitely supported measure μ on G.

Theorem 3.8. *Let $G = H_3(\mathbb{Z})$ and μ a positive measure on G whose finite support S generates the group G. Then every extremal positive μ-harmonic function h on G is proportional either to a character χ of G or to a translate $h_{S_0,\chi_0} \circ \rho_{g_0}$ of a function induced from a harmonic character.*

Corollary 3.9. *Let $G = H_3(\mathbb{Z})$, Z its center, and μ a probability measure on G whose finite support S generates the group G. The following are equivalent.*

(i) *Every positive μ-harmonic function on G is Z-invariant.*

(ii) *G_μ^+ contains two non-central elements whose product is in $Z \smallsetminus \{0\}$.*

Theorem 3.8 and Corollary 3.9 are proven in the sequel paper [2].

We will also see in [2] that on the nilpotent group of rank 4 with cyclic center, there exist extremal positive harmonic functions which are neither a harmonic character nor a function induced from a harmonic character.

References

[1] G. E. Andrews, *The Theory of Partitions*. Encyclopedia Math. Appl. 2, Addison-Wesley, Reading, Mass., 1976 Zbl 0371.10001 MR 0557013

[2] Y. Benoist, Positive harmonic functions on the Heisenberg group II. *J. Éc. polytech. Math.* **8** (2021), 973–1003 Zbl 1472.31012 MR 4257161

[3] E. Breuillard, Local limit theorems and equidistribution of random walks on the Heisenberg group. *Geom. Funct. Anal.* **15** (2005), no. 1, 35–82 Zbl 1083.60008 MR 2140628

[4] E. Breuillard, Equidistribution of dense subgroups on nilpotent Lie groups. *Ergodic Theory Dynam. Systems* **30** (2010), no. 1, 131–150 Zbl 1194.22012 MR 2586348

[5] G. Choquet, Représentations intégrales dans les cônes convexes sans base compacte. *C. R. Acad. Sci. Paris* **253** (1961), 1901–1903 Zbl 0099.31601 MR 131746

[6] G. Choquet and J. Deny, Sur l'équation de convolution $\mu = \mu * \sigma$. *C. R. Acad. Sci. Paris* **250** (1960), 799–801 Zbl 0093.12802 MR 119041

[7] P. Diaconis and R. Hough, Random walk on unipotent matrix groups. *Ann. Sci. Éc. Norm. Supér. (4)* **54** (2021), no. 3, 587–625 Zbl 07452920 MR 4311095

[8] Y. Guivarc'h, Groupes nilpotents et probabilité. *C. R. Acad. Sci. Paris Sér. A-B* **273** (1971), A997–A998 Zbl 0222.43004 MR 296986

[9] J. W. B. Hughes, Lie algebraic proofs of some theorems on partitions. In *Number Theory and Algebra*, pp. 135–155, Acad. Press, 1977 Zbl 0372.10010 MR 0491213

[10] G. A. Margulis, Positive harmonic functions on nilpotent groups. *Soviet Math. Dokl.* **7** (1966), 241–244 Zbl 0187.38902 MR 0222217

[11] S. Melczer, G. Panova, and R. Pemantle, Counting partitions inside a rectangle. *SIAM J. Discrete Math.* **34** (2020), no. 4, 2388–2410 Zbl 1453.05010 MR 4175829

[12] R. A. Proctor, Solution of two difficult combinatorial problems with linear algebra. *Amer. Math. Monthly* **89** (1982), no. 10, 721–734 Zbl 0509.05007 MR 683197

[13] D. Revuz, *Markov Chains*. 2nd edn., North-Holland Mathematical Library 11, North-Holland, Amsterdam, 1984 Zbl 0539.60073 MR 758799

[14] R. P. Stanley, Log-concave and unimodal sequences in algebra, combinatorics, and geometry. In *Graph Theory and its Applications: East and West (Jinan, 1986)*, pp. 500–535, Ann. New York Acad. Sci. 576, New York Acad. Sci., New York, 1989 Zbl 0792.05008 MR 1110850

[15] L. Takács, Some asymptotic formulas for lattice paths. *J. Statist. Plann. Inference* **14** (1986), no. 1, 123–142 Zbl 0616.60016 MR 845921

Yves Benoist
CNRS, Université Paris-Saclay, Bâtiment 307, 91405 Orsay, France; yves.benoist@u-psud.fr

DOI 10.4171/8ECM/23

Kähler–Einstein metrics and Archimedean zeta functions

Robert J. Berman

Abstract. While the existence of a unique Kähler–Einstein metric on a canonically polarized manifold X was established by Aubin and Yau already in the 70s, there are only a few explicit formulas available. In a previous work, a probabilistic construction of the Kähler–Einstein metric was introduced – involving canonical random point processes on X – which yields canonical approximations of the Kähler–Einstein metric, expressed as explicit period integrals over a large number of products of X. Here it is shown that the conjectural extension to the case when X is a Fano variety suggests a zero-free property of the Archimedean zeta functions defined by the partition functions of the probabilistic model. A weaker zero-free property is also shown to be relevant for the Calabi–Yau equation. The convergence in the case of log Fano curves is settled, exploiting relations to the complex Selberg integral in the orbifold case. Some intriguing relations to the zero-free property of the local automorphic L-functions appearing in the Langlands program and arithmetic geometry are also pointed out. These relations also suggest a natural p-adic extension of the probabilistic approach.

1. Introduction

A metric ω on a compact complex manifold X is said to be *Kähler–Einstein* if it has constant Ricci curvature:

$$\operatorname{Ric}\omega = -\beta\omega$$

for some constant β and ω is Kähler (i.e., parallel translation preserves the complex structure on X). Such metrics play a prominent role in current complex differential geometry and the study of complex algebraic varieties, in particular in the context of the Yau–Tian–Donaldson conjecture [39] and the minimal model program (MMP) in birational algebraic geometry [61]. In [7, 8], a probabilistic construction of Kähler–Einstein metrics with negative Ricci curvature on a complex projective algebraic variety X was introduced, where the Kähler–Einstein metric emerges from a canonical random point process on X. The random point process is defined in terms of purely algebro-geometric data. Accordingly, one virtue of this approach is that it generates

2020 Mathematics Subject Classification. Primary 32Q20; Secondary 14J45, 60G55, 11S40, 14G40.
Keywords. Kähler–Einstein metric, Fano variety, random point process, Langlands L-functions, Arakelov geometry.

new links between differential geometry on the one hand and algebraic-geometry on the other. In the present work, it is, in particular, shown that the conjectural extension to Kähler–Einstein metrics with positive Ricci curvature suggests a zero-free property of the Archimedean zeta functions defined by the partition functions of the probabilistic model. The particular case of Kähler–Einstein metrics with conical singularities on the Riemann sphere is settled, which from the algebro-geometric perspective corresponds to the case of log Fano curves.

We start by providing some background on Kähler–Einstein metrics and recapitulating the probabilistic approach to Kähler–Einstein metrics; the reader is referred to the survey [9] for more background and [13] for relations to the Yau–Tian–Donaldson conjecture. See also [20] for connections to quantum gravity in the context of the AdS/CFT correspondence and [11, 41] for connections to polynomial approximation theory and pluripotential theory in $\mathbb{C}^n$.

1.1. Kähler–Einstein metrics

The existence of a Kähler–Einstein metric on X implies that the *canonical line bundle* K_X of X (i.e., the top exterior power of the cotangent bundle of X) has a definite sign:

$$\operatorname{sign}(K_X) = \operatorname{sign}(\beta). \tag{1.1}$$

We will be using the standard terminology of positivity in complex geometry: a line bundle L is said to be *positive*, $L > 0$, if it is ample and *negative*, $L < 0$, if its dual is positive. In analytic terms, $L > 0$ iff L carries some Hermitian metric with strictly positive curvature. The standard additive notation for tensor products of line bundles will be adopted. Accordingly, the dual of L is expressed as $-L$. We will focus on the cases when $\beta \neq 0$. Then X is automatically a complex projective algebraic manifold and after a rescaling of the metric we may as well assume that $\beta = \pm 1$. For example, in the case when X is a hypersurface in $\mathbb{P}^{n+1}$, cut out by a homogeneous polynomial of degree d,

$$K_X > 0 \Leftrightarrow d > n + 2, \quad -K_X > 0 \Leftrightarrow d < n + 2.$$

In the case when $K_X > 0$, the existence of a Kähler–Einstein metric was established in the late seventies [3, 82]. The opposite case $-K_X > 0$ is the subject of the Yau–Tian–Donaldson conjecture, which was settled only recently (see the survey [39]). However, these are abstract existence results and there are very few explicit formulas for Kähler–Einstein metrics on complex algebraic varieties available. For example, even in the simplest case when $K_X > 0$ and X is complex curve, $n = 1$, finding an explicit formula for the Kähler–Einstein metric is equivalent to finding an explicit uniformization map from the curve X to the quotient $\mathbb{H}/G$ of the upper half-plane by a discrete subgroup $G \subset \mathrm{SL}(2, \mathbb{R})$. This has only been achieved for very special curves (such as the Klein quartic and Fermat curves), using techniques originating in the classical works by Weierstrass, Riemann, Fuchs, Schwartz, Klein, Poincaré, etc. Thus one virtue of the probabilistic approach is that it yields canonical approximations of the Kähler–Einstein metric on X, expressed as essentially explicit period-type integrals formulas (see formula (1.4)). These are reminiscent of the aforementioned few explicit formulas for Kähler–Einstein metrics, involving hypergeometric integrals (see [9, Section 2.1]).

1.2. The probabilistic approach

First, recall that, in the case when $\beta \neq 0$, a Kähler–Einstein metric ω_{KE} on X can be readily recovered from its (normalized) volume form dV_{KE}:

$$\omega_{KE} = \frac{1}{\beta}\frac{i}{2\pi}\partial\bar{\partial}\log dV_{KE},$$

where we have identified the volume form dV with its local density, defined with respect to a choice of local holomorphic coordinates z. The strategy of the probabilistic approach is to construct the normalized volume form dV_{KE} by a canonical sampling procedure on X. In other words, after constructing a canonical symmetric probability measure $\mu^{(N)}$ on X^N, the goal is to show that the corresponding *empirical measure*

$$\delta_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i},$$

viewed as a random discrete measure on X, converges in probability as $N \to \infty$ to the volume form dV_{KE} of the Kähler–Einstein metric ω_{KE}.

1.2.1. The case $\beta > 0$. When $K_X > 0$, the canonical probability measure $\mu^{(N)}$ on X^N, introduced in [7], is defined for a specific subsequence of integers N_k tending to infinity, the *plurigenera* of X:

$$N_k := \dim H^0(X, kK_X),$$

where $H^0(X, kK_X)$ denotes the complex vector space of all global holomorphic sections $s^{(k)}$ of the kth tensor power of the canonical line bundle $K_X \to X$ (called pluricanonical forms). The assumption that $K_X > 0$ ensures that $N_k \to \infty$, as $k \to \infty$. In terms of local holomorphic coordinates $z \in \mathbb{C}^n$ on X, a section $s^{(k)}$ of $kK_X \to X$ may be represented by local holomorphic functions $s^{(k)}$ on X, such that $|s^{(k)}|^{2/k}$ transforms as a density on X, i.e., defines a measure on X. The canonical symmetric probability measure $\mu^{(N_k)}$ on X^{N_k} is concretely defined by

$$\mu^{(N_k)} := \frac{1}{\mathcal{Z}_{N_k}}|\det S^{(k)}|^{2/k}, \quad \mathcal{Z}_{N_k} := \int_{X^{N_k}} |\det S^{(k)}|^{2/k}, \tag{1.2}$$

where $\det S^{(k)}$ is the holomorphic section of the canonical line bundle $(kK_{X^{N_k}})$ over X^{N_k}, defined by the Slater determinant

$$(\det S^{(k)})(x_1, x_2, \dots, x_{N_k}) := \det\big(s_i^{(k)}(x_j)\big), \tag{1.3}$$

in terms of a given basis $s_i^{(k)}$ in $H^0(X, kK_X)$. Under a change of bases, the section $\det S^{(k)}$ only changes by a multiplicative complex constant (the determinant of the change of bases matrix on $H^0(X, kK_X)$) and so does the normalizing constant $\mathcal{Z}_{N_k}$. As a result, $\mu^{(N_k)}$ is indeed canonical, i.e., independent of the choice of bases. Moreover, it is completely encoded by algebro-geometric data in the following sense: realizing X as projective algebraic subvariety, the section $\det S^{(k)}$ can be identified with a homogeneous polynomial, determined by the coordinate ring of X (or more precisely, the degree k component of the canonical ring of X).

The following convergence result was shown in [7].

Theorem 1.1. *Let X be a compact complex manifold with positive canonical line bundle K_X. Then the empirical measures δ_{N_k} of the corresponding canonical random point processes on X converge in probability, as $N_k \to \infty$, towards the normalized volume form dV_{KE} of the unique Kähler–Einstein metric ω_{KE} on X.*

In fact, the proof (discussed in Section 2.2) shows that the convergence holds at an exponential rate, in the sense of large deviation theory: for any given $\varepsilon > 0$, there exists a positive constant C_ε such that

$$\text{Prob}\left(d\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, dV_{KE}\right) > \varepsilon\right) \le C_\varepsilon e^{-N\varepsilon},$$

where d denotes any metric on the space $\mathcal{P}(X)$ of probability measures on X compatible with the weak topology. The convergence in probability implies, in particular, that the measures dV_k on X, defined by the expectations $\mathbb{E}(\delta_{N_k})$ of the empirical measure δ_{N_k}, converge towards dV_{KE} in the weak topology of measures on X:

$$dV_k := \mathbb{E}(\delta_{N_k}) = \int_{X^{N_k-1}} \mu^{(N_k)} \to dV_{KE}, \quad k \to \infty.$$

For k sufficiently large (ensuring that kK_X is very ample), the measures dV_k are, in fact, volume forms on X and induce a sequence of canonical Kähler metrics ω_k on X, expressed in terms of period-type integrals:

$$\omega_k := \frac{i}{2\pi}\partial\bar{\partial}\log dV_k = \frac{i}{2\pi}\partial\bar{\partial}\log\int_{X^{N_k-1}} |\det S^{(k)}|^{2/k}, \tag{1.4}$$

whose integrands are encoded by the degree k component of the canonical ring of X. The convergence above also implies that the canonical Kähler metrics ω_k converge, as $k \to \infty$, towards the Kähler–Einstein metric ω_{KE} on X, in the weak topology.

1.2.2. The case $\beta < 0$. When $-K_X > 0$, i.e., X is a Fano manifold, there are obstructions to the existence of a Kähler–Einstein metric. According to the *Yau–Tian–Donaldson conjecture (YTD)*, X admits a Kähler–Einstein metric iff X is *K-polystable*. The non-singular case was settled in [31–33] and the singular case in [68–70], building on the proof of the uniform version of the YTD conjecture on Fano manifolds in [18] (the "only if" direction was previously shown in [6]). In the probabilistic approach, a different type of stability condition naturally appears, dubbed *Gibbs stability* (connections with the YTD conjecture are discussed in [13]). The starting point for the probabilistic approach on a Fano manifold, introduced in [8, Section 6], is the observation that when $-K_X > 0$, one can replace k with $-k$ in the previous constructions concerning the case $K_X > 0$. Thus, given a positive integer k, we set

$$N_k := \dim H^0(X, -kK_X)$$

(which tends to infinity as $k \to \infty$, since $-K_X$ is ample) and define a measure on X^{N_k} by

$$\mu^{(N_k)} := \frac{1}{\mathcal{Z}_{N_k}} |\det S^{(k)}|^{-2/k}, \quad \mathcal{Z}_{N_k} := \int_{X^{N_k}} |\det S^{(k)}|^{-2/k}. \tag{1.5}$$

However, in this case it may happen that the normalizing constant $\mathcal{Z}_{N_k}$ diverges, since the integrand of $\mathcal{Z}_{N_k}$ blows up along the zero-locus in X^{N_k} of $\det S^{(k)}$. Accordingly, a Fano manifold X is called *Gibbs stable at level* k if $Z_{N_k} < \infty$ and *Gibbs stable* if it is Gibbs stable at level k for k sufficiently large. For a Gibbs stable Fano manifold X, the measure $\mu^{(N_k)}$ in formula (1.5) defines a canonical symmetric probability measure on X^{N_k}. We thus arrive at the following probabilistic analog of the YTD conjecture posed in [8, Section 6]:

Conjecture 1.2. *Let X be Fano manifold. Then*

- *X admits a unique Kähler–Einstein metric ω_{KE} if and only if X is Gibbs stable;*
- *if X is Gibbs stable, the empirical measures δ_N of the corresponding canonical point processes converge in probability towards the normalized volume form of ω_{KE}.*

In order to briefly compare with the YTD conjecture, denote by $\mathrm{Aut}(X)_0$ the Lie group of automorphisms (biholomorphisms) of X homotopic to the identity I. Fano manifolds are divided into the two classes, according to whether $\mathrm{Aut}(X)_0$ is *trivial* or *non-trivial*,

$$\mathrm{Aut}(X)_0 = \{I\} \quad \text{or} \quad \mathrm{Aut}(X)_0 \neq \{I\}.$$

In the former case, the Kähler–Einstein metric is uniquely determined (when it exists), while in the latter case, it is only uniquely determined modulo the action of the group $\mathrm{Aut}(X)_0$. This dichotomy is also reflected in the difference between *K-polystability* and the stronger notion of *K-stability*, which implies that $\mathrm{Aut}(X)_0$ is trivial. Similarly, the Gibbs stability of X also implies that the group $\mathrm{Aut}(X)_0$ is trivial [14] and should thus be viewed as the analog of K-stability. Accordingly, we shall focus on the case when $\mathrm{Aut}(X)_0$ is trivial (but see [9, Conjecture 3.8] for a generalization of Conjecture 1.2 to the case when $\mathrm{Aut}(X)_0$ is non-trivial).

There is also a natural analog of the stronger notion of *uniform K-stability* (discussed in more detail in [13]). To see this, first recall that Gibbs stability can be given a purely algebro-geometric formulation, saying that the $\mathbb{Q}$-divisor $\mathcal{D}_{N_k}$ in X^{N_k} cut out by the (multi-valued) holomorphic section $(\det S^{(k)})^{1/k}$ of $-K_{X^{N_k}}$ has mild singularities in the sense of the MMP. More precisely, X is Gibbs stable at level k iff $\mathcal{D}_{N_k}$ is *Kawamata log terminal (klt)*. This means that the *log canonical threshold (lct)* of $\mathcal{D}_{N_k}$ satisfies

$$\mathrm{lct}(\mathcal{D}_{N_k}) > 1 \tag{1.6}$$

(as follows directly from the analytic representation of the lct of a $\mathbb{Q}$-divisor $\mathcal{D}$, recalled in the appendix). Accordingly, X is called *uniformly Gibbs stable* if there exists $\varepsilon > 0$ such that, for k sufficiently large,

$$\mathrm{lct}(\mathcal{D}_{N_k}) > 1 + \varepsilon. \tag{1.7}$$

One is thus led to pose the following purely algebro-geometric conjecture:

Conjecture 1.3. *Let X be a Fano manifold. Then X is (uniformly) K-stable iff X is (uniformly) Gibbs stable.*

The uniform version of the "if" direction was settled in [48], using algebro-geometric techniques (see also [12] for a different direct analytic proof that uniform Gibbs stability implies the existence of a unique Kähler–Einstein metric). However, the converse is still widely open. And even if confirmed, it is a separate analytic problem to prove the convergence towards the Kähler–Einstein metric in Conjecture 1.2. In [9, Section 7], a variational approach to the convergence problem was introduced, which reduces the proof of the convergence towards the volume form dV_{KE} of Kähler–Einstein metric to establishing the following convergence result for the normalization constants $\mathcal{Z}_{N_k}$:

$$\lim_{N_k\to\infty} -\frac{1}{N_k}\log \mathcal{Z}_{N_k} = \inf_{\mu\in\mathcal{P}(X)} F(\mu), \tag{1.8}$$

where $F(\mu)$ is a functional on the space $\mathcal{P}(X)$ of probability measures on X, minimized by dV_{KE}, which may be identified with the Mabuchi functional (see Section 2.2). This variational approach is inspired by a statistical mechanical formulation, where F appears as a free-energy type functional and β appears as the "inverse temperature”. A central role is played by the *partition function*

$$\mathcal{Z}_{N_k}(\beta) := \int_{X^{N_k}} \|\det S^{(k)}\|^{2\beta/k}\, dV^{\otimes N_k}, \quad \beta \in [-1, \infty[\tag{1.9}$$

coinciding with the normalization constant $\mathcal{Z}_N$ when $\beta = -1$. However, for $\beta \neq -1$, $\mathcal{Z}_{N_k}(\beta)$ depends on the choice of a Hermitian metric $\|\cdot\|$ on $-K_X$, which, in turn, induces a volume form dV on X. In order to establish the convergence (1.8), two different approaches were put forth in [9, Section 7], which hinge on establishing either of the following two hypotheses:

- the “upper bound hypothesis” for the mean energy (discussed in Section 2.2),
- the “zero-free hypothesis” (discussed in Section 2.4):

$$\mathcal{Z}_{N_k}(\beta) \neq 0 \text{ on some } N_k\text{-independent neighborhood } \Omega \text{ of }]-1, 0] \text{ in } \mathbb{C}. \tag{1.10}$$

While originally defined for $\beta \in [-1, \infty[$, the partition function $\mathcal{Z}_{N_k}(\beta)$ extends to a meromorphic function of $\beta \in \mathbb{C}$, all of whose poles appear on the negative real axes. Indeed, by taking a covering of X, the function $\mathcal{Z}_{N_k}(\beta)$ may be expressed as a sum of functions of the form

$$Z(\beta) := \int_{\mathbb{C}^m} |f|^{2\beta}\Phi\, d\lambda, \tag{1.11}$$

for a holomorphic function f and a Schwartz function Φ on $\mathbb{C}^m$. One can then invoke classical general results of Atyiah and Bernstein for such meromorphic functions $Z(\beta)$ (recalled in Section A.2 of the appendix). The first negative pole of $\mathcal{Z}_{N_k}(\beta)$ is precisely the negative of the log canonical threshold $\operatorname{lct}(\mathcal{D}_{N_k})$. The zero-free hypothesis referred to above demands that there exists an N-independent neighborhood of $]-1, 0]$ in $\mathbb{C}$, where $\mathcal{Z}_{N_k}(\beta) \neq 0$. As shown in Section 2.4, the virtue of this hypothesis is that it allows one to prove the convergence in formula (1.8) by “analytically continuing” the convergence for $\beta > 0$ to $\beta = -1$. In the statistical mechanics literature, this line of argument goes back to the Lee–Yang theory of phase transitions (see Remark 2.7).

1.3. The partition function $\mathcal{Z}_{N_k}(\beta)$ viewed as local Archimedean zeta function

From an algebro-geometric perspective, the partition function $\mathcal{Z}_{N_k}(\beta)$ (formula (1.9)) is an instance of an *Archimedean zeta function*. More generally, replacing the local field $\mathbb{C}$ and its standard Archimedean absolute value $|\cdot|$ with a local field F and an absolute value $|\cdot|_F$ on F, meromorphic functions $Z(\beta)$ as in formula (1.11) can be attached to any polynomial f defined over the local field F. Such meromorphic functions are usually called *local Igusa zeta functions* [53]. This is briefly recalled in Section A.2 of the appendix. For example, the Riemann zeta function $\zeta(s)$ may be expressed as a Euler product over such local meromorphic functions $Z_p(s)$ as p

ranges over all primes p, i.e., all non-Archimedean places p of the global field $\mathbb{Q}$:

$$\zeta(s) := \sum_{n=1}^{\infty} n^{-s} = \prod_p Z_p(s), \quad Z_p(s) = \int_{\mathbb{Q}_p^\times} |x|_{\mathbb{Q}_p}^s \Phi_p \, d^\times x = (1 - p^{-s})^{-1},$$

where $\mathbb{Q}_p$ is the localization of $\mathbb{Q}$ at p, i.e., the p-adic field $\mathbb{Q}_p$, endowed with its standard normalized non-Archimedean absolute value and multiplicative Haar measure $d^\times x$ on $\mathbb{Q}_p^\times$ and Φ_p denotes the p-adic Gaussian. This is explained in Tate's celebrated thesis [78], where it is shown that the classical procedure of completing the Riemann zeta function amounts to including a factor $Z_p(s)$ corresponding to the standard Archimedean absolute value on $\mathbb{R}$, which is proportional to the Gamma function.[1] In this case, all the local factors $Z_p(s)$ are manifestly non-zero (while the corresponding global zeta function $\zeta(s)$ does have zeros). It should, however, be stressed that it is rare that general local Igusa zeta functions of the form (1.11) and their zeros can be computed explicitly. Still, one might hope that the canonical nature of $\mathcal{Z}_{N_k}(\beta)$ may facilitate the situation. One small step in this direction is taken in Section 5, where some intriguing relations between the partition functions $\mathcal{Z}_{N_k}(\beta)$ and the local L-functions appearing in the Langlands program are pointed out (generalizing the local factors $Z_p(\beta)$ of the Riemann zeta function). In particular, it is shown that in the simplest case when X is n-dimensional complex projective space and N_k is minimal, i.e., $N_k = n + 1$, the partition function $\mathcal{Z}_{N_k}(\beta)$ can be identified with a standard local L-function L_p attached to the group $\mathrm{GL}(n + 1, \mathbb{Q})$ when the place p of the global field $\mathbb{Q}$ is taken to be the one defined by the complex Archimedean absolute. Accordingly, in this particular case, $\mathcal{Z}_{N_k}(\beta)$ has a strong zero-free property as a consequence of the standard zero-free property of local L-functions.

1.4. Main new results in the case of log Fano curves

Here it will be demonstrated that both approaches discussed above are successful in one complex dimension, $n = 1$. The only one-dimensional Fano manifold X is the complex projective line (the Riemann sphere) and its Kähler–Einstein metrics are all biholomorphically equivalent to the standard round metric on the two-sphere. But a geometrically richer situation appears when introducing weighted points (conical singularities) on the Riemann sphere. From the algebro-geometric point of view, this fits into the standard setting of *log pairs* (X, Δ), consisting of complex (normal) projective variety X (here assumed to be non-singular, for simplicity) endowed with a $\mathbb{Q}$-divisor Δ on X, i.e., a sum of irreducible subvarieties Δ_i of X of codimension one, with coefficients w_i in $\mathbb{Q}$. In this log setting, the role of the canonical line bundle

[1] Expressing $d^\times x = x^{-1} dx$ reveals that the role of β is played by $s - 1$; see Section 5.1.

K_X is placed by the *log canonical line bundle*

$$K_{(X,\Delta)} := K_X + \Delta$$

(viewed as a $\mathbb{Q}$-line bundle) and the role of the Ricci curvature $\operatorname{Ric}\omega$ of a metric ω is played by twisted Ricci curvature $\operatorname{Ric}\omega - [\Delta]$, where $[\Delta]$ denotes the current of integration defined by Δ. The corresponding *log Kähler–Einstein equation* thus reads

$$\operatorname{Ric}\omega - [\Delta] = \beta\omega, \quad \beta = \pm 1, \tag{1.12}$$

where $[\Delta]$ denotes the current of integration along Δ. When β is non-zero, existence of a solution ω_{KE} forces

$$\beta(K_X + \Delta) > 0.$$

In general, the equation (1.12) should be interpreted in the weak sense of pluripotential theory [16,42]. However, in case when (X, Δ) is *log smooth*, i.e., the components of Δ have simple normal crossings (which means that they intersect transversally), it follows from [52, 55] that a positive current ω solves the equation (1.12) iff ω is a bona fide Kähler–Einstein metric on $X - \Delta$ and ω has edge-cone singularities along Δ, with cone-angle $2\pi(1 - w_i)$, prescribed by the coefficients w_i of Δ. In particular, in the *orbifold case*

$$\Delta = \sum \left(1 - \frac{1}{m_i}\right)\Delta_i, \quad m_i \in \mathbb{Z}_+, \tag{1.13}$$

the log Kähler–Einstein metrics locally lift to a bona fide Kähler–Einstein metric on local coverings of X (branched along Δ and $K_X + \Delta$ may be identified with the orbifold canonical line bundle) [26, Section 2].

Example 1.4. Let X be the complex hypersurface of weighted projective space $\mathbb{P}(a_0, \dots, a_n)$, cut out by a quasi-homogeneous polynomial F on $\mathbb{C}^{n+1}$ of degree d, whose zero-locus $Y \subset \mathbb{C}^{n+1} - \{0\}$ is assumed non-singular. Then the orbifold (X, Δ) defined by the branching divisor Δ on X of the fibration $Y - \{0\} \to X$, induced by the natural quotient projection

$$\mathbb{C}^{n+1} - \{0\} \to \mathbb{P}(a_0, \dots, a_n),$$

is a Fano orbifold (i.e., $-(K_X + \Delta) > 0$) iff $d < a_0 + a_1 + \dots + a_n$.

The probabilistic approach naturally extends to the setting of log pairs (X, Δ) satisfying $\beta(K_X + \Delta) > 0$ yielding a canonical probability measure on X^{N_k}, that we shall denote by $\mu_\Delta^{(N_k)}$. Indeed, one simply replaces the canonical line bundle K_X with the log canonical line bundle $K_{(X,\Delta)}$ in the previous constructions (cf. [8, Section 5] and [9, Section 3.2.4]).

1.4.1. Log Fano curves. Let now (X, Δ) be a log Fano curve (X, Δ), i.e., X is the complex projective line and

$$\Delta = \sum_{i=1}^{m} w_i p_i$$

for positive weights w_i satisfying $\sum_{i=1}^{m} w_i < 2$. In this case, it turns out that the "upper bound hypothesis" for the mean energy does hold, which leads to the following result announced in [9, Section 3.2.4]:

Theorem 1.5. *Let (X, Δ) be a log Fano curve. Then the following is equivalent:*

- *(X, Δ) is Gibbs stable;*
- *(X, Δ) is uniformly Gibbs stable;*
- *the following weight condition holds:*

$$w_i < \sum_{i \neq j} w_j, \quad \forall i; \tag{1.14}$$

- *there exists a unique Kähler–Einstein metric ω_{KE} for (X, Δ).*

Moreover, if any of the conditions above hold, then the laws of the corresponding empirical measures δ_N satisfy a large deviation principle (LDP) with speed N, whose rate functional has a unique minimizer, namely $\omega_{KE}/\int_X \omega_{KE}$. In particular, for any given $\varepsilon > 0$,

$$\text{Prob}\left(d\left(\frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}, \frac{\omega_{KE}}{\int_X \omega_{KE}} \right) > \varepsilon \right) \leq C_\varepsilon e^{-N\varepsilon}.$$

Existence of solutions to the log Kähler–Einstein equation (1.12) in the one-dimensional setting was first shown in [79], under the weight condition (1.14) and uniqueness in [71]. The weight condition (1.14) is also equivalent to uniform K-stability of (X, Δ) [47, Example 6.6] and thus the previous theorem confirms Conjecture 1.3 for log Fano curves.

We also show that in the case when the support of Δ consists of three points, the following variant of the "zero-free hypothesis" holds:

$$\mathcal{Z}_{N_k,\Delta} \neq 0,$$

when the coefficients of Δ are complexified, so that $\mathcal{Z}_{N_k,\Delta}$ is extended to a meromorphic function on $\mathbb{C}^3$ (the proof exploits that $\mathcal{Z}_{N_k,\Delta}$ can be expressed as the complex Selberg integral, which first appeared in the conformal field theory (CFT)). This leads to an alternative proof of the previous theorem, in this particular case, by "analytically continuing" the convergence result in the case $K_X + \Delta > 0$ to the log Fano case $K_X + \Delta < 0$.

Example 1.6. The case of three points includes, in particular, the case when X is a *Fano orbifold curve*. Such a curve may be embedded into a weighted $\mathbb{P}^2$ and is defined by the zero-locus of explicit quasi-homogeneous polynomial $F(X, Y, Z)$ in $\mathbb{C}^3$ (the du Val singularities). In the case of three orbifold points, there always exists a unique log Kähler–Einstein metric on X, concretely realized as the quotient $\mathbb{P}^1/G$ of the standard SU(2)-invariant metric on $\mathbb{P}^1$ under the action of a discrete subgroup G of SU(2) (branched over the three points in question).

1.5. Organization

In Section 2, conditional convergence results on log Fano varieties are obtained, formulated in terms of either the "upper bound hypothesis" on the mean-energy or the "zero-free hypothesis" of the partition function. Then – after a digression on the Calabi–Yau equation in Section 3 – in Section 4, the hypotheses in question are verified for log Fano curves and Fano orbifolds, respectively. Section 5 is of a speculative nature, comparing the strong form of the zero-free hypothesis with the standard zero-free property of the local L-functions appearing in the Langlands program. The paper is concluded with an appendix, providing background on lct's and Archimedean zeta functions.

2. Conditional convergence results on log Fano varieties

In this section, it is explained how to reduce the proof of the convergence on Fano manifolds X in Conjecture 1.2 to establishing either one of two different hypotheses, building on [9, Section 7]. More generally, we will consider the setup of log Fano varieties (X, Δ), discussed in Section 1.4. For simplicity, X will be assumed to be non-singular. We will be using the standard correspondence between metrics $\|\cdot\|$ on log canonical line bundles $-(K_X + \Delta)$ and volume forms dV_Δ on $X - \Delta$, which are singular when viewed as measures on X (see [9, Section 4.1.7] for background, where the measure dV_Δ is denoted by μ_0).

2.1. Setup

Let (X, Δ) be a log Fano variety. As recalled in Section 1.4, this means that Δ is a divisor with positive coefficients and that $-(K_X + \Delta) > 0$. We will allow Δ to have real coefficients. Set

$$N_k := \dim H^0\big(X, -k(K_X + \Delta)\big),$$

where k ranges over the positive numbers with the property that $-k(K_X + \Delta)$ is a well-defined line bundle on X. To simplify the notation, we will often drop the subscript k in the notation for N_k. Since,

$$k \to \infty \Leftrightarrow N \to \infty,$$

this should, hopefully, not cause any confusion. As discussed in Section 1.4, assuming that (X, Δ) is Gibbs stable, we get a sequence of canonical probability measures $\mu_\Delta^{(N)}$ on X^N. Fixing a smooth Hermitian metric $\|\cdot\|$ on the $\mathbb{R}$-line bundle $-(K_X + \Delta)$ with positive curvature $\mu_\Delta^{(N)}$ may be expressed as

$$\mu_\Delta^{(N)} := \frac{1}{\mathcal{Z}_N} \|\det S^{(k)}\|^{2/k} dV_{(X,\Delta)}^{\otimes N}, \quad \mathcal{Z}_N := \int_{X^N} \|\det S^{(k)}\|^{2/k} dV_{(X,\Delta)}^{\otimes N}, \tag{2.1}$$

where $dV_{(X,\Delta)}$ is the singular volume form on X corresponding to the metric $\|\cdot\|$ on $-(K_X + \Delta)$ and $\det S^{(k)}$ is the Slater determinant of $H^0(X, -k(K_X + \Delta))$ induced by a choice of bases $s_1^{(k)}, \dots, s_N^{(k)}$ for $H^0(X, -k(K_X + \Delta))$, defined as in formula (1.3). Since $\mu_\Delta^{(N)}$ is independent of the choice of bases, we may as well assume that the basis is orthonormal with respect to the Hermitian product induced by $(\|\cdot\|, dV)$. The condition that (X, Δ) is Gibbs stable means that the normalization constant $\mathcal{Z}_N$ is finite. Hence, it implies that the local densities of dV are in L^1_{loc} (which in algebraic terms means that Δ is klt divisor).

From a statistical mechanical point of view, the probability measure $\mu_\Delta^{(N)}$ on X^N may be expressed as the *Gibbs measure*

$$\mu_\beta^{(N)} = \frac{e^{-\beta N E^{(N)}}}{\mathcal{Z}_N(\beta)} dV_\Delta^{\otimes N},$$

$$E^{(N)}(x_1, \dots, x_N) := -\frac{1}{kN} \log\left(\left\| \det S^{(k)}(x_1, \dots, x_N)\right\|^2\right) \tag{2.2}$$

with $\beta = -1$. In physical terms, the Gibbs measure represents the microscopic state of N interacting particles in thermal equilibrium at inverse temperature β, with $E^{(N)}(x_1, \dots, x_N)$ playing the role of the *energy per particle and* the normalizing constant

$$\mathcal{Z}_N(\beta) = \int_{X^N} e^{-\beta N E^{(N)}} dV_{(X,\Delta)}^{\otimes N} = \int_{X^N} \|\det S^{(k)}\|^{2\beta/k} dV_{(X,\Delta)}^{\otimes N} \tag{2.3}$$

is called the *partition function*. It should, however, be stressed that, while the probability measure $\mu_\Delta^{(N)}$ is canonical, i.e., independent of the choice of metric $\|\cdot\|$, this is not so when $\beta \neq -1$. But one advantage of introducing the parameter β is that $\mu_\beta^{(N_k)}$ is a well-defined probability measure as long as $\beta > -\operatorname{lct}(X, \Delta)$, where $\operatorname{lct}(X, \Delta)$ denotes the global lct of (X, Δ) (whose definition is recalled in the appendix). In particular, it is, trivially, well defined when $\beta > 0$.

Fixing $\beta \in [-1,\infty[$, we can view the empirical measure

$$\delta_N := \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i} : \quad X^N \to \mathcal{P}(X)$$

as a random discrete measure on X. To be more precise, δ_N is a random variable on the ensemble $(X^N, \mu_\beta^{(N)})$, taking values in the space $\mathcal{P}(X)$ of probability measures on X. Accordingly, the *law* of δ_N is the probability measure

$$\Gamma_{N,\beta} := (\delta_N)_*\mu_\beta^{(N)} \in \mathcal{P}\big(\mathcal{P}(X)\big)$$

on $\mathcal{P}(X)$, defined as the push-forward of the probability measure $\mu_\beta^{(N)}$ on X^N to $\mathcal{P}(X)$ under the map δ_N.

2.2. The case $\beta > 0$

The following result, which is a special case of [7, Theorem 5.7] (when Δ is trivial) and [8, Theorem 4.3] (when Δ is non-trivial), establishes an LDP for the laws $\Gamma_{N,\beta}$ of δ_N as $N \to \infty$, which may be symbolically expressed as

$$\Gamma_{N,\beta} := (\delta_N)_*\mu_\beta^{(N)} \sim e^{-N(F(\mu)-F(\beta))}, \quad N \to \infty$$

(formally viewing the right-hand side as a density on the infinite dimensional space $\mathcal{P}(X)$; the precise meaning of the LDP is recalled below).

Theorem 2.1. *Let (X,Δ) be a log Fano variety. For $\beta > 0$, the sequence $\Gamma_{N,\beta}$ of probability measures on $\mathcal{P}(X)$ satisfies an LDP speed N and rate functional*

$$F_\beta(\mu) - F(\beta), \quad F(\mu) := \beta E(\mu) + \mathrm{Ent}(\mu), \quad F(\beta) := \inf_{\mathcal{P}(X)} F_\beta(\mu), \tag{2.4}$$

where $E(\mu)$ is the pluricomplex energy of μ relative to the Kähler form ω defined by the curvature of the metric $\|\cdot\|$ on $-(K_X+\Delta)$ and Ent(μ) *is the entropy of μ relative to dV_Δ. In particular, the random measure δ_N converges in probability, as $N \to \infty$, to the unique minimizer μ_β of F_β in $\mathcal{P}(X)$, i.e.,*

$$\lim_{N\to\infty} \Gamma_{N,\beta} = \delta_{\mu_\beta} \quad \textit{in } \mathcal{P}\big(\mathcal{P}(X)\big) \tag{2.5}$$

and the following convergence of the partition functions $\mathcal{Z}_N(\beta)$ holds:

$$\lim_{N\to\infty} -\frac{1}{N}\log \mathcal{Z}_N(\beta) = F(\beta). \tag{2.6}$$

We recall that the *entropy* $\mathrm{Ent}(\mu)$ of μ relative to a given measure ν is defined by

$$\mathrm{Ent}(\mu) = \int_X \log \frac{\mu}{\nu} \mu$$

when μ has a density with respect to ν and otherwise $\mathrm{Ent}(\mu) := \infty$.[2] As for the pluricomplex energy $E(\mu)$ of a measure μ on X, relative to a reference form ω_0, it was first introduced in [17, Theorem 4.3]. From a thermodynamical point of view, the functional $F_\beta(\mu)$, introduced in [4, Theorem 4.3], can be viewed as the *free energy*.[3] The pluricomplex $E(\mu)$ may be defined as the greatest lsc extension to $\mathcal{P}(X)$ of the functional $E(\mu)$ on the space of volume forms μ in $\mathcal{P}(X)$ whose first variation is given by

$$dE(\mu) = -\varphi_\mu, \tag{2.7}$$

where φ_μ is a smooth solution to the complex Monge–Ampère equation (also known as the Calabi–Yau equation):

$$\frac{1}{V}\Big(\omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi_\beta\Big)^n = \mu, \quad V := \int_X \omega^n.$$

This property determines the functional $E(\mu)$ up to an additive constant which is fixed by imposing the normalization condition

$$E(\omega_0^n/V) = 0, \tag{2.8}$$

in the case when the reference form ω_0 is Kähler. Using the property (2.7), it is shown in [9, Proposition 4.1] that the minimizer μ_β of $F_\beta(\mu)$ is the normalized volume form on $X - \Delta$ uniquely determined by the property that

$$\mu_\beta = e^{\beta\varphi_\beta} dV_\Delta,$$

where the function φ_β is the unique smooth bounded Kähler potential on $X - \Delta$ solving the complex Monge–Ampère equation

$$\frac{1}{V}\Big(\omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi_\beta\Big)^n = e^{\beta\varphi_\beta} dV_\Delta. \tag{2.9}$$

It follows that the corresponding Kähler form

$$\omega_\beta := \omega + \frac{1}{\beta}\frac{i}{2\pi}\partial\bar{\partial}\log\frac{\mu_\beta}{dV_\Delta}\Big(= \omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi_\beta\Big)$$

[2] We are using the “mathematical” sign convention for the entropy, which renders $\mathrm{Ent}(\mu)$ non-negative when the reference measure ν is a probability measure and thus $\mathrm{Ent}(\mu)$ coincides with the *Kullback–Leibler divergence* in information theory.

[3] Strictly speaking, it is F_β/β which plays the role of free energy in thermodynamics.

satisfies the twisted Kähler–Einstein equation

$$\operatorname{Ric}\omega_\beta - [\Delta] = -\beta\omega_\beta + (\beta+1)\omega_0, \tag{2.10}$$

on X, coinciding with the (log) Kähler–Einstein equation (1.12) when $\beta = -1$.

Remark 2.2. Incidentally, the functional

$$\mathcal{M}(\varphi) := F_{-1}\Big(\frac{1}{V}\Big(\omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi_\beta\Big)^n\Big)$$

coincides with the *Mabuchi functional* for the log Fano variety (X, Δ), as explained in [9, Section 5.3]. Moreover, the twisted Kähler–Einstein equation (2.10) coincides with the logarithmic version of Aubin's continuity equation with "time-parameter" $t := -\beta$.

The precise definition of an LDP, which goes back to Cramér and Varadhan [37], is recalled in [9, Proposition 4.1]. For the purpose of the present paper, it will be convenient to use the following equivalent ("dual") characterization of the LDP in the previous theorem: for any continuous function $\Phi(\mu)$ on $\mathcal{P}(X)$:

$$\lim_{N\to\infty} -\frac{1}{N}\log\int_{X^N} e^{-N\beta E^{(N)}}e^{-N\Phi(\delta_N)} = \inf_{\mathcal{P}(X)}\big(F(\mu)+\Phi(\mu)\big) \tag{2.11}$$

(as follows from well-known general results of Varadhan and Bryc [37, Theorem 4.4.2]).

2.2.1. Outline of the proof.

Before turning to the case when $\beta < 0$, we briefly recall that a key ingredient in the proof of the previous theorem is the convergence

$$E^{(N)}(x_1,\dots,x_N) \to E(\mu), \quad N\to\infty, \tag{2.12}$$

which holds in the sense of Gamma-convergence (deduced from the convergence and differentiability of weighted transfinite diameters in [15, Theorems A and B]). Combining this convergence with some heuristics going back to Boltzmann suggests that the contribution of the volume form $dV^{\otimes N}$ in the Gibbs measure (2.2) should give rise to the additional entropy term appearing in the rate functional:

$$(\delta_N)_*(e^{-\beta NE^{(N)}}dV^{\otimes N}) \sim e^{-NE(\mu)}(\delta_N)_*(dV^{\otimes N}) \sim e^{-N\beta E(\mu)}e^{-N\operatorname{Ent}(\mu)}.$$

This is made rigorous in [7] using an effective submean property of the density of $\mu_\beta^{(N)}$ on the N-fold symmetric product of X, viewed as a Riemannian orbifold (leveraging results in geometric analysis).

2.3. The case $\beta < 0$

In the case when $\beta < 0$, we may define the free energy functional $F_\beta(\mu)$ by the same expression as in formula (2.4), $F_\beta = \beta E + \mathrm{Ent}(\mu)$, when $E_{\omega_0}(\mu) < \infty$ and otherwise we set $F_\beta(\mu) = \infty$. The definition is made so that we still have $F_\mu(\mu) \in]-\infty, \infty]$ with $F_\mu(\mu) < \infty$ iff both $E(\mu) < \infty$ and $\mathrm{Ent}(\mu) < \infty$.

In order to handle the large N-limit in the case when $\beta < 0$, a variational approach was introduced in [9, Section 7], which reduces the problem to establishing the following *"upper bound hypothesis"* for the mean energy:

$$\limsup_{N\to\infty} \int_{X^N} E^{(N)} \mu_{\Delta,\beta}^{(N)} \leq E(\Gamma_\beta) := \int_{\mathcal{P}(X)} E(\mu)\Gamma_\beta(\mu) \tag{2.13}$$

for any large N-limit point Γ of $\Gamma_{N,\beta}$ in $\mathcal{X}$. This property is independent of the choice of metric $\|\cdot\|$ on $-(K_X + \Delta)$. Moreover, the corresponding lower bound always holds (as follows from the convergence (2.12)). The following theorem is an extension of the results in [9, Section 7] to the case when Δ is non-trivial.

Theorem 2.3. *Let (X, Δ) be a log Fano variety. Assume that (X, Δ) is uniformly Gibbs stable. Then (X, Δ) admits a unique Kähler–Einstein metric ω_{KE}. Moreover, in the following list each statement implies the next one:*

(1) *the "upper bound hypothesis" (2.13) for the mean energy holds when $\beta = -1$;*

(2) *the convergence (2.6) for the partition functions holds when $\beta = -1$;*

(3) *the empirical measures δ_N of the canonical random point process on X converge in law towards the normalized volume form dV_{KE} of ω_{KE}; i.e., the convergence (2.5) holds when $\beta = -1$.*

Furthermore,if the "upper bound hypothesis" (2.13) is replaced by the stronger hypothesis that the convergence holds when $E^{(N)}$ is replaced by $E^{(N)} + \Phi(\delta_N)$ for any continuous functional Φ on $\mathcal{P}(X)$ (and E is replaced by $E + \Phi$), then the LDP in Theorem 2.1 holds for $\beta = -1$.

Proof. The proof in the general case is similar to the case when Δ is trivial. Indeed, the assumption that (X, Δ) is uniformly Gibbs stable implies, by a simple modification of the proof of [48, Theorem 2.5] (concerning the case when Δ is trivial) that $\delta(X, \Delta) > 1$, which by [47] is equivalent to (X, Δ) being uniformly K-stable. Hence, by the solution of the uniform version of the YTD conjecture for log Fano varieties (X, Δ) with X non-singular in [18] (extended to general log Fano varieties in [68, 69]), it follows that (X, Δ) admits a unique Kähler–Einstein metric. Next, we summarize the proof of the convergence in [9, Section 7]; all steps are essentially the same in the case when Δ is non-trivial. Set

$$F_N(\beta) := -\frac{1}{N} \log \mathcal{Z}_N(\beta), \quad F(\beta) := \inf_{\in \mathcal{P}(X)} F_\beta(\mu) \tag{2.14}$$

and consider the mean free energy functional on $\mathcal{P}(X^N)$ defined by

$$F_N(\mu_N) := \beta \int_{X^N} E^{(N)}\mu_N + \frac{1}{N}\operatorname{Ent}(\mu_N),$$

where $\operatorname{Ent}(\mu_N)$ denotes the entropy of μ_N relative to $(dV_\Delta)^{\otimes N}$. By Gibbs variational principle (or Jensen's inequality),

$$F_N(\beta) = \inf_{\mu_N \in \mathcal{P}(X^N)} F_{N,\beta}(\mu_N) = F_{N,\beta}(\mu_{N,\beta}). \tag{2.15}$$

Moreover,

$$F(\beta) = \inf_{\mathcal{P}(\mathcal{P}(X))} F_\beta(\Gamma) = F_\beta(\delta_{\mu_\beta}), \tag{2.16}$$

where $F_\beta(\Gamma)$ denotes the following functional on $\mathcal{P}(\mathcal{P}(X))$:

$$F_\beta(\Gamma) := \int_{\mathcal{P}(X)} F_\beta(\mu)\Gamma$$

and δ_{μ_β} is the unique minimizer of $F(\Gamma)$ in $\mathcal{P}(\mathcal{P}(X))$ (using that $F(\mu)$ is lsc, thanks to the energy/entropy compactness theorem in [16] and hence $F(\Gamma)$ is lsc and linear on $\mathcal{P}(\mathcal{P}(X))$). Now, as shown in the course of the proof of [8, Theorem 6.7] (and refined in Step 1 in the proof of [9, Theorem 7.6]) for *any* β, the following inequality holds:

$$\limsup_{N\to\infty} F_N(\beta) \leq F(\beta) \tag{2.17}$$

(as follows from combining Gibbs variational principle with the Gamma-convergence (2.12) of $E^{(N)}$ towards $E(\mu)$). Combining Gibbs variational principle (2.15) with the variational principle (2.16) for $F(\beta)$, this means that

$$\limsup_{N\to\infty}\Big(\inf_{\mu_N\in\mathcal{P}(X^N)} F_{N,\beta}(\mu_N)\Big) \leq \inf_{\in\mathcal{P}(X)} F_\beta(\mu).$$

Moreover, as shown in [9, Section 7], if the "upper bound hypothesis" on the mean energy holds, then the corresponding lower bound also holds; i.e., the convergence (2.6) of the partition functions holds:

$$\lim_{N\to\infty} F_N(\beta) = F(\beta). \tag{2.18}$$

Indeed, combining the "upper bound hypothesis" with the well-known sub-additivity property of the mean entropy yields

$$F_\beta(\Gamma_\beta) \leq \liminf_{N\to\infty} F_{N,\beta}(\mu_{N,\beta})$$

for any limit point Γ_β of $\Gamma_{N,\beta}$, in the case $\beta = -1$. Combined with the upper bound

(2.17) and formula (2.16) for $F(\beta)$, it then follows that Γ_β minimizes $F_{-1}(\Gamma)$ and hence, by the uniqueness of minimizer, $\Gamma = \delta_{\mu_{-1}}$, as desired. All in all, this shows that "(1)⇒(2)⇒(3)" in the theorem.

Finally, to prove the LDP stated in the theorem, one just repeats the previous argument with $E^{(N)}$ replaced by $E^{(N)}_\Phi := E^{(N)} + \Phi(\delta_N)$. Then $\mathcal{Z}_N(\beta)$ gets replaced with $\int_{X^N} e^{-NE^{(N)}_\Phi} dV^{\otimes N}$ and hence the convergence (2.11) follows, as before, from the implication (1)⇒(2), now applied to $E^{(N)}_\Phi$.

In fact, the implications in the previous theorem may "almost" be reversed, by exploiting that the mean N-particular energy at inverse temperature β is proportional to the logarithmic derivative of $Z_N(\beta)$. More precisely, the following theorem holds, where it is assumed, for technical reasons, that X is a Fano orbifold. ∎

Theorem 2.4. *Let* (X, Δ) *be a Fano orbifold and assume that* (X, Δ) *is uniformly Gibbs stable. Then there exists* $\varepsilon > 0$ *such that* F_β *admits a unique minimizer* μ_β *for any* $\beta \in]-1-\varepsilon, 0[$. *Moreover, the following is equivalent:*

(1) *the "upper bound hypothesis" for the mean energy* (2.13) *holds for any* $\beta \in]-1-\varepsilon, 0[$;

(2) *the convergence* (2.6) *for the partition functions holds for any* $\beta \in]-1-\varepsilon, 0[$;

(3) *the convergence* (2.6) *for the partition functions holds* and *the convergence* (2.5) *of the laws of* δ_N *holds for any* $\beta \in]-1-\varepsilon, 0[$.

Furthermore, If (1), (2) or (3) holds, then

$$\lim_{N\to\infty} \int_{X^N} E^{(N)} \mu^{(N)}_{\Delta,\beta} = E(\mu_\beta). \tag{2.19}$$

Proof. First, assume that (X, Δ) is a log Fano variety. As explained in the proof of the previous theorem, X admits a unique Kähler–Einstein metric. Hence, it follows from [34] (and [18]) that $F_{-1}(\mu)$ is coercive with respect to E; i.e., there exists $\varepsilon > 0$ such that

$$F_{-1} \geq \varepsilon E - 1/\varepsilon$$

on $\mathcal{P}(X)$. Thus F_β is also coercive with respect to E for any $\beta > -1-\varepsilon$. In particular, it follows from the energy-entropy compactness theorem in [16] that F_β admits a minimizer. Moreover, as shown in [16], any minimizer has the property that the corresponding function φ_β satisfies the complex Monge–Ampère equation (2.9). Next assume that (X, Δ) is a Fano orbifold. Then, for β sufficiently close to -1, the equation (2.9) has a unique solution. Indeed, since the Kähler–Einstein metric is unique, the orbifold X admits no non-trivial orbifold holomorphic vector fields, which, in turn, implies that the linearization of the equation (2.9) has a unique solution, defining a smooth function in the orbifold sense (see [36]). It then follows from a standard application of the implicit function theorem on orbifolds that the solution ϕ_β is uniquely determined for β sufficiently close to -1.

By the previous theorem (and its proof), it will be enough to show that (2)⇒(1). Since, trivially, (2)⇒(3), we have that $\Gamma_\beta = \delta_{\mu_\beta}$ and hence it will be enough to show the convergence in formula (2.19). To this end, first note that the functions $F_N(\beta)$ and $F(\beta)$ (defined in formula (2.14)) are concave in β, as follows readily from the definitions. Moreover, $F_N(\beta)$ and $F(\beta)$ are differentiable on $]-1-\varepsilon, 0[$ and

$$\frac{dF_N(\beta)}{d\beta} = \int_{X^N} E^{(N)} \mu^{(N)}_{\Delta,\beta}, \quad \frac{dF(\beta)}{d\beta} = E(\mu_\beta), \tag{2.20}$$

using that μ_β is the unique minimizer of F_β. Hence, if the convergence in item (2) of the theorem holds, then it follows from basic properties of concave functions that the derivative of $F_N(\beta)$ converges towards the derivative of $F(\beta)$ at $\beta = -1$ (see [19, Lemma 3.1]). Applying formula (2.20) thus concludes the proof of the convergence (2.19). ■

Remark 2.5. The reason that we have assumed that (X, Δ) is a Fano *orbifold* is that the proof involves the implicit function theorem in Banach spaces and thus relies on analytic properties of the linearized log Kähler–Einstein equation. We will come back to this point in Section 2.4.3.

2.4. The zero-free hypothesis

An alternative approach towards the case $\beta < 0$ was also introduced in [9, Section 7.1]. In a nutshell, it aims to "analytically continue" the convergence when $\beta > 0$ to $\beta < 0$. Here we formulate the approach in terms of the following *zero-free hypothesis* on the partition function $\mathcal{Z}_N(\beta)$ (defined in formula (2.3)):

$$\mathcal{Z}_N(\beta) \neq 0 \text{ on some } N\text{-independent neighborhood } \Omega \text{ of }]-1,0] \text{ in } \mathbb{C}. \tag{2.21}$$

We also need to assume that $\mathcal{Z}_N(\beta)$ is finite on a neighborhood of $[-1, 0]$ in $\mathbb{R}$ in a quantitative manner depending on N. This is made precise in the following result, which is a refinement of [9, Theorem 7.9]:

Theorem 2.6. *Let (X, Δ) be a Fano orbifold. Assume that there exists $\varepsilon > 0$ such that*

- $\mathcal{Z}_N(\beta) \leq C^N$ *for* $\beta = -(1+\varepsilon)$,
- *the zero-free hypothesis* (2.21) *holds.*

Then (X, Δ) admits a Kähler–Einstein metric ω_{KE} and δ_N converges in law towards the normalized volume form dV_{KE} of ω_{KE}. More precisely, the convergence (2.5) *of laws holds and $-\frac{1}{N}\log \mathcal{Z}_N(\beta)$ converges towards $F(\beta)$ in the C^∞-topology on a* *neighborhood of* $]-1,0]$. *Moreover, if* $[-1,0]\Subset\Omega$, *then the convergence holds on a neighborhood of* $[-1,0]$.

Proof. First, assume that (X,Δ) is a log Fano variety. Then the first point in the theorem implies that F admits a minimizer μ_β for any $\beta\in]-1-\varepsilon,0[$. Indeed, by the bound (2.17), $F(\beta)$ is bounded from below for any $\beta\in]-1-\varepsilon,0]$. Thus, for any $\beta\in]-1-\varepsilon,0[$, there exists $\delta>0$ such that $F_\beta\geq\delta E-\delta^{-1}$, which implies the existence of μ_β (as recalled in the proof of Theorem 2.4). In particular, taking $\beta=-1$ shows that X admits a unique Kähler–Einstein metric. Next, assume that X is a Fano orbifold. Then the argument using the implicit function, employed in the proof of Theorem 2.4, shows that after perhaps replacing ε with a small positive number there exists a unique solution φ_β to the equation (2.9), in the orbifold sense. In the case when X is a Fano manifold, it was shown in the proof of [9, Theorem 7.9] that $F(\beta)$ $(=F(\mu_\beta))$ defines a real-analytic function on $]-(1+\varepsilon),\infty[$. Since the proof only employs the implicit function theorem, it applies more generally when (X,Δ) is a Fano orbifold. Next, first consider the case when $\mathcal{Z}_N(\beta)$ is zero-free on an N-independent neighborhood Ω of $[-1,0]$ in $\mathbb{C}$. By Theorem 2.3, it will be enough to show that $\mathcal{Z}_N(\beta)^{1/N}\to e^{-F(\beta)}$ point-wise on $]-(1+\varepsilon),\varepsilon[$. To this end, first recall that, by Theorem 2.1, the convergence holds when $\beta\geq 0$. Next, by the zero-free hypothesis, $\mathcal{Z}_N(\beta)^{1/N}$ extends from $[-1,0]$ to a holomorphic function defined on a neighborhood Ω of $[-1,0]$ in $\mathbb{C}$. Moreover, by the first point,

$$\left|\mathcal{Z}_N(\beta)^{1/N}\right|\leq C\quad\text{on }\Omega\tag{2.22}$$

(using that $|\mathcal{Z}_N(\beta)^{1/N}|\leq\mathcal{Z}_N(\Re\beta)^{1/N}\leq\mathcal{Z}_N(-1-\varepsilon)^{1/N}$, which is uniformly bounded, by assumption). Hence, after perhaps passing to a subsequence, we may assume that $\mathcal{Z}_{N_j}(\beta)^{1/N_j}$ converges uniformly in the C^∞-topology on any compact subset of Ω to a holomorphic function $\mathcal{Z}(\beta)$, which, in particular, defines a real-analytic function on $]-1-\varepsilon,\varepsilon[$. But when $\beta\geq 0$, we have, as explained above, that $\mathcal{Z}(\beta)=e^{-F(\beta)}$ which extends to a real-analytic function on $]-1-\varepsilon,\varepsilon[$. By the identity principle for real-analytic functions, it thus follows that $\mathcal{Z}_{N_j}(\beta)^{1/N_j}\to e^{-F(\beta)}$ for any β in $]-1-\varepsilon,\varepsilon[$, in the C^∞-topology. Since the limit is uniquely determined, it thus follows that the whole sequence $\mathcal{Z}_N(\beta)^{1/N}$ converges towards $e^{-F(\beta)}$, as desired.

Finally, consider the case when it is only assumed that Ω is a neighborhood of $]-1,0]$ in $\mathbb{C}$. By assumption, the sequence of functions

$$F_N(\beta):=-\log\left(\mathcal{Z}_N(\beta)^{1/N}\right)$$

is uniformly bounded on $[-1-\varepsilon,\varepsilon]$. Since $F_N(\beta)$ is concave in β, it thus follows that $F_N(\beta)$ is uniformly Lipschitz continuous on $[-1,0]$. Hence, by the Arzela–Ascoli theorem, we may, after perhaps passing to a subsequence, assume that $F_N(\beta)$ converges uniformly to continuous function $F_\infty(\beta)$ on $[-1,0]$. By the previous argument, $F_\infty(\beta) = F(\beta)$ on $]-1,0]$. But since F_∞ and F are both continuous on $[-1,0]$, it follows that they also coincide at $\beta = -1$, as desired. ■

Remark 2.7. In statistical mechanical terms, the C^∞-convergence of $N^{-1}\log \mathcal{Z}_N(\beta)$ amounts to the absence of phase transitions [75, Chapter 5]. It seems natural to expect that the zero-free hypothesis (2.21) is satisfied as soon as X admits a Kähler–Einstein metric. Indeed, it can be viewed as a strengthening of the real-analyticity of free energy $F(\beta)$ in some neighborhood of $]0,1]$ in $\mathbb{C}$ (discussed in the proof of the previous theorem). The zero-free hypothesis for general statistical mechanical partition functions was introduced in the Lee–Yang theory of phase transitions (and has been verified for some spin systems and lattice gases [67, 81]). More precisely, originally Lee–Yang considered zeros in the complexified field parameter h called *Lee–Yang zeros*, while zeros with respect to the complexified inverse temperature β are called *Fisher zeros* [43]. The role of h in the present complex geometric setup is discussed in Remark 3.4.

As discussed in [8, Section 6], the bound in first point in the previous theorem – which is independent of the choice of metric $\|\cdot\|$ (up to changing the constant C) – can be viewed as an analytic (stronger) version of uniform Gibbs stability (cf. [8, Theorem 6.7]). As shown in [9, Lemma 7.1], the bound always holds for β sufficiently close to 0. More precisely,

$$\beta > -\operatorname{lct}(-K_X) \Rightarrow \mathcal{Z}_N(\beta) \le C_\beta^N \tag{2.23}$$

for any N $(= N_k)$, where $\operatorname{lct}(L)$ denotes the global lct of a line bundle L (whose definition is recalled in the appendix). The proof exploits that $\operatorname{lct}(-K_X)$ coincides with Tian's analytically defined α-invariant $\alpha(-K_X)$. Accordingly, under the weaker hypothesis that $\mathcal{Z}_N(\beta)$ is zero-free, for β in some ε-neighborhood of $]-\operatorname{lct}(X),0]$ in $\mathbb{C}$, the convergence statements in the theorem hold when $\beta \in\,]-\operatorname{lct}(X),0]$.

Remark 2.8. If $\operatorname{lct}(X) > 1$, the first assumption in Theorem 2.6 is automatically satisfied. Such Fano orbifolds are called *exceptional* (see [30], where two-dimensional exceptional hypersurfaces in three-dimensional weighted projective space are classified). Exceptional Fano orbifolds appear naturally in the MMP as the base of exceptional isolated affine singularities [76].

2.4.1. The strong zero-free hypothesis. The zero-free hypothesis is independent of the choice of basis in $H^0(X,-kK_X)$. Indeed, under a change of basis, $\det S^{(k)}$ gets multiplied by a non-zero scalar $c \in \mathbb{C}$ and hence $\mathcal{Z}_{N_k}(\beta)$ gets multiplied by $c^{\beta/k}$. However, it should be stressed that the zero-free hypothesis depends, a priori, on the choice of metric $\|\cdot\|$. For example, there are reasons to expect that it fails unless

$\|\cdot\|$ has positive curvature. Accordingly, the zero-free hypothesis might be more accessible for special/canonical choices of positively curved metrics, such as the Kähler–Einstein metric itself. This is illustrated by the following example, where $\mathcal{Z}_{N_k}(\beta)$ can be explicitly computed:

Example 2.9. When $X = \mathbb{P}^n_{\mathbb{C}}$ we have that $-K_X = \mathcal{O}(n+1)$ and hence the minimal value for k is $k = 1/(n+1)$, which means that the minimal value for N_k is $N_k = n+1$. Taking $\|\cdot\|$ to be the Fubini–Study metric (which is Kähler–Einstein) the following formula holds in the minimal case $N = n+1$ (where c_n is a computable positive constant), proved in the appendix (see Proposition A.3):

$$\mathcal{Z}_{n+1}(\beta) = c_n \frac{\prod_{j=1}^n \Gamma\big(\beta(n+1)+j\big)}{\big(\Gamma\big(\beta(n+1)+n+1\big)\big)^n}, \quad \Gamma(a) := \int_0^\infty t^a e^{-t}\frac{dt}{t} \tag{2.24}$$

where $\Gamma(a)$ denotes the classical Γ-function, which defines a meromorphic function on $\mathbb{C}$ whose poles are located at $0, -1, -2, \ldots$ (as follows from the functional relation $\Gamma(a+1) = a\Gamma(a)$). Thus the first negative pole of $\mathcal{Z}_N(\beta)$ comes from the first pole of the factor corresponding to $j = 1$ in the nominator above, i.e., when $\beta = -1(n+1)$. Moreover, since $\Gamma(a)$ is zero-free on all of $\mathbb{C}$, $\mathcal{Z}_N(\beta)$ is zero-free in the maximal strip $\{\Re\beta > -1/(n+1)\}$ of holomorphicity (but the meromorphic continuation $\mathcal{Z}_N(\beta)$ does have zeros in $\mathbb{C}$, coming from the poles of the denominator).

In the light of this example, it is tempting to speculate that the following *strong zero-free hypothesis* holds for Kähler–Einstein metrics:

$$\mathcal{Z}(\beta) \neq 0, \quad \text{when } \Re\beta > \max\big\{-\operatorname{lct}(\mathcal{D}_N), -1\big\}.$$

In other words, this means that $\mathcal{Z}_N(\beta)$ is zero-free in the maximal strip inside $\{\Re\beta > -1\}$, where it is holomorphic. To provide some further evidence for the strong zero-free property, we note that if its holds, then the bound (2.23), combined with the proof of Theorem 2.6, shows that, for any given $\varepsilon > 0$, the function $F(\beta)$ on $]-\operatorname{lct}(-K_X)+\varepsilon, \varepsilon[\subset \mathbb{R}$, induced by the Kähler–Einstein metric, is “strongly real-analytic” in the following sense: $F(\beta)$ extends to a bounded holomorphic function on the infinity strip $]-\operatorname{lct}(-K_X)+\varepsilon, \varepsilon[+i\mathbb{R} \subset \mathbb{C}$. This condition is much stronger than ordinary real-analyticity (which only implies holomorphic extension to a finite strip). But it does hold for the Kähler–Einstein metric. Indeed, in this case,

$$F(\beta) \equiv 0, \quad \beta \in]-1, \infty[,$$

which trivially extends to a bounded holomorphic function on the infinity strip. To prove the identity above, first observe that when $\omega_0 = \omega_{KE}$, the twisted Kähler–Einstein equation (2.10) is solved by $\omega_\beta = \omega_{KE}$ for *any* β (equivalently, in the case when $\omega_0 = \omega_{KE}$, we have $\omega_0^n/V = dV_{(X,\Delta)}$ and hence the complex Monge–Ampère equation (2.9) is solved by $\varphi_\beta = 0$). But, as recalled above, for $\beta > -1$, the equation (2.9) admits a unique solution and hence

$$F(\beta) = F_\beta(dV_{KE}) = 0$$

(using the vanishing (2.8) combined with the vanishing $\mathrm{Ent}(\mu) = 0$ when $\mu = dV_{KE} = dV_\Delta$). In fact, this argument shows that $F(\beta) \equiv 0$ on all of $[-1, \infty[$. Moreover, if $\mathrm{Aut}(X)_0$ is trivial, then there exists an $\varepsilon > 0$ such that $F(\beta) \equiv 0$ on all of $]-1-\varepsilon, \infty[$, as follows from the argument using the implicit function theorem, employed in the proof of Theorem 2.4. This argument suggests that when $\mathrm{Aut}(X)_0$ is trivial, one can, perhaps, expect the strong zero-free property to even hold in the larger region where $\Re\beta > \max\{-\mathrm{lct}(\mathcal{D}_N), -1-\varepsilon\}$ for some $\varepsilon > 0$.

Remark 2.10. Coming back to Example 2.9, it is natural to ask if there exists an explicit formula for $\mathcal{Z}_N(\beta)$ when $X = \mathbb{P}^n_{\mathbb{C}}$ for general N, generalizing formula (2.24) (or, more precisely, for any N of the form $N = N_k$). However, as discussed in Remark A.4, this problem appears to be open even when $n = 1$. But one interesting consequence of formula (2.24) is that it reveals that in the case when $X = \mathbb{P}^n_{\mathbb{C}}$ and N is minimal,

$$\mathrm{lct}(\mathcal{D}_N) = \mathrm{lct}(-K_X)$$

since $\mathrm{lct}(-K_X) = 1/(n+1)$. This shows that the estimate in formula (2.23) is sharp (in the sense that there are cases where it fails for $\beta \leq -\mathrm{lct}(-K_X)$). The point of Conjecture 1.2, however, is that it only requires that $\mathrm{lct}(\mathcal{D}_{N_k}) > 1$ when N_k is sufficiently large. Similarly, in the case of $\mathbb{P}^n_{\mathbb{C}}$, where $\mathrm{Aut}(X)_0 \neq \{I\}$, the corresponding conjecture only requires that $\mathrm{lct}(\mathcal{D}_{N_k}) \to 1$, when $N_k \to \infty$ (see [9, Conjecture 3.8]). For example, when $X = \mathbb{P}^1_{\mathbb{C}}$, one has $\mathrm{lct}(\mathcal{D}_N) = (N-1)/N$ (by Theorem 4.5) which indeed tends to 1 as $N \to \infty$ (and equals $1/2$ when $N = 2$, which is the minimal case).

2.4.2. Allowing singular metrics $\|\cdot\|$. Alternatively, when X is a Fano manifold, one can take $\|\cdot\|$ to be the singular metric induced by the anti-canonical $\mathbb{Q}$-divisor Δ_m defined by the zero-locus of a holomorphic section of $-mK_X$, assuming that $m > 0$ and the zero-locus is non-singular (which ensures that the corresponding singular volume form dV has a density in L^p_{loc} for some $p > 1$). In other words, the curvature of $\|\cdot\|$ is given by the positive current $[\Delta_m]$ supported on Δ_m. Then Theorem 2.6 still applies. Indeed, in the proof one can apply the implicit function to the wedge-Hölder spaces appearing in [38, 55], which are independent of β (see, in particular, [55, Corollary 3.5]). In this singular setup, the corresponding equations (2.10) become Donaldson's variant of Aubin's continuity equations

$$\mathrm{Ric}\,\omega_\beta = t\omega_\beta + (1-t)[\Delta_m], \quad t = -\beta, \tag{2.25}$$

that were used in the proof of the YTD conjecture in [31–33], by deforming t from an initial small value, where there always exists a solution (by [4, Theorem 1.5]) to $t = 1$, assuming that X is K-stable. In other words, β is deformed down to -1. In the present probabilistic approach, the (potential) advantage of employing the singular metric on $-K_X$ induced by the $\mathbb{Q}$-divisor Δ_m is that the corresponding partition function $\mathcal{Z}_N(\beta)$ is encoded by purely algebraic data: the divisors $\mathcal{D}_N$ and Δ_m on X^N and X, respectively. In this case, combining [4, Proposition 6.2] with [9, Lemma 7.1] gives

$$\beta > -\min\left\{\operatorname{lct}(-K_X), \operatorname{lct}(-K_{X|\Delta_m})\right\} \Rightarrow \mathcal{Z}_N(\beta) \leq C_\beta^N,$$

where $-K_{X|\Delta_m}$ denotes the restriction of $-K_X$ to the support of Δ_m. More generally, it seems natural to expect that Theorem 2.6 holds for *any* log Fano variety (X, Δ) (when $\|\cdot\|$ is either a smooth metric on $K_X + \Delta$ with positive curvature or the singular metric defined by *any* klt $\mathbb{Q}$-divisor in $-(K_X + \Delta)$). In the case when $\Delta + \Delta_m$ defines a divisor whose components are non-singular and mutually non-intersecting, the aforementioned results in [38, 55] still apply.

2.4.3. Deforming the divisor Δ. Sometimes, it is advantageous to keep $\beta = -1$ and instead deform the divisor Δ as follows. Given a log Fano variety (X, Δ) and a positive real number k such that $-k(K_X + \Delta)$ is a well-defined line bundle $\mathcal{L}$, i.e., defines an element in the integral lattice $H^2(X, \mathbb{Z})$ of $H^2(X, \mathbb{R})$, consider the affine subspace $\mathcal{A}$ of $\mathbb{R}^{M+1}$ of all $(\boldsymbol{w}, s)$ which are "admissible" in the sense that

$$-\big(K_X + \Delta(\boldsymbol{w})\big) = s\mathcal{L}, \tag{2.26}$$

where $\Delta(\boldsymbol{w})$ denotes the divisor with the same M irreducible components as the given divisor Δ and coefficients $\boldsymbol{w} \in \mathbb{R}^M$. In particular, $(\boldsymbol{w}_0, k^{-1})$ is "admissible", where $\boldsymbol{w}_0 \in \mathbb{R}^M$ denotes the coefficients of the initial divisor Δ. If there exists $(\boldsymbol{w}_1, s_1) \in \mathcal{A}$ such that $K_X + \Delta(\boldsymbol{w}_1) > 0$ (and hence $s_1 < 0$), the conclusion of Theorem 2.6 still applies if the corresponding partition function $\mathcal{Z}_N$, viewed as a meromorphic function on $\mathbb{C}^{M+1}$, satisfies

- $\mathcal{Z}_N \leq C_0^N$ in a neighborhood in $\mathbb{R}^{M+1}$ of $(\boldsymbol{w}_0, k^{-1})$,
- $\mathcal{Z}_N \neq 0$ in an N-independent neighborhood of the line-segment in $\mathbb{C}^{M+1}$ connecting $(\boldsymbol{w}_0, k^{-1})$ and $(\boldsymbol{w}_1, s_1)$.

More precisely, as discussed in the previous section, in order to apply the implicit function theorem in Banach spaces, the appropriate linear PDE-theory needs to be in place. For example, by [38, 55], this is the case when the components of Δ are non-singular and mutually non-intersecting (results concerning the case when (X, Δ) is log smooth are announced in [73]). The previous proof can then by applied to the meromorphic function $\mathcal{Z}_N(t)$ on $\mathbb{C}$ defined by the partition functions associated to the line-segment $I \Subset \mathbb{C}^{m+1}$ connecting the initial $(\boldsymbol{w}_0, k^{-1})$ with $(\boldsymbol{w}_1, s_1)$ (where t

denotes the complexification of the standard parametrization of I). In this situation, the estimate (2.22) still holds, i.e., $|\mathcal{Z}_N(t)^{1/N}| \leq C$ on some N-independent neighborhood Ω of $[0,1]$ in $\mathbb{C}$. Indeed, by assumption, the estimate holds with constant C_0 in a neighborhood of $t = 0$ and, moreover, it trivially holds with a constant C_1 when t is close to $t = 1$. Since $\log \mathcal{Z}_N(t)$ is convex with respect to $t \in [0,1]$, one can thus take $C = \max\{C_0, C_1\}$.

3. Intermezzo: A zero-free hypothesis for polarized manifolds (X, L) and the Calabi–Yau equation

Before turning to the case of log Fano curves, we make a digression on general polarized manifolds (X, L), i.e., a compact complex manifold X endowed with an ample line bundle L. To a metric $\|\cdot\|$ on L and a volume form dV on X, we may attach partition functions $\mathcal{Z}_N(\beta)$, by replacing the log canonical line bundle $-(K_X + \Delta)$ with L and dV_Δ with dV in formula (2.3):

$$\mathcal{Z}_N(\beta) := \int_{X^N} \|\det S^{(k)}\|^{2\beta/k} dV^{\otimes N}, \tag{3.1}$$

where k is a given positive integer and N denotes the dimension of $H^0(X, kL)$. This is the general setup considered in [7], where the corresponding free energy functional is of the form

$$F_\beta(\mu) := \beta E(\mu) + \mathrm{End}(\mu),$$

where $E(\mu)$ denotes the pluricomplex energy of μ with respect to the normalized curvature form ω of the metric $\|\cdot\|$ on L and $\mathrm{Ent}(\mu)$ denotes the entropy of μ relative to dV. The minimizers μ_β of $F_\beta(\mu)$ are of the form

$$\mu_\beta = e^{\beta\varphi_\beta} dV$$

for a smooth solution φ_β of the complex Monge–Ampère equation

$$\frac{1}{V}\left(\omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi_\beta\right)^n = e^{\beta\varphi_\beta} dV. \tag{3.2}$$

Remark 3.1. In the case when $\beta = k$ and X is a Riemann surface, the corresponding partition function $\mathcal{Z}_N(\beta)$ coincides with the L^2-norm of the Laughlin wave function for the (integer) Quantum Hall state on X, subject to the magnetic two-form $ik\omega$ [58]. Accordingly, as shown in [5], in this case (and for any dimension of X) the corresponding large N-limit is described by the minimizers $F_\beta(\mu)/\beta$, as $\beta \to \infty$, i.e., of $E(\mu)$. However, here we are concerned with the case when β is fixed, where entropy enters the picture and dominates when β is close to 0.

Consider, in this general setup, the following *weak zero-free hypothesis*:

$$\mathcal{Z}_N(\beta) \neq 0 \text{ on some } N\text{-independent neighborhood } \Omega \text{ of } 0 \text{ in } \mathbb{C}. \tag{3.3}$$

It implies a weaker form of the upper bound hypothesis (2.13) on the mean energy:

Theorem 3.2. *Let (X, L) be a polarized manifold. Given a metric $\|\cdot\|$ on L and a volume form dV on X, assume that the corresponding partition functions $\mathcal{Z}_N(\beta)$ satisfy the weak zero-free hypothesis above. Then $-\frac{1}{N}\log \mathcal{Z}_N(\beta)$ converges towards $F(\beta)$ in the C^∞-topology on a neighborhood of 0 in $\mathbb{R}$. In particular, the mean energy of $dV^{\otimes N}$ converges towards the pluricomplex energy $E(dV)$ of dV:*

$$\begin{aligned} &\lim_{N\to\infty}\int_{X^N} E^{(N)} dV^{\otimes N} = E(dV), \\ &E^{(N)}(x_1,\dots,x_N) := -\frac{1}{kN}\log\left(\|\det S^{(k)}(x_1,\dots,x_N)\|^2\right). \end{aligned} \tag{3.4}$$

Proof. In general, given a metric $\|\cdot\|$ on L and a volume form dV on X, there exists $\varepsilon > 0$ such that $F(\beta)$ is real-analytic on $]-\varepsilon,\varepsilon[$. Indeed, this follows, as before, from an application of the implicit function theorem at $\beta = 0$. Moreover, by the argument discussed in connection to formula (2.23),

$$\beta > -\operatorname{lct}(L) \Rightarrow \mathcal{Z}_N(\beta) \leq C_\beta^N. \tag{3.5}$$

In particular, the estimate holds when $\beta > -\varepsilon$ for ε sufficiently small. The C^∞-convergence of $-\frac{1}{N}\log \mathcal{Z}_N(\beta)$ towards $F(\beta)$ then follows exactly as in the proof of Theorem 2.6. Finally, the convergence of the first derivatives at $\beta = 0$ yields the convergence (3.4). ■

We next show that a variant of the weak zero-free hypothesis yields canonical approximations φ_N of the solution of the *Calabi–Yau equation*, i.e., the equation obtained by setting $\beta = 0$ in equation (3.2):

$$\frac{1}{V}\left(\omega + \frac{i}{2\pi}\partial\bar{\partial}\varphi\right)^n = dV \tag{3.6}$$

for a smooth function φ on X. By Yau's theorem [82], there exists a unique smooth solution φ with vanishing average on (X, dV). Given a volume form dV with unit total volume, the canonical approximation φ_N in question is defined by the integral formula

$$\varphi_N(x) := \int \frac{1}{k}\log\left(\|\det S^{(k)}(x, x_2,\dots,x_N)\|^2\right) dV^{\otimes N-1} - c_N, \tag{3.7}$$

where c_N is the constant ensuring that the average of φ_N on (X, dV) vanishes:

$$c_N := \int_{X^N} \frac{1}{k}\log\left(\|\det S^{(k)}(x, x_2,\dots,x_N)\|^2\right) dV^{\otimes N}.$$

For a given smooth function u on X, denote by $\mathcal{Z}_N(\beta,h)$ the function on $\mathbb{R}^2$ obtained by replacing dV in formula (3.1) with $e^{hu}dV$:

$$\mathcal{Z}_N(\beta,h) := \int_{X^N} \| \det S^{(k)}\|^{2\beta/k}(e^{hu}dV)^{\otimes N}. \tag{3.8}$$

Theorem 3.3. *Let (X,L) be a polarized manifold and $\|\cdot\|$ a metric on L. Given a volume form dV on X with unit total volume, assume that*

$$\mathcal{Z}_N(\beta,h) \neq 0 \text{ on some } N\text{-independent neighborhood } \Omega \text{ of } (0,0) \text{ in } \mathbb{C}^2, \tag{3.9}$$

for any smooth function u on X (where Ω depends on u). Then the functions φ_N, defined by formula (3.7), converge in $L^1(X)$, as $N\to\infty$, to the unique smooth solution φ of the Calabi–Yau equation (3.6) satisfying $\int_X \varphi dV = 0$.

Proof. First, observe that $\varphi_N(x)$ is ω-psh, since it is a superposition of the ω-psh functions $\log(\|\det S^{(k)}(x,x_2,\dots,x_N)\|^2)$. Hence, by standard properties of ω-psh functions, the L^1-convergence in question is equivalent to weak convergence. In other words, it is equivalent to proving that for any given smooth function $u\in C^\infty(X)$

$$\lim_{N\to\infty}\int \varphi_N u\, dV = \int \varphi\, dV.$$

Moreover, since the integrals on both sides of the previous equality vanish for $u=1$, it is enough to prove the convergence for any $u\in C^\infty(X)$ satisfying $\int u dV = 0$. To this end, fix such a function u and consider the corresponding partition functions $\mathcal{Z}_N(\beta,h)$, defined by formula (3.8). A direct calculation reveals that

$$\int \varphi_N u\, dV = \frac{\partial}{\partial h}\frac{\partial}{\partial \beta} N^{-1}\log \mathcal{Z}_N(\beta,h), \quad \text{at } (\beta,h)=(0,0). \tag{3.10}$$

By assumption, there exists a neighborhood Ω of $(0,0)$ in $\mathbb{C}^2$, where $\log\mathcal{Z}_N(\beta,h)$ is holomorphic. Moreover, by Theorem 3.2,

$$-N^{-1}\log\mathcal{Z}_N(\beta,h) \to F(\beta,h) := \inf_{\in\mathcal{P}(X)}\left(\beta E(\mu) - h\int_X u\, dV + \mathrm{Ent}(\mu)\right)$$

in the C^∞_{loc}-topology on Ω, where $\mathrm{Ent}(\mu)$ denotes the entropy of μ relative to dV. In particular, the convergence of the second derivatives at $(0,0)$ yields, by formula (3.10),

$$\lim_{N\to\infty}\int \varphi_N u\, dV = -\frac{\partial}{\partial h}\frac{\partial F(\beta,h)}{\partial\beta} \quad \text{at } (\beta,h)=(0,0).$$

Since dV is the unique minimizer of F_β when $\beta=0$,

$$\frac{\partial F(\beta,h)}{\partial\beta} = E(dV_h), \quad dV_h := dVe^{hu}\Big/\int_X dVe^{hu}.$$

The proof is thus concluded by invoking the property (2.7) of the functional E, which gives

$$-\frac{E(dV_h)}{\partial h}_{|u=0} = \int_X \varphi u \, dV. \qquad \blacksquare$$

In the particular case when X is a Calabi–Yau manifold – i.e., when some power of K_X is trivial – we can apply the previous theorem to the canonical normalized volume form dV on X,

$$dV := \frac{(s_m \wedge \bar{s}_m)^{1/m}}{\int_X (s_m \wedge \bar{s}_m)^{1/m}},$$

where s_m trivializes mK_X for some positive integer m. Then the corresponding convergence implies that the positive $(1,1)$-currents

$$\omega_N := \frac{i}{2\pi k} \int \partial\bar{\partial} \log \left(\left| \det S^{(k)}(\cdot, x_2, \dots, x_N) \right|^2 \right) dV^{\otimes N-1}$$

converge weakly towards the unique Calabi–Yau metric ω_{CY} on X in $c_1(L)$, i.e., towards the unique Ricci flat Kähler metric in $c_1(L)$. Note that, by the Poincaré–Lelong formula, ω_N is the average over X^{N-1} of the currents of integration defined by the zero-loci in X of the holomorphic sections $\det S^{(k)}(\cdot, x_2, \dots, x_N)$.

Remark 3.4. It seems natural to expect that the zero-free hypothesis (3.9) is always satisfied. Indeed, it can be viewed as a strengthening of the real-analyticity of the free energy $F(\beta, h)$ in some neighborhood of $(0,0)$ in $\mathbb{C}^2$ (discussed in the proof of the previous theorem). This expectation is in line with corresponding expectations in the Lee–Yang theory of phase transitions [67, 81], where the role of β and h/β is played by the inverse temperature and the field strength, respectively (see the discussion in the introduction of [64]).

When X is a compact complex curve, i.e., $n = 1$, the convergence in Theorem 3.2 and Theorem 3.3 can, unconditionally, be deduced from the bosonization formula for $\det S^{(k)}(x_1, \dots, x_N)$ [1]. To the leading order, this formula expresses $\| \det S^{(k)}(x, x_2, \dots, x_N) \|$ as a product of $G(x_i, x_j)$, where G is Green's function for the Laplacian $i\partial\bar{\partial}$ (see Lemma 4.3 for the case when $X = \mathbb{P}^1_{\mathbb{C}}$).

4. The case of log Fano curves

Let X be the complex projective line $\mathbb{P}^1_{\mathbb{C}}$. Fix an $\mathbb{R}$-divisor Δ on X, i.e.,

$$\Delta := \sum_{1=1}^{m} p_i w_i$$

for given points $p_1, \dots, p_m$ on X and with real coefficients/weights w_i and assume that

$$w_i < 1.$$

In contrast to Section 1.4, we thus allow w_i to be negative. Assume that (X, Δ) is a log Fano manifold, i.e., the anti-canonical line bundle of (X, Δ) is positive:

$$L := -(K_X + \Delta) > 0.$$

Since X is a complex curve, the assumption that L is positive simply means that its degree d_L is positive:

$$d_L = 2 - \sum w_i > 0. \tag{4.1}$$

Given a positive real number k and assuming that kL defines a line bundle, i.e., kd_L is an integer set,

$$N_k := \dim H^0(X, kL).$$

To the log Fano curve (X, Δ) we attach (as in the beginning of Section 2) the following symmetric probability measure on X^{N_k}:

$$\mu_\Delta^{(N_k)} = \frac{1}{\mathcal{Z}_{N_k}} \left| \det S^{(k)}(z_1, \dots, z_N) \right|^{-2/k} |s_\Delta|^{-2}(z_1) \cdots |s_\Delta|^{-2}(z_{N_k}),$$

which is well defined precisely when $\mathcal{Z}_{N_k} < \infty$. The following result implies Theorem 1.5 (concerning the case when $w_i > 0$):

Theorem 4.1. *Let (X, Δ) be a log Fano curve. Then the following is equivalent:*

- *$\mathcal{Z}_{N_k} < \infty$ for k sufficiently large;*
- *the following weight condition holds:*

$$w_i < \sum_{i \neq j} w_j, \quad \forall i. \tag{4.2}$$

Moreover, if any of the conditions above hold, then the law of the empirical measure δ_N on $(X^{N_k}, \mu_\Delta^{(N_k)})$ satisfies an LDP with speed N and rate functional $F_{-1} - \inf_{\mathcal{P}(X)} F_{-1}$ (where F_{-1} is the free energy functional on $\mathcal{P}(X)$ defined in Section 2.3, which coincides with the Mabuchi functional for (X, Δ)).

Remark 4.2. In particular, if the weight condition above holds, then F_{-1} is lsc on $\mathcal{P}(X)$ (since, in general, any rate functional for an LDP is lsc) and thus admits a minimizer. The existence of a minimizer was first shown in [79] using a different variational argument. By the general results for log Fano varieties (X, Δ) in [16], any minimizer satisfies the Kähler–Einstein equation for (X, Δ). In general, a solution is not uniquely determined (see [71, Remark 2]). However, when $w_i > 0$, the uniqueness in the case of the Riemann sphere was shown in [71] (see [16, 31–33] for the general higher dimensional log Fano case).

To prove the previous theorem, we first recall some standard identifications (see [11, Section 3.7]). Fixing a point p_∞, we identify $X - \{p_\infty\}$ with $\mathbb{C}$. The point p_∞ induces a trivialization e_∞ of the restriction of the hyperplane line bundle $\mathcal{O}(1) \to \mathbb{P}^1_{\mathbb{C}}$ to $\mathbb{C}$ (vanishing at p_∞) and thus the space $H^0(X, d\mathcal{O}(1))$ of all global holomorphic sections of the dth tensor power of the hyperplane line bundle $\mathcal{O}(1) \to X$ may be identified with the space of all polynomials in z of degree at most d. Moreover, the anti-canonical line bundle $-K_X$ of X may be identified with $2\mathcal{O}(1)$ and s_Δ with a (multivalued) holomorphic section of $\sum w_i \mathcal{O}(1)$. In particular, we identify

$$kL \leftrightarrow kd_L\mathcal{O}(1) = k\left(2 - \sum_{i=1}^{m} w_i\right)\mathcal{O}(1),$$

(recall that we are assuming that kd_L is an integer). Thus $H^0(X, kL)$ gets identified with the space of all polynomials in z of degree at most $k(2 - \sum_{i=1}^{m} w_i)$. This identification reveals that

$$N_k = kd_L + 1. \tag{4.3}$$

Fix the standard basis of monomials $1, z, z^2, \ldots$ in $H^0(X, kL)$. Then the corresponding section $\det S^{(k)}$ over X^{N_k} gets identified with the usual Vandermonde determinant on $\mathbb{C}^{N_k}$:

$$\det S^{(k)} \leftrightarrow D(z_1, \ldots, z_{N_k}) := \det_{i,j \le N_k}(z_i^j). \tag{4.4}$$

Next, we identify X with the unit-sphere S^2 in $\mathbb{R}^3$, using the standard stereographic projection, so that the fixed point $p_\infty \in X$ corresponds to the "north-pole" $(0, 0, 1)$ in S^2:

$$z \mapsto x := \left(\frac{z + \bar{z}}{1 + |z|^2}, \frac{z - \bar{z}}{1 + |z|^2}, \frac{-1 + |z|^2}{1 + |z|^2}\right), \quad \mathbb{C} \to \mathbb{R}^3.$$

Denote by dV_X the area form of the standard round metric on S^2 and by G the following lsc function on X:

$$G(x, y) := -\log \|x - y\|,$$

expressed in terms of the Euclidean norm on $\mathbb{R}^3$.

Lemma 4.3. *In terms of the standard identifications over $\mathbb{C}$,*

$$\left|\det S^{(k)}(z_1, \ldots, z_N)\right|^{-2/k} |s_\Delta|^{-2}(z_1) \cdots |s_\Delta|^{-2}(z_{N_k})$$
$$= \frac{1}{\left(\prod_{i \neq j} |z_i - z_j|\right)^{\frac{d_L}{N-1}}} \frac{1}{\prod_i} \frac{1}{|z_i - p_j|^{2w_j}}$$

(where d_L is defined in formula (4.1)). As a consequence, on $X := \mathbb{P}^1_{\mathbb{C}}$ the probability

measure $\mu_\Delta^{(N)}$ *may be expressed as*

$$\mu_\Delta^{(N)} = \frac{1}{Z_N} e^{\frac{d_L}{N-1}\sum_{i\neq j\leq N} G(x_i,x_j)} dV^{\otimes N}, \quad dV := e^{\sum_{i\leq m} w_i G(x,p_i)} dV_X. \tag{4.5}$$

Proof. First, factorizing the Vandermonde determinant $D(z_1,\dots,z_{N_k})$ on $\mathbb{C}^N$ reveals that $D(z_1,\dots,z_{N_k})$ is the product of $(z_i - z_j)$ over all i, j in $\{1,\dots,N\}$ such that $i < j$. Hence,

$$\left|D(z_1,\dots,z_{N_k})\right|^2 = \prod_{i\neq j} |z_i - z_j|. \tag{4.6}$$

Since $N_k = kd_L + 1$, we have that $k = (N-1)/d_L$ and hence the first formula of the lemma follows. To prove the second one, first recall that in the general setting of log Fano manifolds (X, Δ), the measure $\mu_\Delta^{(N)}$ may be expressed as in formula (2.1). In the present case, we take $\|\cdot\|$ to be the metric on L induced from the Fubini–Study metric $\|\cdot\|_{FS}$ on $\mathcal{O}(1)$ under the identification of L with $d_L\mathcal{O}(1)$. Recall that

$$\|e_\infty\|_{FS}^2 = e^{-\phi_{FS}(z)}, \quad \phi_{FS}(z) := \log\left(1 + |z|^2\right).$$

Hence, formula (4.5) follows from the following two facts: first,

$$\|z - w\|_{FS}^2 := |z - w|^2 e^{-\phi_{FS}(z)} e^{-\phi_{FS}(w)} \tag{4.7}$$

is proportional to the squared norm in $\mathbb{R}^3$ under stereographic projection and second,

$$\|dz\|_{FS}^2 := \|e_\infty^{\otimes 2}\|_{FS}^2 := e^{-2\phi_{FS}}$$

is proportional to the density of dV_X. These are well-known relations that can be checked explicitly, but they also follow readily from their invariance under the isometry group of S^2. ∎

Next, we recall the following general LDP [10, Theorem 1.5], generalizing the convergence in probability established in [27, 57] for the point-vortex model in a planar compact domain. Given a symmetric function W on a compact metric space X, a measure μ_0 on X, and $p \in \mathbb{R}$, set

$$\mu^{(N)}[p] = \frac{1}{Z_{N[p]}} e^{-p\frac{1}{N}\sum_{x_i\neq x_j} W(x_i,x_j)} \mu_0^{\otimes N},$$

$$Z_N[p] := \int_{X^N} e^{-p\frac{1}{N}\sum_{x_i\neq x_j} W(x_i,x_j)} \mu_0^{\otimes N},$$

assuming that $Z_N[p] < \infty$.

Theorem 4.4. *Let X be a compact metric space, μ_0 a measure on X, W a lower semi-continuous symmetric measurable function on X^2, and p_0 a negative number such that*

$$\sup_{x\in X}\int_X e^{-p_0 W(x,y)}\mu_0(y) < \infty. \tag{4.8}$$

Then, for any $p > p_0$, the normalizing constant $Z_N[p]$ is finite and the law of the empirical measure δ_N on $(X^N, \mu^{(N)}[p])$ satisfies an LDP with a rate functional

$$F_p - \inf_{\mathcal{P}(X)} F_p, \quad F_p(\mu) := p\int_{X\times X} W\mu\otimes\mu + \mathrm{Ent}_{\mu_0}(\mu).$$

Proof. It may be illuminating to reformulate the proof given in [10] in terms of the conditional convergence result in Theorem 2.3. First, the finiteness of $Z_N[p]$ follows readily from the arithmetic-geometric means inequality, using the integrability condition (4.8). A refinement of this argument also yields a priori estimates on each j-point correlation measure on X^j, building on [9, Section 3.2.4], showing that its density is uniformly bounded in $L^p(\mu_0^{\otimes j})$ for any $p > 1$. Applying this estimate to $j \le 2$ shows that the "upper bound hypothesis" (2.13) of the energy is satisfied. A twist of this argument also yields the stronger form of the upper bound hypothesis with respect to any given continuous function $\Phi(\mu)$, as formulated in Theorem 2.3, and thus also the LDP. ∎

In the present case, we thus have

$$W(z,w) = -d_L \log\|z-w\|, \quad p = \beta\frac{N-1}{N}.$$

Moreover,

$$\int_X W\mu\otimes\mu = E(\mu) + C \tag{4.9}$$

for some constant C. Indeed, by a simple scaling argument, it is enough to consider the case when $d_L = 1$. Then we can write $W(x,y) = G(x,y)/2$, where $G(x,y) = -\log(\|z-w\|^2)$ has the property that $-\frac{i}{2\pi}\partial\bar\partial G(x,\cdot) = \delta_x - \omega_0$, where ω_0 is the normalized curvature of the Fubini–Study metric. Hence, the first variation of the functional $\mu \mapsto \int_X W\mu\otimes\mu$ on $\mathcal{P}(X)$ coincides with the first variation of $E(\mu)$ (formula (2.7)), which proves formula (4.9).

4.1. Conclusion of the proof of Theorem 4.1

Set $p = -t$ and observe that

$$\int_X e^{-pW(x,y)}\mu_0(y) = \int_X e^{-(td_L\log\|x-y\|+\sum_i w_i\log\|x-p_i\|^2)}\,dV_X.$$

For any given $y \in X$, the function $e^{-c \log \|x-y\|^2}$ is locally integrable on X iff $c < 1$. Hence, the right-hand side above is integrable iff for any fixed index i

$$td_L/2 + w_i < 1, \quad \forall i.$$

But this condition holds for some $t > 1$ iff

$$d_L/2 + w_i < 1, \quad \forall i,$$

i.e., iff $1 - \sum w_j/2 + w_i < 1$ for all i, that is, $w_i < \sum_{j \neq i} w_j$, which is equivalent to the weight condition (4.2). Hence, if the weight condition holds, then by Theorem 4.4, the desired LDP follows.

Next, assume that the weight condition is violated. Without loss of generality we may assume that it is violated for the index $i = 1$, which equivalently means that

$$-d_L + 2(1 - w_1) = 0.$$

Set $B_R := \{\|x - p_1\| \leq R\}$. Since $e^{-\log \|x-y\|} \geq R^{-1}$ on B_R, we have

$$\int_{B_R^N} e^{W(x,y)} \mu_0(y) \geq (R^{-1})^{d_L N} \int_{B_R^N} \mu_0^{\otimes R}.$$

Using $\int_{|z| \leq R} e^{-w \log |z|^2} d(r^2) \wedge d\theta = \frac{1}{1-w}(R^2)^{1-w}$, we thus get

$$\int_{B_R} \mu_0 \geq \int e^{-(w_1 \log \|x-p_1\|^2)} \, dV_X \geq C(R^2)^{(1-w_1)}$$

for some constant independent of R. All in all, this means that

$$\left(\int_{B_R^N} e^{W(x,y)} \mu_0(y) \right)^{1/N} \geq CR^{-d_L + 2(1-w_1)} \geq CR^0 \geq C > 0.$$

But the right-hand side is independent of R. Hence, letting $R \to 0$ shows that the density $e^{W(x,y)}$ cannot be in $L^1(X^N \mu_0^{\otimes N})$, which means that $Z_{N,-1} = \infty$, as desired.

4.2. The case of a general divisor Δ

Now consider the case of general coefficients $w_i \in]-\infty, 1[$. By the previous theorem, $Z_{N,-1}$ diverges for large N, unless the weight condition (4.2) holds. But fixing any continuous metric $\| \cdot \|$ on L, we can consider the corresponding probability measures $\mu_{\Delta,\beta}^{(N)}$, defined by formula (2.2), which are well defined when $-\beta$ is sufficiently small.

Theorem 4.5. *$\mathcal{Z}_N(\beta) < \infty$ iff $\beta > -\gamma_N$, where*

$$\gamma_N = \frac{N-1}{N} 2 \frac{1 - \max_i w_i}{2 - \sum_i w_i}.$$

Moreover, if $\mathcal{Z}_N(\beta) < \infty$, then the law of the random variable δ_N on $(X^N, \mu^{(N)}_{\Delta,\beta})$ satisfies an LDP with speed N and rate functional $F_\beta - \inf_{\mathcal{P}(X)} F_\beta$.

Proof. First, consider the case when $\|\cdot\|$ is the metric $\|\cdot\|_{FS}$ induced from the Fubini–Study metric on $\mathcal{O}(1)$. Then we get, as above, that $\mu_\beta^{(N)} = \mu^{(N)}[p]$ for $p = \beta\frac{N-1}{N}$. Hence, by the argument in the beginning of the previous section, the integrability threshold is given by

$$\gamma_N = \frac{N-1}{N}\gamma, \quad \gamma = \sup\{t : td_L/2 + w_i < 1, \ \forall i\} = 2\frac{1-\max_i w_i}{2-\sum_i w_i},$$

and the LDP follows from the general LDP in Theorem 4.4. Finally, writing a general continuous metric $\|\cdot\|$ as $e^{-u/2}\|\cdot\|_{FS}$ for a continuous function u on X, we can express $\mu_\beta^{(N)} = \mu^{(N)}[p]$, where $\mu_0 = e^{-(\beta+1)u}dV$, and again apply Theorem 4.4. ■

As recalled in Section 2.3, any minimizer ω_β of F_β satisfies the twisted Kähler–Einstein equation (2.10) with ω_0 equal to the normalized curvature form of the metric $\|\cdot\|$ on L.

Remark 4.6. In the case when Δ is trivial (i.e., $w_i = 0$), the formula for γ_N in the previous theorem was shown in [46, Section 3], using a different algebro-geometric argument.

4.3. The zero-free hypothesis in the case of three points and the complex Selberg integral

We will next give an alternative proof of Theorem 4.1 in the case when $m = 3$ using the approach in Section 2.4.3. To simplify the notation, we will drop the subscript k in the notation N_k in formula (4.3). In other words, as our data we take a divisor Δ on $\mathbb{P}^1_{\mathbb{C}}$ and an integer N which is strictly greater than one (k can then be recovered from formula (4.3)). First, recall that, by Lemma 4.3, the normalizing constant $\mathcal{Z}_N$ – that we will write as $\mathcal{Z}_N(\Delta)$ to indicate the dependence on Δ – may be expressed by

$$\mathcal{Z}_N(\Delta) = \int_{\mathbb{C}^N} \Big(\prod_{i\neq j}|z_i - z_j|\Big)^{-\frac{d_L}{N-1}} \prod_{i\leq N, j\leq m} |z_i - p_j|^{-2w_i} \prod_i \frac{i}{2} dz_i \wedge d\bar{z}_i.$$

Now specialize to $m = 3$. Then we may, after perhaps applying an automorphism of $\mathbb{P}^1_{\mathbb{C}}$, assume that the points p_1, p_2, and p_3 are given by the points 0, 1, and ∞. Hence,

$$\mathcal{Z}_N(\Delta) = \int_{\mathbb{C}^N} \Big(\prod_{i\neq j}|z_i - z_j|\Big)^{-\frac{d}{N-1}} \prod_i |z_i|^{-2w_0} \prod_i |z_i - 1|^{-2w_1} \prod_i \frac{i}{2} dz_i \wedge d\bar{z}_i,$$

$$d = 2 - (w_0 + w_1 + w_2).$$

This integral is known as the *complex Selberg integral* (when expressed in terms of the parameters w_0, w_1, and $d/(N-1)$). The original Selberg integral is the integral obtained by replacing $\mathbb{C}^N$ with $[0,1]^N$ and generalizes Euler's classical Beta-function to $N > 1$ (see the survey [44]). Its complex version above seems to first have appeared in the CFT, in the context of minimal CFTs, where it is known as one of the *Dotsenko–Fateev integrals* [40] (an equivalent formula was also established in [2], expressed in terms of the original Selberg integral). By [40, formula (B.9)], the integral $\mathcal{Z}_N(\Delta)$ is explicitly given by the following remarkable formula involving the classical Γ-function:

$$\mathcal{Z}_N(\Delta) = N!\left(\frac{\pi}{l\left(-\frac{1}{2}\frac{d}{N-1}\right)}\right)^N \prod_{j=1}^{N} \frac{l\left(-\frac{j}{2}\frac{d}{N-1}\right)}{l\left(w_1+\frac{j}{2}\frac{d}{N-1}\right) l\left(w_2+\frac{j}{2}\frac{d}{N-1}\right) l\left(w_3+\frac{j}{2}\frac{d}{N-1}\right)},$$
$$l(x) := \frac{\Gamma(x)}{\Gamma(1-x)}. \tag{4.10}$$

Remark 4.7. The integral $\mathcal{Z}_N(\Delta)$ also appears in connection to the DOZZ formula of Dorn–Otto and Zamolodchikov–Zamolodchikov for the 3-point structure constants $C_\gamma(\alpha_1, \alpha_2, \alpha_3)$ in Liouville CFT, which has recently been given a rigorous proof in [63] (see also the exposition in [80, Section 2.3]). A general formula for Selberg-type integrals over a local field F of characteristic zero was recently established in [45] (specializing to Selberg's original integral when $F = \mathbb{R}_{>0}$ and its complex generalization when $F = \mathbb{C}$).

We next observe that for any given $\varepsilon \in]0,1[$, $\mathcal{Z}_N(\Delta)$ is zero-free in the convex tube domain Ω in $\mathbb{C}^3$ defined by

$$\Omega = \{\boldsymbol{w} \in \mathbb{C}^3 : \Re w_i < 1,\ \Re w_1 + \Re w_2 + \Re w_3 > 0\}. \tag{4.11}$$

Indeed, by formula (4.10),

$$\mathcal{Z}_N(\Delta) = N!\pi^N \left(\frac{\Gamma\left(1+\frac{1}{2}\frac{d}{N-1}\right)}{\Gamma\left(-\frac{1}{2}\frac{d}{N-1}\right)}\right)^N \prod_{j=1}^{N} \left(\frac{\Gamma\left(-\frac{j}{2}\frac{d}{N-1}\right)}{\Gamma\left(1+\frac{j}{2}\frac{d}{N-1}\right)} \frac{\Gamma\left(1-w_1-\frac{j}{2}\frac{d}{N-1}\right)}{\Gamma\left(w_1+\frac{j}{2}\frac{d}{N-1}\right)} \cdots\right),$$

where the dots indicate similar factors obtained by replacing w_1 with w_2 and w_3. It is a classical fact that $\Gamma(x)$ is a meromorphic zero-free function of $x \in \mathbb{C}$ with poles at $0, -1, -2, \dots$. Hence, the zeros of $\mathcal{Z}_N(\Delta)$ can only come from the poles of the Gamma factors appearing in the denominators above. First, consider the case when $d \neq 0$. Since $N \geq 2$ and $2 > \Re d$, the factor $\Gamma(-\frac{1}{2}\frac{d}{N-1})$ has no poles in Ω. Similarly, since $\Re d > -1$, the factor $\Gamma(1 + \frac{j}{2}\frac{d}{N-1})$ has no poles and since $\Re w_1 < 1$, the factor $\Gamma(w_1 + \frac{j}{2}\frac{d}{N-1})$ has no poles in Ω (using that, for $\boldsymbol{w} \in \mathbb{R}^3$, when $d < 0$, $w_1 + \frac{j}{2}\frac{d}{N-1}$

is minimal when $j = N$ and $N = 2$, i.e., the minimum is $w_1 + d = 2 - w_1 - w_2 > 0$) and likewise when w_1 is replaced by w_2 and w_3. Finally, when $d = 0$, we get

$$\mathcal{Z}_N(\Delta) = N!\pi^N \left(\frac{\Gamma(1 - w_1)}{\Gamma(w_1)} \cdots \right)^N$$

which is non-zero, since $\Re w_i > 0$ (and thus the denominator above has no poles).

This argument also reveals that the "first" negative poles of $\mathcal{Z}_N(\Delta)$ appear when $1 - x = 0$, for $x = w + td/2$ for $w \in \{w_0, w_1, w_2\}$ and $t = i/(N-1)$ for $i = 1, \dots, N$, i.e., when $w + td/2 = 1$. In particular, if $w + td/2 > 1$ for the maximal value of t, i.e., for $t = N/(N-1)$, then $\mathcal{Z}_N(\Delta) < \infty$. This is precisely the condition for the finiteness of $\mathcal{Z}_N(\Delta)$ that came up in the beginning of Section 4.1 which is equivalent to the weight condition (4.2) for w real. The explicit formula (4.10) for $\mathcal{Z}_N(\Delta)$ then also gives

$$\mathcal{Z}_N(\Delta) \leq C^N.$$

4.3.1. Proving Theorem 4.1 by deforming Δ in the case when $m = 3$. We finally explain how to give an alternative proof of Theorem 4.1 in the case $m = 3$ using the zero-free property and the bound on $\mathcal{Z}_N(\Delta)$ established in the previous section, combined with the approach discussed in Section 2.4.3. In this case, the affine space $\mathcal{A}$ of all "admissible" $(s, \boldsymbol{w})$ is defined by the condition

$$d_L^{-1}\left(2 - \left(\sum_{i=1}^{m} w_i\right)\right) = s,$$

where, as before, d_L denotes the degree of the anti-canonical line bundle of the given log Fano variety (whose weight vector is denoted by $\boldsymbol{w}_0$ in Section 2.4.3). In particular, since we consider the case when $m = 3$, we get $s < 0$ by choosing a real weight vector $\boldsymbol{w}_1$ with components sufficiently close to 1 (which can be done as soon as $m > 2$) and, in particular, $\boldsymbol{w}_1 \in \Omega$ (where Ω is the domain in formula (4.11)). Since the components $p_1, \dots, p_m$ of Δ are, trivially, non-singular and mutually non-intersecting, the implicit function theorem does apply. Hence, so does the approach in Section 2.4.3.

5. Speculations on the strong zero-free hypothesis, L-functions, and arithmetic geometry

In this last section, we discuss some intriguing relations between the strong zero-free hypothesis for the partition functions $\mathcal{Z}_N(\beta)$ on Fano manifolds introduced in Section 2.4.1 and the zero-free property of the representation-theoretic (automorphic) local zeta functions $L_p(s)$ appearing in the Langlands program [65]. Conjecturally, the latter zeta functions are related to arithmetic/motivic L-functions [66].

First, recall that given a reductive group G over a global field F together with automorphic representations π and ρ of G and its Langlands dual, respectively, one attaches a local L-function $L_p(s)$ to any place (prime) p of F. By definition, the places p of F correspond to multiplicative (normalized) absolute value $|\cdot|_p$ on F. In the case when $|\cdot|_p$ is non-Archimedean, the local L-function $L_p(s)$ is defined as the inverse of a characteristic polynomial attached to the induced representation of G_p and thus $L_p(s)$ is automatically zero-free. For Archimedean $|\cdot|_p$, the local L-function $L_p(s)$ may be defined as an appropriate product of Γ-functions and is thus also zero-free; see [59, Section 4] for the case $G = \mathrm{GL}(N, \mathbb{C})$ and the relation to the local Langlands correspondence. Conjecturally, any local automorphic L-function $L_p(s)$ is a product of the *standard L-functions* corresponding to the case when $G = \mathrm{GL}(N, F_p)$ and ρ is the standard representation of $\mathrm{GL}(N, \mathbb{C})$ [65] (generalizing the local versions of the classical Hecke L-functions, e.g. the Riemann zeta function when $N = 1$).

5.1. The "minimal" partition function on $\mathbb{P}^n_{\mathbb{C}}$ as a standard local L-function

In the standard case, it was shown in [51] (generalizing Tate's thesis [78] to $N > 1$) that $L_p(s)$ may – for any given admissible irreducible representation π – be realized as a "zeta integral":

$$L_p(s) = \int_{\mathrm{GL}(N,F_p)} \left|\det(g)\right|_p^s \mu_p(g) \tag{5.1}$$

for a distinguished measure μ_p on $\mathrm{GL}(F_p, N)$, depending on π, which is absolutely continuous with respect to Haar measure. As a consequence, for such particular measures $\mu_p(g)$ the zeta integral above is zero-free (since $L_p(s)$ is).

To see the relation to the partition functions $\mathcal{Z}_N(\beta)$ for Fano manifolds, first note that we may, in the zeta integral above, replace the group $\mathrm{GL}(F_p, N)$ with the algebra $\mathrm{Mat}(F_p, N)$ of $N \times N$ matrices A with coefficients in F_p (since μ_P puts no mass on the complement of $\mathrm{GL}(F_p, N)$ in $M(F_p, N)$). Then, after a suitable shift, $s \to s + \lambda$, the measure μ_p is of the form

$$\mu_p = f_\pi \Phi dA,$$

where dA is the additive Haar measure on $\mathrm{Mat}(F_p, N)$, the function f_π is an appropriate matrix element of π, and Φ is a suitable Schwartz–Bruhat function on $\mathrm{Mat}(F_p, N)$. In the "unramified case", f_π is the spherical function attached to π and Φ is its own Fourier transform [51, Proposition 6.12]. In case when p is non-Archimedean, this means that Φ is the characteristic function of $M(O_p, N)$, where O_p denotes the ring of integers of F_p, while in the Archimedean case, Φ is the Gaussian (see [54] for the case $F_p = \mathbb{C}$). Now, when p is taken to be the standard (squared) Archimedean absolute value on $\mathbb{C}(= F_p)$, with π the trivial representation, we get

$$\mathcal{Z}_N(\beta) = c_n \big(\Gamma(s + n + 1)\big)^{-(n+1)} L_p(s), \quad s = \beta(n + 1), \tag{5.2}$$

where $\mathcal{Z}_N(\beta)$ denotes the partition function for the standard Kähler–Einstein metric on the Fano manifold $\mathbb{P}^n_{\mathbb{C}}$ with N the minimal one (i.e., $N = n + 1$) considered in Example 2.9. Indeed, this follows directly from combining formula (5.1) (for $f_\pi = 1$) with formula (A.5) for $\mathcal{Z}_N(\beta)$ in the appendix. Note that the first factor in the right-hand side above is non-vanishing when $\Re\beta > -1$ and thus the zero-free property of $\mathcal{Z}_N(\beta)$ in the strip $\Re\beta > -1$ can be attributed to the zero-free property of the corresponding local L-function $L_p(s)$.

5.2. Zeta integrals associated to Calabi–Yau subvarieties of Mat$(N_k, \mathbb{C})$

It would be interesting to compute $\mathcal{Z}_{N_k}(\beta)$ in more examples to check if it can be expressed as products (and quotients) of Γ-function and related to local Archimedean L-functions as above. For example, if a reductive group G acts holomorphically on X (e.g. if X is a flag variety), one might be able to exploit that the section $\det S^{(k)}$ over X^{N_k} is invariant under the diagonal action of G on X^{N_k}, up to multiplication by the determinant of the induced G-action on $H^0(X, -kK_X)$.

For a general Fano manifold X and N_k, it seems, however, unlikely that $\mathcal{Z}_{N_k}(\beta)$ can be related to an automorphic local L-function. Anyhow, as next explained the integral $\mathcal{Z}_{N_k}(\beta)$ can be expressed in terms of an integral over a Calabi–Yau subvariety of Mat$(N_k, \mathbb{C})$, which has some intriguing structural similarities with the zeta integral for the standard L-function $L_p(s)$ in formula (5.1). We start by lifting the integral $\mathcal{Z}_{N_k}(\beta)$ to an integral where the projective variety X is replaced by the affine variety Y_k of dimension $n+1$ obtained by blowing down of the zero-section in the total space of the line bundle $-kK_X \to X$. To this end, first note that the standard $\mathbb{C}^*$-action on $-kK_X$ induces a $\mathbb{C}^*$-action on the affine variety Y_k with a unique fixed point y_0, i.e., Y_k can be viewed as an affine cone over X:

$$X \simeq \left(Y_k - \{y_0\}\right)/\mathbb{C}^*.$$

On the affine variety Y_k, there is a unique $\mathbb{C}^*$-equivariant holomorphic top form Ω (modulo a multiplicative constant). The Kähler–Einstein metric ω_{KE} on X corresponds to a conical Calabi–Yau metric ω_{CY} on Y_k, i.e., a Ricci-flat Kähler metric with a conical singularity at y_0 [49]. Denote by r the distance to the fixed point y_0 in Y_k with respect to the Calabi–Yau metric ω_{CY}. We may then express

$$\begin{aligned}\mathcal{Z}_{N_k}(\beta) &= c_n\Big(\Gamma\big((n+1)\beta + n + 1\big)\Big)^{-N_k}\tilde{\mathcal{Z}}_{N_k}(\beta),\\ \tilde{\mathcal{Z}}_{N_k}(\beta) &:= \int_{Y_k^{N_k}} |\det \Psi^{(k)}|^{2\beta/k}(e^{-r^2}\Omega\wedge\bar{\Omega})^{\otimes N_k},\end{aligned}$$

where $\Psi^{(k)}$ is the holomorphic function on $Y_k^{N_k}$ corresponding to the section $\det S^{(k)}$ of $-kK_{X^{N_k}}$ and c_n is a (computable) positive constant c_n. This is shown essentially as in the proof of Proposition A.3 in the appendix. Next, assume that k is sufficiently large to ensure that $-kK_X$ is very ample. Then one obtains a holomorphic $(\mathbb{C}^*)^{N_k}$-equivariant embedding

$$Y_k^{N_k} \to \mathrm{Mat}(N_k, \mathbb{C}), \quad (y_1, \dots, y_{N_k}) \mapsto \big(\mathbf{\Psi}^{(k)}(y_1), \dots, \mathbf{\Psi}^{(k)}(y_{N_k})\big),$$

where $\mathbf{\Psi}^{(k)}(y)$ denotes the N_k-tuple of holomorphic functions $\psi_1^{(k)}, \dots, \psi_{N_k}^{(k)}$ on Y_k corresponding to the fixed bases in $H^0(X, -kK_X)$. In geometric terms, the embedding above is just the embedding induced from the Kodaira embedding of X in the projectivization of $H^0(X, -kK_X)^*$. Denoting by $\mathcal{Y}_k$ the image of $Y_k^{N_k}$ in $\mathrm{Mat}(N_k, \mathbb{C})$, we can thus express $\widetilde{\mathcal{Z}}_{N_k}(\beta)$ as a matrix integral:

$$\widetilde{\mathcal{Z}}_{N_k}(\beta) := \int_{\mathcal{Y}_k \Subset \mathrm{Mat}(N_k, \mathbb{C})} |\det A|^{2\beta/k} e^{-r^2} \Omega \wedge \bar{\Omega},$$

where now r denotes the distance to the origin in $\mathrm{Mat}(N_k, \mathbb{C})$ with respect to the Calabi–Yau metric on the subvariety $\mathcal{Y}_k$ and Ω denotes the equivariant holomorphic top form on $\mathcal{Y}_k$ (which can be viewed as a Poincaré-type residue of the standard holomorphic top form on $\mathrm{Mat}(N_k, \mathbb{C})$ along $\mathcal{Y}_k$). This matrix integral is reminiscent of the integral expression (5.1) for the local L-functions $L_p(s)$, if μ_p is taken to be the measure on $\mathrm{Mat}(N_k, \mathbb{C})$ induced by pairing of $\Omega \wedge \bar{\Omega}$ with the subvariety $\mathcal{Y}_k$, weighted by the Gaussian-type factor e^{-r^2} (and $s := \beta/k$). In view of this structural similarity, it is tempting to speculate on a very strong zero-free hypothesis, saying that, in general, the "lifted" partition function $\widetilde{\mathcal{Z}}_{N_k}(\beta)$ is zero-free on all of $\mathbb{C}$, when viewed as a meromorphic function.

Remark 5.1. The same considerations apply when X is a Fano orbifold if K_X is replaced by the orbifold canonical line bundle (coinciding with $-K_X + \Delta$ as $\mathbb{Q}$-line bundle). Then the natural projection from $Y_k - \{y_0\}$ to X is a submersion over the complement of the branching divisor Δ and the orbifold Kähler–Einstein metric on X corresponds to a bona fide Calabi–Yau metric on $Y_k - \{y_0\}$ [49].

One further piece of evidence for the very strong form of the zero-free hypothesis (complementing the "minimal" case on $\mathbb{P}^n$ appearing in Proposition A.3) is provided by the case when $X = \mathbb{P}^1$ and $k = 1$, i.e., $N_k = 3$ (which is the case next to minimal dimension, $N_k = n + 1$). Then identifying $-K_X$ with $2\mathcal{O}(1)$ and $\det S^{(1)}$ with the Vandermonde determinant $D^{(3)}$ on $\mathbb{C}^3$ (as in Lemma 4.3) and using that the Kähler–Einstein metric is explicitly given by the Fubini–Study metric (formula (4.7)), $\mathcal{Z}_{N_k}(\beta)$ may be expressed as

$$\mathcal{Z}_{N_k}(\beta) = \int_{\mathbb{C}^3} \prod_{i<j\leq 3} |z_i - z_j|^{2\beta} \prod_{i<j\leq 3} \big(1 + |z_i|^2\big)^{-(2\beta+2)},$$

integrating with respect to Lebesgue measure. Applying the formula in [22, Theorem 1] (to $\sigma_i = \nu_i = \beta + 1$), which originally appeared in the CFT, thus yields

$$\mathcal{Z}_{N_k}(\beta) = \pi^3\big(\Gamma(2\beta+2)\big)^{-3}\Gamma(3\beta+2)\Gamma(\beta+1)^3. \tag{5.3}$$

This means that the meromorphic function $\widetilde{\mathcal{Z}}_{N_k}(\beta)$ is a product of four Gamma functions and thus zero-free on all of $\mathbb{C}$. The elegant proof in [22] leverages the diagonal action of $\mathrm{GL}(N_k, \mathbb{C})$ on X^{N_k} alluded to above (following the corresponding real case considered in [21] in the context of automorphic triple products).

The general case on $X = \mathbb{P}^1$, when $N_k > 3$, appears to be open. However, a similar formula does hold for any N_k when X is replaced by its *real* points, i.e., when $\mathbb{P}^1_{\mathbb{C}}$ is replaced by $\mathbb{P}^1_{\mathbb{R}}$. Then the role of $\mathcal{Z}_{N_k}(\beta)$ is played by

$$\begin{aligned}\mathcal{Z}_{N_k}(\beta)_{\mathbb{R}} &:= \int_{(\mathbb{P}^1_{\mathbb{R}})^{N_k}} \|\det S^{(k)}\|^{\beta/k}\, dV^{\otimes N_k} \\ &= \int_{(S^1)^{N_k}} \prod_{i<j\leq N_k} |z_i - z_j|^{\frac{2\beta}{N_k-1}}\, d\theta^{\otimes N_k}, \quad N_k = 2k+1,\end{aligned}$$

where $\|\cdot\|$ denotes the Fubini–Study metric and dV denotes the corresponding volume form on $(\mathbb{P}^1_{\mathbb{R}})$. In the second equality above, we have exploited that the integrand is invariant under the diagonal action of $\mathrm{SU}(2)$ to replace the real points $\mathbb{P}^1_{\mathbb{R}}$ of $\mathbb{P}^1_{\mathbb{C}}$ with the unit-circle S^1 in $\mathbb{C} \subset \mathbb{P}^1_{\mathbb{C}}$. The latter integral over $(S^1)^{N_k}$ coincides with the partition function for the 2D Coulomb gas confined to $S^1 \subset \mathbb{C}$ at inverse temperature $2\beta/(N_k - 1)$ (known as the circular ensemble). Applying [44, formula (1.12)] (originally conjectured by Dyson and established by Gunson and Wilson) thus yields

$$\mathcal{Z}_{N_k}(\beta)_{\mathbb{R}} = (2\pi)^{N_k}\Gamma\Big(1+\beta\frac{1}{N_k-1}\Big)^{-N_k}\Gamma\Big(1+\beta\frac{N_k}{N_k-1}\Big), \quad N_k = 2k+1.$$

This formula reveals that the real analog $\mathcal{Z}_{N_k}(\beta)_{\mathbb{R}}$ of the partition function on $\mathbb{P}^1_{\mathbb{C}}$ does satisfy the strong zero-free hypothesis. This real analog may, from the point of view of localization, be obtained by replacing the squared absolute value $|\cdot|^2_{\mathbb{C}}$ corresponding to the complex Archimedean place of the global field $\mathbb{Q}$ with the absolute value $|\cdot|_{\mathbb{R}}$ corresponding to the real Archimedean place of $\mathbb{Q}$. The extension to non-Archimedean places is discussed in Section 5.4. But first we start by a brief detour on arithmetic aspects of the partition function.

5.3. Invariants of arithmetical Fano varieties

Let $\mathcal{X}$ be an arithmetic variety of dimension $n+1$ (i.e., a projective scheme flat over $\mathbb{Z}$, $\mathcal{X} \to \operatorname{Spec}\mathbb{Z}$) such that the corresponding n-dimensional complex variety X (i.e., the complexification of the generic fiber $X_{\mathbb{Q}}$ of $\mathcal{X}$) is Fano. Assume that $\mathcal{X}$

is endowed with a relatively nef line bundle $\mathcal{L}$ such that the induced line bundle on X equals $-K_X$. Then $(\mathcal{X},\mathcal{L})$ induces a section $\det S^{(k)}$ of $-kK_{X^{N_k}} \to X^{N_k}$ which is uniquely determined up to multiplication by ± 1. Indeed, $(\mathcal{X},\mathcal{L})$ induces a lattice $H^0(\mathcal{X},k\mathcal{L})$ of integral sections in $H^0(X,-kK_X)$ and $\det S^{(k)}$ may be defined as in formula (1.3) with respect to any basis in $H^0(\mathcal{X},k\mathcal{L})$ (any two such bases are related by a matrix with integral coefficients, which thus has determinant equal to ± 1). As a consequence, the corresponding partition function $\mathcal{Z}_{N_k}(\beta)$ only depends on $(\mathcal{X},\mathcal{L})$ and the choice of a metric $\|\cdot\|$ on $-K_X$ (and is independent of the metric at $\beta=-1$). In fact, the explicit expression for $\mathcal{Z}_{N_k}(\beta)$ appearing in Proposition A.3 – related to a local L-function in formula (5.2) – was computed with respect to the standard integral model $(\mathcal{X},\mathcal{L})$ for $(\mathbb{P}^n,\mathcal{O}(1))$ (where $H^0(\mathcal{X},k\mathcal{L})$ is the lattice spanned by the sections defined by multinomials). In the light of the speculations in the previous section, this appears to fit well with the arithmetical side of the Langlands program.

In particular, taking $\beta=-1$ yields an invariant $\mathcal{Z}_{N_k}$ of $(\mathcal{X},\mathcal{L})$ (which is finite iff X is Gibbs stable at level k). The following conjecture relates the arithmetic invariants $\mathcal{Z}_{N_k}$ to the arithmetic intersection numbers introduced by Gillet–Soulé in the context of Arakelov geometry (see the book [77]).

Conjecture 5.2. *Let $(\mathcal{X},\mathcal{L})$ be an arithmetic variety as above and assume that the corresponding Fano manifold X admits a unique Kähler–Einstein metric, whose volume form is denoted by dV_{KE}, normalized to have unit total volume. Then, as $k\to\infty$, $\frac{(n+1)!}{k^n}\log\mathcal{Z}_{N_k}$ converges towards the $(n+1)$-fold arithmetic self-intersection number of the line bundle $\mathcal{L}$, metrized by dV_{KE}.*

In fact, using the arithmetic Hilbert–Samuel theorem in [83, Theorem 1.4] (generalizing the relative ample case in [50]), this conjecture is equivalent to the convergence of the partition function appearing in Theorem 2.4, defined with respect to any basis of $H^0(X,kK_X)$ which is orthonormal with respect to the Hermitian product induced by a Kähler metric on X. Thus, by Theorem 2.6, in order to establish the conjecture it would, for example, be enough to show that the lifted partition function $\widetilde{\mathcal{Z}}_{N_k}(\beta)$ may be expressed as a product of $O(N_k)$ shifted Γ-functions all of whose poles are located in the region where $\Re\beta<-1-\varepsilon$ for some $\varepsilon>0$.

Remark 5.3. Other (polarized) arithmetic varieties on arithmetic varieties $\mathcal{X}$, endowed with a relatively ample line bundle $\mathcal{L}$, are introduced in [23, 84] (which are finite precisely when $(X,k\mathcal{L})$ is Chow stable) and related to constant scalar curvature metrics in [74].

The analog of Conjecture 5.2 does hold when $-K_X$ is replaced by K_X (assumed ample) and $\log\mathcal{Z}_{N_k}$ is replaced by the arithmetic invariant $-\log\mathcal{Z}_{N_k}$ (as follows from combining the convergence of $\mathcal{Z}_{N_k}(1)$ in Theorem 2.1 with the arithmetic Hilbert–Samuel theorem).

5.4. Extension to non-Archimedean places

In view of the connections to local L-functions, L_p at the (complex) Archimedean place p, exhibited in Section 5.1, one may wonder if the probabilistic setup can be extended to non-Archimedean places p. The case of the trivial place is discussed in (5.1), in connection to Gibbs stability. What follows are some speculations on the case of non-trivial non-Archimedean places p, inspired by the adelic geometric setup in [29], where geometric Igusa local zeta functions are studied (see Section A.2).

Let X be a non-singular variety defined over $\mathbb{Q}$ and first consider the case when $K_{X(\mathbb{Q})}$ is ample. Given a non-trivial non-Archimedean place p (i.e., a prime number), denote by $X(\mathbb{Q}_p)$ the projective variety over the corresponding p-adic local field $\mathbb{Q}_p$ (the completion of $\mathbb{Q}$ with respect to $|\cdot|_p$), which comes with the structure of a $\mathbb{Q}_p$-analytic manifold. By general principles, any continuous metric on $K_{X(\mathbb{Q}_p)}$ induces a measure on $X(\mathbb{Q}_p)$, which is absolutely continuous with respect to the local Haar measures [29, Section 2.1]. In particular, a section s_k of $kK_{X(\mathbb{Q}_p)}$ induces a measure on $X(\mathbb{Q}_p)$, whose local density may be symbolically expressed as $|s_k|_p^{1/k}$.[4] Hence, replacing the squared Archimedean absolute value appearing in formula (1.2) with $|\cdot|_p$, one arrives at a symmetric probability measure $\mu_p^{(N_k)}$ on $X(\mathbb{Q}_p)^{N_k}$. This construction thus yields a canonical random point process on $X(\mathbb{Q}_p)$. Accordingly, it seems natural to ask if the convergence in Theorem 1.1 can be extended to this non-Archimedean setup, if dV_{KE} is replaced by an appropriate measure $dV_{KE,p}$ on $X(\mathbb{Q}_p)$. In analogy with the Archimedean setup, the measure $dV_{KE,p}$ should be characterized as the unique minimizer of a free-energy type functional F_1 on the space of probability measure μ on $X(\mathbb{Q}_p)$ of the form

$$F_1(\mu) = E(\mu) + \mathrm{Ent}(\mu), \tag{5.4}$$

where $\mathrm{Ent}(\mu)$ denotes the entropy of the measure μ relative to a fixed measure on $X(\mathbb{Q}_p)$, absolutely continuous with respect to the local Haar measure and $E(\mu)$ is a non-Archimedean analog of the energy discussed in Section 2.2. In particular, $dV_{KE,p}$ is then absolutely continuous with respect to the local Haar measure.

Ideally, one might hope that the collection of metrics on $-K_{X(\mathbb{Q}_p)}$ defined by $dV_{KE,p}$, as p ranges over all primes p, is induced by some model $(\mathcal{X}, \mathcal{L})$ for $(X, K_{X(\mathbb{Q})})$ over $\mathbb{Z}$, away from primes p with bad reduction (cf. [29, Section 2.2.3]). This would, loosely speaking, yield a probabilistic construction of a "canonical" integral model attached to $X(\mathbb{Q})$. This is in line with the analogy between the Kähler–Einstein condition of a metric on $X(\mathbb{C})$ (i.e., at $p = \infty$) and the minimality condition of an integral model for $X(\mathbb{Q})$ put forth in [72] and further studied in [74].

[4] One can also consider a field extension F_p of $\mathbb{Q}_p$ and get a measure on the corresponding analytic manifolds $X(F_p)$, as in [56], but here $F_p = \mathbb{Q}_p$, for simplicity.

Remark 5.4. Embedding $X(\mathbb{Q}_p)$ in its Berkovich analytification X_p^{an} and pushing forward a measure μ on $X(\mathbb{Q}_p)$ to X_p^{an}, the functional on $C^0(X_p^{an})$ defined as the Legendre–Fenchel transform of the functional $E(\mu)$ in formula (5.4) should, in analogy to the Archimedean setup [4, 17], be given by the primitive of the non-Archimedean Monge–Ampère operator introduced in [28, 62]. The primitive in question is called the "energy functional" in [24]. In the case of a trivial non-Archimedean absolute value, such an energy $E(\mu)$ appears in [25, formula (6.1)] and plays an important role in the non-Archimedean approach to K-stability.

Similar considerations apply in the Fano case. In particular, to a given metric on $-K_{X(\mathbb{Q}_p)}$ one can associate a lifted partition function $\widetilde{\mathcal{Z}}_{N_k,p}(\beta)$. By general principles [29, Section 4.1], this defines a meromorphic function on $\mathbb{C}$ which in the light of Section 5.1 plays the role of the local L-functions L_p in the Langlands program. More precisely, in order to render $\widetilde{\mathcal{Z}}_{N_k,p}(\beta)$ as canonical as possible, the metric on $-K_{X(\mathbb{Q}_p)}$ should be taken to be defined by a "canonical" integral model $(\mathcal{X}, \mathcal{L})$ for $(X(\mathbb{Q}), -K_{(\mathbb{Q})})$ and det $S^{(k)}$ should be defined with respect to any basis in $H^0(\mathcal{X}, \mathcal{L})$ (as in Section 5.3). Finally, one could then attempt to define a global L-type function as a Euler product of $\widetilde{\mathcal{Z}}_{N_k,p}(\beta)$ over all p, generalizing the Riemann zeta function.

A. Log canonical thresholds and Archimedean zeta functions

In this appendix, we recall the basic notions of lct's, α-invariants, and their connections to Archimedean zeta functions, which are as essentially well known. We conclude with a proof of the formula appearing in Example 2.9.

A.1. Log canonical thresholds

Let X be a compact complex manifold.

A.1.1. The lct of a divisor on X. By definition, an $\mathbb{R}$-divisor D is a finite formal sum of irreducible analytic subvarieties $D_i \subset X$ of complex codimension one:

$$D = \sum_{i=1}^{m} c_i D_i, \quad c_i \in \mathbb{R}.$$

The *log canonical threshold* $\mathrm{lct}_X(D)$ of an $\mathbb{R}$-divisor D has various algebro-geometric formulations (using discrepancies, valuations, multiplier ideal sheaves, etc.) [60], but for the purposes of the present paper, it will be enough to recall its analytic definition as an integrability threshold. First, consider the case when the coefficients D are in $\mathbb{Z}_+$. This equivalently means that there exists a holomorphic line bundle $L_D \to X$ and a holomorphic section s_D such that D is cut-out by s_D, including multiplicities, i.e., s_D vanishes to order c_i along the irreducible varieties D_i. The lct may then be defined as the following integrability index:

$$\operatorname{lct}_X(D) := \sup_{\gamma>0}\left\{\gamma : \int_X \|s_D\|^{-2\gamma}\, dV < \infty\right\}, \tag{A.1}$$

in terms of any Hermitian metric $\|\cdot\|$ on L and volume form dV on X. This definition first extends to the case when $c_i \in \mathbb{Z}$, if s_D is viewed as a meromorphic section, so that the negative coefficients correspond to the poles of s_D, and then to $c_i \in \mathbb{Q}$ by viewing s_D as a multi-valued holomorphic section and noting that $\|s\|$ is still a well-defined function on X (taking values in $[0,\infty]$). Finally, the definition extends, by continuity, to any $\mathbb{R}$-divisor D or, alternatively, by noting that the function $\|s_D\|$ is still well defined (and can be viewed as the norm on an $\mathbb{R}$-line bundle, i.e., a formal sum of the line bundles L_{D_i}).

A.1.2. The lct of a divisor on (X, Δ). More generally, if Δ is a given $\mathbb{Q}$-divisor of X, then the lct of D relative to the *log pair* (X,Δ) [30] may be analytically defined as

$$\operatorname{lct}_{(X,\Delta)}(D) := \sup_{\gamma>0}\left\{\gamma : \int_X \|s\|^{-2\gamma}\, dV_\Delta < \infty\right\},$$

where dV_Δ is a measure on X with singularities encoded by Δ, i.e., locally dV_Δ may be expressed as

$$dV_\Delta = \|s_\Delta\|^{-2} dV_X$$

for some bona fide volume form dV_X on X and metric $\|\cdot\|$ on the $\mathbb{Q}$-line bundle with multivalued holomorphic section s_Δ corresponding to Δ. More generally, as in the previous section, Δ may be taken to be an $\mathbb{R}$-divisor on X.

A.1.3. The lct of a line bundle L and the α-invariant. The log canonical threshold $\operatorname{lct}_X(L)$ of a line bundle $L \to X$ is now defined by

$$\operatorname{lct}_X(L) := \inf_{D\sim L} \operatorname{lct}_X(D),$$

where D ranges over the divisors attached to all the many-valued holomorphic section s of L. By [35], this coincides with Tian's α-invariant of L:

$$\alpha(L) := \sup_{\gamma>0}\left\{\gamma : \exists C \int_X e^{-\gamma(\phi-\phi_0)}\, dV \le C \ \forall \phi \in \mathcal{H}(L)\right\}, \tag{A.2}$$

where $\mathcal{H}(L)$ denotes the space of all metrics on L with positive curvature and ϕ_0 denotes a fixed smooth reference metric on L using additive notation for metrics so that $\phi - \phi_0$ defines a function on X. More generally, the log canonical threshold $\operatorname{lct}_{(X,\Delta)}(L)$ of a line bundle $L \to X$ with respect to a log pair (X,Δ) [30] is defined by

$$\operatorname{lct}_{(X,\Delta)}(L) := \inf_{D\sim L} \operatorname{lct}_{(X,\Delta)}(D).$$

This coincides with the α-invariant defined with respect to the log pair (X,Δ) obtained by replacing dV in formula (A.2) with $dV_{(X,\Delta)}$, as shown in the appendix of [4].

A.2. Archimedean zeta functions

Let μ_0 be a measure on $\mathbb{C}^n$ with compact support and $\psi \in L^1(\mu_0)$. Then we may define the integrability threshold $\mathrm{lct}_{\mu_0}(\psi)$ as in formula (A.1), by replacing $\log \|s\|^2$ with ψ and dV by μ_0. The integral

$$Z(\beta) = \int_{\mathbb{C}^n} e^{2\beta\psi} \mu_0,$$

defines a holomorphic function on the strip $\{\Re\beta > -\mathrm{lct}_{\mu_0}(\psi)\}$ in $\mathbb{C}$ (using that, in this strip, $e^{\beta\psi} \in L^1(\mu_0)$ and that the integrand is holomorphic in β). In the case when $\psi = \log|f|^2$ for f holomorphic, or more precisely,

$$Z(\beta) = \int_{\mathbb{C}^n} |f|^{2\beta} \Phi \, dx, \tag{A.3}$$

for a Schwartz function Φ, the holomorphic function $Z(\beta)$ on the strip $\{\Re\beta > -\mathrm{lct}_{\mu_0}(\psi)\}$ extends to a meromorphic function in $\mathbb{C}$, whose poles are located at the negative real axes.

Remark A.1. This follows from classical results of Atiyah and Bernstein, extended by Igusa to a more general setting of zeta function attached to polynomials defined over local fields [53]. Briefly, meromorphic functions $Z(\beta)$ of the form (A.3) can be defined more generally by replacing $\mathbb{C}$ and its standard Archimedean absolute value $|\cdot|$ with any local field F, endowed with an absolute value $|\cdot|_F$. Such functions $Z(\beta)$ are usually called *Igusa local zeta function* [53] and thus $Z(\beta)$ in formula (A.3) is called an Igusa Archimedean zeta function or simply an *Archimedean zeta function* in the literature on algebraic and arithmetic geometry. In the case when f is a polynomial with integer coefficients and F is the p-adic field, $F = \mathbb{Q}_p$, the meromorphic function $Z(\beta)$ encodes the number of solutions of the equation $f(x_1,\ldots,x_n) = 0$, modulo powers of p, when Φ is taken as the characteristic function of the n-fold product of the ring $\mathbb{Z}_p$ of integers of $\mathbb{Q}_p$,

Similarly, given a holomorphic section s of a line bundle $L \to X$ over a compact complex manifold, a metric $\|\cdot\|$ on L and a singular volume form dV_Δ associated to a log pair (X,Δ)

$$\mathcal{Z}(\beta) := \int_X \|s\|^{2\beta} \, dV_{(X,\Delta)} \tag{A.4}$$

defines a holomorphic function in the strip $\{\Re\beta > -\mathrm{lct}_{(X,\Delta)}(D)\}$ in $\mathbb{C}$, where D denotes the divisor cut out by the section s. More precisely, the function $\mathcal{Z}(\beta)$ extends to a meromorphic function on $\mathbb{C}$, whose poles are located on the negative real axes (using a partition of unity to reduce to the case of $X = \mathbb{C}^n$). The first negative pole is precisely $-\operatorname{lct}_{(X,\Delta)}(D)$.

Remark A.2. Functions of the form (A.4) have previously appeared in a general adelic setup [29] (containing both the Archimedean and the p-adic setup), motivated by number theory and arithmetic geometry on log Fano varieties.

In the present probabilistic setup on Fano manifolds, discussed in Section 2.4.1, the manifold is of the form X^{N_k}, the section is the many-valued holomorphic section $(\det S^{(k)})^{1/k}$ of $-K_{X^{N_k}}$, and the measure is of the form $dV_X^{\otimes N_k}$ (and similarly in the case of log Fano pairs). We conclude by proving the explicit formula for $\mathcal{Z}(\beta)$ stated in Example 2.9.

Proposition A.3. *In the setup of Example 2.9, the following formula holds:*

$$\mathcal{Z}(\beta) = c_n \frac{\prod_{j=1}^{n} \Gamma\big(\beta(n+1)+j\big)}{\big(\Gamma\big(\beta(n+1)+n+1\big)\big)^n}.$$

In particular, the maximal holomorphicity strip of $\mathcal{Z}(\beta)$ is given by $\Omega = \{\Re(\beta) > -1/(n+1)\} \Subset \mathbb{C}$ and $\mathcal{Z}(\beta)$ is zero-free in Ω. More precisely, the zeros of $\mathcal{Z}(\beta)$ are located at $\beta = -1 + j/(n+1)$, where $j = 0, 1, 2, \dots$.

Proof. In this “minimal” case, a basis $s_1, \dots, s_{N_k}$ in the complex vector space

$$H^0(X, -kK_X) = H^0\big(\mathbb{P}^n, \mathcal{O}(1)\big)$$

is obtained from the homogeneous coordinates $Z_0, \dots, Z_n$ on $\mathbb{P}^n$. Denote by $\boldsymbol{Z} := (Z_0, \dots, Z_n)$ the corresponding vector in $\mathbb{C}^{n+1}$. We will represent an element in $(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_N) \in (\mathbb{C}^{n+1})^N$ with an $(n+1) \times N$-matrix, denoted by $[\boldsymbol{Z}]$. Then the corresponding Slater determinant $\det S^{(k)}$ may be identified with the homogeneous polynomial $\det[\boldsymbol{Z}]$ on $\mathbb{C}^{(n+1)^2}$, defined by the determinant of the matrix $[\boldsymbol{Z}]$. Using the $\mathrm{SU}(n+1)$-symmetry of the Fubini–Study metric on $\mathcal{O}(1) \to \mathbb{P}^n$, we may then first lift the integral $Z(\beta)$ on $(\mathbb{P}^n)^{n+1}$ to the product of unit-spheres S in $\mathbb{C}^{n+1}$:

$$\mathcal{Z}(\beta) = c_n \int_{S^{(n+1)}} \big|\det[\boldsymbol{Z}]\big|^{2s} \, d\sigma^{\otimes N}, \quad s := \beta/k,$$

where $d\sigma$ denotes the standard $\mathrm{SU}(n+1)$-invariant measure on S. Next, exploiting that $\det[\boldsymbol{Z}]$ is homogeneous of degree 1 in each column gives

$$\int_{S^{(n+1)}} \big|\det[\boldsymbol{Z}]\big|^{2s} \, d\sigma^{\otimes N} = c_n \frac{\int_{\mathbb{C}^{(n+1)^2}} \big|\det[\boldsymbol{Z}]\big|^{2s} e^{-|\boldsymbol{Z}|^2} \, d\lambda}{\big(\int_0^\infty (r^2)^s e^{-r^2} r^{2(n+1)-1} \, dr\big)^{n+1}}.$$

Hence, making the change of variables $t = r^2$ in the denominator (and rewriting $r^{2(n+1)-1}dr = r^{2(n+1)}r^{-2}d(r^2)/2$) reveals that

$$\mathcal{Z}(\beta) = c_n \frac{\int_{\mathbb{C}^{(n+1)^2}} \left|\det[\boldsymbol{Z}]\right|^{2s} e^{-|\boldsymbol{Z}|^2}\, d\lambda}{\left(\Gamma(s+n+1)\right)^{(n+1)}}, \quad \Gamma(a) := \int_0^\infty t^a e^{-t} \frac{dt}{t}. \tag{A.5}$$

Finally, the proof is concluded by invoking the following formula in [53, Theorem 6.3.1]:

$$Z(s) := \int_{\mathbb{C}^{(n+1)^2}} \left|\det[\boldsymbol{Z}]\right|^{2s} e^{-|\boldsymbol{Z}|^2}\, d\lambda = c_n \prod_{j=1}^{n+1} \Gamma(s+j). \tag{A.6}$$

■

Remark A.4. The proof of formula (A.6) in [53] exploits that the polynomial $f := \det[\boldsymbol{Z}]$ on $\mathbb{C}^{(n+1)^2}$ has the property that

$$P(\partial) f^{s+1} = b(s) f^s \tag{A.7}$$

with

$$b(s) = \prod_{j=1}^{n+1} (s+j),$$

when $P(z) = f(z)$. This leads to the functional relation $b(s)Z(s) = Z(s+1)$, that can then be compared with the classical functional relation for $\Gamma(s)$ to deduce formula (A.6). Recall that in general, given a polynomial $f(z)$ on $\mathbb{C}^m$, the monic polynomial $b(s)$ on $\mathbb{C}$ with minimal degree for which there exists a polynomial $P(z)$ satisfying formula (A.7) is called the *Bernstein–Sato polynomial* attached to f [53]. In general, it is very hard to compute $b(s)$ explicitly (and thus to also find $P(z)$) but the present case, $f(z) = \det[\boldsymbol{Z}]$, fits into Sato's theory of prehomogenuous vector spaces. This is explained in [53]. Alternatively, formula (A.6) follows from the Iwasawa decomposition of $\mathrm{GL}(N, \mathbb{C})$ (as in [54, Section 2]). It would be interesting to see if similar considerations could be applied to $X = \mathbb{P}^n$ when N_k is not assumed to be minimal, i.e., when $N_k > n+1$. However, even the case when $n = 1$ appears to be open (apart from the case when $N_k = 3$ appearing in formula (5.3), where a symmetry argument can be exploited).

Acknowledgments. I am grateful to Sébastien Boucksom, Dennis Eriksson, Mattias Jonsson, Yuji Odaka, Daniel Persson, and Yanir Rubinstein for stimulating discussions. Also thanks to the referee whose comments helped to improve the exposition.

Funding. This work was supported by grants from the Knut and Alice Wallenberg foundation, the Göran Gustafsson foundation, and the Swedish Research Council.

References

[1] L. Alvarez-Gaumé, J.-B. Bost, G. Moore, P. Nelson, and C. Vafa, Bosonization on higher genus Riemann surfaces. *Comm. Math. Phys.* **112** (1987), no. 3, 503–552 Zbl 0647.14019 MR 908551

[2] K. Aomoto, On the complex Selberg integral. *Quart. J. Math. Oxford Ser. (2)* **38** (1987), no. 152, 385–399 Zbl 0639.33002 MR 916224

[3] T. Aubin, Équations du type Monge–Ampère sur les variétés kählériennes compactes. *Bull. Sci. Math. (2)* **102** (1978), no. 1, 63–95 Zbl 0374.53022 MR 494932

[4] R. J. Berman, A thermodynamical formalism for Monge–Ampère equations, Moser–Trudinger inequalities and Kähler–Einstein metrics. *Adv. Math.* **248** (2013), 1254–1297 Zbl 1286.58010 MR 3107540

[5] R. J. Berman, Determinantal point processes and fermions on complex manifolds: large deviations and bosonization. *Comm. Math. Phys.* **327** (2014), no. 1, 1–47 Zbl 1337.60093 MR 3177931

[6] R. J. Berman, K-polystability of $\mathbb{Q}$-Fano varieties admitting Kähler–Einstein metrics. *Invent. Math.* **203** (2016), no. 3, 973–1025 Zbl 1353.14051 MR 3461370

[7] R. J. Berman, Large deviations for Gibbs measures with singular Hamiltonians and emergence of Kähler–Einstein metrics. *Comm. Math. Phys.* **354** (2017), no. 3, 1133–1172 Zbl 1394.32019 MR 3668617

[8] R. J. Berman, Kähler–Einstein metrics, canonical random point processes and birational geometry. In *Algebraic Geometry: Salt Lake City 2015*, pp. 29–73, Proc. Sympos. Pure Math. 97, Amer. Math. Soc., Providence, RI, 2018 Zbl 1446.32017 MR 3821145

[9] R. J. Berman, Kähler–Einstein metrics, canonical random point processes and birational geometry. In *Algebraic Geometry: Salt Lake City 2015*, pp. 29–73, Proc. Sympos. Pure Math. 97, Amer. Math. Soc., Providence, RI, 2018 Zbl 1446.32017 MR 3821145

[10] R. J. Berman, On large deviations for Gibbs measures, mean energy and gamma-convergence. *Constr. Approx.* **48** (2018), no. 1, 3–30 Zbl 1398.82034 MR 3825945

[11] R. J. Berman, Statistical mechanics of interpolation nodes, pluripotential theory and complex geometry. *Ann. Polon. Math.* **123** (2019), no. 1, 71–153 Zbl 1452.32040 MR 4025011

[12] R. J. Berman, The probabilistic vs the quantization approach to Kähler–Einstein geometry. 2021, arXiv:2109.06575

[13] R. J. Berman, Emergent complex geometry. 2022, *The Proceedings of the ICM* (to appear)

[14] R. J. Berman, Measure preserving holomorphic vector fields, invariant anti-canonical divisors and Gibbs stability. *Anal. Math.* **48** (2022), no. 2, 347–375 MR 4440748

[15] R. J. Berman and S. Boucksom, Growth of balls of holomorphic sections and energy at equilibrium. *Invent. Math.* **181** (2010), no. 2, 337–394 Zbl 1208.32020 MR 2657428

[16] R. J. Berman, S. Boucksom, P. Eyssidieux, V. Guedj, and A. Zeriahi, Kähler–Einstein metrics and the Kähler–Ricci flow on log Fano varieties. *J. Reine Angew. Math.* **751** (2019), 27–89 Zbl 1430.14083 MR 3956691

[17] R. J. Berman, S. Boucksom, V. Guedj, and A. Zeriahi, A variational approach to complex Monge–Ampère equations. *Publ. Math. Inst. Hautes Études Sci.* **117** (2013), 179–245 Zbl 1277.32049 MR 3090260

[18] R. J. Berman, S. Boucksom, and M. Jonsson, A variational approach to the Yau–Tian–Donaldson conjecture. *J. Amer. Math. Soc.* **34** (2021), no. 3, 605–652 Zbl 07420176 MR 4334189

[19] R. J. Berman, S. Boucksom, and D. Witt Nyström, Fekete points and convergence towards equilibrium measures on complex manifolds. *Acta Math.* **207** (2011), no. 1, 1–27 Zbl 1241.32030 MR 2863909

[20] R. J. Berman, T. C. Collins, and D. Persson, Emergent Sasaki-Einstein geometry and AdS/CFT. *Nat. Commun.* **13** (2022), Article No. 365

[21] J. Bernstein and A. Reznikov, Estimates of automorphic functions. *Mosc. Math. J.* **4** (2004), no. 1, 19–37, 310 Zbl 1081.11037 MR 2074982

[22] B. V. Binh and V. Schechtman, Invariant functionals and Zamolodchikovs' integral. *Funct. Anal. Appl.* **49** (2015), no. 1, 57–59 Zbl 1323.33005 MR 3382951

[23] J.-B. Bost, Intrinsic heights of stable varieties and abelian varieties. *Duke Math. J.* **82** (1996), no. 1, 21–70 Zbl 0867.14010 MR 1387221

[24] S. Boucksom, C. Favre, and M. Jonsson, Solution to a non-Archimedean Monge–Ampère equation. *J. Amer. Math. Soc.* **28** (2015), no. 3, 617–667 Zbl 1325.32021 MR 3327532

[25] S. Boucksom and M. Jonsson, Global pluripotential theory over a trivially valued field. *Ann. Fac. Sci. Toulouse Math. (6)* **31** (2022), no. 3, 647–836 MR 4452253

[26] C. P. Boyer, K. Galicki, and J. Kollár, Einstein metrics on spheres. *Ann. of Math. (2)* **162** (2005), no. 1, 557–580 Zbl 1093.53044 MR 2178969

[27] E. Caglioti, P.-L. Lions, C. Marchioro, and M. Pulvirenti, A special class of stationary flows for two-dimensional Euler equations: a statistical mechanics description. *Comm. Math. Phys.* **143** (1992), no. 3, 501–525 Zbl 0745.76001 MR 1145596

[28] A. Chambert-Loir, Mesures et équidistribution sur les espaces de Berkovich. *J. Reine Angew. Math.* **595** (2006), 215–235 Zbl 1112.14022 MR 2244803

[29] A. Chambert-Loir and Y. Tschinkel, Igusa integrals and volume asymptotics in analytic and adelic geometry. *Confluentes Math.* **2** (2010), no. 3, 351–429 Zbl 1206.11086 MR 2740045

[30] I. Cheltsov, J. Park, and C. Shramov, Exceptional del Pezzo hypersurfaces. *J. Geom. Anal.* **20** (2010), no. 4, 787–816 Zbl 1211.14047 MR 2683768

[31] X. Chen, S. Donaldson, and S. Sun, Kähler–Einstein metrics on Fano manifolds. I: Approximation of metrics with cone singularities. *J. Amer. Math. Soc.* **28** (2015), no. 1, 183–197 Zbl 1312.53096 MR 3264766

[32] X. Chen, S. Donaldson, and S. Sun, Kähler–Einstein metrics on Fano manifolds. II: Limits with cone angle less than 2π. *J. Amer. Math. Soc.* **28** (2015), no. 1, 199–234 Zbl 1312.53097 MR 3264767

[33] X. Chen, S. Donaldson, and S. Sun, Kähler–Einstein metrics on Fano manifolds. III: Limits as cone angle approaches 2π and completion of the main proof. *J. Amer. Math. Soc.* **28** (2015), no. 1, 235–278 Zbl 1311.53059 MR 3264768

[34] T. Darvas and Y. A. Rubinstein, Tian's properness conjectures and Finsler geometry of the space of Kähler metrics. *J. Amer. Math. Soc.* **30** (2017), no. 2, 347–387 Zbl 1386.32021 MR 3600039

[35] J.-P. Demailly, On tian's invariant and log canonical thresholds. Appendix to I. Cheltsov and c. Shramov's article "Log canonical thresholds of smooth Fano threefolds". *Uspekhi Mat. Nauk* **63** (2008), no. 5(383), 165–172 Zbl 1167.14024 MR 2484031

[36] J.-P. Demailly and J. Kollár, Semi-continuity of complex singularity exponents and Kähler–Einstein metrics on Fano orbifolds. *Ann. Sci. École Norm. Sup. (4)* **34** (2001), no. 4, 525–556 Zbl 0994.32021 MR 1852009

[37] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Jones and Bartlett Publishers, Boston, MA, 1993 Zbl 0793.60030 MR 1202429

[38] S. K. Donaldson, Kähler metrics with cone singularities along a divisor. In *Essays in Mathematics and its Applications*, pp. 49–79, Springer, Heidelberg, 2012 Zbl 1326.32039 MR 2975584

[39] S. K. Donaldson, Stability of algebraic varieties and Kähler geometry. In *Algebraic Geometry: Salt Lake City 2015*, pp. 199–221, Proc. Sympos. Pure Math. 97, Amer. Math. Soc., Providence, RI, 2018 Zbl 1446.32018 MR 3821150

[40] V. S. Dotsenko and V. A. Fateev, Four-point correlation functions and the operator algebra in 2D conformal invariant theories with central charge $C \leq 1$. *Nuclear Phys. B* **251** (1985), no. 5-6, 691–734 MR 789026

[41] R. Dujardin, Théorie globale du pluripotentiel équidistribution et processus ponctuels [d'aprés Berman, Boucksom, Witt Nyström, …]. *Astérisque* **422** (2020), no. Séminaire Bourbaki. Vol. 2018/2019. Exposés 1151–1165, 61–107 Zbl 1470.32098 MR 4224632

[42] P. Eyssidieux, V. Guedj, and A. Zeriahi, Singular Kähler–Einstein metrics. *J. Amer. Math. Soc.* **22** (2009), no. 3, 607–639 Zbl 1215.32017 MR 2505296

[43] M. E. Fisher, The nature of critical points. In *Lectures in Theoretical Physics: Volume VII C – Statistical Physics, Weak Interactions, Field Theory*, edited by W. E. Brittin, pp. 1–159, University of Colorado Press, Boulder, CO, 1965

[44] P. J. Forrester and S. O. Warnaar, The importance of the Selberg integral. *Bull. Amer. Math. Soc. (N.S.)* **45** (2008), no. 4, 489–534 Zbl 1154.33002 MR 2434345

[45] Z. Fu and Y. Zhu, Selberg integral over local fields. *Forum Math.* **31** (2019), no. 5, 1085–1095 Zbl 1428.33041 MR 4000580

[46] K. Fujita, On Berman–Gibbs stability and K-stability of $\mathbb{Q}$-Fano varieties. *Compos. Math.* **152** (2016), no. 2, 288–298 Zbl 1372.14039 MR 3462554

[47] K. Fujita, A valuative criterion for uniform K-stability of $\mathbb{Q}$-Fano varieties. *J. Reine Angew. Math.* **751** (2019), 309–338 Zbl 1435.14039 MR 3956698

[48] K. Fujita and Y. Odaka, On the K-stability of Fano varieties and anticanonical divisors. *Tohoku Math. J. (2)* **70** (2018), no. 4, 511–521 Zbl 1422.14047 MR 3896135

[49] J. P. Gauntlett, D. Martelli, J. Sparks, and S.-T. Yau, Obstructions to the existence of Sasaki–Einstein metrics. *Comm. Math. Phys.* **273** (2007), no. 3, 803–827 Zbl 1149.53026 MR 2318866

[50] H. Gillet and C. Soulé, An arithmetic Riemann–Roch theorem. *Invent. Math.* **110** (1992), no. 3, 473–543 Zbl 0777.14008 MR 1189489

[51] R. Godement and H. Jacquet, *Zeta Functions of Simple Algebras*. Lecture Notes in Math. 260, Springer, Berlin, 1972 Zbl 0244.12011 MR 0342495

[52] H. Guenancia and M. Păun, Conic singularities metrics with prescribed Ricci curvature: general cone angles along normal crossing divisors. *J. Differential Geom.* **103** (2016), no. 1, 15–57 Zbl 1344.53053 MR 3488129

[53] J.-i. Igusa, *An Introduction to the Theory of Local Zeta Functions*. AMS/IP Stud. Adv. Math. 14, American Mathematical Society, Providence, RI; International Press, Cambridge, MA, 2000 Zbl 0959.11047 MR 1743467

[54] T. Ishii, Godement–Jacquet integrals on GL$(n, \mathbf{C})$. *Ramanujan J.* **49** (2019), no. 1, 129–139 Zbl 1462.11045 MR 3942298

[55] T. Jeffres, R. Mazzeo, and Y. A. Rubinstein, Kähler–Einstein metrics with edge singularities. *Ann. of Math. (2)* **183** (2016), no. 1, 95–176 Zbl 1337.32037 MR 3432582

[56] M. Jonsson and J. Nicaise, Convergence of p-adic pluricanonical measures to Lebesgue measures on skeleta in Berkovich spaces. *J. Éc. polytech. Math.* **7** (2020), 287–336 Zbl 1430.14056 MR 4077578

[57] M. K.-H. Kiessling, Statistical mechanics of classical particles with logarithmic interactions. *Comm. Pure Appl. Math.* **46** (1993), no. 1, 27–56 Zbl 0811.76002 MR 1193342

[58] S. Klevtsov, X. Ma, G. Marinescu, and P. Wiegmann, Quantum Hall effect and Quillen metric. *Comm. Math. Phys.* **349** (2017), no. 3, 819–855 Zbl 1358.81179 MR 3602817

[59] A. W. Knapp, Local Langlands correspondence: the Archimedean case. In *Motives (Seattle, WA, 1991)*, pp. 393–410, Proc. Sympos. Pure Math. 55, Amer. Math. Soc., Providence, RI, 1994 Zbl 0811.11071 MR 1265560

[60] J. Kollár, Singularities of pairs. In *Algebraic Geometry—Santa Cruz 1995*, pp. 221–287, Proc. Sympos. Pure Math. 62, Amer. Math. Soc., Providence, RI, 1997 Zbl 0905.14002 MR 1492525

[61] J. Kollár, The structure of algebraic varieties. In *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. 1*, pp. 395–419, Kyung Moon Sa, Seoul, 2014 Zbl 1373.14001 MR 3728477

[62] M. Kontsevich and Y. Tschinkel, Non-Archimedean Kähler geometry. To appear

[63] A. Kupiainen, R. Rhodes, and V. Vargas, Integrability of Liouville theory: proof of the DOZZ formula. *Ann. of Math. (2)* **191** (2020), no. 1, 81–166 Zbl 1432.81055 MR 4060417

[64] D. A. Kurtze and M. E. Fisher, Yang-Lee edge singularities at high temperatures. *Phys. Rev. B* **20** (1979), no. 7, 2785–2796

[65] R. P. Langlands, Problems in the theory of automorphic forms. In *Lectures in Modern Analysis and Applications, III*, pp. 18–61, Lecture Notes in Math. 170, Springer, Berlin, 1970 Zbl 0225.14022 MR 0302614

[66] R. P. Langlands, L-functions and automorphic representations. In *Proceedings of the International Congress of Mathematicians (Helsinki, 1978)*, pp. 165–175, Acad. Sci. Fennica, Helsinki, 1980 Zbl 0426.12007 MR 562605

[67] T. D. Lee and C. N. Yang, Statistical theory of equations of state and phase transitions. II. Lattice gas and Ising model. *Phys. Rev. (2)* **87** (1952), 410–419 Zbl 0048.43401 MR 53029

[68] C. Li, G-uniform stability and Kähler–Einstein metrics on Fano varieties. *Invent. Math.* **227** (2022), no. 2, 661–744 Zbl 07470589 MR 4372222

[69] C. Li, G. Tian, and F. Wang, The uniform version of Yau–Tian–Donaldson conjecture for singular Fano varieties. *Peking Math. J.* **5** (2022), no. 2, 383–426 Zbl 1504.32068 MR 4492658

[70] Y. Liu, C. Xu, and Z. Zhuang, Finite generation for valuations computing stability thresholds and applications to K-stability. *Ann. of Math. (2)* **196** (2022), no. 2, 507–566 MR 4445441

[71] F. Luo and G. Tian, Liouville equation and spherical convex polytopes. *Proc. Amer. Math. Soc.* **116** (1992), no. 4, 1119–1129 Zbl 0806.53012 MR 1137227

[72] Y. I. Manin, New dimensions in geometry. In *Workshop Bonn 1984 (Bonn, 1984)*, pp. 59–101, Lecture Notes in Math. 1111, Springer, Berlin, 1985 Zbl 0579.14002 MR 797416

[73] R. Mazzeo and Y. A. Rubinstein, The Ricci continuity method for the complex Monge–Ampère equation, with applications to Kähler–Einstein edge metrics. *C. R. Math. Acad. Sci. Paris* **350** (2012), no. 13-14, 693–697 Zbl 1273.32043 MR 2971382

[74] Y. Odaka, Canonical Kähler metrics and arithmetics: generalizing Faltings heights. *Kyoto J. Math.* **58** (2018), no. 2, 243–288 Zbl 1411.14026 MR 3799703

[75] D. Ruelle, *Statistical Mechanics. Rigorous Results, Reprint of the 1989 Edition*. World Scientific Publishing, River Edge, NJ; Imperial College Press, London, 1999 Zbl 1016.82500 MR 1747792

[76] V. V. Shokurov, Complements on surfaces. *J. Math. Sci. (New York)* **102** (2000), no. 2, 3876–3932 Zbl 1177.14078 MR 1794169

[77] C. Soulé, *Lectures on Arakelov Geometry*. Cambridge Stud. Adv. Math. 33, Cambridge University Press, Cambridge, 1992 Zbl 0812.14015 MR 1208731

[78] J. T. Tate, Fourier analysis in number fields, and Hecke's zeta-functions. In *Algebraic Number Theory (Proc. Instructional Conf., Brighton, 1965)*, pp. 305–347, Thompson, Washington, DC, 1967 MR 0217026

[79] M. Troyanov, Prescribing curvature on compact surfaces with conical singularities. *Trans. Amer. Math. Soc.* **324** (1991), no. 2, 793–821 Zbl 0724.53023 MR 1005085

[80] V. Vargas, Lecture notes on Liouville theory and the DOZZ formula. arXiv:1712.00829

[81] C. N. Yang and T. D. Lee, Statistical theory of equations of state and phase transitions. I. Theory of condensation. *Phys. Rev. (2)* **87** (1952), 404–409 Zbl 0048.43305 MR 53028

[82] S. T. Yau, On the Ricci curvature of a compact Kähler manifold and the complex Monge–Ampère equation. I. *Comm. Pure Appl. Math.* **31** (1978), no. 3, 339–411 Zbl 0369.53059 MR 480350

[83] S. Zhang, Positive line bundles on arithmetic varieties. *J. Amer. Math. Soc.* **8** (1995), no. 1, 187–221 Zbl 0861.14018 MR 1254133

[84] S. Zhang, Heights and reductions of semi-stable varieties. *Compositio Math.* **104** (1996), no. 1, 77–105 Zbl 0924.11055 MR 1420712

Robert J. Berman
Mathematical Sciences, Chalmers University of Technology and the University of
Gothenburg, 41296 Göteborg, Sweden; robertb@chalmers.se

DOI 10.4171/8ECM/01

Variational regularization in inverse problems and machine learning

Martin Burger

Abstract. This paper discusses basic results and recent developments on variational regularization methods, as developed for inverse problems. In a typical setup we review basic properties needed to obtain a convergent regularization scheme and further discuss the derivation of quantitative estimates respectively the needed ingredients such as Bregman distances for convex functionals.

In addition to the approach developed for inverse problems, we will also discuss variational regularization in machine learning and work out some connections to the classical regularization theory. In particular we will discuss a reinterpretation of machine learning problems in the framework of regularization theory and a reinterpretation of variational methods for inverse problems in the framework of risk minimization. Moreover, we establish some previously unknown connections between error estimates in Bregman distances and generalization errors.

1. Introduction

Regularization methods are an approach of fundamental importance in the solution of ill-posed problems. Their main paradigm is to approximate an ill-posed problem by a parametrized family of well-posed problems, with appropriate convergence properties as the regularization parameter and the so-called noise level tend to zero. The noise level is a measure for the size of deterministic and stochastic errors in the data, which are usually the main cause of concern due to the ill-posedness.

A detailed theory of regularization has been developed in the typical setting of inverse problems, obviously with more precise results in the case of linear forward models than for nonlinear ones (cf. [3, 20, 24, 54, 56] and references therein). Regularization is however not only relevant in inverse problems, similar methods are now routinely used in machine learning, mainly from a practical point of view, with theoretical results often hidden in the statistical theory of generalization (cf. e.g. [27, 33, 40]). The role and objective of regularization is less clear and less developed

2020 Mathematics Subject Classification. Primary 65J20; Secondary 47A52.
Keywords. Regularization theory, variational methods, inverse problems, machine learning.

in the machine learning domain. In this paper we will thus aim to give a unified overview and present some links between the formulations and questions in inverse problems and those in machine learning. We will concentrate on the prominent class of variational regularization methods, which we interpret in a rather broad way.

2. Regularization theory

In order to present the basic ideas of regularization methods in a rather unified way for inverse and machine learning problems, we will first adopt a high-level point of view. Regularization theory is based on the following ingredients.

- An *ideal problem* respectively an *ideal solution* u^*. We can assume that the ideal problem is given by a map $\Phi : \mathcal{V}_D \to \mathcal{U}$, where $\mathcal{V}_D$ is the space of ideal data and $\mathcal{U}$ is the space of admissible solutions. The typical analysis is confined to Banach or at least metric spaces.
- A space $\mathcal{V} \supset \mathcal{V}_D$ of *possible data* and a measure of noise between the ideal data $v^* = \Phi^{-1}(u^*) \in \mathcal{V}_D$ and noisy data $v \in \mathcal{V}$. In the case of an ill-posed problem, the operator Φ is not continuous when considered from (a subset of) $\mathcal{V}$ to $\mathcal{U}$; it may be continuous on bounded subsets of $\mathcal{V}_D$ however. The latter leads to the concept of conditional stability (cf. [57,58]) and corresponding stability estimates.
- A family of continuous, possibly multivalued, maps $\Phi_\alpha : \mathcal{V} \to \mathcal{P}(\mathcal{U})$, $\alpha \in \mathcal{A}$, such that for a sequence $(v_n) \subset \mathcal{V}$ converging to $v^* \in \mathcal{V}_D$, there exists a parameter sequence α_n such that there is $u_n \in \Phi_{\alpha_n}(v_n)$ converging to u^* (in a suitable metrizable topology, possibly weak or weak-star on bounded sets in the Banach space case). Sometimes the notion of convergence is restricted to subsequences.

To make these notions more concise we will discuss them in the setting of inverse problems as well as machine learning subsequently.

2.1. Inverse problems

In the typical case of inverse problems, there is first a (continuous) forward operator $F : \mathcal{U} \to \mathcal{V}$, which is typically not invertible and if it is on a subset of $\mathcal{V}$, the inverse is discontinuous. The set of ideal data is a subset of $F(\mathcal{U})$, and there the multivalued operator

$$\Phi_0 : \mathcal{V}_D \to \mathcal{P}(\mathcal{U}), \quad v \mapsto F^{-1}(v)$$

can be defined. In order to obtain a unique (generalized) inverse, a further selection operator $\Sigma : \mathcal{P}(\mathcal{U}) \to \mathcal{U}$ is defined to obtain $\Phi := \Sigma \circ \Phi_0$. Let us mention that there are standard examples of the selection operator such as the minimum norm solution, but often this issue is treated in a hidden or unprecise way. We refer to [3] for a detailed discussion of selection operators in inverse problems.

The standard notion of noise is the perturbation of the data, i.e. $v - v^*$, either as a deterministic or a stochastic quantity. The norm of $v - v^*$ in the Banach space $\mathcal{V}$ (or the expectation of some power of the norm) serves as a definition of the noise level.

The solution of the inverse problem can then be cast as the solution of the ill-posed operator equation

$$F(u) = v$$

or as the minimization of

$$D(u) = L\big(F(u), v\big), \tag{2.1}$$

where L is some distance measure between the predicted data $F(u)$ and the measured data v. If statistical information about the noise is available or the forward model contains other stochastic elements, L is typically a negative log-likelihood functional.

As mentioned above, regularization methods are families of multivalued operators $\Phi_\alpha : \mathcal{V} \to \mathcal{P}(\mathcal{U})$; in most cases the parameter domain $\mathcal{A}$ is a subset of the positive real numbers. The well-posedness of Φ_α is characterized by some set-valued continuity, e.g. if $u_n \to u$, then $\Phi_\alpha(u_n)$ contains a convergent subsequence and each limit v of a convergent subsequence satisfies $v \in \Phi_\alpha(u)$. In most cases the regularization operator satisfies a stronger stability estimate of the form

$$d_U(u_1, u_2) \leq C_\alpha d_V(v_1, v_2) \quad \forall u_1 \in \Phi_\alpha(v_1),\ u_2 \in \Phi_\alpha(v_2), \tag{2.2}$$

where d_U and d_V are appropriate distance measures (that may be degenerate in the sense that $d_U(u_1, u_2)$ can vanish also if $u_1 \neq u_2$).

Regularization methods are constructed along several different paradigms.

- Data smoothing or mollifier methods, which are of the form $\Phi_\alpha = F^{-1} \circ M_\alpha$, where $M_\alpha : \mathcal{V} \to \mathcal{V}_D$ is a family of mollifying (smoothing) operators into an appropriate subspace of $\mathcal{V}$ on which there exists a continuous inverse of F. In order to obtain suitable regularization methods, a quite detailed characterization of the forward operator is needed in order to be sure to construct a mollification to the right subspace. Consequently, such methods became popular for inverse problems with well-understood forward operators such as tomography (cf. [43,44]).
- Direct approximation of the operator F by continuously invertible operators (cf. [24,37,41,56] and references therein). The construction of approximations is usually done only in the case of linear forward operators based on modifying (small) singular vectors or by approximating the normal equation, i.e. F^*F. The latter is however related to the minimization of the least-squares function $\|F(u) - v\|^2$ and can thus be viewed as a variational method. Another approach modifying the forward operator is discretization, the regularization parameter thus being related to the discretization fineness.

- Variational methods are based on a perturbation of the likelihood minimization, with Φ_α mapping v to the set of minimizers of

$$D_\alpha(u) = L\big(F(u), v\big) + \alpha J(u)$$

  for some regularization functional J that introduces the needed compactness properties for the existence of minimizers and $\alpha \in \mathbb{R}_+$ being the regularization parameter (cf. [3, 54]).

- Iterative regularization methods use a well-defined iteration method such as a fixed-point iteration or some descent scheme for the likelihood minimization to define an approximation of the inverse of F, with the iteration number $\alpha \in \mathbb{N}$ being the regularization parameter (cf. [14, 24, 34, 35, 49]). Since the majority of iterative methods, in particular in the nonlinear case, are iterative methods for variational problems, there is an intimate connection to variational regularization methods.

- Learned regularization methods are of increasing relevance recently (cf. [1,3] and references therein), which are categorized into supervised and semi-supervised approaches. The supervised approach tries to learn the regularization operator Φ_α directly from a collection of pairs of training data (u_i, v_i), e.g. by approximation with a deep neural network. Consistent data pairs are however difficult to obtain in many inverse problems, in particular with realistic input data u_i and realistic noise in v_i. The alternative semi-supervised approach mainly works on suitable solutions u_i, e.g. images for reconstruction tasks, and tries to learn a more conventional regularization approach, e.g. the regularization functional J in variational regularization methods. With certain restrictions such as convex networks those become accessible for theoretical arguments of regularization theory.

Besides providing a well-posed problem for fixed α, which often requires some advanced analysis itself (e.g. existence of minimizers for variational problems), a major goal of regularization theory is to study the convergence of regularized solutions. While a qualitative convergence theory can be developed under generic conditions, it is well known that a quantitative theory will rely on additional assumptions on the ideal solution u^* due to the underlying ill-posedness. To understand the possibility to derive such estimates and the used assumptions from a generic point of view, let us consider a sequence of data $v_n \to v^*$ and a parameter choice α_n, assuming that α_n is a nonnegative scalar sequence converging to zero (e.g. the regularization parameters in a variational regularization method or $\alpha_n = \frac{1}{k_n}$ with k_n the maximal iteration number in an iterative regularization method). Now assume that the stability estimate (2.2) holds and that u^* satisfies a *range condition* for the regularization operator (cf. [3]).

Definition 2.1. An element $u^* \in \mathcal{U}$ is said to satisfy a *range condition for the regularization operator* Φ_α if for all α there exists v_α^* such that $u^* \in \Phi_\alpha(v_\alpha^*)$.

Under a range condition we can write

$$u_n - u^* \in \Phi_{\alpha_n}(v_n) - \Phi_{\alpha_n}(v_{\alpha_n}^*)$$

and exploit the stability estimate (2.2) to obtain

$$d_U(u^*, u_n) \leq C_{\alpha_n} d_V(v_{\alpha_n}^*, v^*).$$

Thus, if we can control the range condition in the sense that we can construct an element $v_{\alpha_n}^*$ out of v^* such that the distance can be estimated, we directly obtain an error estimate. This will be made more precise in the next section on variational regularization methods.

2.2. Learning and risk minimization

In the typical case of machine learning problems (cf. [33,45]) we are given (randomly sampled) input samples $x_i \in \mathcal{X}$ and output samples $y_i \in \mathcal{Y}, i = 1, \dots, N$, and want to infer a parametrized map $f_\theta : \mathcal{X} \to \mathcal{Y}$ reasonably reproducing these training data and generalizing further to other data of the same kind. These properties are frequently obtained from risk minimization arguments. Given a loss ℓ measuring deviations in the output space, the empirical risk is given by

$$\widehat{R}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(x_i), y_i\big)$$

and approximate solutions are constructed as approximate minimizers of $\widehat{R}$, e.g. via variational regularization methods minimizing

$$D_\alpha(\theta) = \widehat{R}(\theta) + \alpha J(\theta)$$

or by iterative methods such as the gradient descent

$$\theta^{k+1} = \theta^k - \tau^k \widehat{R}'(\theta^k)$$

or even more often by stochastic gradient descent, where the term implicit regularization is common (cf. [48]).

Generalization is usually measured by the behavior on the population risk, i.e.

$$R(\theta) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f_\theta(x), y\big)\big);$$

in particular the generalization error defined by

$$G(\theta) = R(\theta) - \widehat{R}(\theta),$$

evaluated at a regularized solution. Note that the generalization error is actually a random variable depending on the samples $(x_i, y_i)_{i=1}^N$, hence it is relevant to consider its distribution among the random sampling.

The ideal model could be defined in two ways, depending on what variable is identified to be the relevant one. In any case the ideal solution is perceived as a minimizer of the population risk, however one could define u^* as the optimal parameter value or the optimal function. Thus we are led to the following cases.

(i) The first case, corresponding to classical approaches in statistics such as regression, is to define $\mathcal{U}$ as the set of possible parameters, genuinely a finite-dimensional space (with few generalizations to infinite-dimensional models recently, cf. [39, 47]). Thus, the ideal solution is given by

$$\theta^* \in \arg\min_{\theta \in \mathcal{U}} R(\theta).$$

(ii) The second case rather corresponds to the perspective of modern learning theory; it extends the population risk to some function class $\mathcal{F}$ and computes for $f \in \mathcal{F}$

$$S(f) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f(x), y\big)\big).$$

The ideal solution is given by

$$f^* \in \arg\min_{f \in \mathcal{F}} S(f).$$

Another obvious question in this case is how to define the ideal and perturbed data. We follow a distributional viewpoint and define the ideal data v^* as the data distribution $\mathbb{P}$. Correspondingly, the perturbed data are given by the empirical distribution

$$\mathbb{P}^N = \frac{1}{N}\sum_{i=1}^N \delta_{(x_i, y_i)},$$

where δ_z denotes the concentrated measure at z. Thus, the noise level becomes a distance between (probability) distributions, standard distances such as the total variation distance or Wasserstein metrics.

The regularization operator Φ_α maps from a space of probability distributions to (a set of) regularized solutions. Take the variational regularization of minimizing D_α as an example. Then in case (i), Φ_α is given by

$$\Phi_\alpha : \mathbb{P}^N \mapsto \arg\min_\theta D_\alpha(\theta),$$

while in the second case (ii) we have

$$\Phi_\alpha : \mathbb{P}^N \mapsto \bigl\{ f_\theta \mid \theta \in \arg\min_\theta D_\alpha(\theta) \bigr\}.$$

We finally mention that these models can obviously be generalized, in particular to the case of further data errors in the samples (x_i, y_i). Then the samples can be considered to be drawn from a distribution $\mathbb{P}'$ and the effective error is not just determined by sampling but also by the distance of $\mathbb{P}$ and $\mathbb{P}'$.

Thus, we see that regularized learning problems can be reformulated in the language of regularization theory for inverse problems (see also [12,53]). In turn we will see that many inverse problems can be reformulated as risk minimization problems, in particular if there is additional sampling of measurement points.

2.3. Risk minimization formulation of inverse problems

Many inverse problems are dealing with data being functions of a variable x, e.g. in integral equations of the first kind or tomography, where x is a set of distances and angles (cf. [46]). Denoting the unknown of the inverse problem by θ, we thus obtain $F(\theta)$ as a function of x and denote $f(x;\theta) = F(\theta)(x)$. Moreover, standard log-likelihood functionals in this setting are of the form

$$L\bigl(F(\theta), v\bigr) = \int_\Omega \ell\bigl(F(\theta)(x), v(x)\bigr)\, dx$$

for some function ℓ. Thus, choosing $\mathcal{P} = \mathcal{L}_\Omega \delta_{v(x)}$, where $\mathcal{L}_\Omega$ denotes the Lebesgue measure on Ω, we obtain

$$L\bigl(F(\theta), v\bigr) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\bigl(\ell\bigl(F(\theta)(x), y\bigr)\bigr) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\bigl(\ell\bigl(f(x;\theta), y\bigr)\bigr).$$

The ideal problem is thus the minimization of the loss for appropriate data v^*.

In a practical setting we have a finite sampling of data with additional noise, which we consider to be additive for simplicity in the following. This means the practical data are a finite number N of samples $y_i = F(\theta)(x_i) + n_i$, where n_i are the noise samples drawn from some distribution. The practical distribution of samples and data is of the form

$$\mathbb{P}^N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i} \otimes \delta_{F(\theta^*)(x_i)+n_i},$$

where the x_i are drawn from a prior distribution (usually a deterministic one) and the n_i are drawn from the noise distribution.

Example 2.2. As a simple example consider the inversion of the Radon transform on a domain $\Omega \subset \mathbb{R}^2$. Then in the standard parametrization we can choose $x \in [0,\pi) \times$

$[0, L]$ as the angle and distance to origin of the lines to be integrated on. Correspondingly $F(\theta)(x)$ is the line integral of the density function θ on the line parametrized by x. Now let x be drawn from the uniform distribution on $[0, \pi) \times [0, L]$ and let each n be drawn from a Gaussian distribution G_σ with zero mean and finite variance. Then the population risk becomes

$$\begin{aligned} R(\theta) &= \frac{1}{2L\pi} \int_{[0,\pi)\times[0,L]} \int_{\mathbb{R}} \left|F(\theta)(x) - F(\theta^*)(x) - n\right|^2 dG_\sigma(n)\, dx \\ &= \frac{1}{2L\pi} \int_{[0,\pi)\times[0,L]} \left|F(\theta)(x) - F(\theta^*)(x)\right|^2 dx + \int_{\mathbb{R}} n^2\, dG_\sigma(n). \end{aligned}$$

Hence, after affine transform with terms independent of θ, the population risk equals the squared L^2-distance of the Radon transforms of θ and θ^*, which is the usual data discrepancy L. The empirical risk on the other hand is of the form

$$\widehat{R}(\theta) = \frac{1}{2N} \sum_{i=1}^{N} \left|F(\theta)(x_i) - y_i\right|^2,$$

which is the standard functional minimized in practice.

For a more general noise model one may construct the conditional distribution for y based on using the appropriate push-forward of the noise distribution based on applying the noise to $F(\theta^*)(x)$ and an appropriately chosen loss function. Moreover, errors in the forward model could be included in the stochastic model, which will imply that even in the ideal model the conditional distribution of y given x is not concentrated.

3. Variational regularization

In the following we present some key steps in the analysis of iterative regularization methods, for the sake of a simpler presentation restricting ourselves to a linear forward model and a quadratic data fidelity in a Hilbert space, i.e.

$$D_\alpha(u) = \frac{1}{2}\|Fu - v\|^2 + \alpha J(u), \tag{3.1}$$

where $J : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ is assumed to be convex and proper. Moreover, we assume $\mathcal{V}$ to be a Hilbert space and $\mathcal{U}$ a Banach space being the dual of some Banach space $\mathcal{W}$, with the additional property that the weak-star topology on $\mathcal{U}$ is metrizable on bounded sets. The operator $F : \mathcal{U} \to \mathcal{V}$ is assumed to be bounded and the adjoint of a bounded linear operator $E : \mathcal{V} \to \mathcal{W}$. With abuse of notation we shall write

$F^* = E$. Finally, we need some additional property of the regularization functional; we assume that it is the convex conjugate of some other functional $H : \mathcal{W} \to \mathbb{R}$, i.e.

$$J(u) = \sup_{w \in \mathcal{W}} \langle u, w \rangle - H(w).$$

Let us mention that convex conjugates are weak-star lower semicontinuous, which is obviously an important property of the functional and can be inferred by similar arguments as the weak lower semicontinuity results in [23]. Finally, a coercivity property is needed to apply weak-star compactness arguments (based on the Banach–Alaoglu theorem); we assume that the sublevel sets

$$M_C = \{u \in \mathcal{U} \mid J(u) \leq C\}$$

are bounded in $\mathcal{U}$ for $C > 0$. The final property we need is that J is bounded below; we can assume directly that J is nonnegative.

There are various important examples in literature motivating the above model and assumptions. A popular and reasonably easy to compute approach is classical Tikhonov–Phillips regularization with $\mathcal{U}$ being a Hilbert space and

$$J(u) = \frac{1}{2}\|u\|^2.$$

Possibly the most prominent example with a variety of applications is total variation regularization (cf. [16, 19]), i.e. $\mathcal{U} = BV(\Omega)$ and

$$J(u) = \sup_{g \in C_0^\infty(\Omega)^d,\, \|g\|_\infty \leq 1} \int_\Omega u \nabla \cdot g \, d\mathcal{L}_\Omega,$$

where $\Omega \subset \mathbb{R}^d$ is the domain on which the function to be reconstructed is defined. There are various variants of total variation, including higher order versions, which received considerable attention. Another class of important regularization methods are sparsity-enforcing priors (cf. [50]), in the simplest setup $\mathcal{U} = \ell^1$ and

$$J(u) = \sum |u_i|.$$

An interesting case in deconvolution problems as well as mean-field approaches to learning with neural networks is the continuum variant, the total variation norm of Radon measures (cf. [7, 22]). Here we have $\mathcal{U} = \mathcal{M}(\Omega)$ and

$$J(u) = \sup_{w \in C_0(\Omega)} \int_\Omega w \, du.$$

3.1. Basic properties of variational regularization methods

A key result, often found for special cases in literature (cf. e.g. [16, 55]), is the existence of a minimizer and some stability, which verifies the well-posedness of the regularization operator $\Phi_\alpha(v) := \arg\min_u D_\alpha(u)$.

Theorem 3.1. *Under the above assumptions on $\mathcal{U}$, $\mathcal{V}$, F, and J there exists a minimizer of $D_\alpha(u)$ for every $v \in \mathcal{V}$ and every $\alpha > 0$. Moreover, if $\alpha > 0$, $v_n \to v$, and $u_n \in \Phi_\alpha(v_n)$, then there exists a weak-star convergent subsequence v_{n_k} and the limit u of every weak-star convergent subsequence satisfies $u \in \Phi_\alpha(v)$.*

In general, no uniqueness can be shown under the above conditions, which is anyway not to be expected for the rather degenerate examples above. However, a weaker type of uniqueness can be inferred from the convexity and optimality condition

$$F^*(Fu - v) + \alpha p = 0, \quad p \in \partial J(u),$$

where $\partial J(u)$ denotes the subdifferential

$$\partial J(u) = \{ w \in \mathcal{U}^* \mid J(u) + \langle w, \tilde{u} - u \rangle \leq J(\tilde{u}) \ \forall \tilde{u} \in \mathcal{U} \}.$$

From the assumptions on F we see that F^* effectively maps to the predual space $\mathcal{W}$, thus the subgradients in the optimality condition effectively satisfy $p \in \mathcal{W}$, which is a weak regularity condition. A key concept needed in the following is the Bregman distance or generalized Bregman distance (cf. [11, 38]).

Definition 3.2. Let $J : \mathcal{U} \to \mathbb{R} \cup \{+\infty\}$ be a convex proper functional and let $u, \tilde{u} \in \mathcal{U}$ with $p \in \partial J(u)$. Then the (generalized) Bregman distance $d_J^p(\tilde{u}, u)$ is defined by

$$d_J^p(\tilde{u}, u) = J(\tilde{u}) - J(u) - \langle p, \tilde{u} - u \rangle.$$

If $\tilde{p} \in \partial J(\tilde{u})$, the symmetric Bregman distance $d_J^{\tilde{p},p}(\tilde{u}, u)$ is defined by

$$d_J^{\tilde{p},p}(\tilde{u}, u) = \langle \tilde{p} - p, \tilde{u} - u \rangle.$$

Now assume that there are two minimizers u_1 and u_2 of the variational regularization problem, then the difference in optimality conditions yields

$$F^*F(u_1 - u_2) + \alpha(p_1 - p_2) = 0$$

and from a duality product with $u_1 - u_2$ we infer

$$\left\| F(u_1 - u_2) \right\|^2 + \alpha d_J^{p_1,p_2}(u_1, u_2) = 0.$$

Hence, by the nonnegativity of both terms we obtain uniqueness of the output value, i.e., $Fu_1 = Fu_2$ as well as a vanishing symmetric Bregman distance between u_1 and u_2.

Finally, we can turn our attention to convergence properties of the regularization method. For this sake we use an exposition based on Γ-convergence (cf. [6]).

Lemma 3.3. *Let $v_n \to v^* = Fu^*$ in $\mathcal{V}$ and $\alpha_n \to 0$. Then the sequence of functionals D_{α_n} defined by*

$$D_{\alpha_n}(u) = \frac{1}{2}\|Fu - v_n\|^2 + \alpha_n J(u)$$

Γ-converges to

$$D_0(u) = \frac{1}{2}\|Fu - v^*\|^2$$

with respect to the weak-star topology in $\mathcal{U}$.

This kind of convergence is not strong enough to infer convergence of minimizers, in particular since there is no equicoercivity property. To achieve this, we need to rescale the functional, i.e. use Γ-convergence by development to the next order.

Lemma 3.4. *Let $v_n \to v^* = Fu^*$ in $\mathcal{V}$ and $\alpha_n \to 0$ such that*

$$\frac{\|v_n - v^*\|^2}{\alpha_n} \to 0.$$

Then the sequence of functionals E_{α_n} defined by

$$E_{\alpha_n}(u) = \frac{1}{2\alpha_n}\|Fu - v_n\|^2 + J(u)$$

Γ-converges to

$$E_0(u) = \begin{cases} J(u) & \text{if } Fu = v^*, \\ +\infty & \text{else}, \end{cases}$$

with respect to the weak-star topology in $\mathcal{U}$.

Let us mention that we obtain divergence, i.e. E_{α_n} converges to the functional identically equal to $+\infty$, if the condition on the parameter choice is violated, i.e. $\liminf \frac{\|v_n - v^*\|^2}{\alpha_n} > 0$. Since $E_\alpha \geq J$ and J is coercive, we immediately conclude the equicoercivity of the sequence E_{α_n}.

Corollary 3.5. *Let $v_n \to v^* = Fu^*$ in $\mathcal{V}$ and $\alpha_n \to 0$ such that*

$$\frac{\|v_n - v^*\|^2}{\alpha_n} \to 0.$$

Moreover, let u_n be a sequence of minimizers of D_{α_n} (or equivalently E_{α_n}), then there exists a subsequence converging with respect to the weak-star topology in $\mathcal{U}$ and the limit u^{**} of each weakly convergent subsequence is a minimizer of E_0. Moreover, $J(u_n) \to J(u^{**})$.

Corollary 3.5 confirms that indeed the regularization operator defined by

$$\Phi_\alpha(v) = \arg\min_u D_\alpha(u)$$

yields a convergent regularization. Let us mention some further direct consequences.

- If the J-minimizing solution is unique, i.e. u^{**} is the unique minimizer of E_0, then the whole sequence u_n converges weakly-star to u^{**}. Moreover, if there is $p^{**} \in \partial J(u) \cap \mathcal{W}$, then due to the convergence of J and the weak-star convergence we conclude
$$d_J^{p^{**}}(u_n, u^{**}) \to 0.$$
- If u^* satisfies $Fu^* = v^*$, but is not a J-minimizing solution (a minimizer of E_0), it cannot be reconstructed by the regularization method, i.e. it is not the limit of minimizers of the variational regularization for positive α. This is related to the question whether the regularization functional introduces the right type of prior knowledge. If we are interested in reconstructing a solution like u^* that is not J-minimizing, then J is not a suitable choice.
- If J is the norm in $\mathcal{U}$ as in many frequent examples and $\mathcal{U}$ satisfies a Radon-Riesz property, the previous result indeed implies a strong convergence of subsequences.

The above analysis was based on a deterministic approach, but in a similar way a stochastic theory can be developed, e.g. for a sequence of random variables v_n with variance $\mathbb{E}(\|v_n - v^*\|^2)$ converging to zero.

3.2. Quantitative estimates

As mentioned above, it is important to derive quantitative estimates between solutions of the regularized problem and ideal solutions, which we present here based on using range conditions as sketched above. In the following we denote by u_α a regularized solution, i.e. a minimizer of D_α. Due to convexity $u_\alpha \in \Phi_\alpha(v)$ is characterized as the solution of the optimality condition

$$F^*(Fu_\alpha - v) + \alpha p_\alpha = 0, \quad p_\alpha \in \partial J(u_\alpha).$$

Taking two such solutions one can establish a stability estimate for the Bregman distance (cf. [3]).

Theorem 3.6. *Let $u_\alpha \in \Phi_\alpha(v)$ and $\tilde{u}_\alpha \in \Phi_\alpha(\tilde{v})$. Then the estimate*

$$\frac{1}{2}\|Fu_\alpha - F\tilde{u}_\alpha\|^2 + \alpha d_J^{p_\alpha,\tilde{p}_\alpha}(u_\alpha, \tilde{u}_\alpha) \leq \frac{1}{2}\|v - \tilde{v}\|^2$$

holds, where p_α respectively $\tilde{p}_\alpha$ are the subgradients appearing in the optimality condition for u_α respectively $\tilde{u}_\alpha$.

Now we turn to the range condition, effectively reformulating a result from [15].

Lemma 3.7. *An element $u^* \in \mathcal{U}$ with $v^* = Fu^*$ satisfies the range condition for the variational regularization operator Φ_α if and only if it satisfies the* source condition

$$\exists z^* \in \mathcal{V} : F^*z^* \in \partial J(u^*).$$

The key part of the proof is the explicit construction $v_\alpha^* = v^* + \alpha z^*$, which allows to obtain an estimate of the right-hand side in the error estimate, due to

$$\|v - v_\alpha^*\| \leq \|v - v^*\| + \|v^* - v_\alpha^*\| = \|v - v^*\| + \alpha\|z^*\|.$$

This leads to the error estimates as derived in [15].

Corollary 3.8. *Let $u_\alpha \in \Phi_\alpha(v)$ and let $v^* = Fu^*$, with u^* satisfying the source condition $p^* = F^*z^* \in \partial J(u^*)$. Then the estimate*

$$\frac{1}{2}\|Fu_\alpha - Fu^*\|^2 + \alpha d_J^{p_\alpha,p^*}(u_\alpha, u^*) \leq \|v - v^*\|^2 + \alpha^2\|z^*\|^2.$$

In the error estimate we see again the condition on the choice of α needed for the convergence of regularization methods. While the estimate on the output error $\|Fu_\alpha - Fu^*\|$ is uniform in α; the effective estimate for the Bregman distance is of the form

$$d_J^{p_\alpha,p^*}(u_\alpha, u^*) \leq \frac{\|v - v^*\|^2}{\alpha} + \alpha\|z^*\|^2,$$

which is small again only if α and the quotient $\frac{\|v-v^*\|^2}{\alpha}$ are small.

One also observes a bias-variance decomposition inherent in the estimate, even more clearly when we assume an underlying stochastic noise model, i.e., v is a random variable. Without systematic errors in the measurements, we have $\mathbb{E}(v) = v^*$ and hence

$$\mathbb{E}\big(d_J^{p_\alpha,p^*}(u_\alpha, u^*)\big) \leq \frac{\mathbb{E}\big(\|v - v^*\|^2\big)}{\alpha} + \alpha\|z^*\|^2.$$

The measure on the left-hand side is the natural generalization of the mean-squared error to the case of convex variational regularization, and the right-hand side is composed of the data variance and the bias term $\|z^*\|^2$, scaled by the regularization parameter.

Let us mention that the above estimates in Bregman distances lead to estimates in norms if J satisfies strong convexity conditions (cf. [54]). In the case of not strictly convex functionals, the Bregman distance can vanish even if $u_\alpha \neq u^*$, e.g. in total variation regularization if they differ by a change of contrast $u_\alpha = h(u)$ with a monotone function h, but rather measures a deviation of the discontinuity sets (cf. [3, 16]). In such cases the multivaluedness of the subdifferential can even be an advantage that needs to be exploited, since we do not have just a single estimate, but actually an estimate for each p^* satisfying a source condition. Estimates for other quantities can then be derived from the Bregman distance estimates by optimizing over the possible p^* and the associated source elements z^* (respectively, their norm appearing in the error estimates). An example are estimates for the total variation regularization for piecewise constant functions; it has been shown already in [15] how the total variation of u_α away from the discontinuity set of u^* can be estimated by choosing appropriate subgradients.

Again the above type of conditions and estimates are the canonical ones, but can be developed much farther (cf. e.g. [2, 25, 26, 28, 30–32, 51, 52, 57]). The first issue is the question of having better estimates under stronger conditions, and a typical example is an improved source condition $p^* = F^*F\eta^* \in \partial J(u^*)$ for some $\eta^* \in \mathcal{U}$. In this case the element η^* can be used to construct an approximate solution $u_\alpha^* = u^* - \alpha\eta^*$ instead of approximate data for a range condition. This was carried out in [51] (see also [29]) to obtain the estimate

$$d_J^{p^*}(u_\alpha, u^*) \leq d_J^{p^*}(u^* - \alpha\eta^*, u^*) + \frac{\|v - v^*\|^2}{2\alpha}.$$

The exact characterization of $d_J^{p^*}(u^* - \alpha\eta^*, u^*)$ depends on the properties of the functional and may be on u^* itself. For J being Fréchet-differentiable with Lipschitz-continuous (or Hölder-continuous) derivative, it is always quadratic in α; hence the estimate is of higher order in α. For the nonsmooth functionals like total variation or the ℓ^1-norm the situation is different; at a first glance it cannot be expected that $d_J^{p^*}(u^* - \alpha\eta^*, u^*)$ is of higher order in α. However, in such situations we can even have $d_J^{p^*}(u^* - \alpha\eta^*, u^*) = 0$ for α small, e.g. in ℓ^1 regularization if the support of η^* is contained in the support of u^*.

The opposite question of weaker estimates arises if u^* does not satisfy the source condition $p^* = F^*z^*$. In this case approximate source conditions are used, which measure the deviation from the source condition. A frequently used concept is the so-called *distance function*

$$D_\rho(p^*) = \inf\left\{\|F^*z - p^*\| \mid z \in \mathcal{V},\ \|z\| \leq \rho\right\},$$

which is useful in particular under strong convexity assumptions and allows to build a theory in a similar way by optimizing the value ρ that finally appears in the error estimate. For functionals not being strictly convex and in particular the one-homogeneous cases like total variation, a reformulation in terms of a dual problem is more suitable as seen in [13]. There the measure

$$e_{\alpha,\nu}(p^*) = \inf_{z \in \mathcal{V}} \nu J^*\left(\frac{F^*z - p^*}{\nu}\right) + \frac{\alpha}{2}\|z\|^2$$

was used to derive estimates. One observes some duality to the concept of distance functions, noticing that for J being a norm in Banach space we just have

$$e_{\alpha,\nu}(p^*) = \alpha \inf\left\{\|z\| \mid z \in \mathcal{V},\ \|F^*z - p^*\|_* \leq \nu\right\},$$

where $\|\cdot\|_*$ is the dual norm to J. It was also shown that approximate source conditions are inherently related to the case of large noise, which is particularly relevant for stochastic models like white noise having non-finite variance (cf. [5, 13, 36]).

While the literature was focused on asymptotic results for a long time, the specific shape of solutions at a fixed positive α became a more attractive topic in the last two decades. In order to understand this issue, a better understanding of the range condition for the regularization method is needed, which means the source condition $p^* = F^*z^*$ in the case of variational regularization. Since F is modeled as a smoothing operator in inverse problems, F^* is smoothing as well, which implies that the source condition is an abstract smoothness condition. However, the smoothness is rather indirect, since it concerns the subgradient p^* and not directly u^*. Various results on the structure of minimizers, from sparsity properties for $J = \ell^1$ or its counterpart in the space of measures to total variation and staircasing phenomena can be found in literature (cf. [18, 19]).

Another issue that found strong recent interest is debiasing, since in the case of large noise the bias caused by the regularization term (and the large value of α that is needed to achieve stability) spoils the possible quality of regularized solutions. The influence of bias can also be seen from the term depending on $\|z^*\|$ in the error estimates, and in practice it is often observed that the reconstruction of the subgradient is better than the one of the primal solution due to bias. First debiasing methods (also called refitting) appeared in ℓ^1 regularization, where in a first step the variational regularization is used and in a second step a simple least-squares problem is used on the support obtained from the first step, sometimes also with a sign constraint as obtained from the subgradient in the first step (cf. [21, 42]). This approach can be translated to a more general two-step approach for debiasing as worked out in [8], which computes

$$\Phi_\alpha(v) = \arg\min\left\{d_J^{p_\alpha}(u, u_\alpha) \mid u_\alpha \in \Phi_\alpha^0(v)\right\},$$

with Φ_α^0 being the regularization operator from the variational regularization method.

Another approach effectively leading to debiasing, but also with other advantages, are iterative regularization methods such as the Bregman iteration (cf. [49]). In the case of a quadratic functional, it can be formulated as an augmented Lagrangian method for computing the J-minimizing solution of $Fu = v$, i.e.

$$\begin{aligned} u^{k+1} &\in \arg\min_u \frac{1}{2}\|Fu - v^k\|^2 + \alpha J(u), \\ v^{k+1} &= v^k + v - Fu^{k+1}, \end{aligned}$$

with $v^0 = v$. To have a suitable generalization also for other loss functionals this can be reformulated as

$$\begin{aligned} u^{k+1} &\in \arg\min_u \frac{1}{2}\|Fu - v\|^2 + \alpha d_J^{p^k}(u, u^k), \\ p^{k+1} &= p^k + \frac{1}{\alpha} F^*(v - Fu^{k+1}) \in \partial J(u^{k+1}). \end{aligned}$$

The regularization parameter in this case is not α, which is to be chosen rather larger in order to achieve good results, but the number of iterations carried out. Due to the variational structure in each iteration step, variational methods can be employed to prove well-definedness of the regularization operator, convergence, and error estimates. We refer to [3, 17, 49] for a detailed discussion of such iterative approaches and their analysis. Let us finally mention that in this respect there is another relation to machine learning, since Bregman iterations for ℓ^1 regularizations have been developed further recently for the training of sparse deep neural networks and their architecture design (cf. [9, 10]).

4. Variational regularization and generalization

In this final part we discuss some possible relations between the setup in machine learning and the above results on variational regularization theory. In particular we highlight some connections between the typical error measures used in the two fields, namely generalization errors on the one hand and Bregman distances on the other.

4.1. Error decomposition and generalization error

Let us return to the setup of machine learning with the minimization of the empirical risk with a convex loss ℓ, taking the viewpoint that the ideal solution is the function f^*. While we have seen that naturally Bregman distances are estimated in the theory of variational regularization, the generalization error

$$G = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f(x;\theta), y\big)\big) - \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\ell\big(f(x;\theta), y\big)\big)$$

is the commonly used quantity in machine learning.

In order to understand the connections to Bregman distances, consider an ideal solution $f^* \in \mathcal{F}$ minimizing the population risk, i.e.

$$f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f(x), y\big)\big) = \arg\min_{f \in \mathcal{F}} R^*(f).$$

Since the population risk is convex with respect to f, we conclude that $0 \in \partial R^*(f^*)$, which implies that

$$d^0_{R^*}\big(f(\cdot,\theta), f^*\big) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f(x;\theta), y\big)\big) - \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f^*(x), y\big)\big).$$

The latter can be decomposed in a similar spirit to the error decomposition in [4]:

$$\begin{aligned} d^0_{R^*}\big(f(\cdot,\theta), f^*\big) &= \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f(x;\theta), y\big)\big) - \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\ell\big(f(x;\theta), y\big)\big) \\ &\quad + \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\ell\big(f(x;\theta), y\big) - \ell\big(f^*(x), y\big)\big) \\ &\quad + \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\ell\big(f^*(x), y\big)\big) - \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\ell\big(f^*(x), y\big)\big). \end{aligned}$$

We see that the Bregman distance is decomposed into three parts: in addition to the generalization error in the first line, we have an approximation error in the second line (or rather a term that can be controlled with an approximation error in standard spaces) and a sampling error in the last line. The approximation error can be estimated beforehand or is often even negligible, since overparametrized models such as deep neural networks can usually be trained to have $\mathbb{E}_{(x,y)\sim\mathbb{P}^N}(\ell(f(x;\theta), y)) \approx 0$ and the second part is nonpositive. Moreover, the last term vanishes on expectation over the sampling if $\mathbb{P}^N$ is obtained from i.i.d. samples. Thus, in order to control the expected Bregman distance, the most important term is indeed the expected generalization error.

4.2. Estimates with operator errors and generalization

Errors due to sampling are effectively related to operator errors in inverse problems, which we see also from Example 2.2, where effectively the operator F is replaced by an operator $\widetilde{F}$ being the concatenation of F with a random sampling operator. Moreover, we assume again a source condition of the form $p^* = F^* z^* \in \partial J(u^*)$.

The generalization error in this notation is given by (noticing that we might need to use different norms for the two terms)

$$G(u) = \|Fu - v\|^2 - \|\widetilde{F}u - \tilde{v}\|^2.$$

Hence, let us start again with the optimality condition of a regularized solution

$$u_\alpha \in \arg\min_u \frac{1}{2}\|\widetilde{F}u - \tilde{v}\|^2 + \alpha J(u),$$

which is given by

$$\widetilde{F}^*(\widetilde{F}u_\alpha - \tilde{v}) + \alpha p_\alpha = 0, \quad p_\alpha \in \partial J(u_\alpha).$$

Rewriting to

$$F^*F(u_\alpha - u^*) + \alpha(p_\alpha - p^*) = F^*(Fu_\alpha - v) - \widetilde{F}^*(\widetilde{F}u_\alpha - \tilde{v}) - \alpha F^* z^*,$$

we are in a position to derive the kind of estimate we are after. A duality product with $u_\alpha - u^*$ and several applications of Young's inequality imply

$$\frac{1}{4}\left\|F(u_\alpha - u^*)\right\|^2 + \alpha d_J^{p_\alpha,p^*}(u_\alpha, u^*) \leq \alpha^2\|z^*\|^2 + \|\widetilde{F}u^* - \tilde{v}\|^2 + \frac{1}{2}G(u_\alpha).$$

In the case of consistent data, such as obtained from sampling F, we further have $\tilde{v} = \widetilde{F}u^*$; i.e., we obtain in particular

$$d_J^{p_\alpha,p^*}(u_\alpha, u^*) \leq \alpha\|z^*\|^2 + \frac{1}{2\alpha}G(u_\alpha).$$

Thus, the error in the Bregman distance is controlled by the systematic error and the generalization error.

4.3. Regularized risk minimization problems

The above arguments can be extended to convex risk minimization problems of the form

$$D_\alpha(\theta) = \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\ell\big(f(x;\theta), y\big) + \alpha J(\theta)\big).$$

For simplicity we assume that the model f is linear, i.e. $f(x;\theta) = (F\theta)(x)$ with a linear operator F mapping to an appropriate function space $\mathcal{F}$, and ℓ is the squared Euclidean norm. Consequently, we will consider F as a bounded linear operator from some parameter space Θ to $L^2_{\mathbb{P}}(\Omega)^m$ for some domain $\Omega \subset \mathbb{R}^d$. The ideal solution θ^* is a minimizer of the population risk

$$R(\theta) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\left\|(F\theta)(x) - y\right\|^2\big).$$

With this setup, the regularization operator is given by

$$\Phi_\alpha(\mathbb{P}^N) = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\left(\frac{1}{2}\left\|(F\theta)(x) - y\right\|^2 + \alpha J(\theta)\right). \tag{4.1}$$

Moreover, the source condition becomes

$$p^* = F^*z^* \in \partial J(\theta^*) \quad \text{with } z^* \in L^2_{\mathbb{P}}(\Omega)^m. \tag{4.2}$$

Similar to the reasoning in the previous section we can use the optimality condition

$$\mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\langle(F\theta_\alpha)(x)-y, F\theta'\rangle\big) = +\alpha p_\alpha = 0, \quad p_\alpha \in \partial J(\theta_\alpha)$$

for all $\theta' \in \Theta$ to derive the following result.

Theorem 4.1. *Let $\theta_\alpha \in \Phi_\alpha(\mathbb{P}^N)$ be defined by* (4.1) *and let the source condition* (4.2) *be satisfied. Then for appropriate $p_\alpha \in \partial J(u_\alpha)$ the estimate*

$$\frac{1}{4}\mathbb{E}_{(x,y)\sim\mathbb{P}}\big(\|(F\theta_\alpha)(x)-(F\theta^*)(x)\|^2\big) + \alpha d_J^{p_\alpha,p^*}(\theta_\alpha,\theta^*) \\ \leq \frac{1}{2}G(\theta_\alpha) + \alpha^2\|z^*\|^2 + \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\big(\|(F\theta^*)(x)-y\|^2\big),$$

with the generalization error

$$G(\theta_\alpha) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\Big(\frac{1}{2}\|(F\theta_\alpha)(x)-y\|^2\Big) - \mathbb{E}_{(x,y)\sim\mathbb{P}^N}\Big(\frac{1}{2}\|(F\theta_\alpha)(x)-y\|^2\Big).$$

Funding. Research supported in part by ERC via Grant EU FP7 ERC Consolidator Grant 615216 LifeInverse, by the German Ministry of Science and Technology (BMBF) under grant 05M2020 - DELETO, and by the EU under grant 2020 NoMADS - DLV-777826.

References

[1] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, Solving inverse problems using data-driven models. *Acta Numer.* **28** (2019), 1–174 Zbl 1429.65116 MR 3963505

[2] M. Benning and M. Burger, Error estimates for general fidelities. *Electron. Trans. Numer. Anal.* **38** (2011), 44–68 Zbl 1285.47015 MR 2871859

[3] M. Benning and M. Burger, Modern regularization methods for inverse problems. *Acta Numer.* **27** (2018), 1–111 Zbl 1431.65080 MR 3826506

[4] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, The modern mathematics of deep learning. 2021, arXiv:2105.04026

[5] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart, Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.* **45** (2007), no. 6, 2610–2636 Zbl 1234.62062 MR 2361904

[6] A. Braides, Γ*-Convergence for Beginners*. Oxford Lecture Ser. Math. Appl. 22, Oxford University Press, Oxford, 2002 Zbl 1198.49001 MR 1968440

[7] K. Bredies and H. K. Pikkarainen, Inverse problems in spaces of measures. *ESAIM Control Optim. Calc. Var.* **19** (2013), no. 1, 190–218 Zbl 1266.65083 MR 3023066

[8] E.-M. Brinkmann, M. Burger, J. Rasch, and C. Sutour, Bias reduction in variational regularization. *J. Math. Imaging Vision* **59** (2017), no. 3, 534–566 Zbl 1385.49002 MR 3712429

[9] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger, A Bregman learning framework for sparse neural networks. 2021, arXiv:2105.04319

[10] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger, Neural architecture search via Bregman iterations. 2021, arXiv:2106.02479

[11] M. Burger, Bregman distances in inverse problems and partial differential equations. In *Advances in Mathematical Modeling, Optimization and Optimal Control*, pp. 3–33, Springer Optim. Appl. 109, Springer, Cham, 2016 Zbl 1346.90007 MR 3526561

[12] M. Burger and H. W. Engl, Training neural networks with noisy data as an ill-posed problem. *Adv. Comput. Math.* **13** (2000), no. 4, 335–354 Zbl 1126.41301 MR 1826332

[13] M. Burger, T. Helin, and H. Kekkonen, Large noise in variational regularization. *Trans. Math. Appl.* **2** (2018), no. 1, tny002 Zbl 1404.49025 MR 3899949

[14] M. Burger, B. Kaltenbacher, and A. Neubauer, Iterative solution methods. In *Handbook of Mathematical Methods in Imaging. Vol. 1, 2, 3*, pp. 431–470, Springer, New York, 2nd edn., 2015 Zbl 1331.65080 MR 3560073

[15] M. Burger and S. Osher, Convergence rates of convex variational regularization. *Inverse Problems* **20** (2004), no. 5, 1411–1421 Zbl 1068.65085 MR 2109126

[16] M. Burger and S. Osher, A guide to the TV zoo. In *Level Set and PDE Based Reconstruction Methods in Imaging*, pp. 1–70, Lecture Notes in Math. 2090, Springer, Cham, 2013 Zbl 1342.94014 MR 3203352

[17] M. Burger, E. Resmerita, and L. He, Error estimation for Bregman iterations and inverse scale space methods in image restoration. *Computing* **81** (2007), no. 2-3, 109–135 Zbl 1147.68790 MR 2354192

[18] V. Caselles, A. Chambolle, and M. Novaga, The discontinuity set of solutions of the TV denoising problem and some extensions. *Multiscale Model. Simul.* **6** (2007), no. 3, 879–894 Zbl 1145.49024 MR 2368971

[19] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, pp. 263–340, Radon Ser. Comput. Appl. Math. 9, Walter de Gruyter, Berlin, 2010 Zbl 1209.94004 MR 2731599

[20] J. Cheng and B. Hofmann, Regularization methods for ill-posed problems. In *Handbook of Mathematical Methods in Imaging. Vol. 1, 2, 3*, pp. 91–123, Springer, New York, 2nd edn., 2015 Zbl 1331.65081 MR 3560066

[21] C.-A. Deledalle, N. Papadakis, J. Salmon, and S. Vaiter, CLEAR: Covariant LEAst-square Refitting with applications to image restoration. *SIAM J. Imaging Sci.* **10** (2017), no. 1, 243–284 Zbl 1365.49034 MR 3615463

[22] Q. Denoyelle, V. Duval, and G. Peyré, Support recovery for sparse super-resolution of positive measures. *J. Fourier Anal. Appl.* **23** (2017), no. 5, 1153–1194 Zbl 1417.65223 MR 3704760

[23] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*. Classics Appl. Math. 28, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999 Zbl 0939.49002 MR 1727362

[24] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*. Math. Appl. 375, Kluwer Academic Publishers Group, Dordrecht, 1996 Zbl 0859.65054 MR 1408680

[25] J. Flemming, Variational smoothness assumptions in convergence rate theory—an overview. *J. Inverse Ill-Posed Probl.* **21** (2013), no. 3, 395–409 Zbl 1293.47058 MR 3069356

[26] J. Flemming and B. Hofmann, A new approach to source conditions in regularization with general residual term. *Numer. Funct. Anal. Optim.* **31** (2010), no. 1-3, 254–284 Zbl 1195.47040 MR 2677255

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA, 2016 Zbl 1373.68009 MR 3617773

[28] M. Grasmair, Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. *Inverse Problems* **27** (2011), no. 7, 075014 Zbl 1219.35359 MR 2817427

[29] M. Grasmair, Variational inequalities and higher order convergence rates for Tikhonov regularisation on Banach spaces. *J. Inverse Ill-Posed Probl.* **21** (2013), no. 3, 379–394 Zbl 1288.47016 MR 3069355

[30] B. Hofmann, B. Kaltenbacher, C. Pöschl, and O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems* **23** (2007), no. 3, 987–1010 Zbl 1131.65046 MR 2329928

[31] B. Hofmann and P. Mathé, Parameter choice in Banach space regularization under variational inequalities. *Inverse Problems* **28** (2012), no. 10, 104006 Zbl 1253.47041 MR 2987901

[32] T. Hohage and F. Weidling, Characterizations of variational source conditions, converse results, and maxisets of spectral regularization methods. *SIAM J. Numer. Anal.* **55** (2017), no. 2, 598–620 Zbl 1432.65070 MR 3623207

[33] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer Texts Statist. 103, Springer, New York, 2013 Zbl 1281.62147 MR 3100153

[34] B. Kaltenbacher, A. Neubauer, and O. Scherzer, *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*. Radon Ser. Comput. Appl. Math. 6, Walter de Gruyter, Berlin, 2008 Zbl 1145.65037 MR 2459012

[35] B. Kaltenbacher, F. Schöpfer, and T. Schuster, Iterative methods for nonlinear ill-posed problems in Banach spaces: convergence and applications to parameter identification problems. *Inverse Problems* **25** (2009), no. 6, 065003 Zbl 1176.65070 MR 2506848

[36] H. Kekkonen, M. Lassas, and S. Siltanen, Analysis of regularized inversion of data corrupted by white Gaussian noise. *Inverse Problems* **30** (2014), no. 4, 045009 Zbl 1287.35101 MR 3178119

[37] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*. 2nd edn., Appl. Math. Sci. 120, Springer, New York, 2011 Zbl 1213.35004 MR 3025302

[38] K. C. Kiwiel, Proximal minimization methods with generalized Bregman functions. *SIAM J. Control Optim.* **35** (1997), no. 4, 1142–1168 Zbl 0890.65061 MR 1453294

[39] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: Learning maps between function spaces. 2021, arXiv:2108.08481

[40] J. Kukačka, V. Golkov, and D. Cremers, Regularization for deep learning: A taxonomy. 2017, arXiv:1710.10686

[41] M. M. Lavrentiev, *Some Improperly Posed Problems of Mathematical Physics*. Springer Tracts Nat. Philos. 11, Springer, New York, 1967 Zbl 0149.41902

[42] J. Lederer, Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. 2013, arXiv:1306.0113

[43] A. K. Louis, Approximate inverse for linear and some nonlinear problems. *Inverse Problems* **12** (1996), no. 2, 175–190 Zbl 0851.65036 MR 1382237

[44] A. K. Louis and P. Maass, A mollifier method for linear operator equations of the first kind. *Inverse Problems* **6** (1990), no. 3, 427–440 Zbl 0713.65040 MR 1057035

[45] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA, 2018 Zbl 1407.68007 MR 3931734

[46] F. Natterer, *The Mathematics of Computerized Tomography*. Classics Appl. Math. 32, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001 Zbl 0973.92020 MR 1847845

[47] N. H. Nelsen and A. M. Stuart, The random feature model for input-output maps between Banach spaces. *SIAM J. Sci. Comput.* **43** (2021), no. 5, A3212–A3243 Zbl 07398767 MR 4313849

[48] B. Neyshabur, R. Tomioka, and N. Srebro, In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International Conference on Learning Representations (ICLR)*, 2015

[49] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration. *Multiscale Model. Simul.* **4** (2005), no. 2, 460–489 Zbl 1090.94003 MR 2162864

[50] R. Ramlau and G. Teschke, Sparse recovery in inverse problems. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, pp. 201–262, Radon Ser. Comput. Appl. Math. 9, Walter de Gruyter, Berlin, 2010 Zbl 1210.65107 MR 2731600

[51] E. Resmerita, Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems* **21** (2005), no. 4, 1303–1314 Zbl 1082.65055 MR 2158110

[52] E. Resmerita and O. Scherzer, Error estimates for non-quadratic regularization and the relation to enhancement. *Inverse Problems* **22** (2006), no. 3, 801–814 Zbl 1103.65062 MR 2235638

[53] L. Rosasco, A. Caponnetto, E. Vito, F. Odone, and U. Giovannini, Learning, regularization and ill-posed inverse problems. In *Advances in Neural Information Processing Systems*, edited by L. Saul, Y. Weiss, and L. Bottou, pp. 1145–1152, 17, 2005

[54] T. Schuster, B. Kaltenbacher, B. Hofmann, and K. S. Kazimierski, *Regularization Methods in Banach Spaces*. Radon Ser. Comput. Appl. Math. 10, Walter de Gruyter, Berlin, 2012 Zbl 1259.65087 MR 2963507

[55] T. I. Seidman and C. R. Vogel, Well-posedness and convergence of some regularisation methods for nonlinear ill posed problems. *Inverse Problems* **5** (1989), no. 2, 227–238 Zbl 0691.35090 MR 991919

[56] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Scripta Series in Mathematics, John Wiley & Sons, New York, 1977 Zbl 0354.65028 MR 0455365

[57] F. Weidling, *Variational source conditions and conditional stability estimates for inverse problems in PDEs*. Ph.D. thesis, Göttingen University, 2019

[58] F. Werner and B. Hofmann, Convergence analysis of (statistical) inverse problems under conditional stability estimates. *Inverse Problems* **36** (2020), no. 1, 015004 Zbl 07153871 MR 4047904

Martin Burger
Department of Mathematics and Research Center for Mathematics of Data,
Friedrich-Alexander-Universität Erlangen-Nürnberg, Cauerstr. 11, 91058 Erlangen, Germany;
martin.burger@fau.de

DOI 10.4171/8ECM/06

Some minimization problems for mean field models with competing forces

Rupert L. Frank

Abstract. We review recent results on three families of minimization problems, defined on subsets of nonnegative functions with fixed integral. The competition between attractive and repulsive forces leads to transitions between parameter regimes, where minimizers exist and where they do not. The problems considered are generalized liquid drop models, swarming models, and generalized Keller–Segel models.

1. Introduction

In this survey we discuss three families of minimization problems. They are simple mathematical toy models for physical or biological phenomena. While their origins are rather different, they share some mathematical similarities and differences and we think it is worthwhile to look at them side by side.

The common feature of all three problems is that they are of mean-field type. They involve an "energy" functional that is defined on a subset of nonnegative functions ("densities") whose integral is fixed ("total mass"). They are, at least on a heuristic level, derived from microscopic, many-body models. The densities in the mean-field models describe the distribution of the microscopic particles in the limit of a large number of particles, and similarly the energy functionals in our models are obtained as macroscopic approximations to microscopic energy functionals.

Another common feature of the problems discussed here is that the energy functionals have two contributions that compete with each other. There are attractive forces that keep the particles together and try to concentrate them and there are repulsive forces that push them apart and try to spread them out. Typically, these forces act on different length scales and one is of short range and the other one of long range type. The existence of a minimizer can be understood as the forces being in a local

2020 Mathematics Subject Classification. Primary 49Q10; Secondary 49Q20, 49S05, 81V35, 92C17.
Keywords. Minimization problems, mean field models, liquid drop model, swarming, Keller–Segel model.

equilibrium, while the nonexistence typically means that one of the forces dominates the other.

We are particularly interested in situations where, as a parameter of the problem is varied continuously, there is either a transition between existence and nonexistence of minimizers, or a sharp change in the properties of minimizers. A typical parameter that is varied is the total mass, but in one of the models it is also a parameter describing the shape of the forces acting between the particles.

The models. Let us be more specific about the three families of models that we will consider. Throughout, $N \geq 1$ is the dimension of the underlying Euclidean space.

For the *generalized liquid drop model*, depending on a parameter $\lambda \in (0, N)$, we define, for any measurable set $\Omega \subset \mathbb{R}^N$,

$$\mathcal{E}_\lambda^{\mathrm{gld}}[\Omega] := \operatorname{Per}\Omega + \frac{1}{2}\iint_{\Omega\times\Omega} \frac{dx\,dy}{|x-y|^\lambda}. \tag{1.1}$$

Here $\operatorname{Per}\Omega$ denotes the perimeter in the sense of De Giorgi; see, e.g., [45]. The corresponding minimization problem is, for $m \in (0, \infty)$,

$$E_\lambda^{\mathrm{gld}}(m) := \inf\{\mathcal{E}_\lambda^{\mathrm{gld}}[\Omega] : \Omega \subset \mathbb{R}^N \text{measurable}, |\Omega| = m\}. \tag{1.2}$$

The original liquid drop model, suggested by Gamow [35] for the description of atomic nuclei, corresponds to $\lambda = 1$ in dimension $N = 3$.

For the *flocking model*, depending on parameters $\lambda \in (0, N)$ and $\alpha \in (0, \infty)$, we define, for any nonnegative, measurable function ρ on $\mathbb{R}^N$,

$$\mathcal{E}_{\lambda,\alpha}^{\mathrm{f}}[\rho] := \frac{1}{2}\iint_{\mathbb{R}^N\times\mathbb{R}^N} \rho(x)\big(|x-y|^{-\lambda} + |x-y|^\alpha\big)\rho(y)\,dx\,dy. \tag{1.3}$$

The corresponding minimization problem is, for $m \in (0, \infty)$,

$$E_{\lambda,\alpha}^{\mathrm{f}}(m) := \inf\left\{\mathcal{E}_{\lambda,\alpha}^{\mathrm{f}}[\rho] : \rho \in L^1(\mathbb{R}^N),\ 0 \leq \rho \leq 1,\ \int_{\mathbb{R}^N} \rho\,dx = m\right\}. \tag{1.4}$$

This model was suggested by Burchard, Choksi, and Topaloglu [7]. It is a simple model to describe the flocking behavior in stable states of a large group of animals such as fish or birds.

For the *generalized Keller–Segel model*, depending on parameters $q \in (0, 1)$ and $\alpha \in (0, \infty)$, we define, for any nonnegative function $\rho \in L^q(\mathbb{R}^N)$,

$$\mathcal{E}_{q,\alpha}^{\mathrm{gKS}}[\rho] := -\int_{\mathbb{R}^N} \rho^q\,dx + \frac{1}{2}\iint_{\mathbb{R}^N\times\mathbb{R}^N} \rho(x)|x-y|^\alpha\rho(y)\,dx\,dy. \tag{1.5}$$

The corresponding minimization problem is

$$E_{q,\alpha}^{\mathrm{gKS}} := \inf\left\{\mathcal{E}_{q,\alpha}^{\mathrm{gKS}}[\rho] : 0 \leq \rho \in L^q(\mathbb{R}^N),\ \int_{\mathbb{R}^N} \rho\,dx = 1\right\}. \tag{1.6}$$

Note that here, in contrast to the two previous problems, we fix the integral of ρ to be one. The more general case, where it is fixed to be equal to m, can be reduced to the present one by scaling. The generalized Keller–Segel model was introduced in [8] and generalizes the standard Keller–Segel model, which corresponds (after some rescaling) to the limit cases $q = 1$ and $\alpha = 0$ in dimension $N = 2$.

Competing forces. Let us discuss in which sense in the above models two forces compete with each other.

In the generalized liquid drop model, the perimeter term corresponds to an attractive short range force, whereas the double integral term corresponds to a repulsive long range force. Note that by the isoperimetric inequality (see, e.g., [45])

$$\inf\left\{\operatorname{Per}\Omega : \Omega\subset\mathbb{R}^N \text{ measurable}, |\Omega| = m\right\} = N^{\frac{N-1}{N}}|\mathbb{S}^{N-1}|^{\frac{1}{N}} m^{\frac{N-1}{N}}$$

with equality if and only if Ω is a ball (up to sets of measure zero). On the other hand, it is easy to see that

$$\inf\left\{\frac{1}{2}\iint_{\Omega\times\Omega}\frac{dx\,dy}{|x-y|^\lambda} : \Omega\subset\mathbb{R}^N \text{ measurable}, |\Omega| = m\right\} = 0$$

and the infimum is not attained. A minimizing sequence is given, for instance, by taking Ω as a union of a large number of small balls placed very far apart from each other. Next, we note that, by scaling,

$$E_\lambda^{\mathrm{gld}}(m) = \inf\left\{m^{\frac{N-1}{N}}\operatorname{Per}\omega + m^{\frac{2N-\lambda}{N}}\frac{1}{2}\iint_{\omega\times\omega}\frac{dx\,dy}{|x-y|^\lambda} : \omega\subset\mathbb{R}^N \text{ meas.}, |\omega| = 1\right\}.$$

Since $(N-1)/N < (2N-\lambda)/N$, the perimeter term is dominant for small m, whereas the double integral is dominant for large m. We therefore expect existence of minimizers for small m, whereas for large m we might have nonexistence of minimizers.

In the flocking model, the α-term corresponds to an attractive force, while the λ-term corresponds to a repulsive force. Moreover, the α-term is relevant on large distances and the λ-term on short ones. By rearrangement inequalities and the bathtub principle (see, e.g., [41, Theorems 1.14 and 3.7])

$$\inf\left\{\frac{1}{2}\iint_{\mathbb{R}^N\times\mathbb{R}^N}\rho(x)\,|x-y|^\alpha\rho(y)\,dx\,dy : \rho\in L^1(\mathbb{R}^N),\ 0\le\rho\le 1,\ \int_{\mathbb{R}^N}\rho\,dx = m\right\}$$

is attained if and only if ρ is the characteristic function of a ball of volume m. Moreover, as a consequence of what we said in the generalized liquid drop model,

$$\inf\left\{\frac{1}{2}\iint_{\mathbb{R}^N\times\mathbb{R}^N}\frac{\rho(x)\rho(y)}{|x-y|^\lambda}\,dx\,dy : \rho\in L^1(\mathbb{R}^N),\ 0\le\rho\le 1,\ \int_{\mathbb{R}^N}\rho\,dx = m\right\} = 0$$

and the infimum is not attained. Next, we note that, by scaling,

$$E^{\mathrm{f}}_{\lambda,\alpha}(m) = \inf\Big\{ m^{\frac{2N-\lambda}{N}} \frac{1}{2} \iint_{\mathbb{R}^N\times\mathbb{R}^N} \frac{\sigma(x)\,\sigma(y)}{|x-y|^{\lambda}}\,dx\,dy + m^{\frac{2N+\alpha}{N}} \frac{1}{2} \iint_{\mathbb{R}^N\times\mathbb{R}^N} \sigma(x)\,|x-y|^{\alpha}\sigma(y)\,dx\,dy : \sigma\in L^1(\mathbb{R}^N),\ 0\le\sigma\le 1,\ \int_{\mathbb{R}^N}\sigma\,dx = 1\Big\}.$$

Since $(2N-\lambda)/N < (2N+\alpha)/N$, the α-term is dominant for large m and in this regime we expect existence of minimizers and closeness to the characteristic function of a ball. We also have

$$E^{\mathrm{f}}_{\lambda,\alpha}(m) = m^2 \inf\Big\{ \frac{1}{2} \iint_{\mathbb{R}^N\times\mathbb{R}^N} \sigma(x)\big(|x-y|^{-\lambda} + |x-y|^{\alpha}\big)\sigma(y)\,dx\,dy : \sigma\in L^1(\mathbb{R}^N),\ 0\le\sigma\le m^{-1},\ \int_{\mathbb{R}^N}\sigma\,dx = 1\Big\}.$$

For small m, we expect that the constraint $\sigma \le m^{-1}$ is irrelevant and that the minimizer is m times the minimizer of the problem

$$\inf\Big\{ \frac{1}{2} \iint_{\mathbb{R}^N\times\mathbb{R}^N} \sigma(x)\big(|x-y|^{-\lambda} + |x-y|^{\alpha}\big)\sigma(y)\,dx\,dy : 0\le\sigma\in L^1(\mathbb{R}^N),\ \int_{\mathbb{R}^N}\sigma\,dx = 1\Big\},$$

provided that a minimizer for the latter problem exists and is bounded.

Finally, in the generalized Keller–Segel model, the L^q term corresponds to a repulsive short range force, whereas the double integral term corresponds to an attractive long range force. Note that

$$\inf\Big\{ -\int_{\mathbb{R}^N}\rho^q\,dx : 0\le\rho\in L^q(\mathbb{R}^N),\ \int_{\mathbb{R}^N}\rho\,dx = 1\Big\} = -\infty.$$

A minimizing sequence is given, for instance, by a sequence that spreads out like $\ell^{-N}\sigma(x/\ell)$ with $\ell\to\infty$. On the other hand,

$$\inf\Big\{ \frac{1}{2} \iint_{\mathbb{R}^N\times\mathbb{R}^N} \rho(x)\,|x-y|^{\alpha}\rho(y)\,dx\,dy : 0\le\rho\in L^q(\mathbb{R}^N),\ \int_{\mathbb{R}^N}\rho\,dx = 1\Big\} = 0$$

and the infimum is not attained. A minimizing sequence is given, for instance, by a delta sequence $\ell^{-N}\sigma(x/\ell)$ with $\ell\to 0$. Since, as we already mentioned, in this model the dependence on the total mass is trivial, we are looking here for a transition in terms of the parameters q and α. Intuitively, the repulsive force is stronger the smaller q and the attractive force is stronger the larger α. The above examples suggest that two mechanisms for the nonexistence of a minimizer are conceivable, namely both spreading out and concentration of minimizing sequences.

Structure of the paper. In the following three sections we summarize what is known about the three families of minimization problems. The presentation will be rather compact and we refer to the original papers for the proofs. We do, however, emphasize several open questions concerning each model. In a short appendix, we provide details for a simple, unpublished result in the one-dimensional generalized liquid drop model.

2. The generalized liquid drop model

In this section, we consider the energy functional (1.1) and the corresponding minimization problem (1.2). We assume throughout that $0 < \lambda < N$.

Let us set, for fixed λ and N,

$$m_* := \left(\frac{2^{1/N} - 1}{1 - 2^{-(N-\lambda)/N}} \frac{\operatorname{Per} B_1}{\frac{1}{2} \iint_{B_1 \times B_1} |x-y|^{-\lambda}\, dx\, dy} \right)^{N/(N-\lambda+1)} |B_1|,$$

where B_1 denotes the unit ball in $\mathbb{R}^N$. The number m_* is the unique solution $m > 0$ of the equation

$$\mathcal{E}_\lambda^{\mathrm{gld}}\left[\left(\frac{m}{|B_1|} \right)^{1/N} B_1 \right] = 2\mathcal{E}_\lambda^{\mathrm{gld}}\left[\left(\frac{m}{2|B_1|} \right)^{1/N} B_1 \right]. \tag{2.1}$$

Thus, the energy of a ball of mass m_* is equal to the energy of two balls, each of mass $m_*/2$, placed infinitely far apart. For $m < m_*$ one has $<$ instead of $=$ in (2.1) and for $m > m_*$ one has $>$.

In the physics literature, it is typically taken for granted that in the special case $\lambda = 1$ and $N = 3$, balls are minimizers for $E_\lambda^{\mathrm{gld}}(m)$ for $m \leq m_*$ and there is no minimizer for $m > m_*$. In the mathematics literature, this appears explicitly as a conjecture in the work of Choksi and Peletier [12, 13].

One may wonder whether the analogous conjecture is valid in the general case $0 < \lambda < N$. In dimension $N = 1$, this is indeed the case, as can be verified by elementary computations; see Appendix A. It is shown in [37, 3] that for any $N \geq 2$ there is a $\lambda_c > 0$ such that for all $0 < \lambda < \lambda_c$ the conjecture is true; see [46] for an explicit lower bound on λ_c for $N = 2$. In the remaining cases, the validity or invalidity of the conjecture is open.

Existence. As a first step towards this conjecture, before asking whether minimizers for $E_\lambda^{\mathrm{gld}}(m)$ are balls for all $m \leq m_*$, it is natural to ask whether minimizers exist for all $m \leq m_*$. This is indeed the case, as shown in [31]. Moreover, it is shown there as well that if there are no minimizers for $m > m_*$, then balls are minimizers for $m \leq m_*$.

The proof of [31] proceeds by verifying that for any $m < m_*$ one has the strict binding inequality

$$E_\lambda^{\mathrm{gld}}(m) < E_\lambda^{\mathrm{gld}}(m') + E_\lambda^{\mathrm{gld}}(m-m') \quad \text{for all } 0 < m' < m.$$

According to a compactness result in [26] this implies the existence of a minimizer for $E_\lambda^{\mathrm{gld}}(m)$ for $m \leq m_*$.

Uniqueness. We address the question of whether balls are minimizers. A convexity argument due to Bonacini and Cristoferi [3, Theorem 2.10] shows that there is a number $m_c^{\mathrm{ball}} \in [0,\infty) \cup \{\infty\}$ (depending on λ and N) such that for $m < m_c^{\mathrm{ball}}$ balls are the unique minimizers of $E_\lambda^{\mathrm{gld}}(m)$, for $m = m_c^{\mathrm{ball}} > 0$ balls are minimizers of $E_\lambda^{\mathrm{gld}}(m)$, and for $m > m_c^{\mathrm{ball}}$ balls are not minimizers of $E_\lambda^{\mathrm{gld}}(m)$. (This part of [3] does not use the assumption $\lambda < N-1$.)

An important result is that $m_c^{\mathrm{ball}} > 0$, that is, for small $m > 0$ balls are minimizers for $E_\lambda^{\mathrm{gld}}(m)$. In the full parameter regime, this result is due to [21], extending earlier results in [37, 38, 36, 3]. The proofs in these papers are based directly or indirectly on the quantitative form of the isoperimetric inequality (see [33] and also [22, 15]) and the regularity theory for quasiminimizers of the perimeter (see, e.g., [45, Part III]). As far as we are aware, these proofs use compactness arguments and do not give a numerical lower bound on m_c^{ball}.

On the other hand, one can show that $m_c^{\mathrm{ball}} < \infty$, that is, for large $m > 0$ balls are not minimizers for $E_\lambda^{\mathrm{gld}}(m)$. Indeed, setting

$$m_c^{\mathrm{stab}} := \left(\frac{N+1}{\lambda(N-\lambda)} \, \frac{\operatorname{Per} B_1}{\frac{1}{2}\iint_{B_1\times B_1} |x-y|^{-\lambda}\,dx\,dy} \right)^{N/(N-\lambda+1)} |B_1|,$$

one finds that for $m < m_c^{\mathrm{stab}}$ the ball is stable against small volume-preserving perturbations and for $m > m_c^{\mathrm{stab}}$ it is unstable. (Stability here means that the Hessian is positive definite except for zero modes coming from translations. Instability means that the Hessian is not positive semidefinite.) This computation goes back to Bohr and Wheeler [2] for $N = 3$, $\lambda = 1$ and can be found in the general case in [3, 21]. Clearly, $m_c^{\mathrm{ball}} \leq m_c^{\mathrm{stab}}$, so the former quantity is indeed finite.

Nonexistence. Let us discuss the nonexistence of minimizers for $E_\lambda^{\mathrm{gld}}(m)$. For fixed λ and N we set

$$m_c^{\mathrm{n.e.}} := \sup \left\{ m > 0 : \text{there is a minimizer for } E_\lambda^{\mathrm{gld}}(m) \right\}.$$

Then, if $\lambda \leq 2$ (and $\lambda < N$, as always), one can show that $m_c^{\mathrm{n.e.}} < \infty$, that is, there is no minimizer for large m. This is due to [37, 38, 43, 31]. It seems to be unknown whether $m_c^{\mathrm{n.e.}}$ is finite or not for $2 < \lambda < N$.

In [25], it is shown that for $\lambda = 1$, $N = 3$, one has $m_c^{\text{n.e.}} \le 8$. This is to be compared with $m_c^{\text{stab}} = 10$ for these values of λ and N. Thus there is a regime $8 < m < 10$, where balls are stable local minimizers, but not global minimizers. For comparison, for these values of λ and N one has $m_* = 5(2^{1/3} - 1)/(1 - 2^{-2/3}) \approx 3.512$.

Problem 2.1. For $N = 3$ and $\lambda = 1$, show that balls are minimizers for $m \le m_*$ and there are no minimizers for $m > m_*$. In which parameter region of λ's and N's is the analogous conjecture valid?

The following two problems are special cases of the previous one.

Problem 2.2. Do there exist minimizers for $E_\lambda^{\text{gld}}(m)$ for arbitrarily large m in case $2 < \lambda < N$?

Problem 2.3. Find an explicit numerical lower bound on m_c^{ball}, in particular, in the case $N = 3$ and $\lambda = 1$.

We conclude this section by briefly mentioning two further, related models.

The first one concerns the liquid drop model in the presence of a neutralizing background. This problem is motivated, for instance, by the physics of neutron stars and there are interesting mathematical questions; see, e.g., [39]. For simplicity we focus here on the case $\lambda = N - 2$ in dimension $N \ge 3$, although there are similar versions in dimensions $N = 1, 2$ [29]. For a (large) parameter $L > 0$ one sets $\Lambda_L := (0, L)^N$ and considers the minimization problem

$$E_L(\rho) := \inf \Bigg\{ \operatorname{Per}\Omega + \frac{1}{2} \iint_{\Lambda_L \times \Lambda_L} \frac{\big(\mathbb{1}_\Omega(x) - \rho\big)\big(\mathbb{1}_\Omega(y) - \rho\big)}{|x-y|^{N-2}}\, dx\, dy : \\ \Omega \subset \Lambda_L,\ |\Omega| = \rho|\Lambda_L| \Bigg\}.$$

(Sometimes, the kernel $|x-y|^{-N+2}$ is replaced by a constant multiple of the periodic or Neumann Green's function of the Laplacian and the perimeter is replaced by its periodic version or a relative perimeter, but this does not qualitatively change the results discussed below.)

A major open problem is to prove that (for $N = 3$, for simplicity) there are $0 < \rho_{c1} < \rho_{c2} < 1/2$ such that the following holds approximately for minimizers for $E_L(\rho)$ for large $L > 0$ "in the bulk": for $0 < \rho < \rho_{c1}$, minimizers are periodic with respect to a three-dimensional lattice, for $\rho_{c1} < \rho < \rho_{c2}$, minimizers are periodic with respect to a two-dimensional lattice, and for $\rho_{c2} < \rho \le 1/2$, minimizers are periodic with respect to a one-dimensional lattice. For $1/2 < \rho < 1$, the situation reverses, with $1 - \rho$ replacing ρ. This would correspond to what is known as "nuclear pasta" phases in astrophysics.

A fundamental result by Alberti, Choksi, and Otto [1] gives precise bounds on the energy distribution of minimizers that are indicative of the emergence of a regular (e.g., periodic) structure. More precise results about the structure of minimizers are restricted only to the dilute regime. The case $\rho \sim L^{-3}$ is treated in [12] (see also [16] and references therein), the case $\rho \sim L^{-2}$ in [39], and the case $\rho \ll 1$ (independently of L) in [20].

The second generalization of the generalized liquid drop model concerns the addition of an external potential V,

$$\inf\left\{\mathcal{E}_\lambda^{\mathrm{gld}}[\Omega] + \int_\Omega V\,dx : \Omega \subset \mathbb{R}^N \text{ measurable}, |\Omega| = m\right\}.$$

Lu and Otto [44] suggested this model with $V(x) = -Z|x|^{-1}$ in $N = 3$, $\lambda = 1$ as a toy problem for the ionization conjecture in Thomas–Fermi–Dirac–von Weizsäcker theory and proved that there is no minimizer for $m \geq Z + C\max\{1, Z^{2/3}\}$. Nonexistence for $m \geq Z + C\max\{1, Z^{1/3}\}$, as well as the ionization conjecture in Thomas–Fermi–Dirac–von Weizsäcker theory were proved in [32]. For more on the ionization conjecture, also for more complicated models, we refer to [47].

Finally, returning to the standard liquid drop model with $\lambda = 1$ and $N = 3$, we mention the open problem to make the global bifurcation picture of Bohr and Wheeler [2] rigorous. For an initial local bifurcation result, see [23].

3. A simple model for flocking

In this section, we consider the energy functional (1.3) and the corresponding minimization problem (1.4). We assume throughout that $0 < \lambda < N$ and $\alpha > 0$.

It is easy to see that there is a minimizer of $E_{\lambda,\alpha}^{\mathrm{f}}(m)$ for any $m > 0$ [11]. We would like to understand properties of minimizers and, in particular, qualitative changes in these properties as m varies. For instance, one is interested in the existence of the following three "phases" [27]. A first, "liquid" phase occurs when any minimizer ρ for $E_{\lambda,\alpha}^{\mathrm{f}}(m)$ satisfies $\rho < 1$ almost everywhere. A second, "intermediate" phase occurs when there is a minimizer ρ for $E_{\lambda,\alpha}^{\mathrm{f}}(m)$ such that $\{0 < \rho < 1\}$ has positive measure strictly less than m. A third, "solid" phase occurs when any minimizer ρ for $E_{\lambda,\alpha}^{\mathrm{f}}(m)$ satisfies $\rho = 1$ almost everywhere.

Some initial results. The case $N \geq 3$, $\lambda = N - 2$, and $\alpha = 2$ can be solved explicitly [7] and one finds that there is an explicit $m_N \in (0, \infty)$ such that the unique (up to translations) minimizer for $E_{\lambda,\alpha}^{\mathrm{f}}(m)$ is a multiple of the characteristic function of a ball of measure m_N if $m \leq m_N$ and the characteristic function of a ball of measure m if $m > m_N$. In particular, in this special case, the second, intermediary phase does not occur.

In the case $2 \le \alpha \le 4$ (and any $N \ge 1$ and $0 < \lambda < N$), one can show that for any $m > 0$ minimizers of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ are unique up to translations [42] and, in particular, radially symmetric. This relies on an interesting convexity argument. Moreover, the case $N = 3$, $\lambda = 1$, and $\alpha = 4$ is explicitly solved in [42]. In particular, there are critical constants $0 < m' < m'' < \infty$ such that the system is in phase one for $m \le m'$, in phase two for $m' < m < m''$, and in phase three for $m \ge m''$.

Small *m* regime. In [27], it is shown that for $N = 3$ and $\lambda = 1$ (and any $\alpha \ge 1$) there is an $m_* > 0$, depending on α, such that for all $m < m_*$ any minimizer ρ of $E^{\mathrm{f}}_{1,\alpha}(m)$ satisfies $\rho < 1$ almost everywhere. This result extends, with the same proof, to the case $\lambda = N - 2$ in arbitrary dimension $N \ge 3$.

The proof relies on the fact, due to [10], that for $\lambda = N - 2$ minimizing measures of the problem

$$E_{\lambda,\alpha} := \inf \left\{ \frac{1}{2} \iint_{\mathbb{R}^N \times \mathbb{R}^N} \left(|x-y|^{-\lambda} + |x-y|^{\alpha} \right) d\mu(x)\, d\mu(y) : \mu \in P(\mathbb{R}^N) \right\} \quad (3.1)$$

are absolutely continuous with respect to Lebesgue measure with a bounded density. Here $P(\mathbb{R}^N)$ denotes the set of Borel probability measures on $\mathbb{R}^N$. More precisely, one needs a bound on the density depending only on N and α.

There are also results in [10] concerning the problem $E_{\lambda,\alpha}$ for $0 \le N-2 < \lambda < N$ and certain assumptions on α. Using these results, one should be able to prove that for certain N, λ, α, there is an $m'_* > 0$, depending on N, λ, α, such that for all $m < m'_*$ there are minimizers ρ of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ satisfying $\rho < 1$.

Large *m* regime. Under the assumption $\lambda < N - 1$, it is shown in [30] that there is an $m^* < \infty$, depending on N, λ, α, such that for $m > m^*$ the only minimizers of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ are characteristic functions of balls. The assumption on λ is optimal in the sense that for $N - 1 \le \lambda < N$ and any $m > 0$, balls are not even critical points for the problem $E^{\mathrm{f}}_{\lambda,\alpha}(m)$.

The results in [30] improve earlier results in [7] for $\alpha = 2$ and in [27] for $\lambda = N - 2$, obtained by different methods.

The technique used in [30] is that of symmetric decreasing rearrangement and, more precisely, a quantitative version of the Riesz rearrangement inequality. This quantitative version is due to M. Christ [14], with some minor extensions and a partially alternate proof in [28]. As an aside, we mention that from the quantitative Riesz rearrangement inequality one can derive quantitative rearrangement inequalities for Riesz potentials. Those were proved, simultaneously and independently, in a restricted range in [34]; see also [4, 48, 5].

Let us conclude this section by mentioning some open problems. Relatively little seems to be known about minimizers of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ outside of the asymptotic regimes $m \to 0$ and $m \to \infty$.

Problem 3.1. Study qualitative properties of minimizers of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$.

Concrete questions to be studied are, for instance, the following. Known examples of minimizers are radially symmetric. Can symmetry breaking occur? For arguments in favor of this, see [6]. Is the support of a minimizer convex? As m increases, do the regions $\{\rho > 0\}$ and $\{\rho = 1\}$ increase (fixing the center of mass, for instance), where ρ is a minimizer? Are minimizers concave or convex on their supports for $\alpha < 2$ and $\alpha > 2$, respectively?

In view of the above small m results, it would be interesting to better understand the case $0 < \lambda < N - 2$. We consider the minimization problem (3.1) and wonder whether the result from [10] extends to $0 < \lambda < N - 2$. An affirmative answer would be related to the existence, for small m, of minimizers ρ for $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ with $\rho < 1$ almost everywhere. Examples, however, suggest that the answer might be negative.

Problem 3.2. For $0 < \lambda < N - 2$, are minimizers μ of $E_{\lambda,\alpha}$ absolutely continuous with respect to Lebesgue measure with a bounded density?

In view of the large m results for $\lambda < N - 1$, it seems interesting to investigate in more detail the case $N - 1 \leq \lambda < N$. We expect that minimizers for large m have values close to one in a large core region and then drop down to zero in a relatively small region. It would be interesting to find the scaling behavior of these regions and, if possible, the transition profile.

Problem 3.3. For $N - 1 \leq \lambda < N$ study the shape of minimizers of $E^{\mathrm{f}}_{\lambda,\alpha}(m)$ for large m.

The dynamical problem. The energy function $\mathcal{E}^{\mathrm{f}}_{\lambda,\alpha}$ considered on functions $0 \leq \rho \leq 1$ leads via a formal Wasserstein-2 gradient flow to an evolution equation called the constrained aggregation equation; see [17, 18]. It would be interesting to understand the long time behavior of solutions to this equation. In particular, for $\lambda < N - 1$ and large m such that characteristic functions of balls are the only optimizers for $E^{\mathrm{f}}_{\lambda,\alpha}(m)$, one might wonder whether the solution approaches the characteristic function of a ball for large times.

4. The generalized Keller–Segel model

In this section, we consider the energy functional (1.5) and the corresponding minimization problem (1.6). We assume throughout that $0 < q < 1$ and $\alpha > 0$. We summarize the results from [8, 9].

The basic fact is that $E^{\mathrm{gKS}}_{q,\alpha} = -\infty$ for $0 < q \leq N/(N + \alpha)$ and $E^{\mathrm{gKS}}_{q,\alpha} > -\infty$ for $N/(N + \alpha) < q < 1$ [8, Proposition 20]. Thus, in the following discussion we will always assume that $q > N/(N + \alpha)$.

It is known and elementary that the case $\alpha = 2$ (and any $N/(N+2) < q < 1$) can be solved explicitly by expanding the square $|x-y|^2$ and setting the center of mass to zero; see [8, Corollary 6 and Proposition 20]. We comment below on the case $\alpha = 4$, which can also be solved to some extent.

It is deeper that the case $q = 2N/(2N+\alpha)$ can be solved explicitly as well. This was observed by Dou and Zhu [19], who discovered a conformal symmetry in this case, similarly as in Lieb's work on the Hardy–Littlewood–Sobolev inequality [40]. The case $q = 2N/(2N+\alpha)$ is also of some conceptual importance. If we reinstate the mass in the variational problem (1.6) and define $E_{q,\alpha}^{\mathrm{gKS}}(m)$ in the natural way, then

$$E_{q,\alpha}^{\mathrm{gKS}}(m) = m^{\frac{2N-(2N+\alpha)q}{N-\alpha-Nq}}\, E_{q,\alpha}^{\mathrm{gKS}}.$$

Thus, for $q = 2N/(2N+\alpha)$, $E_{q,\alpha}^{\mathrm{gKS}}(m)$ is independent of m. As we will see, there are differences between the cases $q > 2N/(2N+\alpha)$ and $q < 2N/(2N+\alpha)$.

Existence in the superconformal case. In the case $2N/(2N+\alpha) < q < 1$, there is a minimizer for $E_{q,\alpha}^{\mathrm{gKS}}$ [8, Proposition 8], and any minimizer is radially symmetric with respect to some point, nonincreasing with respect to the distance from this point and positive almost everywhere [8, Lemma 9]. Symmetric decreasing rearrangment plays an important role in the proof of existence and in the derivation of the properties of minimizers.

Existence and nonexistence in the subconformal case. The case $N/(N+\alpha) < q < 2N/(2N+\alpha)$ is less understood and there are some open questions about the existence of minimizers. A brief summary of the results in this case is as follows. Either there is a minimizer or there is no minimizer, but instead a generalized minimizer. The latter consists of a symmetric nonincreasing function together with a Dirac delta measure at the center of symmetry. Moreover, sufficient conditions for the existence of a "proper" minimizer were given in [8]. The fact that in some cases there are no minimizers, but only generalized minimizers, was shown in [9]. The existence of a generalized minimizer can be understood as a partial mass concentration phenomenon. We find the appearance of this phenomenon in such a model rather surprising.

Let us be more specific. For $N/(N+\alpha) < q < 2N/(2N+\alpha)$, we consider the *relaxed functional*, defined on pairs (ρ, M), where $0 \leq \rho \in L^q(\mathbb{R}^N)$ and $M > 0$,

$$\begin{aligned}\mathcal{E}_{q,\alpha}^{\mathrm{rgKS}}[\rho, M] := &-\int_{\mathbb{R}^N} \rho^q\, dx + \frac{1}{2}\iint_{\mathbb{R}^N\times\mathbb{R}^N} \rho(x)|x-y|^\alpha \rho(y)\, dx\, dy \\ &+ M\int_{\mathbb{R}^N} |x|^\alpha \rho(x)\, dx. \end{aligned} \tag{4.1}$$

The corresponding minimization problem is

$$E_{q,\alpha}^{\mathrm{rgKS}} := \inf \left\{ \mathcal{E}_{q,\alpha}^{\mathrm{rgKS}}[\rho] : 0 \le \rho \in L^q(\mathbb{R}^N),\ M \ge 0,\ \int_{\mathbb{R}^N} \rho\, dx + M = 1 \right\}. \quad (4.2)$$

Intuitively, the energy $\mathcal{E}_{q,\alpha}^{\mathrm{rgKS}}[\rho, M]$ corresponds to the energy functional $\mathcal{E}_{q,\alpha}^{\mathrm{gKS}}$ evaluated at ρ plus a Dirac delta measure of mass M at the origin. Making this intuition rigorous, one finds that [8, equation (5)]

$$E_{q,\alpha}^{\mathrm{rgKS}} = E_{q,\alpha}^{\mathrm{gKS}}$$

and that $E_{q,\alpha}^{\mathrm{gKS}}$ has a minimizer if and only if $E_{q,\alpha}^{\mathrm{rgKS}}$ has a minimizer (ρ_*, M_*) with $M_* = 0$. Moreover, the same arguments as those applied for $q > 2N/(2N + \alpha)$ imply that $E_{q,\alpha}^{\mathrm{rgKS}}$ has a minimizer [8, Proposition 10] and that for any minimizer (ρ_*, M_*) the function ρ_* is radially symmetric with respect to some point, nonincreasing with respect to the distance from this point and positive almost everywhere [8, Lemma 9].

In view of the above discussion, for $N/(N + \alpha) < q < 2N/(2N + \alpha)$, the problem of existence of minimizers for $E_{q,\alpha}^{\mathrm{gKS}}$ is equivalent to the existence of a minimizer (ρ_*, M_*) for the problem $E_{q,\alpha}^{\mathrm{rgKS}}$ with $M_* = 0$. In [8], we gave sufficient conditions for this. Namely, for $N = 1, 2$, there is always a minimizer for $E_{q,\alpha}^{\mathrm{gKS}}$. The same is true for $N \ge 3$ and $\alpha \le 2N/(N-2)$. If $N \ge 3$ and $\alpha > 2N/(N-2)$, this is true provided $q \ge 1 - 2/N$ [8, Proposition 11].

In [9], the case $\alpha = 4$ was analyzed and an example of a minimizer for $E_{q,\alpha}^{\mathrm{rgKS}}$ with $M_* > 0$ was given. More precisely, it was shown that, for $N \ge 6$, the problem $E_{q,4}^{\mathrm{rgKS}}$ has a minimizer with $M_* > 0$ if $q < (N-2)(3N+4)/((N+2)(3N))$. Moreover, this result is optimal, in the sense that, for $N \ge 6$ and $q \ge (N-2)(3N+4)/((N+2)(3N))$, as well as for $N \le 5$, every minimizer of the problem $E_{q,4}^{\mathrm{rgKS}}$ has $M_* = 0$. The proof is based on a semiexplicit solution.

The paper [9] contains also numerical experiments that are consistent with the appearance of minimizers with $M_* > 0$ for $E_{q,4}^{\mathrm{rgKS}}$. This concentration phenomenon seems to be more pronounced for larger N, smaller q, and larger α.

Problem 4.1. Prove the existence of a "large" region of parameters q, α for which $E_{q,\alpha}^{\mathrm{rgKS}}$ has a minimizer (ρ_*, M_*) with $M_* > 0$.

Uniqueness. Uniqueness (up to translations) of minimizers, including minimizers of the relaxed functional, is known in two regimes, namely for $2 \le \alpha \le 4$ and for $\alpha \ge 1$ and $q \ge 1 - 1/N$ [8, Theorem 27]. The first result follows by a small generalization of a proof by Lopes [42], and the latter by the standard tool of displacement convexity in optimal mass transport.

The dynamical problem. The energy functional $\mathcal{E}^{\mathrm{gKS}}_{q,\alpha}$ or, more precisely, its rescaled version

$$-\frac{1}{1-q}\int_{\mathbb{R}^N}\rho^q\,dx+\frac{1}{2\alpha}\iint_{\mathbb{R}^N\times\mathbb{R}^N}\rho(x)|x-y|^\alpha\rho(y)\,dx\,dy \tag{4.3}$$

appears in connection with the aggregation-diffusion equations

$$\partial_t\rho=\Delta\rho^q+\nabla\cdot\big(\rho\nabla(W*\rho)\big),\quad W(x)=\alpha^{-1}|x|^\alpha. \tag{4.4}$$

Indeed, this time-dependent equation is the formal gradient flow with respect to the Wasserstein-2 distance of the free energy functional (4.3). Minimizers or, more generally, critical points of the free energy functional, restricted to probability densities, should play an important role for the long time behavior of solutions of (4.4). It seems particularly interesting to investigate whether in the dynamical setting there is a concentration effect similar to what we have seen for minimizing sequences for $E^{\mathrm{gKS}}_{q,\alpha}$ in case there is no minimizer, or equivalently there is a minimizer for $E^{\mathrm{rgKS}}_{q,\alpha}$ with $M_* > 0$.

Problem 4.2. Investigate the long time behavior of solutions of (4.4) in the case where $E^{\mathrm{rgKS}}_{q,\alpha}$ has a minimizer with $M_* > 0$.

To conclude this section, we mention that while we have focused on the free energy functional (4.3) in the case $\alpha > 0$ and $0 < q < 1$, it has been studied for all $q > 0$ and $\alpha > -N$. (Here we use the convention that $\alpha^{-1}|x-y|^\alpha$ is understood as $\ln|x-y|$ for $\alpha = 0$ and $(1-q)^{-1}\rho^q$ is understood as $-\rho\ln\rho$ for $q = 1$.) The nonexistence phenomenon via partial mass concentration that we discussed above, however, appears at most in the region $\alpha > 0$ and $0 < q < 1$. The case $\alpha > 0$ and $q \geq 1$ is treated in [8, Appendix B]. For $N = 2$, $q = 1$, and $\alpha = 0$ one obtains the original Keller–Segel free energy functional.

A. The generalized liquid drop model in 1D

In this appendix, we consider the minimization problem $E^{\mathrm{gld}}_\lambda(m)$ in the generalized liquid drop model for $0 < \lambda < 1$ in dimension $N = 1$. We will show that for $m \leq m_*$, single intervals are the unique (up to sets of measure zero) minimizers and for $m > m_*$ there are no minimizers. The computations are elementary.

It is well known (see, e.g., [45, Proposition 12.13]) that any set in $\mathbb{R}$ of finite measure and finite perimeter coincides, up to sets of measure zero, with a finite number of bounded intervals with disjoint closures. Moreover, the perimeter is twice the number of intervals. Clearly, if there is more than one interval, these intervals want to be infinitely far apart. Therefore,

$$\begin{aligned} E_\lambda^{\mathrm{gld}}(m) &= \inf\left\{ 2K + \frac{1}{2}\sum_{k=1}^K \int_{-m_k/2}^{m_k/2}\int_{-m_k/2}^{m_k/2} \frac{dx\,dy}{|x-y|^\lambda} : K\in\mathbb{N},\ \sum_{k=1}^K m_k = m \right\} \\ &= \inf\left\{ 2K + \frac{1}{(1-\lambda)(2-\lambda)}\sum_{k=1}^K m_k^{2-\lambda} : K\in\mathbb{N},\ \sum_{k=1}^K m_k = m \right\} \\ &= \inf_{K\in\mathbb{N}} \left(2K + \frac{1}{(1-\lambda)(2-\lambda)} K^{-1+\lambda} m^{2-\lambda} \right) \end{aligned}$$

and there is a minimizer if and only if the infimum occurs at $K=1$. Here we used

$$\sum_{k=1}^K m_k^{2-\lambda} \geq K^{-1+\lambda}\left(\sum_{k=1}^K m_k\right)^{2-\lambda}$$

(with equality if and only if all m_k are equal). The infimum is attained at $K=1$ if and only if $2 + (1-\lambda)^{-1}(2-\lambda)^{-1}m^{2-\lambda} \leq 2K + (1-\lambda)^{-1}(2-\lambda)^{-1}K^{-1+\lambda}m^{2-\lambda}$ for all $K\geq 2$, which is the same as

$$\begin{aligned} m &\leq \left(2(1-\lambda)(2-\lambda) \inf_{K\geq 2} \frac{K-1}{1-K^{-1+\lambda}} \right)^{1/(2-\lambda)} \\ &= \left(\frac{2(1-\lambda)(2-\lambda)}{1-2^{-1+\lambda}} \right)^{1/(2-\lambda)} = m_*. \end{aligned}$$

Here we used the fact that $\kappa \mapsto (\kappa-1)/(1-\kappa^{-1+\lambda})$ is increasing on $(1,\infty)$. This proves the claimed result.

Acknowledgments. The author would like to thank the organizers of the 8th European Congress of Mathematics for the organization of the meeting and for the invitation to speak. Since the topic of his invited talk was recently and rather exhaustively reviewed in [24], this contribution is based on a talk in a minisymposium at the congress, organized by L. Pick, to whom the author is very grateful. The results reviewed here were obtained in collaboration with many researchers and it is a pleasure to thank, in particular, José Carrillo, Matías Delgadino, Jean Dolbeault, Franca Hoffmann, Rowan Killip, Mathieu Lewin, Elliott Lieb, and Phan Thành Nam for many stimulating discussions.

Funding. This research was supported in part by NSF grants DMS-1363432 and DMS-1954995 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC-2111-390814868.

References

[1] G. Alberti, R. Choksi, and F. Otto, Uniform energy distribution for an isoperimetric problem with long-range interactions. *J. Amer. Math. Soc.* **22** (2009), no. 2, 569–605 Zbl 1206.49046 MR 2476783

[2] N. Bohr and J. A. Wheeler, The mechanism of nuclear fission. *Phys. Rev. (2)* **56** (1939), 426–450 Zbl 0022.19003

[3] M. Bonacini and R. Cristoferi, Local and global minimality results for a nonlocal isoperimetric problem on $\mathbb{R}^N$. *SIAM J. Math. Anal.* **46** (2014), no. 4, 2310–2349 Zbl 1301.49114 MR 3226747

[4] A. Burchard and G. R. Chambers, Geometric stability of the Coulomb energy. *Calc. Var. Partial Differential Equations* **54** (2015), no. 3, 3241–3250 Zbl 1331.26034 MR 3412409

[5] A. Burchard and G. R. Chambers, A stability result for Riesz potentials in higher dimensions. 2020, arXiv:2007.11664

[6] A. Burchard, R. Choksi, and E. Hess-Childs, On the strong attraction limit for a class of nonlocal interaction energies. *Nonlinear Anal.* **198** (2020), 111844 Zbl 1443.49021 MR 4083145

[7] A. Burchard, R. Choksi, and I. Topaloglu, Nonlocal shape optimization via interactions of attractive and repulsive potentials. *Indiana Univ. Math. J.* **67** (2018), no. 1, 375–395 Zbl 1402.49033 MR 3776026

[8] J. A. Carrillo, M. G. Delgadino, J. Dolbeault, R. L. Frank, and F. Hoffmann, Reverse Hardy–Littlewood–Sobolev inequalities. *J. Math. Pures Appl. (9)* **132** (2019), 133–165 Zbl 1442.35011 MR 4030251

[9] J. A. Carrillo, M. G. Delgadino, R. L. Frank, and M. Lewin, Fast diffusion leads to partial mass concentration in Keller-Segel type stationary solutions. *Math. Models Methods Appl. Sci.* **32** (2022), no. 4, 831–850 Zbl 07544556 MR 4421218

[10] J. A. Carrillo, M. G. Delgadino, and A. Mellet, Regularity of local minimizers of the interaction energy via obstacle problems. *Comm. Math. Phys.* **343** (2016), no. 3, 747–781 Zbl 1337.49066 MR 3488544

[11] R. Choksi, R. C. Fetecau, and I. Topaloglu, On minimizers of interaction functionals with competing attractive and repulsive potentials. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **32** (2015), no. 6, 1283–1305 Zbl 1329.49019 MR 3425263

[12] R. Choksi and M. A. Peletier, Small volume fraction limit of the diblock copolymer problem: I. Sharp-interface functional. *SIAM J. Math. Anal.* **42** (2010), no. 3, 1334–1370 Zbl 1210.49050 MR 2653253

[13] R. Choksi and M. A. Peletier, Small volume-fraction limit of the diblock copolymer problem: II. Diffuse-interface functional. *SIAM J. Math. Anal.* **43** (2011), no. 2, 739–763 Zbl 1223.49056 MR 2784874

[14] M. Christ, A sharpened Riesz–Sobolev inequality. 2017, arXiv:1706.02007

[15] M. Cicalese and G. P. Leonardi, A selection principle for the sharp quantitative isoperimetric inequality. *Arch. Ration. Mech. Anal.* **206** (2012), no. 2, 617–643 Zbl 1257.49045 MR 2980529

[16] M. Cicalese and E. Spadaro, Droplet minimizers of an isoperimetric problem with long-range interactions. *Comm. Pure Appl. Math.* **66** (2013), no. 8, 1298–1333 Zbl 1269.49085 MR 3069960

[17] K. Craig, I. Kim, and Y. Yao, Congested aggregation via Newtonian interaction. *Arch. Ration. Mech. Anal.* **227** (2018), no. 1, 1–67 Zbl 1384.35136 MR 3740370

[18] K. Craig and I. Topaloglu, Aggregation-diffusion to constrained interaction: minimizers & gradient flows in the slow diffusion limit. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **37** (2020), no. 2, 239–279 Zbl 1436.49015 MR 4072808

[19] J. Dou and M. Zhu, Reversed Hardy–Littewood–Sobolev inequality. *Int. Math. Res. Not. IMRN* **2015** (2015), no. 19, 9696–9726 Zbl 1329.26033 MR 3431607

[20] L. Emmert, R. L. Frank, and T. König, Liquid drop model for nuclear matter in the dilute limit. *SIAM J. Math. Anal.* **52** (2020), no. 2, 1980–1999 Zbl 1439.81090 MR 4089505

[21] A. Figalli, N. Fusco, F. Maggi, V. Millot, and M. Morini, Isoperimetry and stability properties of balls with respect to nonlocal energies. *Comm. Math. Phys.* **336** (2015), no. 1, 441–507 Zbl 1312.49051 MR 3322379

[22] A. Figalli, F. Maggi, and A. Pratelli, A mass transportation approach to quantitative isoperimetric inequalities. *Invent. Math.* **182** (2010), no. 1, 167–211 Zbl 1196.49033 MR 2672283

[23] R. L. Frank, Non-spherical equilibrium shapes in the liquid drop model. *J. Math. Phys.* **60** (2019), no. 7, 071506, 19 Zbl 1416.81223 MR 3981098

[24] R. L. Frank, The Lieb–Thirring inequalities: recent results and open problems. In *Nine Mathematical Challenges*, pp. 45–86, Proc. Sympos. Pure Math. 104, Amer. Math. Soc., Providence, RI, 2021 MR 4337417

[25] R. L. Frank, R. Killip, and P. T. Nam, Nonexistence of large nuclei in the liquid drop model. *Lett. Math. Phys.* **106** (2016), no. 8, 1033–1036 Zbl 1347.49069 MR 3520116

[26] R. L. Frank and E. H. Lieb, A compactness lemma and its application to the existence of minimizers for the liquid drop model. *SIAM J. Math. Anal.* **47** (2015), no. 6, 4436–4450 Zbl 1332.49042 MR 3425373

[27] R. L. Frank and E. H. Lieb, A "liquid-solid" phase transition in a simple model for swarming, based on the "no flat-spots" theorem for subharmonic functions. *Indiana Univ. Math. J.* **67** (2018), no. 4, 1547–1569 Zbl 1420.49040 MR 3853918

[28] R. L. Frank and E. H. Lieb, A note on a theorem of M. Christ. 2019, arXiv:1909.04598

[29] R. L. Frank and E. H. Lieb, Periodic energy minimizers for a one-dimensional liquid drop model. *Lett. Math. Phys.* **109** (2019), no. 9, 2069–2081 Zbl 1428.82062 MR 3996003

[30] R. L. Frank and E. H. Lieb, Proof of spherical flocking based on quantitative rearrangement inequalities. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **22** (2021), no. 3, 1241–1263 Zbl 07417802 MR 4334319

[31] R. L. Frank and P. T. Nam, Existence and nonexistence in the liquid drop model. *Calc. Var. Partial Differential Equations* **60** (2021), no. 6, Paper No. 223 Zbl 07414044 MR 4314139

[32] R. L. Frank, P. T. Nam, and H. Van Den Bosch, The ionization conjecture in Thomas–Fermi–Dirac–von Weizsäcker theory. *Comm. Pure Appl. Math.* **71** (2018), no. 3, 577–614 Zbl 1386.49063 MR 3762278

[33] N. Fusco, F. Maggi, and A. Pratelli, The sharp quantitative isoperimetric inequality. *Ann. of Math. (2)* **168** (2008), no. 3, 941–980 Zbl 1187.52009 MR 2456887

[34] N. Fusco and A. Pratelli, Sharp stability for the Riesz potential. *ESAIM Control Optim. Calc. Var.* **26** (2020), Paper No. 113 Zbl 1473.26036 MR 4185064

[35] G. Gamow, Mass defect curve and nuclear constitution. *Proc. Roy. Soc. Lond. Ser. A* **126** (1930), 632–644 Zbl 56.0762.02

[36] V. Julin, Isoperimetric problem with a Coulomb repulsive term. *Indiana Univ. Math. J.* **63** (2014), no. 1, 77–89 Zbl 1311.49110 MR 3218265

[37] H. Knüpfer and C. B. Muratov, On an isoperimetric problem with a competing nonlocal term I: The planar case. *Comm. Pure Appl. Math.* **66** (2013), no. 7, 1129–1162 Zbl 1269.49087 MR 3055587

[38] H. Knüpfer and C. B. Muratov, On an isoperimetric problem with a competing nonlocal term II: The general case. *Comm. Pure Appl. Math.* **67** (2014), no. 12, 1974–1994 Zbl 1302.49064 MR 3272365

[39] H. Knüpfer, C. B. Muratov, and M. Novaga, Low density phases in a uniformly charged liquid. *Comm. Math. Phys.* **345** (2016), no. 1, 141–183 Zbl 1346.49017 MR 3509012

[40] E. H. Lieb, Sharp constants in the Hardy–Littlewood–Sobolev and related inequalities. *Ann. of Math. (2)* **118** (1983), no. 2, 349–374 Zbl 0527.42011 MR 717827

[41] E. H. Lieb and M. Loss, *Analysis*. 2nd edn., Grad. Stud. Math. 14, Amer. Math. Soc., Providence, RI, 2001 Zbl 0966.26002 MR 1817225

[42] O. Lopes, Uniqueness and radial symmetry of minimizers for a nonlocal variational problem. *Commun. Pure Appl. Anal.* **18** (2019), no. 5, 2265–2282 MR 3962176

[43] J. Lu and F. Otto, Nonexistence of a minimizer for Thomas–Fermi–Dirac–von Weizsäcker model. *Comm. Pure Appl. Math.* **67** (2014), no. 10, 1605–1617 Zbl 1301.49002 MR 3251907

[44] J. Lu and F. Otto, An isoperimetric problem with Coulomb repulsion and attraction to a background nucleus. 2015, arXiv:1508.07172

[45] F. Maggi, *Sets of Finite Perimeter and Geometric Variational Problems. An Introduction to Geometric Measure Theory*. Cambridge Stud. Adv. Math. 135, Cambridge University Press, Cambridge, 2012 Zbl 1255.49074 MR 2976521

[46] C. B. Muratov and A. Zaleski, On an isoperimetric problem with a competing non-local term: quantitative results. *Ann. Global Anal. Geom.* **47** (2015), no. 1, 63–80 Zbl 1312.49053 MR 3302176

[47] P. T. Nam, The ionization problem. *Eur. Math. Soc. Newsl.* (2020), no. 118, 22–27 Zbl 1458.81046 MR 4226843

[48] X. Yan and Y. Yao, Sharp stability for the interaction energy. 2020, arXiv:2008.07502

Rupert L. Frank
Mathematisches Institut, Ludwig–Maximilans Universität München, Theresienstr. 39,
80333 München; Munich Center for Quantum Science and Technology, Schellingstr. 4,
80799 München, Germany; and Mathematics 253-37, Caltech, Pasadena, CA 91125, USA;
r.frank@lmu.de

DOI 10.4171/8ECM/02

Laplacians on infinite graphs: Discrete vs. continuous

Aleksey Kostenko and Noema Nicolussi

Abstract. There are two main notions of a Laplacian operator associated with graphs: discrete graph Laplacians and continuous Laplacians on metric graphs (widely known as quantum graphs). Both objects have a venerable history as they are related to several diverse branches of mathematics and mathematical physics. The existing literature usually treats these two Laplacian operators separately. In this overview, we will focus on the relationship between them (spectral and parabolic properties). Our main conceptual message is that these two settings should be regarded as complementary (rather than opposite) and exactly their interplay leads to important further insight on both sides.

1. Introduction

Laplacian operators on graphs have a long history and enjoy deep connections to several branches of mathematics and mathematical physics. There are two different notions of Laplacians appearing in this context: the key features of (continuous) *Laplacians on metric graphs*, which are also known as *quantum graphs*, include their use as simplified models of complicated quantum systems (see, e.g., [4, 19, 21, 56]) and the appearance of metric graphs in tropical and algebraic geometry, where they serve as non-Archimedean analogues of Riemann surfaces (see, e.g., [1, 17]). The subject of *discrete Laplacians on graphs* is even wider, and a partial overview of the immense literature can be found in [2, 9, 10, 43, 70].

The study of both types of graph Laplacians is heavily influenced by the corresponding investigations in the manifold setting (e.g., spectral geometry of manifolds). In fact, one can also put Laplacians on manifolds, metric graphs, and discrete graphs under the overarching umbrella of *Dirichlet forms*, which provides the systematic framework for studying heat and diffusion processes. From this perspective, metric graph Laplacians have much in common with Laplacians on manifolds since both can be treated in the framework of strongly local Dirichlet forms. Moreover, the notion of an intrinsic metric, first mentioned by E. B. Davies and later emphasized by M. Biroli,

2020 Mathematics Subject Classification. Primary 34B45; Secondary 47B25, 05C63, 81Q35, 35R02, 60J27.
Keywords. Graph, metric graph, Laplacian, intrinsic metric, random walk.

U. Mosco, and K.-T. Sturm (see, e.g., [65]), allows to directly transfer many important results from manifolds to the abstract setting of strongly local Dirichlet forms (and hence metric graph Laplacians). In contrast to this, discrete graph Laplacians are difference operators and hence provide examples of nonlocal operators (e.g., no Leibniz rule). In particular, difficulties in analyzing random walks on graphs often stem from exactly this fact. On the other hand, this area has seen a tremendous progress in the last decade. In our opinion, the successful introduction and systematic use of the notion of *intrinsic metrics on graphs* played (and continues to play!) a major role in this breakthrough (see the fresh monograph [43]).

Despite a vast interest in both types of graph Laplacians, the existing literature usually treats them separately. In the present overview, we mainly focus on the relationship between them and survey connections on different levels (spectral and parabolic properties). This leads to a systematic way of connecting the settings and several applications. Our main conceptual message is that discrete and continuous graph Laplacians should be regarded as complementary (rather than opposite) and exactly their interplay leads to important further insight on both sides. This relationship can also be formulated in the language of intrinsic metrics. Indeed, a large class of intrinsic metrics on discrete graphs is obtained as restrictions to vertices of intrinsic metrics on (weighted) metric graphs. In particular, from this perspective metric graphs indeed serve as a bridge between graphs and manifolds, a heuristic principle which is often mentioned in context with graph Laplacians. Let us also mention that the stochastic side of these connections, namely the approach of using Brownian motion on metric graphs to study random walks on discrete graphs, has been employed several times in the literature [3, 22, 23, 33, 36, 67] (see also references therein).

Most of the results presented here are carefully explained in the recent monograph [49], which also contains many other results not mentioned in this text.

2. Preliminaries

2.1. Graphs

Let us recall basic notions (we mainly follow the terminology in [16]). Let $\mathcal{G}_d = (\mathcal{V}, \mathcal{E})$ be an undirected *graph*; that is, $\mathcal{V}$ is a finite or countably infinite set of vertices and $\mathcal{E}$ is a finite or countably infinite set of edges. Two vertices $u, v \in \mathcal{V}$ are called *neighbors*, and we shall write $u \sim v$ if there is an edge $e_{u,v} \in \mathcal{E}$ connecting u and v. For every $v \in \mathcal{V}$, we define $\mathcal{E}_v$ as the set of edges incident to v. We stress that we allow *multigraphs*; that is, we allow *multiple edges* (two vertices can be joined by several edges) and *loops* (edges from a vertex to itself). Graphs without loops and multiple edges are called *simple*.

Example 2.1 (Cayley graphs). Let G be a finitely generated group and let S be a generating set of G. We shall always assume that

- S is symmetric, $S = S^{-1}$ and finite, $\#S < \infty$,
- the identity element of G does not belong to S (this excludes loops).

The *Cayley graph* $\mathcal{G}_C = \mathcal{C}(\mathsf{G}, S)$ of G w.r.t. S is the simple graph whose vertex set coincides with G and two vertices $x, y \in \mathsf{G}$ are neighbors if and only if $xy^{-1} \in S$.

Sometimes it is convenient to assign an *orientation* on $\mathcal{G}_d$: to each edge $e \in \mathcal{E}$ one assigns the pair (e_ι, e_τ) of its *initial* e_ι and *terminal* e_τ vertices. We shall denote the corresponding oriented graph by $\vec{\mathcal{G}}_d = (\mathcal{V}, \vec{\mathcal{E}})$, where $\vec{\mathcal{E}}$ denotes the set of oriented edges. Notice that for an oriented loop we do distinguish between its initial and terminal vertices. Next, for every vertex $v \in \mathcal{V}$, set

$$\mathcal{E}_v^+ = \big\{(e_\iota, e_\tau) \in \vec{\mathcal{E}} \mid e_\iota = v\big\}, \qquad \mathcal{E}_v^- = \big\{(e_\iota, e_\tau) \in \vec{\mathcal{E}} \mid e_\tau = v\big\}, \tag{2.1}$$

and let $\vec{\mathcal{E}}_v$ be the disjoint union of outgoing $\mathcal{E}_v^+$ and incoming $\mathcal{E}_v^-$ edges,

$$\vec{\mathcal{E}}_v := \mathcal{E}_v^+ \sqcup \mathcal{E}_v^- = \vec{\mathcal{E}}_v^+ \cup \vec{\mathcal{E}}_v^-, \qquad \vec{\mathcal{E}}_v^\pm := \big\{(\pm, e) \mid e \in \mathcal{E}_v^\pm\big\}. \tag{2.2}$$

The *(combinatorial) degree* of $v \in \mathcal{V}$ is

$$\deg(v) := \#(\vec{\mathcal{E}}_v) = \#(\vec{\mathcal{E}}_v^+) + \#(\vec{\mathcal{E}}_v^-) = \#(\mathcal{E}_v) + \#\{e \in \mathcal{E}_v \mid e \text{ is a loop}\}. \tag{2.3}$$

Notice that if $\mathcal{E}_v$ contains no loops, then $\deg(v) = \#(\mathcal{E}_v)$. The graph $\mathcal{G}_d$ is called *locally finite* if $\deg(v) < \infty$ for all $v \in \mathcal{V}$.

A sequence of (unoriented) edges $\mathcal{P} = (e_{v_0,v_1}, e_{v_1,v_2}, \dots, e_{v_{n-1},v_n})$, where $e_{v_i,v_{i+1}}$ connects the vertices v_i and v_{i+1}, is called a *path* of (combinatorial) length $n \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$. Notice that for simple graphs each path $\mathcal{P}$ can be identified with its sequence of vertices $\mathcal{P} = (v_k)_{k=0}^n$. A graph $\mathcal{G}_d$ is called *connected* if for any two vertices there is a path connecting them.

We shall always make the following assumptions on the geometry of $\mathcal{G}_d$.

Hypothesis 2.2. *$\mathcal{G}_d$ is connected and locally finite.*

Remark. We assume connectivity for convenience reasons only (one can always consider each connected component of a graph separately). However, the assumption that a graph is locally finite is indeed important in our considerations.

2.2. Metric graphs

Assigning each edge $e \in \mathcal{E}$ a finite length $|e| \in (0, \infty)$, we can naturally associate with $(\mathcal{G}_d, |\cdot|) = (\mathcal{V}, \mathcal{E}, |\cdot|)$ a metric space $\mathcal{G}$. First, we identify each edge $e \in \mathcal{E}$ with a copy of the interval $\mathcal{I}_e = [0, |e|]$, which also assigns an orientation on $\mathcal{E}$ upon identification of $e_\imath$ and e_τ with the left, respectively, right endpoint of $\mathcal{I}_e$. The topological space $\mathcal{G}$ is then obtained by "glueing together" the ends of edges corresponding to the same vertex v (in the sense of a topological quotient; see, e.g., [6, Chap. 3.2.2]). The topology on $\mathcal{G}$ is metrizable by the *length metric* ϱ_0—the distance between two points $x, y \in \mathcal{G}$ is defined as the arc length of the "shortest path" connecting them (such a path does not necessarily exist and one needs to take the infimum over all paths connecting x and y).

A *metric graph* is a (locally compact) metric space $\mathcal{G}$ arising from the above construction for some collection $(\mathcal{G}_d, |\cdot|) = (\mathcal{V}, \mathcal{E}, |\cdot|)$. More specifically, $\mathcal{G}$ is then called the *metric realization* of $(\mathcal{G}_d, |\cdot|)$, and a pair $(\mathcal{G}_d, |\cdot|)$ whose metric realization coincides with $\mathcal{G}$ is called a *model* of $\mathcal{G}$. For a thorough discussion of metric graphs as topological and metric spaces we refer to [31, Chap. I].

Remark. Let us stress that a metric graph $\mathcal{G}$ equipped with the length metric ϱ_0 (or with any other path metric) is a *length space* (see [6, Chap. 2.1] for definitions and further details). Notice also that complete, locally compact length spaces are *geodesic*; that is, every two points can be connected by a shortest path.

Clearly, different models may give rise to the same metric graph. Moreover, any metric graph has infinitely many models (e.g., they can be constructed by subdividing edges using vertices of degree two). A model $(\mathcal{V}, \mathcal{E}, |\cdot|)$ is called *simple* if the corresponding graph $(\mathcal{V}, \mathcal{E})$ is simple. In particular, every metric graph has a simple model, and this indicates that restricting to simple graphs, that is, assuming in addition to Hypothesis 2.2 that $\mathcal{G}_d$ has no loops or multiple edges, would not be a restriction at all when dealing with metric graphs.

Remark. In most parts of our paper, we will consider a metric graph together with a fixed choice of its model. In this situation, we will usually be slightly imprecise and do not distinguish between these two objects. In particular, we will denote both objects by the same letter $\mathcal{G}$ and write $\mathcal{G} = (\mathcal{V}, \mathcal{E}, |\cdot|)$ or $\mathcal{G} = (\mathcal{G}_d, |\cdot|)$.

Remark (Metric graph as a 1d manifold with singularities). Sometimes it is useful to consider metric graphs as 1d manifolds with singularities. Since every point $x \in \mathcal{G}$ has a neighborhood isomorphic to a star-shaped set

$$\mathcal{E}\big(\deg(x), r_x\big) := \big\{z = r\mathrm{e}^{2\pi\mathrm{i}k/\deg(x)} \mid r \in [0, r_x),\ k = 1, \dots, \deg(x)\big\} \subset \mathbb{C}, \quad (2.4)$$

one may introduce the set of *tangential directions* $T_x(\mathcal{G})$ at x as the set of unit vectors $\mathrm{e}^{2\pi\mathrm{i}k/\deg(x)}$, $k = 1, \dots, \deg(x)$. Then all vertices $v \in \mathcal{V}$ with $\deg(v) \geq 3$ are considered as *branching points/singularities* and vertices $v \in \mathcal{V}$ with $\deg(v) = 1$ as *boundary points*. Notice that for every vertex $v \in \mathcal{V}$ the set of tangential directions

$T_v(\mathcal{G})$ can be identified with $\vec{\mathcal{E}}_v$. If there are no loop edges at the vertex $v \in \mathcal{V}$, then $T_v(\mathcal{G})$ is identified with $\mathcal{E}_v$ in this way.

3. Graph Laplacians

When speaking about graph Laplacians, one often considers one of the operators in the next two examples.

Example 3.1 (Combinatorial Laplacian). For a simple graph $\mathcal{G}_d = (\mathcal{V}, \mathcal{E})$ satisfying Hypothesis 2.2, the so-called *combinatorial Laplacian* is defined on $C(\mathcal{V})$ by

$$\begin{aligned}(L_{\mathrm{comb}} f)(v) &= \sum_{u \sim v} f(v) - f(u) \\ &= \deg(v) f(v) - \sum_{u \sim v} f(u), \quad v \in \mathcal{V}.\end{aligned} \tag{3.1}$$

Here $C(\mathcal{V})$ is the set of complex-valued functions on a countable set $\mathcal{V}$. Notice that the second summand on the RHS,

$$(\mathcal{A} f)(v) = \sum_{u \sim v} f(u), \quad v \in \mathcal{V},$$

is nothing but the operator generated by the adjacency matrix of $\mathcal{G}_d$, which explains the name of L_{comb}. The combinatorial Laplacian plays a crucial role in many areas of mathematics, physics, and engineering. In particular, the relationship between the spectral properties of L_{comb} and various graph parameters is one of the core topics within the field of *Spectral Graph Theory* (see [9, 10] for further details).

Example 3.2 (Normalized Laplacian). Assuming again that $\mathcal{G}_d = (\mathcal{V}, \mathcal{E})$ is a simple graph satisfying Hypothesis 2.2, consider another operator defined on $C(\mathcal{V})$ by

$$\begin{aligned}(L_{\mathrm{norm}} f)(v) &= \frac{1}{\deg(v)} \sum_{u \sim v} f(v) - f(u) \\ &= f(v) - \frac{1}{\deg(v)} \sum_{u \sim v} f(u)\end{aligned} \tag{3.2}$$

for every $v \in \mathcal{V}$. The second summand on the RHS,

$$(\mathcal{M} f)(v) = \frac{1}{\deg(v)} \sum_{u \sim v} f(u), \quad v \in \mathcal{V},$$

is the so-called *Markov (averaging) operator*. Notice that due to our assumptions on $\mathcal{G}_d$, $\mathcal{M}$ is a stochastic matrix known as the *transition matrix* for the simple random walk on the graph. The normalized Laplacian serves as the generator of the simple random walk on $\mathcal{G}_d$ (see, e.g., [2, 70]).

Remark. If the underlying graph $\mathcal{G}_d$ is *regular* ($\deg \equiv c$ is constant on $\mathcal{V}$; for instance, Cayley graphs are regular), then $L_{\mathrm{comb}} = c \cdot L_{\mathrm{norm}} = c \cdot \mathrm{I} - \mathcal{A}$. However, in general these two Laplacians may have very different properties. For instance, L_{norm} generates a bounded operator in $\ell^2(\mathcal{V}; \deg)$ and L_{comb} gives rise to a bounded operator in $\ell^2(\mathcal{V})$ only if $\mathcal{G}_d$ has bounded geometry, i.e., deg is bounded on $\mathcal{V}$ (see (3.6)).

The above two examples can be put into a much more general framework. Namely, let $\mathcal{V}$ be a countable set. A function $m\colon \mathcal{V} \to (0,\infty)$ defines a measure of full support on $\mathcal{V}$ in an obvious way. A pair $(\mathcal{V}, m)$ is called a *discrete measure space*. The set of square summable (w.r.t. m) functions

$$\ell^2(\mathcal{V}; m) = \Big\{ f \in C(\mathcal{V}) \mid \|f\|^2_{\ell^2(\mathcal{V};m)} := \sum_{v\in\mathcal{V}} \big|f(v)\big|^2 m(v) < \infty \Big\}$$

has a natural Hilbert space structure.

Suppose that $b\colon \mathcal{V}\times\mathcal{V} \to [0,\infty)$ satisfies the following conditions:

(i) *symmetry*: $b(u,v) = b(v,u)$ for each pair $(u,v) \in \mathcal{V}\times\mathcal{V}$,

(ii) *vanishing diagonal*: $b(v,v) = 0$ for all $v \in \mathcal{V}$,

(iii) *locally finite*: $\#\{u \in \mathcal{V} \mid b(u,v) \neq 0\} < \infty$ for all $v \in \mathcal{V}$,[1]

(iv) *connected*: for any $u, v \in \mathcal{V}$ there is a finite collection $(v_k)_{k=0}^n \subset \mathcal{V}$ such that $u = v_0$, $v = v_n$ and $b(v_{k-1}, v_k) > 0$ for all $k \in \{1,\dots,n\}$.

Following [41,43], b is called a *(weighted) graph* over $\mathcal{V}$ or over $(\mathcal{V}, m)$ if in addition a measure m of full support on $\mathcal{V}$ is given (b is also called an *edge weight*). To simplify notation, we shall denote a graph b over $(\mathcal{V}, m)$ by $(\mathcal{V}, m; b)$.

Remark. To any graph b over $\mathcal{V}$, we can naturally associate a simple combinatorial graph $\mathcal{G}_b$. Namely, the vertex set of $\mathcal{G}_b$ is $\mathcal{V}$ and its edge set $\mathcal{E}_b$ is defined by calling two vertices $u, v \in \mathcal{V}$ neighbors, $u \sim v$, exactly when $b(u,v) > 0$. Clearly, $\mathcal{G}_b = (\mathcal{V}, \mathcal{E}_b)$ is an undirected graph in the sense of Section 2.1. Let us stress, however, that the constructed *graph $\mathcal{G}_b$ is always simple*.

The *(formal) Laplacian* $L = L_{m,b}$ associated to a graph b over $(\mathcal{V}, m)$ is given by

$$(Lf)(v) = \frac{1}{m(v)} \sum_{u\in\mathcal{V}} b(v,u)\big(f(v) - f(u)\big), \quad v \in \mathcal{V}. \tag{3.3}$$

[1] In fact, using the form approach, one can considerably relax this condition by replacing it with the *local summability*: $\sum_{v\in\mathcal{V}} b(u,v) < \infty$ for all $u \in \mathcal{V}$.

It acts on functions $f \in C(\mathcal{V})$ and this naturally leads to the *maximal* Laplacian $\mathbf{h}$ in $\ell^2(\mathcal{V}; m)$ defined by

$$\mathbf{h} = L \restriction \operatorname{dom}(\mathbf{h}), \quad \operatorname{dom}(\mathbf{h}) = \big\{ f \in \ell^2(\mathcal{V}; m) \mid Lf \in \ell^2(\mathcal{V}; m) \big\}. \tag{3.4}$$

This operator is closed; however, if $\mathcal{V}$ is infinite, it is not symmetric in general (cf. [41, Thm. 6]). Taking into account that b is locally finite, it is clear that $C_c(\mathcal{V}) \subseteq \operatorname{dom}(\mathbf{h})$, where $C_c(\mathcal{V})$ is the space of compactly supported functions in $C(\mathcal{V})$ (w.r.t. the discrete topology on $\mathcal{V}$). Therefore, we can introduce the *minimal* Laplacian $\mathbf{h}^0$ as the closure in $\ell^2(\mathcal{V}; m)$ of the *pre-minimal* Laplacian

$$\mathbf{h}' = L \restriction \operatorname{dom}(\mathbf{h}'), \quad \operatorname{dom}(\mathbf{h}') = C_c(\mathcal{V}). \tag{3.5}$$

Then $\mathbf{h}' \subseteq \mathbf{h}^0 \subseteq \mathbf{h}$ and $(\mathbf{h}')^* = (\mathbf{h}^0)^* = \mathbf{h}$. If $\mathbf{h}^0 = \mathbf{h}$, then $\mathbf{h}$ is *self-adjoint* as an operator in the Hilbert space $\ell^2(\mathcal{V}; m)$ (and $\mathbf{h}'$ is called *essentially self-adjoint*). The problem of self-adjointness is a classical topic, which is of central importance in quantum mechanics (see, e.g., [58, Chap. VIII.11]). We shall return to this issue in Section 8.1. Let us now only mention that the self-adjointness takes place whenever $L = L_{m,b}$ gives rise to a bounded operator on $\ell^2(\mathcal{V}; m)$. It is rather well known (see, e.g., [13, Lem. 1], [40, Thm. 11], and [66, Rem. 1]) that the Laplacian $L = L_{m,b}$ is bounded on $\ell^2(\mathcal{V}; m)$ if and only if the weighted degree function $\operatorname{Deg}\colon \mathcal{V} \to [0, \infty)$ given by

$$\operatorname{Deg}\colon v \mapsto \frac{1}{m(v)} \sum_{u \in \mathcal{V}} b(u, v) \tag{3.6}$$

is bounded on $\mathcal{V}$. In this case, $\mathbf{h}^0 = \mathbf{h}$ and $\|\operatorname{Deg}\|_\infty \le \|\mathbf{h}\|_{\ell^2(\mathcal{V};m)} \le 2\|\operatorname{Deg}\|_\infty$.

Remark. For the combinatorial Laplacian L_{comb}, we have $\operatorname{Deg}_{\mathrm{comb}}(v) = \deg(v)$ and hence L_{comb} is bounded exactly when $\mathcal{G}_d$ has bounded geometry. For the normalized Laplacian L_{norm}, $\operatorname{Deg}_{\mathrm{norm}}(v) \le 1$ for all $v \in \mathcal{V}$ and hence $\|L_{\mathrm{norm}}\| \le 2$.

There is another way to associate a self-adjoint operator in $\ell^2(\mathcal{V}; m)$ with the Laplacian L. With each graph b one can associate the *energy form* $\mathfrak{q}\colon C(\mathcal{V}) \to [0, \infty]$ defined by

$$\mathfrak{q}[f] = \mathfrak{q}_b[f] := \frac{1}{2} \sum_{u,v \in \mathcal{V}} b(v, u) \big| f(v) - f(u) \big|^2. \tag{3.7}$$

Functions $f \in C(\mathcal{V})$ such that $\mathfrak{q}[f] < \infty$ are called *finite energy functions*. Clearly,[2] $C_c(\mathcal{V})$ belongs to the set $\mathcal{D}(\mathfrak{q})$ of finite energy functions and $\langle \mathbf{h} f, f \rangle_{\ell^2(m)} = \mathfrak{q}[f]$

[2] Actually, it suffices to assume that b satisfies the local summability condition; see [41,43].

for all $f \in C_c(\mathcal{V})$. If b is a graph over $(\mathcal{V}, m)$, introduce the graph norm

$$\|f\|_{\mathfrak{q}}^2 := \mathfrak{q}[f] + \|f\|_{\ell^2(\mathcal{V};m)}^2 \tag{3.8}$$

for all $f \in \mathcal{D}(\mathfrak{q}) \cap \ell^2(\mathcal{V}; m) =: \operatorname{dom}(\mathfrak{q})$. Clearly, $\operatorname{dom}(\mathfrak{q})$ is the maximal domain of definition of the form $\mathfrak{q}$ in the Hilbert space $\ell^2(\mathcal{V}; m)$; let us denote this form by $\mathfrak{q}_N$. Restricting further to compactly supported functions and then taking the graph norm closure, we get another form:

$$\mathfrak{q}_D := \mathfrak{q} \restriction \operatorname{dom}(\mathfrak{q}_D), \quad \operatorname{dom}(\mathfrak{q}_D) := \overline{C_c(\mathcal{V})}^{\|\cdot\|_{\mathfrak{q}}}.$$

It turns out that both $\mathfrak{q}_D$ and $\mathfrak{q}_N$ are *Dirichlet forms* (for definitions see [26]) and $\mathfrak{q}_D$ is a *regular Dirichlet form*. Moreover, the converse is also true: *"every (irreducible) regular Dirichlet form over* $(\mathcal{V}, m)$ *arises as the energy form* $\mathfrak{q}_D$ *for some (connected) graph b over* $(\mathcal{V}, m)$*"* (this claim is wrong as stated; however, to make it correct one needs to replace locally finite by the local summability condition on b and also to allow killing terms; see [41, Thm. 7]).

Remark. The notion of *irreducibility* for Dirichlet forms on graphs is closely connected with the notion of *connectivity*. Recall that a graph b is called *connected* if the corresponding graph $\mathcal{G}_b$ is connected. Then the regular Dirichlet form $\mathfrak{q}_D$ is irreducible exactly when the underlying graph b is connected (e.g., [43, Chap. 1.4]).

Now using the representation theorems for quadratic forms (see, e.g., [38]), one can associate in $\ell^2(\mathcal{V}; m)$ the self-adjoint operators $\mathbf{h}_D$ and $\mathbf{h}_N$, the so-called *Dirichlet* and *Neumann Laplacians* over $(\mathcal{V}, m)$, with, respectively, $\mathfrak{q}_D$ and $\mathfrak{q}_N$. Usually, it is a rather nontrivial task to provide an explicit description of the operators $\mathbf{h}_D$ and, especially, $\mathbf{h}_N$.[3] However, the following abstract description always holds:

$$\mathbf{h}_D = \mathbf{h} \restriction \operatorname{dom}(\mathbf{h}_D), \quad \operatorname{dom}(\mathbf{h}_D) = \operatorname{dom}(\mathbf{h}) \cap \operatorname{dom}(\mathfrak{q}_D), \tag{3.9}$$

which also implies that $\mathbf{h}_D$ is the *Friedrichs extension* of the adjoint $\mathbf{h}^0 = \mathbf{h}^*$ to $\mathbf{h}$.

4. Laplacians on metric graphs

4.1. Function spaces on metric graphs

Let $\mathcal{G}$ be a metric graph with a fixed model $(\mathcal{V}, \mathcal{E}, |\cdot|)$. Let also $\mu\colon \mathcal{E} \to (0, \infty)$ be a weight function assigning a positive weight $\mu(e)$ to each edge $e \in \mathcal{E}$. We shall assume that edge weights are orientation independent and we set $\mu(\vec{e}) = \mu(e)$ for all

[3] In fact, to decide whether $\mathbf{h}_N$ and $\mathbf{h}_D$ coincide for given b and m, or equivalently that $\mathfrak{q}_N = \mathfrak{q}_D$, is already a highly nontrivial problem. This property is related to the uniqueness of a *Markovian extension*. For further details we refer to [43, 46], [49, Chap. 7.2].

$\vec{e} \in \vec{\mathcal{E}}_v$, $v \in \mathcal{V}$. Notice that μ can be identified with an edgewise constant function on $\mathcal{G}$ in an obvious way. Identifying every edge $e \in \mathcal{E}$ with a copy of $\mathcal{I}_e = [0, |e|]$, we can introduce Lebesgue and Sobolev spaces on edges and also on $\mathcal{G}$. First of all, with the weight μ we associate the measure μ on $\mathcal{G}$ defined as the edgewise scaled Lebesgue measure such that $\mu(\mathrm{d}x) = \mu(e)\mathrm{d}x_e$ on every edge $e \in \mathcal{E}$. Thus we can define the Hilbert space $L^2(\mathcal{G};\mu)$ of measurable functions $f : \mathcal{G} \to \mathbb{C}$ which are square integrable w.r.t. the measure μ on $\mathcal{G}$. Similarly, one defines the Banach spaces $L^p(\mathcal{G};\mu)$ for $p \in [1,\infty]$. In fact, if $p \in [1,\infty)$, then

$$L^p(\mathcal{G};\mu) \cong \Big\{ f = (f_e)_{e\in\mathcal{E}} \mid f_e \in L^p(e;\mu), \sum_{e\in\mathcal{E}} \|f_e\|^p_{L^p(e;\mu)} < \infty \Big\},$$

where

$$\|f_e\|^p_{L^p(e;\mu)} = \int_e |f_e(x_e)|^p \mu(\mathrm{d}x_e) = \mu(e) \int_e |f_e(x_e)|^p \,\mathrm{d}x_e.$$

If $\mu(e) = 1$, then we shall simply write $L^p(e)$. Next, the subspace of compactly supported L^p functions will be denoted by $L^p_c(\mathcal{G};\mu)$. The space $L^p_{\mathrm{loc}}(\mathcal{G};\mu)$ of locally L^p functions consists of all measurable functions f such that $fg \in L^p_c(\mathcal{G};\mu)$ for all $g \in C_c(\mathcal{G})$. Notice that both L^p_{loc} and L^p_c are independent of the weight μ.

For edgewise locally absolutely continuous functions on $\mathcal{G}$, let us denote by ∇ the edgewise first derivative,

$$\nabla : f \mapsto f'. \tag{4.1}$$

Then for every edge $e \in \mathcal{E}$,

$$H^1(e) = \{ f \in AC(e) \mid \nabla f \in L^2(e) \}, \quad H^2(e) = \{ f \in H^1(e) \mid \nabla f \in H^1(e) \}$$

are the usual Sobolev spaces (upon the identification of e with $\mathcal{I}_e = [0, |e|]$), and $AC(e)$ is the space of absolutely continuous functions on e. Let us denote by $H^1_{\mathrm{loc}}(\mathcal{G} \setminus \mathcal{V})$ and $H^2_{\mathrm{loc}}(\mathcal{G} \setminus \mathcal{V})$ the spaces of measurable functions f on $\mathcal{G}$ such that their edgewise restrictions belong to H^1, respectively, H^2; that is,

$$H^j_{\mathrm{loc}}(\mathcal{G} \setminus \mathcal{V}) = \{ f \in L^2_{\mathrm{loc}}(\mathcal{G}) \mid f|_e \in H^j(e) \ \forall e \in \mathcal{E} \}$$

for $j \in \{1,2\}$. Clearly, for each $f \in H^2_{\mathrm{loc}}(\mathcal{G} \setminus \mathcal{V})$ the quantities

$$f(e_\iota) := \lim_{x_e \to e_\iota} f(x_e), \quad f(e_\tau) := \lim_{x_e \to e_\tau} f(x_e) \tag{4.2}$$

and the normal derivatives

$$\partial f(e_\iota) := \lim_{x_e \to e_\iota} \frac{f(x_e) - f(e_\iota)}{|x_e - e_\iota|}, \quad \partial f(e_\tau) := \lim_{x_e \to e_\tau} \frac{f(x_e) - f(e_\tau)}{|x_e - e_\tau|} \tag{4.3}$$

are well defined for all $e \in \mathcal{E}$. We also need the notation

$$\partial_{\vec{e}} f(v) := \begin{cases} \partial f(e_\imath), & \vec{e} \in \vec{\mathcal{E}}_v^+, \\ \partial f(e_\tau), & \vec{e} \in \vec{\mathcal{E}}_v^-, \end{cases} \tag{4.4}$$

for every $v \in \mathcal{V}$ and $\vec{e} \in \vec{\mathcal{E}}_v$. In the case of a loopless graph, the above notation simplifies since we can identify $\vec{\mathcal{E}}_v$ with $\mathcal{E}_v$ for all $v \in \mathcal{V}$.

4.2. Kirchhoff Laplacians

Let $\mathcal{G}$ be a metric graph together with a fixed model $(\mathcal{V}, \mathcal{E}, |\cdot|)$ and $\mu,\ \nu\colon \mathcal{E} \to (0,\infty)$ two edge weights on $\mathcal{G}$ (for this model). For every $e \in \mathcal{E}$ consider the maximal operator $\mathrm{H}_{e,\max}$ defined in $L^2(e;\mu)$ by

$$\mathrm{H}_{e,\max} f = \tau_e f, \quad \tau_e = -\frac{1}{\mu(x_e)}\frac{\mathrm{d}}{\mathrm{d}x_e}\nu(x_e)\frac{\mathrm{d}}{\mathrm{d}x_e}, \tag{4.5}$$

$$\operatorname{dom}(\mathrm{H}_{e,\max}) = \big\{f \in L^2(e;\mu) \mid f, \nu f' \in AC(e),\ \tau_e f \in L^2(e;\mu)\big\}. \tag{4.6}$$

Since μ and ν are constant on e, $\operatorname{dom}(\mathrm{H}_{e,\max})$ coincides with the Sobolev space $H^2(e)$. The maximal operator on $\mathcal{G}$ is then defined in $L^2(\mathcal{G};\mu)$ as

$$\mathbf{H}_{\max} = \bigoplus_{e\in\mathcal{E}} \mathrm{H}_{e,\max}. \tag{4.7}$$

Clearly, for each $f \in \operatorname{dom}(\mathbf{H}_{\max})$ the quantities (4.2), (4.3), and hence (4.4) are well defined for all $e \in \mathcal{E}$. Now, in order to reflect the underlying graph structure, we impose at each vertex $v \in \mathcal{V}$ the *Kirchhoff boundary conditions*

$$\begin{cases} f \text{ is continuous at } v, \\ \displaystyle\sum_{\vec{e}\in\vec{\mathcal{E}}_v} \nu(\vec{e})\partial_{\vec{e}} f(v) = 0. \end{cases} \tag{4.8}$$

To motivate our definition, consider ∇ as the differentiation operator on $\mathcal{G}$ acting on functions which are edgewise locally absolutely continuous and also continuous at the vertices. Notice that when considering ∇ as an operator acting from $L^2(\mathcal{G};\mu)$ to $L^2(\mathcal{G};\nu)$, its formal adjoint $\nabla^\dagger$ acting from $L^2(\mathcal{G};\nu)$ to $L^2(\mathcal{G};\mu)$ acts edgewise as

$$\nabla^\dagger\colon f \mapsto -\frac{1}{\mu}(\nu f)'. \tag{4.9}$$

Thus the weighted Laplacian Δ acting in $L^2(\mathcal{G};\mu)$, written in the divergence form $\Delta\colon f \mapsto -\nabla^\dagger(\nabla f)$, acts edgewise as the following divergence form Sturm–Liouville operator:

$$\Delta\colon f \mapsto \frac{1}{\mu}(\nu f')'. \tag{4.10}$$

The continuity assumption imposed on f results for Δ in a one-parameter family of symmetric boundary conditions at the vertices (the so-called δ-coupling). In the present text, with the Laplacian Δ acting on $\mathcal{G}$ we shall always associate the *Kirchhoff* vertex conditions (4.8). In particular, imposing these boundary conditions on the maximal domain yields the *(maximal) Kirchhoff Laplacian*:

$$\begin{aligned} \mathbf{H} &= -\Delta \restriction \operatorname{dom}(\mathbf{H}), \\ \operatorname{dom}(\mathbf{H}) &= \big\{f \in \operatorname{dom}(\mathbf{H}_{\max}) \mid f \text{ satisfies (4.8) on } \mathcal{V}\big\}. \end{aligned} \tag{4.11}$$

4.3. Energy forms

Restricting further to compactly supported functions, we end up with the pre-minimal operator

$$\mathbf{H}' = -\Delta \restriction \operatorname{dom}(\mathbf{H}'), \quad \operatorname{dom}(\mathbf{H}') = \operatorname{dom}(\mathbf{H}) \cap C_c(\mathcal{G}). \tag{4.12}$$

Integrating by parts, one obtains for all $f \in \operatorname{dom}(\mathbf{H}')$

$$\langle \mathbf{H}' f, f\rangle_{L^2} = \int_{\mathcal{G}} |\nabla f(x)|^2 \nu(\mathrm{d}x) =: \mathfrak{Q}[f], \tag{4.13}$$

which implies that $\mathbf{H}'$ is a nonnegative symmetric operator in $L^2(\mathcal{G};\mu)$. We define $\mathbf{H}^0$ as the closure of $\mathbf{H}'$ in $L^2(\mathcal{G};\mu)$. It is standard to show that

$$(\mathbf{H}')^* = \mathbf{H}. \tag{4.14}$$

In particular, the equality $\mathbf{H}^0 = \mathbf{H}$ holds if and only if $\mathbf{H}$ is self-adjoint (or, equivalently, $\mathbf{H}'$ is essentially self-adjoint).

With the form $\mathfrak{Q}$ we associate two spaces: first, the Sobolev space $H^1(\mathcal{G}) = H^1(\mathcal{G};\mu,\nu)$ is defined as the subspace of $L^2(\mathcal{G};\mu)$ consisting of continuous functions, which are edgewise absolutely continuous and have finite energy $\mathfrak{Q}[f] < \infty$. Equipping $H^1(\mathcal{G})$ with the standard graph norm turns it into a Hilbert space. Also, we define the space $H^1_0(\mathcal{G}) = H^1_0(\mathcal{G};\mu,\nu)$ as the closure of compactly supported H^1 functions,

$$H^1_0 = H^1_0(\mathcal{G};\mu,\nu) := \overline{H^1_c(\mathcal{G})}^{\|\cdot\|_{H^1(\mathcal{G};\mu,\nu)}},$$

where $H^1_c(\mathcal{G}) := H^1(\mathcal{G}) \cap C_c(\mathcal{G})$. Restricting $\mathfrak{Q}$ to these spaces, we end up with two closed forms in $L^2(\mathcal{G};\mu)$:

$$\mathfrak{Q}_D = \mathfrak{Q} \restriction H^1_0, \quad \mathfrak{Q}_N = \mathfrak{Q} \restriction H^1. \tag{4.15}$$

According to the representation theorem, they give rise to two self-adjoint nonnegative operators $\mathbf{H}_D$ and $\mathbf{H}_N$ in $L^2(\mathcal{G};\mu)$, the *Dirichlet* and *Neumann Laplacians*, respectively. Notice also that $\mathbf{H}_D$ coincides with the Friedrichs extension of $\mathbf{H}'$:

$$\mathrm{dom}(\mathbf{H}_D) = \mathrm{dom}(\mathbf{H}) \cap H_0^1(\mathcal{G}).$$

Remark. Following the analogy with the Friedrichs extension, it might be tempting to think that the domain of the Neumann Laplacian $\mathbf{H}_N$ is given by $\mathrm{dom}(\mathbf{H}) \cap H^1(\mathcal{G})$. However, the operator defined on this domain has a different name—the *Gaffney Laplacian*—and it is not symmetric in general. Moreover, this operator is not always closed (see [48]).

5. Connections

One of the immediate ways to relate Laplacians on metric and discrete graphs is by noticing a connection between their harmonic functions. Despite being elementary, this observation lies at the core of many of our considerations and hence we briefly sketch it here. Every harmonic function f on a weighted metric graph $(\mathcal{G},\mu,\nu)$ (i.e., f satisfies $\Delta f = 0$ and the Kirchhoff conditions (4.8)) must be edgewise affine. The Kirchhoff conditions (4.8) imply that f is continuous and, moreover, satisfies

$$\sum_{\vec{e}\in\vec{\mathcal{E}}_v} \nu(\vec{e})\partial_{\vec{e}} f(v) = \sum_{u\sim v} \sum_{\vec{e}\in\vec{\mathcal{E}}_u : e\in\mathcal{E}_v} \frac{\nu(e)}{|e|}\big(f(u) - f(v)\big) = 0$$

at each vertex $v \in \mathcal{V}$. This suggests to consider a discrete Laplacian (3.3) with edge weights given by

$$b(u,v) = \begin{cases} \sum_{\vec{e}\in\vec{\mathcal{E}}_u : e\in\mathcal{E}_v} \frac{\nu(e)}{|e|}, & u \neq v, \\ 0, & u = v. \end{cases} \tag{5.1}$$

Indeed, then for every Δ-harmonic function f on the weighted metric graph $(\mathcal{G},\mu,\nu)$, its restriction to vertices $\mathbf{f} := f|_{\mathcal{V}}$ is an L-harmonic function; that is, $L\mathbf{f} = 0$. Moreover, the converse is also true. Phrased in a more formal way, the map

$$\begin{aligned} \imath_{\mathcal{V}} \colon C(\mathcal{G}) &\to C(\mathcal{V}) \\ f &\mapsto f|_{\mathcal{V}}, \end{aligned} \tag{5.2}$$

when restricted further to the space of continuous, edgewise affine functions on $\mathcal{G}$ becomes bijective and establishes a bijective correspondence between Δ-harmonic and L-harmonic functions. This indicates a possible connection between the corresponding Laplacians on $\mathcal{G}$ and $\mathcal{G}_d$ (this immediately connects, for instance, the corresponding Poisson and Martin boundaries). However, one also has to take into account the measures μ and m; that is, the vertex weight m should be chosen in a way which connects the corresponding Hilbert spaces $L^2(\mathcal{G};\mu)$ and $\ell^2(\mathcal{V};m)$. The desired connection is given by the choice

$$m\colon v \mapsto \sum_{\vec{e}\in\vec{\mathcal{E}}_v} |e|\mu(e), \quad v \in \mathcal{V}, \tag{5.3}$$

under the additional assumption that $(\mathcal{G},\mu,\nu)$ has *finite intrinsic size*:

$$\eta^*(\mathcal{E}) := \sup_{e\in\mathcal{E}} |e|\sqrt{\frac{\mu(e)}{\nu(e)}} < \infty. \tag{5.4}$$

The quantity $\eta(e) := |e|\sqrt{\frac{\mu(e)}{\nu(e)}}$ is the *intrinsic length* of the edge $e \in \mathcal{E}$ (see Section 7.1 for further details).

Remark. In at least two special cases, the correspondence between the Kirchhoff Laplacian for $(\mathcal{G},\mu,\nu)$ and the discrete Laplacian for the above weights b and m has been known for a quite long time. First of all, in the case of so-called unweighted *equilateral metric graphs* (i.e., $\mu = \nu = 1$ on $\mathcal{G}$ and $|e| = 1$ for all edges e), (3.3) with the weights (5.1), (5.3) turns into the normalized Laplacian (3.2). Connections between their spectral properties have been established in [53, 69] for finite metric graphs and then extended in [5, 7, 18] to infinite metric graphs, and in fact one can even prove some kind of local unitary equivalence [55]. Thus these results allow to reduce the study of Laplacians on equilateral metric graphs to a widely studied object—the normalized Laplacian L_{norm}, the generator of the simple random walk on $\mathcal{G}_d$ (see [2, 10, 70]). The second well-studied case is a slight generalization of the above setting: again, $|e| = 1$ for all edges e; however, $\mu = \nu$ on $\mathcal{G}$ (these are named *cable systems* in the work of Varopoulos [67]). The corresponding Laplacian L with the coefficients (5.1), (5.3) is the generator of a discrete time random walk on $\mathcal{G}_d$ with the probability of jumping from v to u given by

$$p(u,v) = \frac{\mu(e_{u,v})}{\sum_{w\sim v}\mu(e_{v,w})} \quad \text{when } u \sim v,$$

and 0 otherwise. There is a close connection between this random walk and the Brownian motion on the cable system, and exactly this link has been exploited several times in the literature (see [67] and some recent works [3, 22, 23]).

If the underlying model of $(\mathcal{G},\mu,\nu)$ has finite intrinsic size (5.4), it turns out that the maximal Kirchhoff Laplacian $\mathbf{H}$ in $L^2(\mathcal{G};\mu)$ and $\mathbf{h}(\mathcal{G},\mu,\nu)$, the corresponding maximal Laplacian with the weights (5.1), (5.3), share many basic properties.

Spectral properties.

- *Self-adjoint uniqueness*; see [20, Sec. 4] and [49, Chap. 3].
- *Positive spectral gap*; see [20, Sec. 4], [47], and [49, Chap. 3].
- *Ultracontractivity estimates*; see [20, Sec. 5.2], [49, Chap. 4.8], and [60].

Parabolic properties.

- *Markovian uniqueness*; see [49, Chap. 4.4].
- *Recurrence/transience*; see [31, Chap. 4] and [49, Chap. 4.5].
- *Stochastic completeness*; see [23,33,35,36], and [49, Chap. 4.6].

The above lists are by no means complete and we refer to the recent monograph [49] for further details, results, and literature.

Remark. In fact, the idea to relate the properties of Δ and L by taking into account the relationship between their kernels has its roots in the fundamental works of M. G. Krein, M. I. Vishik, and M. Sh. Birman in the 1950s. Indeed, it turns out that L serves as a "boundary operator" for Δ and exactly this fact allows to connect basic spectral properties of these two operators. However, in order to make all that precise one needs to use the machinery of boundary triplets and corresponding Weyl functions, a modern language of extension theory of symmetric operators in Hilbert spaces, which can be seen as far-reaching development of the Birman–Krein–Vishik theory (see [14,15,62]). First applications of this approach to finite and infinite metric graphs can be traced back to the 2000s (see, e.g., [5,19,56]). One of its advantages is the fact that the boundary triplets approach allows to treat metric graphs avoiding restrictive assumptions on the edge lengths [20,44].

6. Cable systems for graph Laplacians

The above considerations naturally lead to the following question: *which graph Laplacians may arise as "boundary operators" for a Kirchhoff Laplacian on a weighted metric graph?* Let us be more precise. Suppose a vertex set $\mathcal{V}$ is given. Each graph Laplacian (3.3) is determined by the vertex weight $m\colon \mathcal{V} \to (0,\infty)$ and the edge weight function $b\colon \mathcal{V}\times\mathcal{V} \to [0,\infty)$ having the properties (i)–(iv) of Section 3. With each such b we can associate a locally finite simple graph $\mathcal{G}_b = (\mathcal{V}, \mathcal{E}_b)$ as described in Section 3.

Definition 6.1. A *cable system* for a graph b over $(\mathcal{V}, m)$ is a model of a weighted metric graph $(\mathcal{G}, \mu, \nu)$ having $\mathcal{V}$ as its vertex set and such that the functions defined by (5.3) and (5.1) coincide with m and, respectively, b. If in addition the underlying graph $(\mathcal{V}, \mathcal{E})$ of the model coincides with $\mathcal{G}_b = (\mathcal{V}, \mathcal{E}_b)$, then the cable system is called *minimal*.

Thus the problem stated at the very beginning now can be formulated as follows: *Which locally finite graphs* $(\mathcal{V}, m; b)$ *have a (minimal) cable system?* It turns out that the existence of a minimal cable system is a nontrivial issue already in the case of a path graph (see [49, Chap. 5.3]). Let us also present the following example.

Example 6.2 (Cable systems for L_{comb}). Consider the combinatorial Laplacian L_{comb} on a simple, connected, locally finite graph $\mathcal{G}_d$, that is, $m \equiv 1$ on $\mathcal{V}$, $b(u, v) = 1$ exactly when $u \sim v$ and $u \neq v$, and $b(u, v) = 0$ otherwise. It turns out that in this case $(\mathcal{V}, m; b)$ *admits a minimal cable system if and only if for each* $e \in \mathcal{E}$ *there is a disjoint cycle cover of* $\mathcal{G}_d$ *containing* e *in one of its cycles.*[4]

However, we stress that a general cable system may have loops and multiple edges and thus the simplicity assumption on the model of $(\mathcal{G}, \mu, \nu)$ (that is, the minimality of a cable system for $(\mathcal{V}, m; b)$) might be too restrictive. Moreover, the underlying combinatorial graph $(\mathcal{V}, \mathcal{E})$ of a cable system for b can always be obtained from the simple graph $\mathcal{G}_b = (\mathcal{V}, \mathcal{E}_b)$ by adding loops and multiple edges. The next result was essentially proved in [23] (see also [49, Chap. 6.3]).

Theorem 6.3. *Every locally finite graph* $(\mathcal{V}, m; b)$ *has a cable system.*

After establishing existence of cable systems, the next natural question is their uniqueness. In fact, every locally finite graph b over $(\mathcal{V}, m)$ has a large number of cable systems. In particular, the construction in [23, p. 2107] is a special case of a general construction using different metrizations of discrete graphs. These connections will be discussed in the next section.

7. Intrinsic metrics on graphs

7.1. Intrinsic metrics on metric graphs

We define the intrinsic metric ϱ of a weighted metric graph $(\mathcal{G}, \mu, \nu)$ as the intrinsic metric of its Dirichlet Laplacian $\mathbf{H}_D$ (in particular, note that $\mathfrak{Q}_D$ is a strongly local, regular Dirichlet form). By [65, eq. (1.3)] (see also [24, Thm. 6.1]), ϱ_{intr} is given by

$$\varrho_{\mathrm{intr}}(x, y) = \sup\left\{f(x) - f(y) \mid f \in \widehat{\mathfrak{D}}_{\mathrm{loc}}\right\}, \quad x, y \in \mathcal{G},$$

where the function space $\widehat{\mathfrak{D}}_{\mathrm{loc}}$ is defined as

$$\widehat{\mathfrak{D}}_{\mathrm{loc}} = \left\{f \in H^1_{\mathrm{loc}}(\mathcal{G}) \mid \nu(x)\left|\nabla f(x)\right|^2 \leq \mu(x) \text{ for a.e. } x \in \mathcal{G}\right\}.$$

[4] https://mathoverflow.net/questions/59117/ (2011): *Assigning positive edge weights to a graph so that the weight incident to each vertex is* 1.

In fact, ϱ_{intr} admits an explicit description: define the *intrinsic weight*

$$\eta = \eta_{\mu,\nu} := \sqrt{\frac{\mu}{\nu}} \quad \text{on } \mathcal{G}. \tag{7.1}$$

This weight gives rise to a new measure on $\mathcal{G}$ whose density w.r.t. the Lebesgue measure is exactly η (we abuse the notation and denote with η both the edge weight and the corresponding measure).

Recall that a path $\mathcal{P}$ in $\mathcal{G}$ is a continuous and piecewise injective map $\mathcal{P}\colon I \to \mathcal{G}$ defined on an interval $I \subseteq \mathbb{R}$. If $\mathcal{I} = [a,b]$ is compact, we call $\mathcal{P}$ a path with starting point $x := \mathcal{P}(a)$ and endpoint $y := \mathcal{P}(b)$, and its *(intrinsic) length* is defined as

$$|\mathcal{P}|_\eta := \sum_j \int_{\mathcal{P}((t_j,t_{j+1}))} \eta(\mathrm{d}x), \tag{7.2}$$

where $a = t_0 < \cdots < t_n = b$ is any partition of $\mathcal{I} = [a,b]$ such that $\mathcal{P}$ is injective on each interval (t_j, t_{j+1}) (clearly, $|\mathcal{P}|_\eta$ is well defined).

Lemma 7.1. *The metric ϱ_η defined by*

$$\varrho_\eta(x,y) := \inf_{\mathcal{P}} |\mathcal{P}|_\eta, \quad x,y \in \mathcal{G}, \tag{7.3}$$

where the infimum is taken over all paths $\mathcal{P}$ from x to y, coincides with the intrinsic metric on $(\mathcal{G},\mu,\nu)$ (w.r.t. $\mathfrak{Q}_D$); that is, $\varrho_{\mathrm{intr}} = \varrho_\eta$.

The proof is straightforward and can be found in, e.g., [31, Prop. 2.21] (see also [45, Lem. 4.3]). Notice that in the case $\mu = \nu$, η coincides with the Lebesgue measure and hence ϱ_η is nothing but the length metric ϱ_0 on $\mathcal{G}$ (see Section 2.2).

Remark. If a path $\mathcal{P}_e$ consists of a single edge $e \in \mathcal{E}$, then

$$|\mathcal{P}_e|_\eta = \int_e \eta(\mathrm{d}x) = |e|\sqrt{\frac{\mu(e)}{\nu(e)}} = \eta(e),$$

which connects $\varrho_{\mathrm{intr}} = \varrho_\eta$ on $(\mathcal{G},\mu,\nu)$ with the intrinsic edge length (see (5.4)).

7.2. Intrinsic metrics on discrete graphs

The idea to use different metrics on graphs can be traced back at least to [12] and versions of metrics adapted to weighted discrete graphs have appeared independently in several works; see, e.g., [22,23,29,52]. In our exposition we follow [24,39].

For a connected graph b over $(\mathcal{V},m)$, a symmetric function $p\colon \mathcal{V}\times\mathcal{V} \to [0,\infty)$ such that $p(u,v) > 0$ exactly when $b(u,v) > 0$ is called a *weight function* for $(\mathcal{V},m;b)$.

Every weight function p generates a *path metric* ϱ_p on $\mathcal{V}$ w.r.t. b via

$$\varrho_p(u,v) := \inf_{\mathcal{P}=(v_0,\dots,v_n):u=v_0,\, v=v_n} \sum_k p(v_{k-1},v_k). \tag{7.4}$$

Here the infimum is taken over all paths in b connecting u and v, that is, all sequences $\mathcal{P} = (v_0,\dots,v_n)$ such that $v_0 = u$, $v_n = v$ and $b(v_{k-1},v_k) > 0$ for all k. Since we assume b to be locally finite, $\varrho_p(u,v) > 0$ whenever $u \neq v$.

Example 7.2 (Combinatorial distance). Let $p\colon \mathcal{V}\times\mathcal{V} \to \{0,1\}$ be given by

$$p(u,v) = \begin{cases} 1, & b(u,v) \neq 0, \\ 0, & b(u,v) = 0. \end{cases} \tag{7.5}$$

Then the corresponding metric ϱ_p is nothing but the combinatorial distance ϱ_{comb} (a.k.a. the *word metric* in the context of Cayley graphs) on a graph b over $\mathcal{V}$.

Definition 7.3. A metric ϱ on $\mathcal{V}$ is called *intrinsic* w.r.t. $(\mathcal{V},m;b)$ if

$$\sum_{u\in\mathcal{V}} b(u,v)\varrho(u,v)^2 \le m(v) \tag{7.6}$$

holds for all $v \in \mathcal{V}$. Similarly, a weight function $p\colon \mathcal{V}\times\mathcal{V} \to [0,\infty)$ is called an *intrinsic weight* for $(\mathcal{V},m;b)$ if

$$\sum_{u\in\mathcal{V}} b(u,v)p(u,v)^2 \le m(v), \quad v \in \mathcal{V}.$$

If p is an intrinsic weight, then the path metric ϱ_p is called *strongly intrinsic*.

For any given locally finite graph $(\mathcal{V},m;b)$ an intrinsic metric always exists (see [34, Ex. 2.1], [39], and also [11]).

Remark. It is straightforward to check that the combinatorial distance ϱ_{comb} is not intrinsic for the combinatorial Laplacian L_{comb} ($m \equiv 1$ on $\mathcal{V}$ in this case). On the other hand, ϱ_{comb} is equivalent to an intrinsic path metric if and only if deg is bounded on $\mathcal{V}$; that is, the corresponding graph has bounded geometry. If $\sup_{\mathcal{V}} \deg(v) = \infty$, then L_{comb} is unbounded in $\ell^2(\mathcal{V})$, and it turned out that ϱ_{comb} is not a suitable metric on $\mathcal{V}$ to study the properties of L_{comb} (in particular, this has led to certain controversies in the past; see [42,71]).

7.3. Connections between discrete and continuous

Consider a weighted metric graph $(\mathcal{G},\mu,\nu)$ and its intrinsic metric ϱ_η. With each model of $(\mathcal{G},\mu,\nu)$ we can associate the vertex set $\mathcal{V}$ together with the vertex weight $m\colon \mathcal{V} \to (0,\infty)$ and the graph $b\colon \mathcal{V}\times\mathcal{V} \to [0,\infty)$; see (5.3), (5.1). The next result shows that the intrinsic metric ϱ_η of $(\mathcal{G}, \mu, \nu)$ gives rise to a particular intrinsic metric for $(\mathcal{V}, m; b)$.

Lemma 7.4 ([49]). *Fix a model of $(\mathcal{G}, \mu, \nu)$ having finite intrinsic size and define the metric $\varrho_{\mathcal{V}}$ on $\mathcal{V}$ as a restriction of ϱ_η onto $\mathcal{V} \times \mathcal{V}$,*

$$\varrho_{\mathcal{V}}(u, v) := \varrho_\eta(u, v), \quad (u, v) \in \mathcal{V} \times \mathcal{V}. \tag{7.7}$$

Then $\varrho_{\mathcal{V}}$ is an intrinsic metric for $(\mathcal{V}, m; b)$. Moreover, $(\mathcal{G}, \varrho_\eta)$ is complete as a metric space exactly when $(\mathcal{V}, \varrho_{\mathcal{V}})$ is complete.

Let us mention that Lemma 7.4 also has an interpretation in terms of *quasi-isometries* (see, e.g., [2, Def. 1.12] and [54, Sec. 1.3]).

Definition 7.5. A map $\phi\colon X_1 \to X_2$ between metric spaces (X_1, ϱ_1) and (X_2, ϱ_2) is called *a quasi-isometry* if there are constants a, $R > 0$, and $b \geq 0$ such that

$$a^{-1}\big(\varrho_1(x, y) - b\big) \leq \varrho_2\big(\phi(x), \phi(y)\big) \leq a\big(\varrho_1(x, y) + b\big), \tag{7.8}$$

for all $x, y \in X_1$ and, moreover,

$$\bigcup_{x \in X_1} B_R\big(\phi(x); \varrho_2\big) = X_2. \tag{7.9}$$

Here and below $B_R(x; \varrho) = \{y \in X \mid \varrho(x, y) < R\}$ is a ball in a metric space (X, ϱ).

It turns out that the map $\imath_{\mathcal{V}}$ defined in Section 5 is closely related with a quasi-isometry between weighted graphs and metric graphs.

Lemma 7.6. *Assume the conditions of Lemma 7.4. Then the map*

$$\varphi\colon \mathcal{V} \to \mathcal{G}, \quad \varphi(v) = v \tag{7.10}$$

defines a quasi-isometry between the metric spaces $(\mathcal{G}, \varrho_\eta)$ and $(\mathcal{V}, \varrho_{\mathcal{V}})$.

Proof. The proof is a straightforward check of (7.8) and (7.9) for the map ϕ with $a = 1$, $b = 0$, and $R = \eta^*(\mathcal{E})$ (notice that the finite intrinsic size (5.4) is necessary for the net property (7.9) to hold). ■

Remark. The notion of quasi-isometries was introduced in the works of M. Gromov and M. Kanai in the 1980s. It is well known in context with Riemannian manifolds and (combinatorial) graphs that roughly isometric spaces share many important properties: e.g., geometric properties (such as volume growth and isoperimetric inequalities), parabolicity/transience, Liouville-type theorems for harmonic functions of finite energy, and many more. However, we stress that most of these connections also require additional (rather restrictive) conditions on the local geometry of the spaces.

Some of our conclusions are reminiscent of this notion, but in fact our results go beyond this framework. For instance, the strong/weak Liouville property (i.e., all positive/bounded harmonic functions are constant) is not preserved under bi-Lipschitz maps in general [51]. However, the equivalence holds true for our setting (see [49, Lem. 6.46]). In addition, we stress that we do not require any additional local conditions (e.g., bounded geometry). On the other hand, our results connect only two particular roughly isometric spaces $(\mathcal{G}, \varrho_\eta)$ and $(\mathcal{V}, \varrho_\mathcal{V})$ and not the whole equivalence class of roughly isometric weighted graphs or weighted metric graphs.

By Lemma 7.4, each cable system having finite intrinsic size[5] gives rise to an intrinsic metric $\varrho_\mathcal{V}$ for $(\mathcal{V}, m; b)$ using a simple restriction to vertices. It is natural to ask which intrinsic metrics on graphs can be obtained as restrictions of intrinsic metrics on weighted metric graphs. Due to the lack of space we omit the description of these results, which roughly speaking state that *to construct an intrinsic metric on a graph b over $(\mathcal{V}, m)$ is almost equivalent to constructing a cable system*. Let us state only one result here (see [49, Lem. 6.27 and Thm. 6.30]).

Theorem 7.7 ([49]). *Let b be a locally finite, connected graph over $(\mathcal{V}, m)$ equipped with a strongly intrinsic path metric ϱ. Assume also that ϱ has finite jump size,*

$$s(\varrho) = \sup\big\{\varrho(u,v) \mid u, v \in \mathcal{V},\ b(u,v) > 0\big\} < \infty.$$

Then there exists a weighted metric graph $(\mathcal{G}, \mu, \nu)$ together with a model such that (5.4) is satisfied, m and b have the form (5.3) and (5.1), respectively, and, moreover, ϱ coincides with the induced path metric $\varrho_\mathcal{V} = \varrho_\eta|_{\mathcal{V}\times\mathcal{V}}$.

Remark. It is hard to overestimate the role of intrinsic metrics in the progress achieved for weighted graph Laplacians during the last decade. Surprisingly, the above-described procedure to construct an intrinsic metric for $(\mathcal{V}, m; b)$ in fact provides a way to obtain all finite jump size intrinsic path metrics on $(\mathcal{V}, m; b)$. Moreover, upon normalization assumptions on cable systems (e.g., restricting to weighted metric graphs with equal weights, i.e., $\mu = \nu$, and also assuming no multiple edges and that all loops have the same length 1), the correspondence in Theorem 7.7 becomes in a certain sense bijective (see [49, Thm. 6.34]).

Let us mention that some versions of this result have been used earlier in [23, 33].

8. Applications

Our main goal in this final section is to demonstrate the established connections between discrete graph Laplacians and metric graph Laplacians. We will describe

[5] Since by definition a cable system is a model of a weighted metric graph, the notion of intrinsic size immediately extends to cable systems.

some applications to the self-adjointness problem and to the problem of recurrence. For further results as well as applications (Markovian uniqueness, spectral gap estimates, stochastic completeness, etc.) we refer to [49, Chap. 7–8].

8.1. Self-adjointness

The first mathematical problem arising in any quantum mechanical model is *self-adjointness* (see, e.g., [58, Chap. VIII.11]); that is, usually a formal symmetric expression for the Hamiltonian has some natural domain of definition in a given Hilbert space (e.g., pre-minimally defined Laplacians) and then one has to verify that it gives rise to an (essentially) self-adjoint operator. Otherwise,[6] there are infinitely many self-adjoint extensions (or restrictions in the maximally defined case) and one has to determine the right one which is the observable.

There are several ways to introduce the notion of self-adjointness. For the Kirchhoff Laplacian as well as for the graph Laplacian (take into account the locally finite assumption), the self-adjointness means that the minimal Laplacian coincides with the maximal Laplacian in the corresponding L^2 space. On the other hand, considering the associated Schrödinger or wave equation, the self-adjointness actually means its L^2-solvability (see, e.g., [63, Sec. 1.1]). Perhaps, the most convenient way for us would be to define the self-adjointness via solutions to the Helmholtz equation

$$\Delta u = \lambda u, \quad \lambda \in \mathbb{R}. \tag{8.1}$$

Since Δ is nonpositive, the maximally defined operator is self-adjoint if and only if for some (and hence for all) $\lambda > 0$ equation (8.1) admits a unique solution $u \in L^2(\mathcal{G}; \mu)$, which is clearly identically zero in this case (see, e.g., [57, Thm. X.26]). Recalling that, in the context of both manifolds and graphs, functions satisfying (8.1) are called λ-harmonic, the self-adjoint uniqueness can be seen as some kind of a Liouville-type property of $\mathcal{G}$[7], and this indicates its close connections with the geometry of the underlying metric space. We begin with the following result, which is widely known in the context of Riemannian manifolds.[8]

Theorem 8.1. *Let $(\mathcal{G}, \mu, \nu)$ be a weighted metric graph and let ϱ_η be the corresponding intrinsic metric. If $(\mathcal{G}, \varrho_\eta)$ is complete as a metric space, then the Kirchhoff Laplacian* $\mathbf{H}$ *is self-adjoint.*

[6] Of course, one needs to check whether the corresponding symmetric operator has equal deficiency indices, which is always the case for Laplacians or, more generally, for symmetric operators which are bounded from below or from above.

[7] Under the positivity of the spectral gap, one can in fact replace $\lambda > 0$ by $\lambda = 0$ and hence in this case one is led to harmonic functions.

[8] M. P. Gaffney [27] noticed the importance of completeness of the manifold in question and the essential self-adjointness in this case was established later [59] (see also [8, 64]).

In the context of metric graphs, the above result seems to be a folklore; however, it is not an easy task to find its proof in the literature. In fact, the above considerations enable us to provide a rather short one.

Proof. Assume that $\mathbf{H}$ is not self-adjoint. Since the minimal Kirchhoff Laplacian $\mathbf{H}^0 = \mathbf{H}^*$ is nonnegative, this means that $\ker(\mathbf{H} + \mathrm{I}) \neq \{0\}$; that is, there is $0 \neq f \in \operatorname{dom}(\mathbf{H})$ such that $\Delta f = f$ (see [57, Thm. X.26]). Moreover, we can choose such an f real-valued and hence $|f|$ is subharmonic. Moreover, $|f| \in L^2(\mathcal{G}; \mu)$ since $f \in \operatorname{dom}(\mathbf{H})$. On the other hand, if $(\mathcal{G}, \varrho_\eta)$ is complete as a metric space, then applying Yau's L^p-Liouville theorem [65, Cor. 1(a)], we conclude that $f \equiv 0$. This contradiction completes the proof. ■

Remark. A few remarks are in order.

(i) Simple examples (e.g., $\mathcal{G}$ is a path graph) show that the completeness w.r.t. ϱ_η is not necessary.

(ii) By the Hopf–Rinow theorem (a metric graph $\mathcal{G}$ equipped with ϱ_η is a length space), completeness of $(\mathcal{G}, \varrho_\eta)$ is equivalent to bounded compactness (compactness of distance balls), as well as to geodesic completeness.

As an immediate corollary of Theorem 8.1 and the above discussed connections, we obtain a version of the Gaffney theorem for graph Laplacians.

Theorem 8.2 ([34]). *Let b be a locally finite graph over $(\mathcal{V}, m)$ and let ϱ be an intrinsic metric which generates the discrete topology on $\mathcal{V}$. If $(\mathcal{V}, \varrho)$ is complete as a metric space, then $\mathbf{h}^0$ is self-adjoint and $\mathbf{h}^0 = \mathbf{h}$.*

Proof. Let us only sketch the proof (missing details can be found in [49, Chap. 7.1]). By Theorem 7.7, there is a cable system for $(\mathcal{V}, m; b)$. Moreover, the corresponding Kirchhoff Laplacian $\mathbf{H}$ is self-adjoint if and only if so is $\mathbf{h}$ (see [20, Sec. 4], [49, Thm. 3.1 (i)]). Taking into account Lemma 7.4 and applying Theorem 8.1, we complete the proof. ■

Remark. A few remarks are in order.

(i) Theorem 8.2 was first established in [34, Thm. 2].

(ii) Both Theorem 8.1 and Theorem 8.2 are not optimal. For instance, Theorem 8.2 does not imply the self-adjointness of the combinatorial Laplacian L_{comb} when it is unbounded (see [37], [41, Thm. 6]). However, Theorems 8.1 and 8.2 enjoy a certain stability property under additive perturbations, which preserve semiboundedness ([30, Thm. 2.16], [45]).

(iii) We refer for further results and details to [45], [49, Chap. 7.1], and [61].

8.2. Recurrence and transience

Recurrence of a random walk/Brownian motion means that a particle returns to its initial position infinitely often (see, e.g., [26] for a formal definition). In fact, recurrence appears (quite often under different names) in many different areas of mathematics and mathematical physics and enjoys deep connections to various important problems (e.g., the type problem for simply connected Riemann surfaces).

The famous theorem of G. Pólya states that the simple random walk on $\mathbb{Z}^d$ is recurrent if and only if either $d = 1$ or $d = 2$. Intuitively, one may explain recurrence of a random walk/Brownian motion as insufficiency of volume in the state space (volume of a ball of radius R in $\mathbb{Z}^d$ or $\mathbb{R}^d$ grows faster as $R \to \infty$ for larger d). The qualitative form of this heuristic statement in the manifold context has a venerable history (we refer to the excellent exposition of A. Grigor'yan [28] for further details), and in the case of complete Riemannian manifolds, it was proved in the 1980s independently by L. Karp, N. Th. Varopoulos, and A. Grigor'yan (see [28, Thm. 7.3]) that

$$\int^{\infty} \frac{r\,dr}{\operatorname{vol}\big(B_r(x)\big)} = \infty$$

guarantees recurrence. Moreover, this condition is close to be necessary. This result was extended to strongly local Dirichlet forms by K.-T. Sturm [65] and hence it also holds in the setting of weighted metric graphs. Again, using the obtained connections between metric graph and weighted graph Laplacians, we can proceed as in the previous subsection and establish the corresponding volume growth test for weighted graph Laplacians, which was originally obtained by B. Hua and M. Keller [32]. Due to the lack of space we only refer to [49, Chap. 7.4] for further details.

We would like to finish this article by reflecting on another interesting topic. Perhaps, the most studied class of graphs are Cayley graphs of finitely generated groups (Example 2.1). Random walks on groups is a classical and very rich subject (the literature is enormous and we only refer to the classic text [70]). Recall that a group G is called *recurrent* if the simple random walk on its Cayley graph $\mathscr{C}(\mathsf{G}, S)$ is recurrent for some (and hence for all) S. The classification of recurrent groups was accomplished in the 1980s by proving the famous *Kesten conjecture*. It is a combination of two seminal theorems—relationship between decay of return probabilities and growth in groups established by N. Th. Varopoulos and the characterization of groups of polynomial growth by M. Gromov (see, e.g., [68, Chap. VI.6], [70, Thm. 3.24]).

Theorem 8.3 (N. Th. Varopoulos). *G is recurrent if and only if G contains a finite index subgroup isomorphic either to $\mathbb{Z}$ or to $\mathbb{Z}^2$.*

It turns out that the problem of recurrence on weighted metric graphs can be reduced to the study of recurrence of random walks on groups (see [49, Thm. 7.49]).

Let $(\mathcal{G}_C, \mu, \nu)$ be a weighted metric graph where $\mathcal{G}_C = \mathcal{C}(\mathsf{G}, S)$ is a Cayley graph of a finitely generated group G. Also, let $\mathbf{H}_D$ be the corresponding Dirichlet Laplacian. Define

$$b_\nu(u,v) = \begin{cases} \frac{\nu(e_{u,v})}{|e_{u,v}|}, & u^{-1}v \in S, \\ 0, & u^{-1}v \notin S, \end{cases} \qquad u, v \in \mathsf{G}. \tag{8.2}$$

Theorem 8.4. *The heat semigroup* $(\mathrm{e}^{-t\mathbf{H}_D})_{t>0}$ *is recurrent if and only if the discrete time random walk on* G *with transition probabilities*

$$p_\nu(u,v) = P(X_{n+1} = v \mid X_n = u) = \frac{b_\nu(u,v)}{\sum_{g\in S} b_\nu(u,ug)}, \qquad u,v \in \mathsf{G}, \tag{8.3}$$

is recurrent.

Combining Theorem 8.4 with Theorem 8.3, we arrive at the following result.

Corollary 8.5. *Assume the conditions of Theorem 8.4.*

(i) *If* G *contains a finite index subgroup isomorphic either to* $\mathbb{Z}$ *or to* $\mathbb{Z}^2$ *and the edge weight* ν *satisfies*

$$\sup_{e\in\mathcal{E}} \frac{\nu(e)}{|e|} < \infty, \tag{8.4}$$

then the heat semigroup $(\mathrm{e}^{-t\mathbf{H}_D})_{t>0}$ *is recurrent.*

(ii) *If* G *is transient (i.e.,* G *does not contain a finite index subgroup isomorphic either to* $\mathbb{Z}$ *or to* $\mathbb{Z}^2$*) and the edge weight* ν *satisfies*

$$\inf_{e\in\mathcal{E}} \frac{\nu(e)}{|e|} > 0, \tag{8.5}$$

then the heat semigroup $(\mathrm{e}^{-t\mathbf{H}_D})_{t>0}$ *is transient.*

In fact, the above result has numerous consequences and actually can be improved. Let us finish by its applications to ultracontractivity estimates. To simplify the exposition we restrict now to unweighted metric graphs.

Theorem 8.6 ([20,49]). *Assume the conditions of Theorem 8.4 and let also* $\mu = \nu \equiv 1$. *Suppose that* G *is not recurrent and the edge lengths satisfy*

$$\sup_{e\in\mathcal{E}} |e| < \infty. \tag{8.6}$$

Then $(\mathrm{e}^{-t\mathbf{H}_D})_{t>0}$ *is ultracontractive and, moreover,*

(i) *if $\gamma_{\mathsf{G}}(n) \approx n^N$ as $n \to \infty$ with some $N \in \mathbb{Z}_{\geq 3}$, then*[9]

$$\|\mathrm{e}^{-t\mathbf{H}_D}\|_{1\to\infty} \leq C_N t^{-N/2}, \quad t > 0; \tag{8.7}$$

(ii) *if G is not virtually nilpotent (i.e., γ_{G} has superpolynomial growth*[10]*), then* (8.7) *holds true for all $N > 2$.*

Remark. Notice that applying Theorems 1.2 and 1.3 of [50] to the Dirichlet Laplacian $\mathbf{H}_D$ and then using Theorem 8.6, we arrive at the Cwikel–Lieb–Rozenblum estimates for additive perturbations, that is, for Schrödinger operators $-\Delta + V(x)$. It is also well known (see [25]) that ultracontractivity estimates and Sobolev-type inequalities lead to Lieb–Thirring bounds ($\mathfrak{S}_p$ estimates on the negative spectra); however, we are not going to pursue this goal here. For further details and historical remarks we refer to [49, Chap. 8.2].

Acknowledgments. Some of the results reviewed here were obtained in collaboration with Pavel Exner, Mark Malamud, and Delio Mugnolo. It is our great pleasure to thank them. We are also grateful to Omid Amini, Matthias Keller, and Wolfgang Woess for numerous useful and stimulating discussions.

Funding. Research of the first author supported by the Austrian Science Fund (FWF) under grants nos. P 28807 and I 4600, and by the Slovenian Research Agency (ARRS) under grant no. N1-0137. Research of the second author supported by the Austrian Science Fund (FWF) under grant no. J 4497.

References

[1] M. Baker and R. Rumely, *Potential Theory and Dynamics on the Berkovich Projective Line*. Math. Surveys Monogr. 159, American Mathematical Society, Providence, RI, 2010 Zbl 1196.14002 MR 2599526

[2] M. T. Barlow, *Random Walks and Heat Kernels on Graphs*. London Math. Soc. Lecture Note Ser. 438, Cambridge University Press, Cambridge, 2017 Zbl 1365.05002 MR 3616731

[3] M. T. Barlow and R. F. Bass, Stability of parabolic Harnack inequalities. *Trans. Amer. Math. Soc.* **356** (2004), no. 4, 1501–1533 Zbl 1034.60070 MR 2034316

[9] Here $\gamma_{\mathsf{G}}\colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{>0}$ is the growth function defined by $\gamma_{\mathsf{G}}(n) = \#\{g \in \mathsf{G} \mid \varrho_{\mathrm{comb}}(g, o) \leq n\}$, where o is the identity element of G.

[10] This means that for each $N > 0$ there is $c > 0$ such that $\gamma_{\mathsf{G}}(n) \geq cn^N$ for all large n.

[4] G. Berkolaiko and P. Kuchment, *Introduction to Quantum Graphs*. Math. Surveys Monogr. 186, American Mathematical Society, Providence, RI, 2013 Zbl 1318.81005 MR 3013208

[5] J. Brüning, V. Geyler, and K. Pankrashkin, Spectra of self-adjoint extensions and applications to solvable Schrödinger operators. *Rev. Math. Phys.* **20** (2008), no. 1, 1–70 Zbl 1163.81007 MR 2379246

[6] D. Burago, Y. Burago, and S. Ivanov, *A Course in Metric Geometry*. Grad. Stud. Math. 33, American Mathematical Society, Providence, RI, 2001 Zbl 0981.51016 MR 1835418

[7] C. Cattaneo, The spectrum of the continuous Laplacian on a graph. *Monatsh. Math.* **124** (1997), no. 3, 215–235 Zbl 0892.47001 MR 1476363

[8] P. R. Chernoff, Essential self-adjointness of powers of generators of hyperbolic equations. *J. Functional Analysis* **12** (1973), 401–414 Zbl 0263.35066 MR 0369890

[9] F. R. K. Chung, *Spectral Graph Theory*. CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, Providence, RI, 1997 Zbl 0867.05046 MR 1421568

[10] Y. Colin de Verdière, *Spectres de Graphes*. Cours Spéc. 4, Société Mathématique de France, Paris, 1998 Zbl 0913.05071 MR 1652692

[11] Y. Colin de Verdière, N. Torki-Hamza, and F. Truc, Essential self-adjointness for combinatorial Schrödinger operators II—metrically non complete graphs. *Math. Phys. Anal. Geom.* **14** (2011), no. 1, 21–38 Zbl 1244.05155 MR 2782792

[12] E. B. Davies, Analysis on graphs and noncommutative geometry. *J. Funct. Anal.* **111** (1993), no. 2, 398–430 Zbl 0793.46043 MR 1203460

[13] E. B. Davies, Large deviations for heat kernels on graphs. *J. London Math. Soc. (2)* **47** (1993), no. 1, 65–72 Zbl 0799.58086 MR 1200978

[14] V. A. Derkach and M. Malamud, Generalized resolvents and the boundary value problems for Hermitian operators with gaps. *J. Funct. Anal.* **95** (1991), no. 1, 1–95 Zbl 0748.47004 MR 1087947

[15] V. A. Derkach and M. Malamud, *The Theory of Extensions of Symmetric Operators and Boundary Value Problems*. Proc. Inst. Math. NAS of Ukraine, **104**, Kiev, 2017

[16] R. Diestel, *Graph Theory*. 5th edn., Grad. Texts in Math. 173, Springer, Berlin, 2017 Zbl 1375.05002 MR 3644391

[17] J. Draisma and A. Vargas, On the gonality of metric graphs. *Notices Amer. Math. Soc.* **68** (2021), no. 5, 687–695 Zbl 1473.05069 MR 4249425

[18] P. Exner, A duality between Schrödinger operators on graphs and certain Jacobi matrices. *Ann. Inst. H. Poincaré Phys. Théor.* **66** (1997), no. 4, 359–371 Zbl 0949.34073 MR 1459512

[19] P. Exner, J. P. Keating, P. Kuchment, T. Sunada, and A. Teplyaev (eds.), *Analysis on Graphs and its Applications*. Proc. Sympos. Pure Math. 77, American Mathematical Society, Providence, RI, 2008 Zbl 1143.05002 MR 2459860

[20] P. Exner, A. Kostenko, M. Malamud, and H. Neidhardt, Spectral theory of infinite quantum graphs. *Ann. Henri Poincaré* **19** (2018), no. 11, 3457–3510 Zbl 06968477 MR 3869419

[21] P. Exner and H. Kovařík, *Quantum Waveguides*. Theoret. and Math. Phys., Springer, Cham, 2015 Zbl 1314.81001 MR 3362506

[22] M. Folz, Volume growth and spectrum for general graph Laplacians. *Math. Z.* **276** (2014), no. 1-2, 115–131 Zbl 1288.47001 MR 3150195

[23] M. Folz, Volume growth and stochastic completeness of graphs. *Trans. Amer. Math. Soc.* **366** (2014), no. 4, 2089–2119 Zbl 1325.60069 MR 3152724

[24] R. L. Frank, D. Lenz, and D. Wingert, Intrinsic metrics for non-local symmetric Dirichlet forms and applications to spectral theory. *J. Funct. Anal.* **266** (2014), no. 8, 4765–4808 Zbl 1304.60081 MR 3177322

[25] R. L. Frank, E. H. Lieb, and R. Seiringer, Equivalence of Sobolev inequalities and Lieb-Thirring inequalities. In *XVIth International Congress on Mathematical Physics*, pp. 523–535, World Sci. Publ., Hackensack, NJ, 2010 Zbl 1203.81072 MR 2730819

[26] M. Fukushima, Y. Oshima, and M. Takeda, *Dirichlet Forms and Symmetric Markov Processes*. extended edn., De Gruyter Stud. Math. 19, Walter de Gruyter, Berlin, 2011 Zbl 0838.31001 MR 2778606

[27] M. P. Gaffney, Hilbert space methods in the theory of harmonic integrals. *Trans. Amer. Math. Soc.* **78** (1955), 426–444 Zbl 0064.34303 MR 68888

[28] A. Grigor'yan, Analytic and geometric background of recurrence and non-explosion of the Brownian motion on Riemannian manifolds. *Bull. Amer. Math. Soc. (N.S.)* **36** (1999), no. 2, 135–249 Zbl 0927.58019 MR 1659871

[29] A. Grigor'yan, X. Huang, and J. Masamune, On stochastic completeness of jump processes. *Math. Z.* **271** (2012), no. 3-4, 1211–1239 Zbl 1408.60076 MR 2945605

[30] B. Güneysu, M. Keller, and M. Schmidt, A Feynman-Kac-Itô formula for magnetic Schrödinger operators on graphs. *Probab. Theory Related Fields* **165** (2016), no. 1-2, 365–399 Zbl 1341.81027 MR 3500274

[31] S. Haeseler, *Analysis of Dirichlet forms on graphs*. Ph.D. thesis, Jena, 2014

[32] B. Hua and M. Keller, Harmonic functions of general graph Laplacians. *Calc. Var. Partial Differential Equations* **51** (2014), no. 1-2, 343–362 Zbl 1298.31009 MR 3247392

[33] X. Huang, A note on the volume growth criterion for stochastic completeness of weighted graphs. *Potential Anal.* **40** (2014), no. 2, 117–142 Zbl 1282.05061 MR 3152158

[34] X. Huang, M. Keller, J. Masamune, and R. K. Wojciechowski, A note on self-adjoint extensions of the Laplacian on weighted graphs. *J. Funct. Anal.* **265** (2013), no. 8, 1556–1578 Zbl 1435.35400 MR 3079229

[35] X. Huang, M. Keller, and M. Schmidt, On the uniqueness class, stochastic completeness and volume growth for graphs. *Trans. Amer. Math. Soc.* **373** (2020), no. 12, 8861–8884 Zbl 07301843 MR 4177278

[36] X. Huang and Y. Shiozawa, Upper escape rate of Markov chains on weighted graphs. *Stochastic Process. Appl.* **124** (2014), no. 1, 317–347 Zbl 1284.60144 MR 3131296

[37] P. E. T. Jorgensen, Essential self-adjointness of the graph-Laplacian. *J. Math. Phys.* **49** (2008), no. 7, 073510, 33 Zbl 1152.81496 MR 2432048

[38] T. Kato, *Perturbation Theory for Linear Operators*. 2nd edn., Grundlehren Math. Wiss. 132, Springer, Berlin, 1976 Zbl 0342.47009 MR 0407617

[39] M. Keller, Intrinsic metrics on graphs: a survey. In *Mathematical Technology of Networks*, pp. 81–119, Springer Proc. Math. Stat. 128, Springer, Cham, 2015 Zbl 3375157 MR 3375157

[40] M. Keller and D. Lenz, Unbounded Laplacians on graphs: basic spectral properties and the heat equation. *Math. Model. Nat. Phenom.* **5** (2010), no. 4, 198–224 Zbl 1207.47032 MR 2662456

[41] M. Keller and D. Lenz, Dirichlet forms and stochastic completeness of graphs and subgraphs. *J. Reine Angew. Math.* **666** (2012), 189–223 Zbl 1252.47090 MR 2920886

[42] M. Keller, D. Lenz, and R. K. Wojciechowski, Volume growth, spectrum and stochastic completeness of infinite graphs. *Math. Z.* **274** (2013), no. 3-4, 905–932 Zbl 1269.05051 MR 3078252

[43] M. Keller, D. Lenz, and R. K. Wojciechowski, *Graphs and Discrete Dirichlet Spaces*. Grundlehren Math. Wiss. 358, Springer, Cham, 2021 Zbl 07414798 MR 4383783

[44] A. Kostenko and M. Malamud, 1-D Schrödinger operators with local point interactions on a discrete set. *J. Differential Equations* **249** (2010), no. 2, 253–304 Zbl 1195.47031 MR 2644117

[45] A. Kostenko, M. Malamud, and N. Nicolussi, A Glazman-Povzner-Wienholtz theorem on graphs. *Adv. Math.* **395** (2022), Paper No. 108158 Zbl 07456627 MR 4356814

[46] A. Kostenko, D. Mugnolo, and N. Nicolussi, Self-adjoint and Markovian extensions of infinite quantum graphs. *J. Lond. Math. Soc. (2)* **105** (2022), no. 2, 1262–1313 MR 4400947

[47] A. Kostenko and N. Nicolussi, Spectral estimates for infinite quantum graphs. *Calc. Var. Partial Differential Equations* **58** (2019), no. 1, Paper No. 15 Zbl 1404.81111 MR 3891807

[48] A. Kostenko and N. Nicolussi, A note on the Gaffney Laplacian on infinite metric graphs. *J. Funct. Anal.* **281** (2021), no. 10, Paper No. 109216 Zbl 07401192 MR 4308056

[49] A. Kostenko and N. Nicolussi, *Laplacians on Infinite Graphs*. Mem. Eur. Math. Soc., EMS Press, Berlin, to appear

[50] D. Levin and M. Solomyak, The Rozenblum-Lieb-Cwikel inequality for Markov generators. *J. Anal. Math.* **71** (1997), 173–193 Zbl 0910.47017 MR 1454250

[51] T. Lyons, Instability of the Liouville property for quasi-isometric Riemannian manifolds and reversible Markov chains. *J. Differential Geom.* **26** (1987), no. 1, 33–66 Zbl 0599.60011 MR 892030

[52] J. Masamune and T. Uemura, Conservation property of symmetric jump processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** (2011), no. 3, 650–662 Zbl 1230.60090 MR 2841069

[53] S. Nicaise, Some results on spectral theory over networks, applied to nerve impulse transmission. In *Orthogonal Polynomials and Applications (Bar-le-Duc, 1984)*, pp. 532–541, Lecture Notes in Math. 1171, Springer, Berlin, 1985 Zbl 0572.00007 MR 839024

[54] P. W. Nowak and G. Yu, *Large Scale Geometry*. EMS Textbk. Math., EMS Press, Zürich, 2012 Zbl 1264.53051 MR 2986138

[55] K. Pankrashkin, Unitary dimension reduction for a class of self-adjoint extensions with applications to graph-like structures. *J. Math. Anal. Appl.* **396** (2012), no. 2, 640–655 Zbl 1260.47010 MR 2961258

[56] O. Post, *Spectral Analysis on Graph-Like Spaces*. Lecture Notes in Math. 2039, Springer, Heidelberg, 2012 Zbl 1247.58001 MR 2934267

[57] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-Adjointness*. Academic Press, New York, 1975 Zbl 0308.47002 MR 0493420

[58] M. Reed and B. Simon, *Methods of Modern Mathematical Physics. I. Functional Analysis*. 2nd edn., Academic Press, New York, 1980 Zbl 0459.46001 MR 751959

[59] W. Roelcke, Über den Laplace-Operator auf Riemannschen Mannigfaltigkeiten mit diskontinuierlichen Gruppen. *Math. Nachr.* **21** (1960), 131–149 Zbl 0197.36401 MR 151927

[60] G. V. Rozenblyum and M. Z. Solomyak, On spectral estimates for Schrödinger-type operators: the case of small local dimension. *Funktsional. Anal. i Prilozhen.* **44** (2010), no. 4, 21–33 Zbl 1272.47055 MR 2768562

[61] M. Schmidt, On the existence and uniqueness of self-adjoint realizations of discrete (magnetic) Schrödinger operators. In *Analysis and Geometry on Graphs and Manifolds*, pp. 250–327, London Math. Soc. Lecture Note Ser. 461, Cambridge University Press, Cambridge, 2020 Zbl 07399421

[62] K. Schmüdgen, *Unbounded Self-Adjoint Operators on Hilbert Space*. Grad. Texts in Math. 265, Springer, Dordrecht, 2012 Zbl 1257.47001 MR 2953553

[63] M. A. Shubin, Spectral theory of elliptic operators on noncompact manifolds. *Astérisque* **207** (1992), 35–108 MR 1205177

[64] R. S. Strichartz, Analysis of the Laplacian on the complete Riemannian manifold. *J. Functional Analysis* **52** (1983), no. 1, 48–79 Zbl 0515.58037 MR 705991

[65] K.-T. Sturm, Analysis on local Dirichlet spaces. I. Recurrence, conservativeness and L^p-Liouville properties. *J. Reine Angew. Math.* **456** (1994), 173–196 Zbl 0806.53041 MR 1301456

[66] P. W. Sy and T. Sunada, Discrete Schrödinger operators on a graph. *Nagoya Math. J.* **125** (1992), 141–150 Zbl 0773.35046 MR 1156908

[67] N. T. Varopoulos, Long range estimates for Markov chains. *Bull. Sci. Math. (2)* **109** (1985), no. 3, 225–252 Zbl 0583.60063 MR 822826

[68] N. T. Varopoulos, L. Saloff-Coste, and T. Coulhon, *Analysis and Geometry on Groups*. Cambridge Tracts in Math. 100, Cambridge University Press, Cambridge, 1992 Zbl 0813.22003 MR 1218884

[69] J. von Below, A characteristic equation associated to an eigenvalue problem on c^2-networks. *Linear Algebra Appl.* **71** (1985), 309–325 Zbl 0617.34010 MR 813056

[70] W. Woess, *Random Walks on Infinite Graphs and Groups*. Cambridge Tracts in Math. 138, Cambridge University Press, Cambridge, 2000 Zbl 0951.60002 MR 1743100

[71] R. K. Wojciechowski, Stochastically incomplete manifolds and graphs. In *Random Walks, Boundaries and Spectra*, pp. 163–179, Progr. Probab. 64, Birkhäuser, Basel, 2011 Zbl 1221.39008 MR 3051698

Aleksey Kostenko
Faculty of Mathematics and Physics, University of Ljubljana, Jadranska ul. 19,
1000 Ljubljana, Slovenia; and Faculty of Mathematics, University of Vienna,
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; aleksey.kostenko@fmf.uni-lj.si

Noema Nicolussi
Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna,
Austria; noema.nicolussi@univie.ac.at

DOI 10.4171/8ECM/33

Uniqueness results for solutions of continuous and discrete PDE

Eugenia Malinnikova

Abstract. We give an overview of some recent results on unique continuation property “at infinity” for solutions of elliptic and dispersive PDE and their discrete counterparts. The proofs of most of the results are given in previous works written with coauthors.

1. Introduction

Let L be a differential operator. We say that L has the (weak) unique continuation property if any solution u to the equation $Lu = 0$ in some domain Ω which vanishes on an open subset of Ω equals zero on Ω. For the case of a linear operator, we conclude that two solutions which coincide on an open subset should coincide on the whole domain. The unique continuation property holds for the class of holomorphic functions, this corresponds to the first-order differential operator $\bar{\partial}$, and, more interestingly, for a large class of second-order elliptic operators. The operator L has the strong unique continuation property if any solution u to the equation $Lu = 0$ in Ω that vanishes at some point $x \in \Omega$ to an infinite order is identically zero in Ω.

In this survey, we consider versions of the uniqueness property at infinity. Let $Lu = 0$ on $\mathbb{R}^d$, assuming some decay or growth restriction condition for u, we want to conclude that u is a trivial solution. The simplest example of such result is the classical Liouville theorem for harmonic functions. If a harmonic function on $\mathbb{R}^d$ is bounded, then it is constant. This theorem has a very short and elegant proof; see [30]. It also has numerous generalizations, which include the analogous statement for harmonic functions on $\mathbb{Z}^d$; see for example [21]. The first topic of this note is a surprising improvement of the Liouville theorem for discrete harmonic functions on $\mathbb{Z}^2$ obtained in [3]. We discuss some follow up questions and very deep related results on Anderson localization for the Anderson–Bernoulli model.

2020 Mathematics Subject Classification. Primary 35J10; Secondary 35B53, 39A12, 31C20.
Keywords. Unique continuation, Schrödinger equation, exponential decay.

In the next part of the note, we consider the stationary Schrödinger operator with a bounded potential, $Lu = -\Delta u + Vu$. We suggest an elementary analysis of the decay properties of solutions to the corresponding equation on the lattice $\mathbb{Z}^d$ and then describe a recent progress on the continuous question, known as the Landis conjecture. The result is proved in [28] and answers the question on the plane; the problem is open in higher dimensions.

Finally, we describe uniqueness results for the operator $Lu = \partial_t u + i(-\Delta + V)u$, obtained by Luis Escauriaza, Carlos Kenig, Gustavo Ponce, and Luis Vega in a remarkable series of articles [12–15], and discuss the semi-discrete operator, citing the results of [1, 16, 17, 22].

2. Uniqueness results for discrete harmonic functions

2.1. Harmonic functions on $\mathbb{Z}^d$

For each point $x = (x_1, \dots, x_d) \in \mathbb{Z}^d$, the $2d$ points $y = (y_1, \dots, y_d)$ such that $\sum_j |x_j - y_j| = 1$ are called the neighbors of x; we write $x \sim y$. Let $V \subset \mathbb{Z}^d$. We define the interior of V as the set of all $x \in V$ such that all neighbors of x also lie in V. Then a function $h : V \to \mathbb{R}$ is called harmonic in V if for any point x in the interior we have

$$h(x) = \frac{1}{2d} \sum_{y \sim x} h(y).$$

This definition easily extends to graphs with finite degrees of vertices. The systematic study of harmonic functions on $\mathbb{Z}^d$ started about a century ago with the classical works of Phillips and Wiener [31], and of Courant, Friedrichs, and Lewy, [5]. It is interesting to note that the first classical articles on the discrete potential theory already mentioned its connections to the probability and random walks. The motivation for these works was the approximation of continuous harmonic functions by discrete ones. One of the results, that can be obtained using such approximation, is the solvability of the Dirichlet problem for bounded domains in $\mathbb{R}^d$ with sufficiently smooth boundary. One might argue that motivation now is reversed; we think that the real world is discrete and study the discrete mathematical models in their own right.

2.2. Weak unique continuation

We start with some simple examples that show the absence of the weak unique continuation property for harmonic functions on $\mathbb{Z}^d$.

Example 2.1. First we consider $\mathbb{Z}^2$. It is easy to see that if h is a harmonic function on $\mathbb{Z}^2$ and $h(x) = 0$ when $x = (x_1, 0)$ and $x = (x_1, 1)$ for all $x_1 \in \mathbb{Z}$, then $h = 0$

on $\mathbb{Z}^2$. On the other hand, we construct a non-trivial harmonic function h on $\mathbb{Z}^2$ such that $h(x) = 0$ when $x = (x_1, x_2)$ with $x_1 + x_2 < 0$. We define $h(x_1, -x_1) = (-1)^{x_1}$ and notice then that one can choose freely the values $h(0, n)$ for $n = 1, 2, \dots$ and all other values of h are then uniquely determined. We note also that this large region of zeros enforces a rigid structure to the values of the harmonic function nearby. On each next diagonal, the harmonic function $h(x_1, n - x_1) = (-1)^{x_1} p_n(x_1)$, where p_n is a polynomial of degree n.

The situation is even more counter-intuitive in higher dimensions.

Example 2.2. We consider the function h_0 on $\mathbb{Z}^2$ defined by

$$h_0(x) = \begin{cases} 0, & \text{when } x = (x_1, x_2),\ x_1 + x_2 \neq 0, \\ (-1)^{x_1}, & \text{when } x = (x_1, -x_1). \end{cases}$$

Then we extend h_0 to the function on $\mathbb{Z}^3 = \mathbb{Z}^2 \times \mathbb{Z}$ as

$$H(x_1, x_2, x_3) = c^{x_3} h_0(x_1, x_2),$$

where $c + c^{-1} = 6$. The resulting harmonic function H equals zero everywhere on $\mathbb{Z}^3$ except for the hyperplane $x_1 + x_2 = 0$.

These examples demonstrate that some of our continuous intuition does not work for discrete harmonic functions.

Nevertheless, there is a trace of the unique continuation property for discrete harmonic functions on $\mathbb{Z}^d$. We denote by Q_N^d the discrete cube $[-N, N]^d \cap \mathbb{Z}^d$.

Lemma 2.3 ([20]). *There exist $C = C(d) > 0$, $c = c(d) > 0$, and $\alpha = \alpha(d) \in (0, 1)$ such that for any discrete harmonic function U on Q_{4N}^d the following inequality holds:*

$$\max_{Q_{2N}^d} |U| \le C\Big(\max_{Q_N^d} |U|^\alpha \max_{Q_{4N}^d} |U|^{1-\alpha} + e^{-cN} \max_{Q_{4N}^d} |U| \Big).$$

A similar result was also proven by Lippner and Mangoubi in [26] using a different method. We remark that the error term $e^{-cN} \max_{Q_{4N}^d} |U|$ cannot be omitted, as Example 2.1 shows, and that the decay of this term as N grows to infinity is sharp. In the continuous setting, the corresponding estimate (without the error term) is known as the three-ball inequality; see for example [24]. This estimate serves as a quantitative version of the weak unique continuation property.

The inequality of Lemma 2.3 was generalized in [3], where we showed that there exist C, c, α as above such that

$$\max_{Q_{2N}^d} |U| \le C\Big(\max_{E} |U|^\alpha \max_{Q_{4N}^d} |U|^{1-\alpha} + e^{-cN} \max_{Q_{4N}^d} |U| \Big) \tag{2.1}$$

holds for any $E \subset Q_N^d$ with $|E| > |Q_N^d|/2$. The proof is based on the fact that discrete harmonic function is a restriction to the lattice of a real analytic function with controlled speed of convergence. On the other hand, it is known that the three-ball inequality and its generalizations concerning propagation of smallness from sets of positive measure hold for a large class of elliptic equations with non-analytic coefficients; see [27]. Recently, interesting three balls inequalities were obtained for solutions of the discrete magnetic Schrödinger equation on the lattice using new discrete Carleman estimates [19].

2.3. Discrete harmonic functions bounded on a large portion of $\mathbb{Z}^d$

Let U be a discrete harmonic function on $\mathbb{Z}^d$, we say that it is bounded by one on a ρ-portion of $\mathbb{Z}^d$ if

$$\left|\left\{x \in Q_N^d : \left|U(x)\right| \leq 1\right\}\right| \geq \rho |Q_N^d|$$

for all N large enough. The inequality (2.1) shows that discrete harmonic functions behave similar to continuous ones and we expect a discrete harmonic function which is bounded on a large portion of $\mathbb{Z}^d$ to grow fast at infinity. More precisely, the following result holds.

Theorem 2.4 ([3]). *There exist $\varepsilon = \varepsilon(d) > 0$ and $b = b(d) > 0$ such that for any sufficiently large N and any discrete harmonic function U on Q_{2N}^d which satisfies $\max_{Q_M^d} |U| \geq 2$ and*

$$\left|\left\{x \in Q_k : \left|U(x)\right| \leq 1\right\}\right| \geq (1-\varepsilon)|Q_K|$$

for every $K \in [M, 2N]$, where $M \leq \sqrt{N}$, we have

$$\max_{Q_N^d} |U| \geq e^{bN}.$$

Example 2.2 shows that for $d \geq 3$ there are discrete harmonic functions bounded on $(1-\varepsilon)$ portion of $\mathbb{Z}^d$, which grow exponentially at infinity. We remark that the continuous intuition would predict for very small ε even faster growth at infinity.

A new uniqueness result for harmonic functions on $\mathbb{Z}^2$ found in [3] says that a discrete harmonic function which vanishes on a $(1-\varepsilon)$ portion of $\mathbb{Z}^2$ for sufficiently small ε is zero. The key observation, exploited in [3], is that near a tilted rectangle of zeros, the restrictions of a discrete harmonic function to diagonals have polynomial structure and thus either vanish or have a few zeros. This result follows from a more general statement.

Theorem 2.5 ([3]). *There exist $\varepsilon_0 > 0$ and $a(\varepsilon) > 0$ such that if U is a discrete harmonic function on Q_{2N}^2, N is sufficiently large, and U is bounded by one on*

$(1-\varepsilon)$ *portion of* Q^2_{2N}, $\varepsilon < \varepsilon_0$, *then*

$$\max_{Q^2_N} |U| \leq e^{a(\varepsilon)N}.$$

Moreover, $a(\varepsilon) \to 0$ *as* $\varepsilon \to 0$.

Theorems 2.4 and 2.5 imply that any discrete harmonic function that is bounded on a $(1-\varepsilon)$ portion of $\mathbb{Z}^2$ with ε small enough is constant.

Theorem 2.5 also implies that *there exist constants* a *and* $\varepsilon < 1$ *such that for any discrete harmonic function on* Q^2_{2N}, *for* N *large enough, we have*

$$\left|\left\{|U| > e^{-aN} \max_{Q^2_N} |U|\right\} \cap Q^2_{2N}\right| \geq \varepsilon N^2. \tag{2.2}$$

It would be interesting to obtain sharp generalizations of this result to harmonic functions on higher dimensional lattices. For example, a toy statement in $\mathbb{Z}^3$ is the following:

> *Suppose that* U *is a discrete harmonic function on* Q^3_{2N} *such that*
>
> $$\left|\{U \neq 0\}\right| \leq cN^2,$$
>
> *where* c *is sufficiently small and* N *is sufficiently large. Then* $U = 0$ *on* Q^3_N.

The interest in the uniqueness theorems for discrete harmonic functions and more general solutions to the Schrödinger equation on lattices is partly due to its connections to the problem of the exponential decay of eigenfunctions of the Schrödinger operator with a random Bernoulli potential, known as the Anderson localization. This connection is discovered and exploited by Bourgain and Kenig in [2], where the continuous model is studied. Recently, Ding and Smart [10], combining the approach developed in [2] with ideas introduced in [3], obtained new results on localization near the edge for the Anderson–Bernoulli model on $\mathbb{Z}^2$. One of the tools developed in [10] is a probabilistic version of (2.2) for solutions of the equation $\Delta U + VU = \lambda U$ with random Bernoulli potential V. It is worth mentioning, that in dimension three the following deterministic statement holds (see [25]):

> *There exists constant* $p > 3/2$ *such that for each* $K > 0$, *there is* $C > 0$, *such that if* $\Delta U + VU = 0$ *on* Q^3_N, N *is large enough, and* $|V| \leq K$, *then*
>
> $$\left|\left\{|U| > e^{-CN}\left|U(0)\right|\right\}\right| \geq N^p.$$

This result is due to Li and Zhang, who generalized the Anderson localization near the edge of the spectrum to the Anderson–Bernoulli model on $\mathbb{Z}^3$ [25].

3. Landis conjecture on decay of solutions to Schrödinger equations

3.1. Decay at infinity

In this section, we consider bounded solutions to the stationary Schrödinger equation with bounded potential, $\Delta u + Vu = 0$, $|V| \leq 1$. Landis conjectured that a solution to this equation cannot decay faster than exponential at infinity. An example of a function that decays exponentially is $u(x) = \exp(-(1+x^2)^{1/2})$.

We assume that there is a bounded solution to the Schrödinger equation with a bounded potential, and we are interested in the possible decay of the quantity $m_u(R) = \sup_{|x|>R} |u(x)|$. A local version of the Landis conjecture, which appeared in [2] in connection to the Anderson–Bernoulli model, is about the possible decay of the quantity $\mu_u(R) = \inf_{|x|=R} \sup_{B(x,1)} |u(x)|$.

For solutions of the continuous Schrödinger equation, the Landis conjecture was disproved by Meshkov, [29]. He gave an example of a complex valued function $u(x)$ which decays as $C \exp -c|x|^{4/3}$ and satisfies the inequality $|\Delta u| \leq |u|$ everywhere. The proof is based on a Carleman inequality. Bourgain and Kenig proved the following local version of the estimate.

Theorem. *Let $\Delta U + Vu = 0$, let $u(0) = 1$, and let u and V be bounded on $\mathbb{R}^d$. Then*

$$\mu_u(R) \geq c \exp(-CR^{4/3} \log R).$$

The proof also exploits a Carleman-type inequality. The remaining question is whether the original Landis conjecture holds for the class of real-valued potentials. For this case one may consider only real-valued solutions. This question is open in dimension $d \geq 3$.

3.2. Discrete equation

First, we consider the corresponding equation on the lattice $\mathbb{Z}^d$, here there is no difference between the real-valued and complex-valued cases, to the best of my knowledge.

Suppose that $\Delta U + VU = 0$, $U : \mathbb{Z}^d \to \mathbb{R}$, $|V| \leq C_0$, and $U \neq 0$, where

$$\Delta U(x) = \sum_{y \sim x} \big(U(y) - U(x)\big).$$

We also refer the reader to [1] for the discussion of this problem. Let

$$m_U(N) = \sup_{x \notin Q_N^d} \big|U(x)\big|.$$

We consider any $x \in Q_{N+1}^d \setminus Q_N^d$. Then there is one of its neighbors y such that $y \in Q_{N+2}^d \setminus Q_{N+1}^d$ and all neighbors of y except x are not in Q_{N+1}^d. Then the equation $\Delta U(y) + V(y)U(y) = 0$ can be written as

$$U(x) = U(y) + \sum_{z\sim y, z\neq x} \big(U(y) - U(z)\big) - V(y)U(y).$$

This implies that $m_U(N) \leq (2^{d+1} + 1 + C_0)m_U(N+1)$. Thus $m_U(N)$ does not decay faster than e^{-CN} as $N \to \infty$, where $C = C(d, C_0)$.

On the other hand, simple example shows that

$$\mu_U(N,k) = \inf_{x\in Q_N^d\setminus Q_{N-1}^d} \max_{|y-x|\leq k} \big|U(y)\big|$$

may be equal to zero for a non-trivial function U and bounded V; see [1]. Let us describe this example on $\mathbb{Z}^2$. We consider a function U which is zero on a tilted square

$$\tilde{Q}_N^2 = \big\{x = (x_1,x_2) \in \mathbb{Z}^2 : |x_1 + x_2| \leq 2N,\ |x_1 - x_2| \leq 2N\big\}$$

and takes non-zero values everywhere else. On the four diagonals $x_1 \pm x_2 = \pm 2N$, we define $U(x_1,x_2) = (-1)^{x_1}$, so that the function is harmonic at each point of $\tilde{Q}_N^2$. Then the values are arbitrary such that, for any $x \sim y$, we have $|U(x)| \leq (1+\varepsilon)|U(y)|$. Then we define $V(x) = -(\Delta U(x))/U(x)$ when $x \notin \tilde{Q}_N^2$. We see that $|V| \leq 8 + 4\varepsilon$. The example shows that there is no local version of the Landis conjecture when the potential is bounded but large enough. It would be interesting to obtain a local version for the case of the small potential.

3.3. Landis conjecture for real-valued potentials on the plane

The question of the estimates for the $m_u(R)$ and $\mu_u(R)$ for real-valued solutions of the Schrödinger equations in $\mathbb{R}^2$ is considered in [7–9,23], where local estimates were obtained under some assumptions on the potential. The decay estimate of the solution for the case of a periodic (in all but one variables) potential in $\mathbb{R}^2$ and $\mathbb{R}^3$ is discussed in [11].

The global and local versions of the result for solution of the Schrödinger equation with general bounded potential on $\mathbb{R}^2$ were recently obtained in [28]. It turns out that the Landis conjecture holds for this case (up to a logarithmic factor). More precisely, the following theorem holds.

Theorem 3.1 ([28]). *Let $u : \mathbb{R}^2 \to \mathbb{R}$ be a C^2 function which satisfies $|\Delta u| \leq |u|$. Then*

(i) *if $|u(x)| \leq \exp(-C|x|(\log|x|)^{1/2})$ and C is large enough, then $u = 0$;*

(ii) *if $\inf_{|x|=R} \sup_{B_1(x)} |u(x)| \leq \exp(-CR(\log R)^{3/2})$, then $u = 0$.*

There are three main steps in the proof. First, one constructs a family of separated D_j disks of equal radii r such that $\operatorname{dist}(D_j, \{u = 0\}) \geq 10r$ and each connected component Ω_k of $\{u \neq 0\} \setminus \bigcup_j D_j$ has the small first Laplace eigenvalue. Then, constructing an auxiliary solution of the equation $\Delta f + Vf = 0$ in Ω_k with boundary values $f = 1$ on $\partial\Omega_k$, one considers the ration $v = u/f$. This reduces the problem to the following one:

> *Let* $v : \mathbb{R}^2 \setminus \bigcup_j D_j \to \mathbb{R}$ *be a solution to the equation* $\operatorname{div}(f^2 \nabla v) = 0$ *and let* v *not change sign in each set* $10D_j \setminus D_j$*. Then if* v *decays as* $\exp(-C|x|(\log|x|)^{1/2})$ *with large* C*, then* $v = 0$.

The second step uses quasiconformal mappings to replace the general elliptic equation in divergence form by the Laplace equation; the factor $\log|x|^{1/2}$ in the exponent appears on this step. This step uses the specifics of dimension two. Finally, the above statement is proved for harmonic functions defined on $\mathbb{R}^2 \setminus \bigcup \tilde{D}_j$. The version of the last step for harmonic functions in higher dimensions is also discussed in [28].

4. Uncertainty principle and uniqueness for Schrödinger evolutions

4.1. Hardy's uncertainty principle

The Hardy uncertainty principle says that if $f \in L^2(\mathbb{R})$, $|f(x)| \leq Ce^{-a|x|^2}$, $|\widehat{f}(\xi)| \leq Ce^{-b|\xi|^2}$, and $ab > 1/4$, then $f = 0$. If $ab = 1/4$, then $f(x) = ce^{-a|x|^2}$. Its dynamical interpretation was found in [4, 12], where it is shown that the principle is equivalent to the following statement.

Theorem. *Let* $u(t,x)$ *be a solution to the free Schrödinger equation*

$$\partial_t u = i\Delta u(t,x).$$

Suppose that $u \in C^1([0,T], W^{2,2}(\mathbb{R}^d))$ *and*

$$\left|u(0,x)\right| \leq Ce^{-\alpha|x|^2} \quad \textit{and} \quad \left|u(T,x)\right| \leq Ce^{-\beta|x|^2},$$

where $\alpha, \beta > 0$*. Then the following hold.*

(i) *If* $\alpha\beta > (16T^2)^{-1}$*, then* $u(t,x) = 0$.

(ii) *If* $\alpha\beta = (16T^2)^{-1}$*, then* $u(t,x) = ce^{-(\alpha + i/(4T))|x|^2}$.

A real-variable proof of this result is given by Cowling, Escauriaza, Kenig, Ponce, and Vega in [6]. The last theorem was generalized to a large class of Schrödinger evolutions of the form $\partial_t u = i(\Delta u + Vu)$ in the series of articles [12–14].

4.2. Uniqueness results for discrete Schrödinger evolutions

Let Δ be again the discrete Laplacian on $\mathbb{Z}^d$. We consider the equation

$$\partial_t U(t,n) = i\big(\Delta_U(t,n) + V(t,n)U(t,n)\big),$$

where V is a bounded potential. We are interested in uniqueness results which says that if a solution to the discrete Schrödinger equation decays fast on $\mathbb{Z}^d$ at two distinct times, then it is trivial. First, we consider the free evolution with $V = 0$. In dimension $d = 1$, there is a solution $U_0(t,n) = i^{-n}e^{-2it}J_n(1-2t)$, where J_n is the Bessel function. This solution has optimal decay at $t = 0$ and $t = 1$. The role of the Gaussian is now played by the Bessel function. We get the following result for the free evolution:

Let $U(t,n)$ be a solution to $\partial_t U(t,n) = i\Delta U(t,n)$ on $[0,1]\times\mathbb{Z}$. Suppose that

$$\big|U(0,n)\big| + \big|U(1,n)\big| \le \frac{C}{\sqrt{|n|}}\left(\frac{e}{2|n|}\right)^{|n|}, \quad n \in \mathbb{Z}\setminus\{0\}.$$

Then $U(t,n) = C i^{-n}e^{-2it}J_n(1-2t)$.

This result was generalized to general bounded potentials in [22] (in dimension $d = 1$) and [1] (in arbitrary dimension). The result is as follows.

Theorem 4.1. *Let $U(t,n) \in C^1([0,1] : \ell^2(\mathbb{Z}^d))$ be a solution to*

$$\partial_t U(t,n) = i\big(\Delta U(t,n) + V(t,n)U(t,n)\big),$$

on $[0,1]\times\mathbb{Z}^d$. Suppose that $\|V\|_\infty \le 1$. There exists constant γ such that if

$$\big|U(0,n)\big| + \big|U(1,n)\big| \le C\exp\big(-\gamma|n|\log|n|\big), \quad n \in \mathbb{Z}^d\setminus\{0\},$$

then $U = 0$.

This result is not precise; we expect the same decay bounds as for the case of the free Schrödinger equation. One of the interesting applications of the uniqueness theorem with general potential which may depend on time is to the nonlinear Schrödinger equation. For this case, we have the same decay result as for the free equation. Let $U : [0,1]\times\mathbb{Z} \to \mathbb{R}$ be a solution to the equation

$$\partial_t U = i\big(\Delta U + c|U|^2 U\big).$$

Suppose that

$$\big|U(0,n)\big| + \big|U(1,n)\big| \le \left(\frac{c}{|n|}\right)^{|n|}, \quad n \in \mathbb{Z}\setminus\{0\},$$

where $c < e/2$. Then $U = 0$. We refer the reader to a recent survey [18] for detailed discussions of the uniqueness results for discrete and continuous Schrödinger evolutions.

Acknowledgments. The main results mentioned in this survey were obtained in [3, 22, 28]; it is my pleasure to thank all the coauthors for many interesting discussions. This note is a write up of my talk at the 8th ECM, and it is my pleasure to thank the organizers for the invitation and for making the Congress possible in these difficult circumstances.

Funding. This work was supported by NSF grant DMS-1956294 and by the Research Council of Norway, Project 275113.

References

[1] A. F. Bertolin and L. Vega, Uniqueness properties for discrete equations and Carleman estimates. *J. Funct. Anal.* **272** (2017), no. 11, 4853–4869 Zbl 1372.35260 MR 3630642

[2] J. Bourgain and C. E. Kenig, On localization in the continuous Anderson–Bernoulli model in higher dimension. *Invent. Math.* **161** (2005), no. 2, 389–426 Zbl 1084.82005 MR 2180453

[3] L. Buhovsky, A. Logunov, E. Malinnikova, and M. Sodin, A discrete harmonic function bounded on a large portion of $\mathbb{Z}^2$ is constant. *Duke Math. J.* **171** (2022), no. 6, 1349–1378 Zbl 07513362 MR 4408120

[4] S. Chanillo, Uniqueness of solutions to Schrödinger equations on complex semi-simple Lie groups. *Proc. Indian Acad. Sci. Math. Sci.* **117** (2007), no. 3, 325–331 Zbl 1129.22006 MR 2352052

[5] R. Courant, K. Friedrichs, and H. Lewy, Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.* **100** (1928), no. 1, 32–74 Zbl 54.0486.01 MR 1512478

[6] M. Cowling, L. Escauriaza, C. E. Kenig, G. Ponce, and L. Vega, The Hardy uncertainty principle revisited. *Indiana Univ. Math. J.* **59** (2010), no. 6, 2007–2025 Zbl 1242.35017 MR 2919746

[7] B. Davey, On Landis' conjecture in the plane for some equations with sign-changing potentials. *Rev. Mat. Iberoam.* **36** (2020), no. 5, 1571–1596 Zbl 1464.35069 MR 4161296

[8] B. Davey, C. Kenig, and J.-N. Wang, On Landis' conjecture in the plane when the potential has an exponentially decaying negative part. *Algebra i Analiz* **31** (2019), no. 2, 204–226 Zbl 1439.35109 MR 3937504

[9] B. Davey and J.-N. Wang, Landis' conjecture for general second order elliptic equations with singular lower order terms in the plane. *J. Differential Equations* **268** (2020), no. 3, 977–1042 Zbl 1453.35060 MR 4028997

[10] J. Ding and C. K. Smart, Localization near the edge for the Anderson Bernoulli model on the two dimensional lattice. *Invent. Math.* **219** (2020), no. 2, 467–506 Zbl 1448.60148 MR 4054258

[11] D. M. Elton, Decay rates at infinity for solutions to periodic Schrödinger equations. *Proc. Roy. Soc. Edinburgh Sect. A* **150** (2020), no. 3, 1113–1126 Zbl 1436.35094 MR 4091054

[12] L. Escauriaza, C. E. Kenig, G. Ponce, and L. Vega, On uniqueness properties of solutions of Schrödinger equations. *Comm. Partial Differential Equations* **31** (2006), no. 10-12, 1811–1823 Zbl 1124.35068 MR 2273975

[13] L. Escauriaza, C. E. Kenig, G. Ponce, and L. Vega, Hardy's uncertainty principle, convexity and Schrödinger evolutions. *J. Eur. Math. Soc. (JEMS)* **10** (2008), no. 4, 883–907 Zbl 1158.35018 MR 2443923

[14] L. Escauriaza, C. E. Kenig, G. Ponce, and L. Vega, The sharp Hardy uncertainty principle for Schrödinger evolutions. *Duke Math. J.* **155** (2010), no. 1, 163–187 Zbl 1220.35008 MR 2730375

[15] L. Escauriaza, C. E. Kenig, G. Ponce, and L. Vega, Uniqueness properties of solutions to Schrödinger equations. *Bull. Amer. Math. Soc. (N.S.)* **49** (2012), no. 3, 415–442 Zbl 1268.35112 MR 2917065

[16] A. Fernández-Bertolin, A discrete Hardy's uncertainty principle and discrete evolutions. *J. Anal. Math.* **137** (2019), no. 2, 507–528 Zbl 1421.35334 MR 3938012

[17] A. Fernández-Bertolin, Convexity properties of discrete Schrödinger evolutions. *J. Evol. Equ.* **20** (2020), no. 1, 257–278 Zbl 1436.39005 MR 4072656

[18] A. Fernández-Bertolin and E. Malinnikova, Dynamical versions of Hardy's uncertainty principle: a survey. *Bull. Amer. Math. Soc. (N.S.)* **58** (2021), no. 3, 357–375 Zbl 1468.42002 MR 4273105

[19] A. Fernández-Bertolin, L. Roncal, A. Rüland, and D. Stan, Discrete Carleman estimates and three balls inequalities. *Calc. Var. Partial Differential Equations* **60** (2021), no. 6, Paper No. 239 Zbl 1481.39004 MR 4328433

[20] M. Guadie and E. Malinnikova, On three balls theorem for discrete harmonic functions. *Comput. Methods Funct. Theory* **14** (2014), no. 4, 721–734 Zbl 1312.65177 MR 3274896

[21] H. A. Heilbronn, On discrete harmonic functions. *Proc. Cambridge Philos. Soc.* **45** (1949), 194–206 Zbl 0033.06303 MR 30051

[22] P. Jaming, Y. Lyubarskii, E. Malinnikova, and K.-M. Perfekt, Uniqueness for discrete Schrödinger evolutions. *Rev. Mat. Iberoam.* **34** (2018), no. 3, 949–966 Zbl 1401.31019 MR 3850274

[23] C. Kenig, L. Silvestre, and J.-N. Wang, On Landis' conjecture in the plane. *Comm. Partial Differential Equations* **40** (2015), no. 4, 766–789 Zbl 1320.35119 MR 3299355

[24] E. M. Landis, *Second Order Equations of Elliptic and Parabolic Type*. Transl. Math. Monogr. 171, American Mathematical Society, Providence, RI, 1998 Zbl 0895.35001 MR 1487894

[25] L. Li and L. Zhang, Anderson–Bernoulli localization on the three-dimensional lattice and discrete unique continuation principle. *Duke Math. J.* **171** (2022), no. 2, 327–415 Zbl 07500552 MR 4375618

[26] G. Lippner and D. Mangoubi, Harmonic functions on the lattice: absolute monotonicity and propagation of smallness. *Duke Math. J.* **164** (2015), no. 13, 2577–2595 Zbl 1337.31015 MR 3405594

[27] A. Logunov and E. Malinnikova, Quantitative propagation of smallness for solutions of elliptic equations. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. III. Invited Lectures*, pp. 2391–2411, World Sci. Publ., Hackensack, NJ, 2018 Zbl 1453.35061 MR 3966855

[28] A. Logunov, E. Malinnikova, N. Nadirashvili, and F. Nazarov, The Landis conjecture on exponential decay. 2020, arXiv:2007.07034

[29] V. Z. Meshkov, On the possible rate of decrease at infinity of the solutions of second-order partial differential equations. *Mat. Sb.* **182** (1991), no. 3, 364–383 Zbl 0782.35010 MR 1110071

[30] E. Nelson, A proof of Liouville's theorem. *Proc. Amer. Math. Soc.* **12** (1961), 995 Zbl 0124.31203 MR 259149

[31] H. B. Phillips and N. Wiener, Nets and the Dirichlet problem. *J. Math. Phys.* **2** (1923), 105–124 Zbl 49.0340.03

Eugenia Malinnikova
Department of Mathematical Sciences, Norwegian University of Science and Technology,
7491 Trondheim, Norway; and Department of Mathematics, Stanford University, Stanford,
CA 94305, USA; eugeniam@stanford.edu

DOI 10.4171/8ECM/15

Some recent developments on the geometry of random spherical eigenfunctions

Domenico Marinucci

Abstract. A lot of efforts have been devoted in the last decade to the investigation of the high-frequency behaviour of geometric functionals for the excursion sets of random spherical harmonics, i.e., Gaussian eigenfunctions for the spherical Laplacian $\Delta_{\mathbb{S}^2}$. In this survey, we shall review some of these results, with particular reference to the asymptotic behaviour of variances, phase transitions in the nodal case (*Berry's cancellation phenomenon*), the distribution of the fluctuations around the expected values, and the asymptotic correlation among different functionals. We shall also discuss some connections with the Gaussian kinematic formula, with Wiener chaos expansions, and with recent developments in the derivation of quantitative central limit theorems (the so-called Stein–Malliavin approach).

1. Introduction

Spherical eigenfunctions are defined as the solutions of the Helmholtz equation

$$\Delta_{\mathbb{S}^2} f_\ell + \lambda_\ell f_\ell = 0, \quad f_\ell : \mathbb{S}^2 \to \mathbb{R}, \quad \ell = 1, 2, \dots,$$

where $\Delta_{\mathbb{S}^2}$ is the spherical Laplacian and $\{-\lambda_\ell = -\ell(\ell+1)\}_{\ell=1,2,\dots}$ is the set of its eigenvalues. A random structure can be constructed easily by assuming that the eigenfunctions $\{f_\ell(\cdot)\}$ follow a Gaussian isotropic random process on $\mathbb{S}^2$. More precisely, for each $x \in \mathbb{S}^2$, we take $f_\ell(x)$ to be a Gaussian random variable defined on a suitable probability space $\{\Omega, \Im, \mathbb{P}\}$; without loss of generality, we assume $\{f_\ell(\cdot)\}$ to have mean zero, unit variance, and covariance function given by

$$\mathbb{E}\big[f_\ell(x) f_\ell(y)\big] = P_\ell\big(\langle x, y\rangle\big), \quad x, y \in \mathbb{S}^2,$$
$$P_\ell(t) := \frac{1}{2^\ell \ell!} \frac{d^\ell}{dt^\ell}(t^2 - 1), \quad t \in [-1, 1],$$

where $\{P_\ell(\cdot)\}$ denotes the family of Legendre polynomials: this is the only covariance

2020 Mathematics Subject Classification. Primary 60G60; Secondary 62M15, 53C65, 42C10, 33C55.
Keywords. Random eigenfunctions, spherical harmonics, Lipschitz–Killing curvatures, kinematic formulae, nodal lines, Wiener–Itô expansions.

structure to ensure that the random eigenfunctions are isotropic, that is, invariant in law with respect to the action of the group of rotations SO(3). Random spherical eigenfunctions, also known as random spherical harmonics, arise in a huge number of applications, especially in connection with mathematical physics: in particular, their role in quantum chaos has drawn strong interest in the last two decades, starting from the seminal papers [7, 8, 43, 61]; also, they represent the Fourier components of isotropic spherical random fields, whose analysis has an extremely important role in cosmology (see, e.g., [35]). Of course, random spherical harmonics are just a special case of a much richer literature on random eigenfunctions on general manifolds; special interest has been drawn for instance by *arithmetic random waves*, i.e., random eigenfunctions on the torus $\mathbb{T}^d$, which were introduced in [52] and then studied, among others, in [9, 10, 19, 26, 27, 33, 36, 53, 54]; see also [17, 55] and the references therein. Although some of the results that we shall discuss have related counterparts on the torus, on the higher-dimensional spheres, on more general compact manifolds, and in the Euclidean case, we will stick mainly to $\mathbb{S}^2$ for brevity and simplicity.

A lot of efforts have been spent in the last decade to characterize the geometry of the excursion sets of random spherical harmonics, which are defined as

$$A_u(f_\ell; \mathbb{S}^2) := \{x \in \mathbb{S}^2 : f_\ell(x) \geq u\}, \quad u \in \mathbb{R}. \tag{1.1}$$

A classical tool for the investigation of these sets is given by the so-called Lipschitz–Killing curvatures (or, equivalently, by Minkowski functionals; see [1]), which in dimension 2 correspond to the Euler–Poincaré characteristic, (half of) the boundary length and the excursion area. A general expression for their expected values (covering much more general Gaussian fields than random eigenfunctions) is given by the *Gaussian kinematic formula* (see [1, 58]). Over the last decade, more refined characterizations for random spherical harmonics have been obtained, including neat analytic expressions (in the high-energy limit $\lambda_\ell \to \infty$) for the fluctuations around their expected values and the correlation among these different functionals; much of the literature has been concerned with the *nodal* case, corresponding to $u = 0$, to which we shall devote special attention. In this survey, we shall review some of these results and present some open issues for future research.

2. The Gaussian kinematic formula for Lipschitz–Killing curvatures on excursions sets

2.1. The Kac–Rice formula and the expectation metatheorem

The first modern attempt to investigate the geometry of random processes and fields can probably be traced back to the groundbreaking work by Kac (1943) and Rice

(1945) [25, 49] on the zeroes of stochastic processes. Their pioneering argument can be introduced as follows: let $f(\cdot,\cdot) : \Omega \times \mathbb{R} \to \mathbb{R}$ be a continuous stochastic process satisfying regularity conditions; our aim is to derive the expected cardinality of its zero set in some finite interval (say $[0, T]$), i.e., the mean of

$$N_0\big([0,T]\big) := \operatorname{Card}\big\{t \in [0,T] : f(t) = 0\big\}.$$

Now assume that $\{f(\cdot)\}$ is C^1 with probability one, such that $f(0), f(T) \neq 0$ and

$$\big\{t : f(t) = 0,\ f'(t) = 0\big\} = \varnothing;$$

then the following result (*Kac's counting lemma*) can be established easily (see [3, p. 69]):

$$N_0\big([0,T]\big) = \lim_{\varepsilon \to 0} \int_0^T \frac{1}{2\varepsilon} \mathbb{I}_{(-\varepsilon,\varepsilon)}\big(f(t)\big)\big|f'(t)\big|\,dt,$$

where as usual $\mathbb{I}_A$ denotes the indicator function of the set A. With further efforts and assuming that all exchanges of integrals and limits can be justified, one obtains also

$$\mathbb{E}\big[N_0\big([0,T]\big)\big] = \int_0^T \mathbb{E}\big[\big|f'(t)\big| \mid f(t) = 0\big]p_{f(t)}(0)\,dt, \tag{2.1}$$

where $\mathbb{E}[\cdot|\cdot]$ denotes as usual the conditional expected value and $p_f(\cdot)$ the marginal density of $f(\cdot)$, which is assumed to exist and admit enough regularity conditions (in the overwhelming majority of the literature and in this whole survey, $f(\cdot)$ will indeed be assumed to be Gaussian); (2.1) is the simplest example of the *Kac–Rice formula*.

The basic idea behind the Kac–Rice approach has proved to be extremely fruitful, leading to an enormous amount of applications and generalizations. In particular, in the research monographs [1, 3], (slightly different) versions of a general *expectation metatheorem* (in the terminology of [1]) are proved. More precisely, let us take M to be a compact, d-dimensional oriented C^1 manifold with a C^1 Riemannian metric g. Assume that $f : M \to \mathbb{R}^d$ and $h : M \to \mathbb{R}^k$ are vector-valued random fields which satisfy suitable regularity conditions (see [1,3] for more details and [56] for some very recent developments). Let $B \subset \mathbb{R}^k$ be a subset with boundary dimension smaller than or equal to $k-1$; then define

$$N_u(f, h, M, B) = \big\{t \in M : f(t) = u,\ h(t) \in B\big\}, \quad u \in \mathbb{R}^d.$$

The following extension of the Kac–Rice formula holds.

Theorem 2.1 ([1, 3]). *It holds that*

$$\mathbb{E}\big[N_u(f, h, M, B)\big] = \int_M \mathbb{E}\big[\big|\det\{\nabla f(t)\}\big|\mathbb{I}_B\big(h(t)\big) \mid f(t) = u\big]p_{f(t)}(u)\sigma_g(dt),$$

where as before $\mathbb{I}_B(\cdot)$ denotes the indicator function, $\nabla f(\cdot)$ the (covariant) gradient of $f(\cdot)$, and $\sigma_g(\cdot)$ the volume form induced by the metric g.

Remark. By taking $k = 1$, $f := \nabla h$ the gradient of h (and hence $\nabla f = \nabla^2 h$ its Hessian) and $u = (0, \dots, 0)$, Theorem 2.1 yields the expected number of critical points with values in B for the scalar random field h. Simple modifications similarly yield the expected values for maxima, minima, and saddle points.

The previous results have all been restricted to vector-valued random fields whose image space has co-dimension zero. However, the results can be similarly generalized to strictly positive co-dimensions. Indeed, under the same setting as before assume instead that $f : M \to \mathbb{R}^{d'}$ is such that $d' < d$; then ∇X is a $d \times d'$ rectangular matrix, and the following generalization of the expectation metatheorem holds (see [1,3]).

Theorem 2.2 ([1,3]). *It holds that*

$$\begin{aligned}\mathbb{E}\big[\mathcal{H}_u(f, h, M, B)\big]\\ = \int_M \mathbb{E}\Big[\big|\det\big\{(\nabla f(t))^T(\nabla f(t))\big\}\big|^{1/2}\mathbb{I}_B(h) \mid f(t) = u\Big]p_{f(t)}(u)\sigma_g(dt),\end{aligned}$$

where $\mathcal{H}_u(f, h, M, B)$ denotes the $d - d'$ dimensional Hausdorff measure of the set $\{t \in M : f(t) = u \text{ and } h(t) \in B\}$.

Example 2.3. Let $M = \mathbb{S}^2$ be the standard unit-dimensional sphere in $\mathbb{R}^3$, let $f : \mathbb{S}^2 \times \Omega \to \mathbb{R}$ be a random field, and let

$$\operatorname{Len}(f) := \mathcal{H}_0(f, \mathbb{S}^2, 0) = \operatorname{meas}\big\{t \in \mathbb{S}^2 : f(t) = 0\big\},$$

i.e., the length of the nodal lines of $f(\cdot)$. Then

$$\begin{aligned}\mathbb{E}\big[\operatorname{Len}(f)\big] &= \int_{\mathbb{S}^2} \mathbb{E}\Big[\big|\det\big\{(\nabla f(t))^T(\nabla f(t))\big\}\big|^{1/2} \mid f(t) = 0\Big]p_{f(t)}(0)\sigma(dt)\\ &= \int_{\mathbb{S}^2} \mathbb{E}\big[\|\nabla f(t)\| \mid f(t) = 0\big]p_{f(t)}(0)\sigma(dt),\end{aligned}$$

where $\|\cdot\|$ denotes Euclidean norm and $\sigma(\cdot)$ the standard Lebesgue measure on the unit sphere. In particular, assuming that the law of $f(\cdot)$ is isotropic (that is, invariant with respect to the action of the group of rotations SO(3)), we get

$$\mathbb{E}\big[\operatorname{Len}(f)\big] = 4\pi \times \mathbb{E}\big[\|\nabla f(t)\| \mid f(t) = 0\big]p_{f(t)}(0).$$

2.2. Intrinsic volumes and Lipschitz–Killing curvatures

In the sequel, as mentioned earlier we will restrict our attention only to Gaussian processes, which have driven the vast majority of research in this area. We need now to introduce the Gaussian kinematic formula (see [1,58]); to this aim, let us first recall the notion of *Lipschitz–Killing curvatures*. In the simplest setting of convex subsets of the Euclidean space $\mathbb{R}^d$, Lipschitz–Killing curvatures (also known as intrinsic volumes) can be defined implicitly by means of *Steiner's tube formula*; to recall the latter, for any convex d-dimensional set $A \subset \mathbb{R}^d$ define the Tube of radius ρ around A as

$$\operatorname{Tube}(A, \rho) := \left\{x \in \mathbb{R}^d : d(x, A) \leq \rho\right\}, \quad d(x, A) = \inf_{y \in A} d(x, y),$$

where $d(\cdot, \cdot)$ is the standard Euclidean distance. Then the following expansion holds:

$$\mu_d\left\{\operatorname{Tube}(A, \rho)\right\} = \sum_{j=0}^{d} \omega_{d-j} \rho^{d-j} \mathcal{L}_j(A),$$

where $\mathcal{L}_j(A)$ denotes the jth Lipschitz–Killing curvatures, $\mu_d(\cdot)$ denotes the d-dimensional Lebesgue measure, and

$$\omega_j := \frac{\pi^{j/2}}{\Gamma\left(\frac{j}{2} + 1\right)}$$

is the volume of the j-dimensional unit ball ($\omega_0 = 1, \omega_1 = 2, \omega_2 = \pi, \omega_3 = \frac{4}{3}\pi$).

Lipschitz–Killing curvatures can be shown to be additive and to scale with dimensionality, in the sense that

$$\mathcal{L}_j(\lambda A) = \lambda^j \mathcal{L}_j(A) \quad \forall \lambda > 0,$$
$$\mathcal{L}_j(A_1 \cup A_2) = \mathcal{L}_j(A_1) + \mathcal{L}_j(A_2) - \mathcal{L}_j(A_1 \cap A_2).$$

For $j = d$, it is immediately seen that $\mathcal{L}_d(A)$ is just the Hausdorff measure of A, whereas for $j = 0$ we obtain $\mathcal{L}_0(A) = \varphi(A)$, the (integer-valued) Euler–Poincaré characteristic of A. A more general definition of $\mathcal{L}_j(\cdot)$ can be given for basic complexes (i.e., disjoint union of complex sets), for which the following characterization (due to Hadwiger, see [1]) holds:

$$\mathcal{L}_j(A) = \frac{\omega_d}{\omega_{d-j}\omega_j} \binom{d}{j} \int_{\mathcal{G}_d} \varphi(A \cap gE_{d-j}) \mu(dg), \tag{2.2}$$

where $\mathcal{G}_d = \mathbb{R}^d \times O(n)$ is the group of rigid motions, E_{d-j} is any $d - j$ dimensional affine subspace, and the volume form $\mu(dg)$ is normalized so that

$$\text{for all } x \in \mathbb{R}^d,\ A \subset \mathbb{R}^d, \quad \mu\{g : gx \in A\} = \mathcal{H}(A),$$

where as before $\mathcal{H}(\cdot)$ denotes the Hausdorff measure. For instance, for $A = \mathbb{S}^2$ it is well known and easy to check that (2.2) gives

$$\mathcal{L}_0(\mathbb{S}^2) = 2, \quad \mathcal{L}_1(\mathbb{S}^2) = 0, \quad \mathcal{L}_2(\mathbb{S}^2) = 4\pi,$$

which represent, respectively, the Euler–Poincaré characteristic, (half) the boundary length, and the area of the 2-dimensional unit sphere.

2.3. The Gaussian kinematic formula

From now on, we shall restrict our attention to Gaussian processes $f : M \to \mathbb{R}$, which we shall take to be zero-mean and isotropic, meaning as usual that $\mathbb{E}[f(t)] = 0$ and $f(gt) \overset{d}{=} f(t)$ for all $t \in M \subset \mathbb{R}^d$ and $g \in SO(d)$; more explicitly, the law of the field $f(\cdot)$ will always be taken to be invariant to rotations. In order to present the Gaussian kinematic formula, let us first introduce a Riemannian structure governed by the covariance function of the field $\{f(\cdot)\}$; more precisely, consider the metric induced on the tangent plan $T_t M$ by the following inner product [1, p. 305]:

$$g^f(X_t, Y_t) := \mathbb{E}[X_t f \cdot Y_t f], \quad X_t, Y_t \in T_t M.$$

This metric takes a particular simple form in case the field $f(\cdot)$ is isotropic; in these circumstances, $g^f(\cdot, \cdot)$ is simply the standard Euclidean metric, rescaled by a factor that corresponds to the square root of (minus) the derivative of the covariance density at the origin.

Example 2.4. Consider the random spherical eigenfunction satisfying

$$\Delta f_\ell = -\lambda_\ell f_\ell, \quad f_\ell : \mathbb{S}^2 \to \mathbb{R}, \quad \ell = 0, 1, 2, \ldots,$$

with

$$\mathbb{E}\big[f_\ell(x)\big] = 0, \quad \mathbb{E}\big[f_\ell(x_1) f_\ell(x_2)\big] = P_\ell\big(\langle x_1, x_2 \rangle\big), \quad P'_\ell(1) = -\frac{\ell(\ell+1)}{2}.$$

Then the induced inner product is simply

$$g^{f_\ell}(X, Y) = \sqrt{\frac{\ell(\ell+1)}{2}} \langle X, Y \rangle_{\mathbb{R}^3};$$

this change of metric can of course be realized by transforming $\mathbb{S}^2$ into

$$\mathbb{S}^2_{\sqrt{\lambda_\ell/2}} := \sqrt{\lambda_\ell/2}\,\mathbb{S}^2.$$

Let us now write $\mathcal{L}_j^f(A)$ for the jth Lipschitz–Killing curvatures of the set A under the metric induced by the zero-mean Gaussian field f; for instance, in the case of spherical random eigenfunctions we get immediately

$$\mathcal{L}_0^{f_\ell}(\mathbb{S}^2) = \mathcal{L}_0\big(\mathbb{S}^2_{\sqrt{\lambda_\ell/2}}\big) = 2, \quad \mathcal{L}_1^{f_\ell}(\mathbb{S}^2) = 0, \quad \mathcal{L}_2^{f_\ell}(\mathbb{S}^2) = 4\pi \frac{\lambda_\ell}{2}.$$

For further notation, as in [1] we shall write

$$\rho_j(u) := \frac{1}{(2\pi)^{1/2+j/2}} \exp(-u^2/2) H_{j-1}(u), \quad j \geq 1,$$
$$\rho_0(u) := 1 - \Phi(u) = \int_u^\infty \varphi(t)\, dt,$$

where as usual $\varphi(t) = (2\pi)^{-1/2} \exp(-t^2/2)$ denotes the standard Gaussian density and we introduced the Hermite polynomials

$$H_k(u) := (-1)^k \exp\left(\frac{u^2}{2}\right) \frac{d^k}{du^k} \exp\left(-\frac{u^2}{2}\right), \quad k = 0, 1, 2, \ldots, \ u \in \mathbb{R}; \tag{2.3}$$

for instance $H_0(u) = 1$, $H_1(u) = u$, $H_2(u) = u^2 - 1, \ldots$. Finally, we shall introduce the *flag coefficients*

$$\begin{bmatrix} d \\ k \end{bmatrix} := \binom{d}{k} \frac{\omega_d}{\omega_k \omega_{d-k}}, \quad k = 0, 1, \ldots, d. \tag{2.4}$$

We are now in the position to state the following.

Theorem 2.5 (Gaussian kinematic formula, [1, Theorem 13.4.1] and [58]). *Under regularity conditions, for all $j = 0, 1, \ldots, n$ one has that*

$$\mathbb{E}\big[\mathcal{L}_j^f\big(A_u(f; M)\big)\big] = \sum_{k=0}^{d-j} \begin{bmatrix} k+j \\ k \end{bmatrix} \rho_k(u) \mathcal{L}_{k+j}^f(M). \tag{2.5}$$

Before we proceed with some examples, it is worth discussing formula (2.5). We are evaluating the expected value of a complex geometric functional on a complicated excursion set, in very general circumstances (under minimal regularity conditions on the field and on the manifold on which it is defined). It is clear that the expected value should depend on the manifold, on the threshold level, and on the field one considers, and one may expect these three factors to be intertwined in a complicated manner. On the contrary, formula (2.5) shows that their role is completely decoupled; more precisely

- the threshold u enters the formula merely through the functions $\rho_j(u)$ which are very simple and fully universal (i.e., they do not depend neither on the field nor on the manifold);
- on the left-hand side Lipschitz–Killing curvatures appear, but they are computed on the original manifold, not on the excursion sets, and they are therefore again extremely simple to compute;

- the role of the field f is confined to the new metric $g^f(\cdot,\cdot)$ that it induces and under which the Lipschitz–Killing curvatures are computed on both sides; under the (standard) assumption of isotropy, this implies only a rescaling of the manifold by means of a factor depending only on the derivative of the covariance function at the origin.

Example 2.6. Let us consider a zero-mean isotropic Gaussian field f defined on $\mathbb{S}^d$ (the unit sphere in $\mathbb{R}^{d+1}$); its covariance function can be written as

$$\mathbb{E}\big[f(x_1)f(x_2)\big] = \sum_{\ell=0}^{\infty} \frac{n_{\ell,d}}{s_{d+1}} C_\ell G_{\ell;\frac{d}{2}}\big(\langle x_1, x_2\rangle\big),$$

where $s_{d+1} = (d+1)\omega_{d+1}$ is the surface measure of $\mathbb{S}^d$, $G_{\ell;\alpha}(\cdot)$ denotes the normalized Gegenbauer polynomials of order α, whereas

$$n_{\ell,d} = \frac{2\ell + d - 1}{\ell}\binom{\ell + d - 2}{\ell - 1} \sim \frac{2}{(d-1)!}\ell^{d-1}, \quad \text{as } \ell \to \infty,$$

is the dimension of the eigenspace corresponding to the ℓth eigenvalue $\lambda_{\ell;d} := \ell(\ell + d - 1)$; here $\{C_\ell\}$ is a sequence of non-negative weights which represent the so-called angular power spectrum of the random field. The derivative of the covariance function at the origin is

$$\mu := \sum_{\ell=0}^{\infty} \frac{n_{\ell,d}}{s_{d+1}} C_\ell \frac{\lambda_{\ell;d}}{d}.$$

Recall that the Lipschitz–Killing curvatures of the manifold $\mathbb{S}^d_\lambda := \lambda \mathbb{S}^d$ are given by

$$\mathcal{L}_j\big(\lambda \mathbb{S}^d\big) = 2\binom{d}{j}\frac{s_{d+1}}{s_{d+1-j}}\lambda^j,$$

for $d - j$ even, and 0 otherwise, see [1, p. 179]. Then the Gaussian kinematic formula reads

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}_j^f\big(A_u(f;\mathbb{S}^d)\big)\big] &= \sum_{k=0}^{d-j} \rho_k(u) \begin{bmatrix} k + j \\ k \end{bmatrix} \mathcal{L}_{k+j}(\sqrt{\mu}\mathbb{S}^d) \\ &= \sum_{k=0}^{d-j} \rho_k(u) \begin{bmatrix} k + j \\ k \end{bmatrix} \mathcal{L}_{k+j}(\mathbb{S}^d)\mu^{(k+j)/2}.\end{aligned}$$

Example 2.7. As a special case of the previous example, assume that $f = f_\ell$ is actually a unit variance random eigenfunction on $\mathbb{S}^2$ corresponding to the eigenvalue $-\ell(\ell+1)$, $\ell = 0, 1, 2, \ldots$.

Then the Gaussian kinematic formula gives

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}_0^{f_\ell}\big(A_u(f_\ell;\mathbb{S}^2)\big)\big] &= \mathbb{E}\big[\mathcal{L}_0\big(A_u(f_\ell;\mathbb{S}^2)\big)\big]\\ &= 2\big\{1-\Phi(u)\big\} + \frac{1}{2\pi}u\phi(u)(4\pi)\frac{\ell(\ell+1)}{2},\\ \mathbb{E}\big[\mathcal{L}_1^{f_\ell}\big(A_u(f_\ell;\mathbb{S}^2)\big)\big] &= \rho_1(u)\begin{bmatrix}2\\1\end{bmatrix}\mathcal{L}_2(\mathbb{S}^2)\left\{\frac{\ell(\ell+1)}{2}\right\}\end{aligned}$$

so that

$$\mathbb{E}\big[\mathcal{L}_1\big(A_u(f_\ell;\mathbb{S}^2)\big)\big] = \pi\exp\left(-\frac{u^2}{2}\right)\left\{\frac{\ell(\ell+1)}{2}\right\}^{1/2},$$

and finally

$$\mathbb{E}\big[\mathcal{L}_2\big(A_u(f_\ell;\mathbb{S}^2)\big)\big] = \big\{1-\Phi(u)\big\}\mathcal{L}_2(\mathbb{S}^2) = \big\{1-\Phi(u)\big\}4\pi.$$

Example 2.8. In the special case of the nodal volume $\mathcal{L}_{d-1}(A_0(\mathbb{S}^d), f_\ell)$ of random eigenfunctions, i.e., half the Hausdorff measure of the zero-set of the eigenfunction, the Gaussian kinematic formula gives

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}_{d-1}^{f}\big(A_u(f_\ell;\mathbb{S}^d)\big)\big] &= \left(\frac{\lambda_\ell}{d}\right)^{(d-1)/2}\mathbb{E}\big[\mathcal{L}_{d-1}\big(A_u(f_\ell;\mathbb{S}^d)\big)\big]\\ &= \rho_1(u)\frac{d\omega_d}{\omega_1\omega_{d-1}}\mathcal{L}_d(\mathbb{S}^d)\left(\frac{\lambda_\ell}{d}\right)^{d/2}\end{aligned}$$

so that, recalling $\omega_j = \frac{\pi^{j/2}}{\Gamma(\frac{j}{2}+1)}$ and $\mathcal{L}_d(\mathbb{S}^d) = (d+1)\omega_{d+1}$, we have

$$\begin{aligned}\mathbb{E}\big[\mathcal{L}_{d-1}\big(A_u(f_\ell;\mathbb{S}^d)\big)\big] &= \frac{1}{2\pi}\exp\left(-\frac{u^2}{2}\right)\frac{d\omega_d}{\omega_1\omega_{d-1}}\mathcal{L}_d(\mathbb{S}^d)\left(\frac{\lambda_\ell}{d}\right)^{1/2}\\ &= \exp\left(-\frac{u^2}{2}\right)\frac{\pi^{d/2}}{\Gamma(\frac{d}{2})}\left(\frac{\lambda_\ell}{d}\right)^{1/2}.\end{aligned}\tag{2.6}$$

For $u=0$ equation (2.6) was derived for instance in [6] (see [61]) and it is consistent with a celebrated conjecture in [63], which states that for C^∞ manifolds the nodal volume of any eigenfunction corresponding to the eigenvalue E should belong to the interval $[c_1\sqrt{E}, c_2\sqrt{E}]$ for some constants $0 < c_1 \le c_2 < \infty$. The conjecture was settled for real analytic manifolds in [22]; for smooth manifolds the lower bound was established much more recently; see [29–31] while the upper bound is addressed in [32]. As a consequence of the results in the next two sections below in the case of the sphere in a probabilistic sense, the upper and lower constants can be taken nearly coincident, in the limit of diverging eigenvalues.

3. Wiener chaos expansions, variances, and correlations

In view of the results detailed in Section 2, the question related to the expectation of intrinsic volumes in the case of Gaussian fields can be considered completely settled. The next step of interest is the computation of the corresponding variances, and the asymptotic laws of fluctuations around the expected values, in the high-frequency regime. The first rigorous results in this area can be traced back to a seminal paper by Igor Wigman [61], where the variance of the nodal length (i.e., $\mathrm{Len}(f_\ell, \mathbb{S}^2) := 2\mathscr{L}_1(A_0(f_\ell, \mathbb{S}^2)))$ for random spherical harmonics in dimension 2 is computed and shown to be asymptotic to

$$\mathrm{Var}\left[\mathrm{Len}(f_\ell, \mathbb{S}^2)\right] = \frac{\log \ell}{32} + O_{\ell\to\infty}(1). \tag{3.1}$$

We shall start instead from the derivation of variances and central limit theorems for Lipschitz–Killing curvatures of excursion sets at $u \neq 0$, although these results were actually obtained more recently than (3.1).

Let us recall first the notion of Wiener chaos expansions. In the simplest setting, consider $Y = G(Z)$, i.e., the transform of a zero mean, unit variance Gaussian random variable Z, such that $\mathbb{E}[G(Z)^2] < \infty$; it is well known that the following expansion holds, in the $L^2(\Omega)$ sense:

$$G(Z) = \sum_{q=0}^{\infty} \frac{J_q(G)}{q!} H_q(Z), \tag{3.2}$$

where $\{H_q(\cdot)\}_{q=0,1,2,\dots}$ denotes the family of Hermite polynomials that we introduced earlier in (2.3), and $J_q(G)$ are projection coefficients given by $J_q(G) := \mathbb{E}[G(Z)H_q(Z)]$ (see, e.g., [24, 46]). The summands in (3.2) are orthogonal, because when evaluated on pairs of standard Gaussian variables Z_1, Z_2, Hermite polynomials enjoy a very simple formula for the computation of covariances:

$$\mathbb{E}\left[H_{q_1}(Z_1)H_{q_2}(Z_2)\right] = \delta_{q_1}^{q_2} q_1! \{\mathbb{E}[Z_1 Z_2]\}^{q_1}, \tag{3.3}$$

where $\delta_{q_1}^{q_2}$ denotes the Kronecker delta. Equation (3.3) is just a special case of the celebrated *diagram (or Wick's) formula*; see [46] for much more discussion and details. We thus have immediately

$$\mathrm{Var}\left\{G(Z)\right\} = \sum_{q=0}^{\infty} \frac{J_q^2(G)}{q!}.$$

More generally, let $\{Z_1, \dots, Z_j, \dots\}$ be any array of independent standard Gaussian variables, and consider elements of the form

$$H_{q_1}(Z_1) \cdots H_{q_p}(Z_p), \quad q_1 + \cdots + q_p = q;$$

the linear span (in the $L^2(\Omega)$ sense) of these random variables is usually written as $\mathcal{C}_q$ (denoted by the qth-order Wiener chaos; see again [46]) and we have the orthogonal decomposition

$$L^2(\Omega) = \bigoplus_{q=0}^{\infty} \mathcal{C}_q.$$

3.1. Wiener chaos expansions for random eigenfunctions

Let us now explain how these techniques can be pivotal for the investigation of fluctuations of geometric functionals. We start from the simplest case, the excursion volume/area for the 2-dimensional sphere, which we can write as

$$\mathcal{L}_2\big(A_u(f_\ell;\mathbb{S}^2)\big) = \int_{\mathbb{S}^2} \mathbb{I}_{[u,\infty)}\big(f_\ell(x)\big)\,dx,$$

$\mathbb{I}_{[u,\infty)}(\cdot)$ denoting the indicator function of the semi-interval $[u,\infty)$. It is not difficult to show that

$$\begin{aligned} J_q\big(\mathbb{I}_{[u,\infty)}(\cdot)\big) &= \mathbb{E}\big[\mathbb{I}_{[u,\infty)}(Z)H_q(Z)\big] \\ &= \int_u^{\infty} H_q(z)\phi(z)dz = (-1)^q H_{q-1}(u)\phi(u), \end{aligned}$$

the last result following by integration by parts, under the convention that

$$(-1)H_{-1}(u)\phi(u) := 1 - \Phi(u).$$

In view of (3.2), we thus have [40, 41]

$$\begin{aligned} \mathcal{L}_2\big(A_u(f_\ell;\mathbb{S}^2)\big) &= \int_{\mathbb{S}^2} \sum_{q=0}^{\infty} (-1)^q H_{q-1}(u)\phi(u) \frac{H_q\big(f_\ell(x)\big)}{q!}\,dx \\ &= \sum_{q=0}^{\infty} \frac{(-1)^q}{q!} H_{q-1}(u)\phi(u) h_{\ell;q}, \end{aligned}$$

where

$$h_{\ell;q} = \int_{\mathbb{S}^2} H_q\big(f_\ell(x)\big)\,dx;$$

as a consequence, we have also

$$\operatorname{Var}\big\{\mathcal{L}_2\big(A_u(f_\ell;\mathbb{S}^2)\big)\big\} = \sum_{q=0}^{\infty} \frac{1}{(q!)^2} H_{q-1}^2(u)\phi^2(u)\operatorname{Var}\{h_{\ell;q}\}. \tag{3.4}$$

The crucial observation to be drawn at this stage is that the variances of the components $\{h_{\ell;q}\}$ exhibit a form of phase transition with respect to their order q, in the high-frequency/high-energy limit $\ell \to \infty$. In particular, a simple application of the diagram formula (3.3), isotropy, and a change of variable yield

$$\begin{aligned}\operatorname{Var}\{h_{\ell;q}\} &= \int_{\mathbb{S}^2\times\mathbb{S}^2} \mathbb{E}\{H_q(f_\ell(x))H_q(f_\ell(y))\}\,dx\,dy\\ &= 8\pi^2 q! \int_0^\pi \{P_\ell(\cos\theta)\}^q \sin\theta\,d\theta;\end{aligned}$$

for instance, for $q = 2$ we obtain exactly

$$\operatorname{Var}\{h_{\ell;q}\} = 2\times 8\pi^2 \int_0^\pi P_\ell^2(\cos\theta)\sin\theta\,d\theta = 16\pi^2\frac{2}{2\ell+1}.$$

Given two sequences of positive numbers a_n, b_n, we shall write $a_n \approx b_n$ when we have that $a_n/b_n \to c$ as $n\to\infty$, $c>0$. By means of the so-called Hilb asymptotics [57,61], it is possible to show that, as $\ell\to\infty$ [42],

$$\operatorname{Var}\{h_{\ell;q}\} \approx \frac{1}{\ell^2}\times\int_0^{\ell\pi}\frac{1}{\psi^{q/2}}\psi\,d\psi \approx \begin{cases}\ell^{-1} & \text{for } q=2\\ \ell^{-2}\log\ell & \text{for } q=4\\ \ell^{-2} & \text{for } q=3,5,\dots.\end{cases}$$

Note that $h_{\ell;1}\equiv 0$ for all $\ell = 1,2,\dots$, whereas the term for $q=3$ requires an ad-hoc argument given in [34,40]. As a consequence, the dominant terms in the variance expansion correspond to $q=2$ when $H_1(u)$ is non-zero, i.e., for $u\neq 0$; for $u=0$ the even-order chaoses vanish and all the remaining terms contribute by the same order of magnitude with respect to ℓ. In conclusion, we have that

$$\begin{aligned}&\mathcal{L}_2(A_u(f_\ell;\mathbb{S}^2)) - \mathbb{E}[\mathcal{L}_2(A_u(f_\ell;\mathbb{S}^2))]\\ &\quad= \frac{1}{2}H_1(u)\phi(u)h_{\ell;2} + O_p(\sqrt{\log\ell/\ell^2}),\end{aligned} \tag{3.5}$$

and for $u\neq 0$

$$\operatorname{Var}\{\mathcal{L}_2(A_u(f_\ell;\mathbb{S}^2))\} \sim \left\{\frac{1}{2}H_1(u)\phi(u)\right\}^2 \operatorname{Var}\{h_{\ell;2}\}, \quad \text{as } \ell\to\infty.$$

Because

$$h_{\ell;2} = \int_{\mathbb{S}^2}\{f_\ell^2(x)-1\}\,dx = \|f_\ell\|^2_{L^2(\mathbb{S}^2)} - \mathbb{E}[\|f_\ell\|^2_{L^2(\mathbb{S}^2)}],$$

equation (3.5) is basically stating that the fluctuations in the excursion area for $u\neq 0$ are dominated by the fluctuations in the random norm of the eigenfunctions.

Interestingly, the same behaviour characterizes also the other Lipschitz–Killing curvatures; for the boundary length we have the expansion

$$2\mathcal{L}_1\big(A_u(f_\ell;\mathbb{S}^2)\big) = \lim_{\varepsilon\to 0}\int_{\mathbb{S}^2} \|\nabla f_\ell(x)\|\delta_\varepsilon\big(f_\ell(x)-u\big)\,dx$$

which holds both ω almost surely and in $L^2(\Omega)$; here we write $\delta_\varepsilon(\cdot) = \frac{1}{2\varepsilon}\mathbb{I}(\cdot)$. Similarly for the Euler–Poincaré characteristic we have

$$\mathcal{L}_0\big(A_u(f_\ell;\mathbb{S}^2)\big) = \lim_{\varepsilon\to 0}\int_{\mathbb{S}^2} \det\big\{\nabla^2 f_\ell(x)\big\}\delta_\varepsilon\big(\nabla f_\ell(x)\big)\mathbb{I}_{[u,\infty)}\big(f_\ell(x)\big)\,dx.$$

Similar arguments can be developed, expanding the integrand function into polynomials evaluated on the random vectors $\{\nabla^2 f_\ell(\cdot), \nabla f_\ell(\cdot), f_\ell(\cdot)\}$; algebraic simplifications occur and the expansions read as follows.

Theorem 3.1. *As $\ell \to \infty$, for $j = 0, 1, 2$*

$$\begin{aligned}&\mathcal{L}_j\big(A_u(f_\ell,\mathbb{S}^2)\big) - \mathbb{E}\big[\mathcal{L}_j\big(A_u(f_\ell;\mathbb{S}^2)\big)\big]\\ &\quad= -\frac{1}{2}\begin{bmatrix}2\\2-j\end{bmatrix} u\rho'_{2-j}(u)(\lambda_\ell/2)^{(2-j)/2}\int_{\mathbb{S}^2} H_2\big(f_\ell(x)\big)\,dx + R_{\ell;j},\end{aligned}\tag{3.6}$$

where

$$\mathbb{E}[R^2_{\ell;j}] = o_{\ell\to\infty}\big(\ell^{3-2j}\big);$$

as a consequence, one has also the variance asymptotics

$$\begin{aligned}\operatorname{Var}\big\{\mathcal{L}_j\big(A_u(f_\ell;\mathbb{S}^2)\big)\big\} &= \frac{1}{4}\left\{\begin{bmatrix}2\\2-j\end{bmatrix} u\rho'_{2-j}(u)\big(\lambda_\ell/2\big)^{(2-j)/2}\right\}^2\\ &\qquad\times\frac{32\pi^2}{2\ell+1} + o_{\ell\to\infty}(\lambda_\ell^{2-j-1}).\end{aligned}\tag{3.7}$$

Some features of the previous result are worth discussing.

- The asymptotic behaviour of all the Lipschitz–Killing curvatures is proportional to a sequence of scalar random variables $\{h_{\ell;2}\}_{\ell\in\mathbb{N}}$. As a consequence, these geometric functionals are fully correlated in the high-energy limit $\ell \to \infty$.
- For the same reasons, these functionals are also fully correlated, in the high-energy limit, when evaluated across different levels u_1, u_2: for the boundary length, this correlation phenomenon was first noted in [62].
- The leading terms all disappear in the "nodal" case $u = 0$, where the variances are hence an order of magnitude smaller. This is an instance of the so-called Berry cancellation phenomenon [61], to which we shall return in Section 4. We noted before that the leading terms are proportional to the centred random norm;

it is thus natural that these terms should disappear in the nodal case, which is independent of scaling factors. Note that for $j = 0$ the cancellation of the leading term occurs also at $u = 1$.

Remark. The proof of Theorem 3.1 was given in [11], in the case of the 2-dimensional sphere $\mathbb{S}^2$. However, we conjecture the result to hold as stated for spherical eigenfunctions in arbitrary dimension; see below for more details. Extensions have also been given to cover for instance the 2-dimensional torus (see [14]), for which a formula completely analogous to (3.1) holds.

Similar results can be shown to hold for other geometric functionals; let us consider for instance critical values, defined by

$$\mathcal{N}_u(f_\ell; \mathbb{S}^2) = \#\{x \in \mathbb{S}^2 : \nabla f_\ell(x) = 0 \text{ and } f_\ell(x) \geq u\}.$$

The asymptotic variance of $\{\mathcal{N}_u(f_\ell; \mathbb{S}^2)\}_{\ell=1,2,\dots}$ was established in [15, 16], and in particular we have

$$\begin{aligned}
\mathbb{E}\big[\mathcal{N}_u(f_\ell; \mathbb{S}^2)\big] &= \lambda_\ell g_1(u),\\
g_1(u) &= \frac{1}{\sqrt{2\pi}} \int_u^\infty \big(2e^{-t^2} + (t^2 - 1)e^{-t^2/2}\big)\, dt\\
&= u\phi(u) + \sqrt{2}\big(1 - \Phi(\sqrt{2}u)\big),\\
\operatorname{Var}\big[\mathcal{N}_u(f_\ell; \mathbb{S}^2)\big] &= \frac{1}{4}\lambda_\ell^2 g_2^2(u) \operatorname{Var}\left\{\int_{\mathbb{S}^2} H_2\big(f_\ell(x)\big)\, dx\right\} + o_{\ell\to\infty}(\ell^3)\\
&= \frac{1}{4}\lambda_\ell^2 g_2^2(u) \frac{2(4\pi)^2}{2\ell+1} + o_{\ell\to\infty}(\ell^3),
\end{aligned}$$

where

$$g_2(u) = \int_u^\infty \frac{1}{\sqrt{8\pi}} e^{-3t^2/2}\big(2 - 6t^2 - e^{-t^2}(1 - 4t + t^4)\big)\, dt.$$

Later in [12] it was shown that the critical values above the threshold level u satisfy the asymptotic

$$\begin{aligned}
&\mathcal{N}_u(f_\ell; \mathbb{S}^2) - \mathbb{E}\big[\mathcal{N}_u(f_\ell; \mathbb{S}^2)\big]\\
&\quad = \frac{1}{2}\lambda_\ell g_2(u) \int_{\mathbb{S}^2} H_2\big(f_\ell(x)\big)\, dx + o_p\Big(\sqrt{\operatorname{Var}\big[\mathcal{N}_u(f_\ell; \mathbb{S}^2)\big]}\Big),
\end{aligned}$$

As a consequence, one has also, for all $u \neq 0, 1$, the correlation result

$$\begin{aligned}
&\operatorname{Corr}^2\big\{\mathcal{N}_u(f_\ell; \mathbb{S}^2), \mathcal{L}_j\big(A_u(f_\ell; \mathbb{S}^2)\big)\big\}\\
&\quad := \frac{\operatorname{Cov}^2\big\{\mathcal{N}_u(f_\ell; \mathbb{S}^2), \mathcal{L}_j\big(A_u(f_\ell; \mathbb{S}^2)\big)\big\}}{\operatorname{Var}\big\{\mathcal{N}_u(f_\ell; \mathbb{S}^2)\big\} \operatorname{Var}\big\{\mathcal{L}_j\big(A_u(f_\ell; \mathbb{S}^2)\big)\big\}} \to 1, \quad \text{as } \ell \to \infty;
\end{aligned}$$

the value $u = 1$ has to be excluded only for $j = 0$. We also have that

$$\operatorname{Corr}^2\left\{\mathcal{N}_{u_1}(f_\ell;\mathbb{S}^2), \mathcal{N}_{u_2}(f_\ell;\mathbb{S}^2)\right\} \to 1, \quad \text{as } \ell \to \infty,$$

that is, asymptotically full correlation between the number of critical values above any two non-zero thresholds u_1, u_2.

As for the Lipschitz–Killing curvatures, a form of Berry's cancellation occurs at $u = 0$ and $u \to \pm\infty$; the total number of critical points has then a lower-order variance (see [16]), as we shall discuss in Section 4.

3.2. Quantitative central limit theorems

The results reviewed in Section 3.1 can be considered as following from a *reduction principle* (see [20]), where the limiting behaviour of $\{\mathcal{N}_u(f_\ell;\mathbb{S}^2), \mathcal{L}_j(A_u(f_\ell;\mathbb{S}^2))\}$ is dominated by a deterministic function of the threshold level u, times a sequence of random variables $\{h_{\ell;2}\}$ which do not depend on u. To derive the asymptotic law of these fluctuations, it is hence enough to investigate the convergence in distribution of $\{h_{\ell;2}\}$, as $\ell \to \infty$. In fact, it is possible to show a stronger result, namely a *quantitative central limit theorem*; to this aim, let us recall that the Wasserstein distance between two random variables X and Y is defined by

$$d_W(X,Y) := \sup_{h\in \operatorname{Lip}(1)} \left|\mathbb{E}h(X) - \mathbb{E}h(Y)\right|,$$

where Lip(1) denotes the class of Lipschitz functions of constant 1; i.e., $|h(x) - h(y)| \leq |x-y|$ for all $x, y \in \mathbb{R}$. $D_W(\cdot,\cdot)$ defines a metric on the space of probability distributions (for more details and other examples of probability metrics; see [46, Appendix C]). Taking $Z \sim N(0,1)$ to be a standard Gaussian random variable, a quantitative central limit theorem is defined as a result of the form

$$\lim_{n\to\infty} d_W\left(\frac{X_n - \mathbb{E}X_n}{\sqrt{\operatorname{Var}(X_n)}}, Z\right) = 0.$$

The field of quantitative central limit theorems has been very active in the last few decades; more recently, a breakthrough has been provided by the discovery of the so-called *Stein–Malliavin approach* by Nourdin, Peccati, and Nualart [45,46,48]. These results entail that for sequences of random variables belonging to a Wiener chaos, say $\mathcal{C}_q$, a quantitative central limit theorem for the Wasserstein distance can be given simply controlling the fourth-moment of X_n, as follows:

$$d_W\left(\frac{X_n - \mathbb{E}X_n}{\sqrt{\operatorname{Var}(X_n)}}, Z\right) \leq \sqrt{\frac{2q-2}{3\pi q}} \sqrt{\mathbb{E}\left[\left(\frac{X_n - \mathbb{E}X_n}{\sqrt{\operatorname{Var}(X_n)}}\right)^4\right] - 3}. \tag{3.8}$$

Similar results hold for other probability metrics, for instance the Kolmogorov and total variation distances; see again [46].

Quantitative central limit theorems lend themselves to an immediate application for the sequences $\{h_{\ell;q}\}$ that we introduced above. It should be noted indeed that by construction all these random variables belong to the qth-order Wiener chaos; it is then possible to exploit (3.8) to obtain quantitative central limit theorems for these polyspectra at arbitrary orders: their fourth moment can be computed by means of the diagram formula. These results were first given in [40] and then refined in [37], yielding the following.

Theorem 3.2. *As $\ell \to \infty$, one has*

$$d_W\left(\frac{h_{\ell;q}-\mathbb{E}[h_{\ell;q}]}{\sqrt{\operatorname{Var}(h_{\ell;q})}}, Z\right) = \begin{cases} O\left(\frac{1}{\sqrt{\ell}}\right) & \text{for } q = 2,3, \\ O\left(\frac{1}{\log \ell}\right) & \text{for } q = 4, \\ O(\ell^{-1/4}) & \text{for } q = 5,6,\dots . \end{cases}$$

Now, we have just shown that for nonzero thresholds $u \neq 0$ the Lipschitz–Killing curvatures and the critical values are indeed proportional to a term belonging to the second-order chaos, plus a remainder that it is asymptotically negligible. The following quantitative central limit theorem then follows immediately (see [11, 40, 50]).

Theorem 3.3. *As $\ell \to \infty$, for $u \neq 0$ ($j = 1, 2$) and for $u \neq 0, 1$ (for $j = 0$) one has that*

$$d_W\left(\frac{\mathcal{L}_j(A_u(f_\ell;\mathbb{S}^2)) - \mathbb{E}\left[\mathcal{L}_j(A_u(f_\ell;\mathbb{S}^2))\right]}{\sqrt{\operatorname{Var}\left(\mathcal{L}_j(A_u(f_\ell;\mathbb{S}^2))\right)}}, Z\right) = O(\ell^{-1/2}).$$

3.3. A higher-dimensional conjecture

The results we discussed so far have been limited to random-spherical harmonics on the 2-dimensional sphere $\mathbb{S}^2$. Research in progress suggests however that further generalizations should hold: to this aim, let us define the set of singular points $P_j := \{u \in \mathbb{R} : u\rho_j'(u) = 0\}$ (for instance, $P_0 = P_1 = \{0\}$, $P_2 = \{0,1\}$, $P_3 = \{0, \pm\sqrt{3}\}, \dots$). Let us now consider Gaussian random eigenfunctions on the higher-dimensional unit sphere $\mathbb{S}^d$; e.g.,

$$\Delta_{\mathbb{S}^d} f_{\ell;d} = -\lambda_{\ell;d} f_{\ell;d}, \quad \lambda_{\ell;d} := \ell(\ell+d-1);$$

these eigenfunctions are normalized so that (see [37, 51])

$$\mathbb{E}[f_{\ell;d}] = 0, \quad \mathbb{E}[f_{\ell;d}^2] = 1, \quad \mathbb{E}\left[f_{\ell;d}(x) f_{\ell;d}(y)\right] = G_{\ell;d/2}(\langle x, y\rangle),$$

where as before $G_{\ell;d/2}(\cdot)$ is the standardized ℓth Gegenbauer polynomial of order $\frac{d}{2}$

(normalized with $G_{\ell;d/2}(1) = 1$); it is convenient to recall that

$$G'_{\ell;d/2}(1) = \frac{\lambda_{\ell;d}}{d}.$$

We recall also that the dimension of the corresponding eigenspaces is

$$n_{\ell;d} = \frac{2\ell + d - 1}{\ell}\binom{\ell + d - 2}{\ell - 1} \sim \frac{2}{(d-1)!}\ell^{d-1}, \quad \text{as } \ell \to \infty.$$

By means of Parseval's equality we have also as a consequence

$$\begin{aligned}\operatorname{Var}\left[\int_{\mathbb{S}^d} H_2\big(f_{\ell;d}(x)\big)\,dx\right] &= \frac{2s_d^2}{n_{\ell;d}} = \frac{2(d+1)^2\omega_{d+1}^2}{n_{\ell;d}}\\ &\sim \frac{(d+1)^2\omega_{d+1}^2(d-1)!}{\ell^{d-1}} \quad \text{as } \ell \to \infty.\end{aligned}$$

We then propose the following.

Conjecture 3.4. *As $\ell \to \infty$, for all $k = 0, 1, \dots, d$ one has that*

$$\begin{aligned}&\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big) - \mathbb{E}\big[\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big)\big]\\ &\quad = -\frac{1}{2}\begin{bmatrix} d \\ k \end{bmatrix}\rho'_{d-k}(u)u\left(\frac{\lambda_{\ell;d}}{d}\right)^{(d-k)/2}\int_{\mathbb{S}^d} H_2\big(f_{\ell;d}(x)\big)\,dx\\ &\qquad + o_p\left(\sqrt{\ell^{d-2k+1}}\right).\end{aligned}$$

Remark. An immediate consequence of this conjecture would be

$$\frac{\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big) - \mathbb{E}\big[\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big)\big]}{\sqrt{\operatorname{Var}\big[\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big)\big]}} = \frac{h_{\ell;q}}{\sqrt{\operatorname{Var}\big[h_{\ell;d}(2)\big]}} + o_p(1),$$

$$h_{\ell;q} = \int_{\mathbb{S}^d} H_2\big(f_{\ell;d}(x)\big)\,dx.$$

Remark. The remainder term in Conjecture 3.4 is expected to be $O(\sqrt{\ell^{d-2k}})$, in the $L^2(\Omega)$ sense.

Three further consequences of Conjecture 3.4 would be the following.

- *(Variance asymptotics).* As $\ell \to \infty$, for all $k = 0, 1, \dots, d$ and for non-singular points $u \notin P_{d-k}$, one has

$$\begin{aligned}&\operatorname{Var}\big\{\mathcal{L}_k\big(A_u(f_\ell; \mathbb{S}^d)\big)\big\}\\ &\quad = \frac{H_{d-k}^2(u)\phi^2(u)u^2}{(2\pi d)^{(d-k)}}\frac{d!}{(d-k)!k!}\frac{\omega_d^2\omega_{d+1}^2}{\omega_k^2\omega_{d-k}^2}\frac{(d+1)^2\lambda_{\ell;d}^{d-k}}{2n_{\ell;d}} + o(\ell^{d-2k+1}).\end{aligned}$$

- *(Central limit theorem).* As $\ell \to \infty$, for all $k = 0, 1, \dots, d$ and for non-singular points $u \notin P_{d-k}$, one has

  $$d_W\left(\frac{\mathcal{L}_k\big(A_u(f_\ell;\mathbb{S}^d)\big) - \mathbb{E}\big[\mathcal{L}_k\big(A_u(f_\ell;\mathbb{S}^d)\big)\big]}{\sqrt{\operatorname{Var}\big[\mathcal{L}_k\big(A_u(f_\ell;\mathbb{S}^d)\big)\big]}}, Z\right) = o(1),$$

  where $Z \sim \mathcal{N}(0,1)$.
- *(Correlation asymptotics).* As $\ell \to \infty$, for all $k_1, k_2 = 0, 1, \dots, d$ and all u_1, u_2 such that $u_1 u_2 H_{d-k_1}(u_1) H_{d-k_2}(u_2) \neq 0$, one has

  $$\lim_{\ell\to\infty} \operatorname{Corr}^2\left(\mathcal{L}_{k_1}\big(A_u(f_\ell;\mathbb{S}^d)\big), \mathcal{L}_{k_2}\big(A_u(f_\ell;\mathbb{S}^d)\big)\right) = 1.$$

The driving rationale behind these conjectures is the *ansatz* that the asymptotic variance of the geometric functionals should be governed by fluctuations in the random $L^2(\mathbb{S}^d)$ norm of the eigenfunctions, for non-singular points $u \notin P_j$. In this sense, we believe the result has even greater applicability, for instance to cover combinations of random eigenfunctions defined on more general submanifolds of $\mathbb{R}^n$, such as *Berry's random waves* or "short windows" averages of isotropic random eigenfunctions on general manifolds (see [7, 8, 18, 21, 47, 64]). These issues are the object of currently ongoing research.

4. Nodal cases: Berry cancellation and the role of the fourth-order chaos

Section 4 has discussed the behaviour of geometric functionals for non-zero threshold levels $u \neq 0$; under isotropy, it has been shown that all these functionals are asymptotically proportional, in the $L^2(\Omega)$ sense, to a single random variable representing the (centred) random $L^2(\mathbb{S}^2)$-norm of the eigenfunction. This dominant term has been shown to disappear in the nodal case $u = 0$ (and, more generally, for $\rho'_{d-k}(u)u = 0$, i.e., for the singular points $u \in P_j$); the asymptotic behaviour must then be derived by a different route in these circumstances.

As mentioned above, the first paper to investigate the variance of the nodal length for random spherical harmonics was the seminal work by Igor Wigman [61], which made rigorous an *ansatz* by Michael Berry in the physical literature [8]. In particular, by using a higher-order version of the expectation metatheorem (see again [1, 3]) the following representation for the second moment of the nodal length can be given:

$$\mathbb{E}\left[\left\{\operatorname{Len}(f_\ell;\mathbb{S}^2)\right\}^2\right] = \int_{\mathbb{S}^2\times\mathbb{S}^2} \mathbb{E}\left[\left\|\{\nabla f_\ell(t_1)\}\right\|\left\|\{\nabla f_\ell(t_2)\}\right\| \mid f_\ell(t_1) = 0,\ f_\ell(t_2) = 0\right]$$
$$\times\, p_{f_\ell(t_1), f_\ell(t_2)}(0,0)\sigma_g(dt_1)\sigma_g(dt_2),$$

where as before we write $\mathrm{Len}(f_\ell;\mathbb{S}^2) = 2\mathcal{L}_1(A_0(f_\ell;\mathbb{S}^2))$ for the nodal length. The integrand in the previous formula is denoted by the 2-*point correlation function of the nodal length* and generalizes the Kac–Rice argument to second-order moments; analogous generalizations are possible for the other geometric functionals we considered and for higher-order moments as well (see [1]). By means of a challenging and careful expansion of this correlation function and a deep investigation of its behaviour for $\ell \to \infty$, Wigman was able to investigate the asymptotic for the variance of the nodal length and to show that (3.1) holds.

A natural question which was investigated shortly after this seminal paper was the possibility to derive the asymptotic variances of nodal statistics, and further characterizations such as the law of the asymptotic fluctuations, in terms of the Wiener chaos expansions that we discussed in Section 3. The first efforts were devoted to the analysis of the "nodal area" $\mathcal{L}_2(A_0(f_\ell;\mathbb{S}^2))$, for which it is easily shown that all even-order terms vanish at $u = 0$; from (3.4) we are then left with (see [42])

$$\mathrm{Var}\left\{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\right\} = \frac{1}{\ell^2}\sum_{q=1}^{\infty}\frac{c_{2q+1}}{2\pi q!}H_{2q}^2(0) + o(\ell^{-2}),$$

where

$$\begin{aligned} c_{2q+1} &= \lim_{\ell\to\infty} \ell^2 \int_0^\pi P_\ell^{2q+1}(\cos\theta)\sin\theta d\theta \\ &= \int_0^\infty J_0^{2q+1}(\psi)\psi d\psi, \quad J_0(\psi) := \sum_{k=0}^{\infty}\frac{(-1)^{k+1}(x/2)^{2k}}{(k!)^2}. \end{aligned}$$

The computation of the variance and the results in Theorem 3.2 lead easily also to a central limit theorem, which was given first in [40] and then extended to higher dimensions in [50].

Theorem 4.1 ([40]). *As $\ell \to \infty$, one has*

$$d_W\left(\frac{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big) - \mathbb{E}\big[\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\big]}{\sqrt{\mathrm{Var}\left\{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\right\}}}, Z\right) = o(1),$$

and hence

$$\frac{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big) - \mathbb{E}\big[\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\big]}{\sqrt{\mathrm{Var}\left\{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\right\}}} \to_d \mathcal{N}(0,1).$$

The proof of the previous result is standard; in short, the idea is to write

$$\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big) - \mathbb{E}\big[\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\big] = \sum_{k=1}^{M}\frac{(-1)^{2k+1}}{(2k+1)!}H_{2k}(u)\phi(u)h_{\ell;2k+1} + R_M,$$

where the remainder term is such that, as $M \to \infty$,

$$R_M = \sum_{k=M+1}^{\infty} \frac{(-1)^{2k+1}}{(2k+1)!} H_{2k}(u)\phi(u)h_{\ell;2k+1} = o_p\Big(\sqrt{\mathrm{Var}\left\{\mathcal{L}_2\big(A_0(f_\ell;\mathbb{S}^2)\big)\right\}}\Big).$$

It is then enough to show that the central limit theorem holds for M (sufficiently large but) finite; this can be achieved by an application of the multivariate fourth moment theorem to the terms $(h_{\ell;3},\dots,h_{\ell;2M+1})$ (see [46]). It should be noted that in the case of the defect the limiting behaviour depends on the full sequence $\{h_{\ell;2k+1}\}_{k=1,2,\dots}$; this is due to the exact disappearance of the two natural candidates to be leading terms, that is, $\{h_{\ell;2}\}$ and $\{h_{\ell;4}\}$, both whose coefficients vanish for $u=0$.

It is thus even more remarkable that for the nodal lines the situation simplifies drastically to yield the following result.

Theorem 4.2 ([39]). *As $\ell \to \infty$, one has*

$$\mathrm{Len}(f_\ell;\mathbb{S}^2) - \mathbb{E}\big[\mathrm{Len}(f_\ell;\mathbb{S}^2)\big] = -\frac{1}{4}\sqrt{\frac{\lambda_\ell}{2}}\frac{1}{4!}h_{\ell;4} + o_p\big(\sqrt{\mathrm{Var}\{h_{\ell;4}\}}\big), \tag{4.1}$$

and hence, in view of (3.2)

$$d_W\left(\frac{\mathrm{Len}(f_\ell;\mathbb{S}^2) - \mathbb{E}\big[\mathrm{Len}(f_\ell;\mathbb{S}^2)\big]}{\sqrt{\mathrm{Var}\left\{\mathrm{Len}(f_\ell;\mathbb{S}^2)\right\}}}, Z\right) = o(1).$$

The most notable aspect of Theorem 4.2 is that the limiting behaviour of nodal lines is asymptotically fully correlated with the sequence of random variables $\{h_{\ell;4}\}$, so that in principle it would be possible to "predict" nodal lengths by simply computing the integral of a fourth-order polynomial of the eigenfunctions over the sphere.

A natural question that arises is the structure of correlation among functionals evaluated at different thresholds and those considered for the nodal case $u=0$. Focussing for instance on the boundary length, it is immediate to understand that the latter, which is dominated by the second-order chaos term $\{h_{\ell;2}\}$ when $u \neq 0$, must be independent from the nodal length, which is asymptotically proportional to $\{h_{\ell;4}\}$. A more refined analysis, however, should take into account the fluctuations of the boundary length when the effects of the random norm $\|f_\ell\|_{L^2(\mathbb{S}^2)}$ is subtracted, that is, dropping the second-order chaos term from the Wiener expansion. This corresponds to the evaluation of the so-called partial correlation coefficients Corr*, for which it was shown in [38] that

$$\lim_{\ell\to\infty} \mathrm{Corr}^*\big(\mathrm{Len}(f_\ell;\mathbb{S}^2), \mathcal{L}_1\big(A_u(f_\ell;\mathbb{S}^2)\big)\big) = 1.$$

More explicitly, when compensating the effect of random norm fluctuations, the boundary length at any threshold $u \neq 0$ can be fully predicted on the basis of the

	$\mathcal{L}_j(u_1)$	$\mathcal{L}_j(u_2)$	Len(0)	$\text{Len}^*(u)$	$\mathcal{L}_2(0)$	$\mathcal{N}_u$	$\mathcal{N}_{-\infty}$
$\mathcal{L}_j(u_1)$	1	1	0	0	0	1	0
$\mathcal{L}_j(u_2)$	1	1	0	0	0	1	0
Len(0)	0	0	1	1	0	0	1
$\text{Len}^*(u)$	0	0	1	1	0	0	1
$\mathcal{L}_2(0)$	0	0	0	0	1	0	0
$\mathcal{N}_u$	1	1	0	0	0	1	0
$\mathcal{N}_{-\infty}$	0	0	1	1	0	0	1

Table 1. The limiting value of $\text{Corr}^2(\cdot,\cdot)$, as $\ell \to \infty$.

knowledge of the nodal length, up to a remainder term which is asymptotically negligible in the limit $\ell \to \infty$. It is interesting to note that a similar phenomenon occurs also for the total number of critical points, for which (building on earlier computations in [16]) it was shown in [13] that

$$\mathcal{N}_{-\infty}(f_\ell; \mathbb{S}^2) - \mathbb{E}\big[\mathcal{N}_{-\infty}(f_\ell; \mathbb{S}^2)\big] = -\frac{\lambda_\ell}{2^3 3^2 \sqrt{3}\pi} h_{\ell;4} + o_p(\ell^2 \log \ell);$$

as a consequence, the nodal length of random spherical harmonics and the number of their critical points are perfectly correlated in the high-energy limit:

$$\lim_{\ell\to\infty} \text{Corr}^2\left(\text{Len}(f_\ell; \mathbb{S}^2), \mathcal{N}_{-\infty}(f_\ell; \mathbb{S}^2)\right) = 1.$$

Let us now denote by $\text{Len}^*(u)$ the boundary length at level u after the fluctuations induced by the random norm have been subtracted (e.g., after removing its projection on the second-order chaos); moreover, for brevity's sake we write

$$\mathcal{L}_j\left(A_u(f_\ell; \mathbb{S}^2)\right) = \mathcal{L}_j(u), \quad j = 0, 1, 2,$$
$$\mathcal{N}_u(f_\ell; \mathbb{S}^2) = \mathcal{N}_u, \quad \text{Len}(f_\ell; \mathbb{S}^2) = \text{Len}(0),$$

so that $\mathcal{N}_{-\infty}$ is the total number of critical points and $\mathcal{L}_2(0)$ is the excursion area for $u = 0$. The correlation results that we discussed so far can be summarized in Table 1; here, we denote by $u_1, u_2 \neq 0, 1$ any two non-singular threshold values.

5. Eigenfunctions on different domains

For brevity and simplicity's sake, this survey has focussed only on the behaviour of random eigenfunctions on the sphere. Of course, as mentioned in Section 1, this is just a special case of a much broader research area, including for instance eigenfunctions on $\mathbb{R}^d$ and on the standard flat torus $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$. We do not even attempt to do justice to these developments, but it is important to mention some of them which are particularly close to the results we discussed for $\mathbb{S}^2$.

5.1. Eigenfunctions on the torus: arithmetic random waves

Eigenfunctions on the torus were first introduced in [52] and have then been studied by several other authors; see for instance [10, 23, 26, 33, 36, 53, 54] and the references therein. In dimension 2 these eigenfunctions (*arithmetic random waves*) are defined by the equations

$$\Delta_{\mathbb{T}^2} f_n + E_n f_n = 0, \quad E_n = 4\pi n, \quad n = a^2 + b^2,$$

for $a, b \in \mathbb{Z}$; the dimension of the nth eigenspace is $\mathcal{N}_n := \operatorname{Card}\{a, b \in \mathbb{Z} : a^2 + b^2 = n\}$, while the expected value of nodal lengths is [52]

$$\mathbb{E}\big[\operatorname{Len}(f_n; \mathbb{T}^2)\big] = \frac{\sqrt{E_n}}{2\sqrt{2}}.$$

A major breakthrough was then obtained with the derivation of the variance in [26]. In this paper, the authors introduce a probability measure on $\mathbb{S}^1$ defined by

$$\mu_n(\cdot) := \frac{1}{\mathcal{N}_n} \sum_{a,b:a^2+b^2=n} \delta_{(a,b)}(\cdot),$$

$\delta_{(a,b)}(\cdot)$ denoting the Dirac measure; its kth-order Fourier coefficients are defined by $\widehat{\mu}_n(k) := \int_{\mathbb{S}^1} \exp(ik\theta)\mu_n(d\theta)$. In [26] it is then shown that the variance of nodal lengths has a non-universal behaviour and is proportional to

$$\operatorname{Var}\big\{\operatorname{Len}(f_n; \mathbb{T}^2)\big\} = \frac{1 + \widehat{\mu}_n(4)^4}{512} \frac{E_n}{N_n^2} + o\left(\frac{E_n}{N_n^2}\right), \quad \text{as } n \to \infty \text{ s.t. } \mathcal{N}_n \to \infty.$$

It was later shown in [36] that the behaviour of $\operatorname{Len}(\mathbb{T}^2, f_n)$ is dominated by its fourth-order chaos component, similarly to what we observed above for random spherical harmonics (the result on the torus was actually established earlier than the corresponding case for the sphere). More precisely, we have that

$$\begin{aligned}
&\operatorname{Len}(f_n; \mathbb{T}^2) - \mathbb{E}\big[\operatorname{Len}(f_n; \mathbb{T}^2)\big] \\
&\quad = \sum_{q=2}^{\infty} \operatorname{Proj}\big[\operatorname{Len}(f_n; \mathbb{T}^2)|2q\big] \\
&\quad = \operatorname{Proj}\big[\operatorname{Len}(f_n; \mathbb{T}^2)|4\big] + o_p\left(\sqrt{\operatorname{Var}\big\{\operatorname{Len}(f_n; \mathbb{T}^2)\big\}}\right),
\end{aligned}$$

where $\mathrm{Proj}[\cdot|q]$ denotes projection on the qth-order chaos. On the contrary of what we observed for the case of the sphere, here it is not possible to express the fourth-order chaos as a polynomial functional of the random eigenfunctions $\{f_n\}$ alone. Moreover, the limiting distribution is non-Gaussian and non-universal; i.e., it depends on the asymptotic behaviour of $\lim_{j\to\infty} \widehat{\mu}_{n_j}(4)$ which varies along different subsequences $\{n_j\}_{j=1,2,\ldots}$ (the attainable measures for the weak convergence of the sequences $\{\mu_{n_j}(\cdot)\}_{n\in\mathbb{N}}$ have been investigated in [26, 27]). Further results in this area include [10, 44] for arithmetic random waves in higher dimension and [28] for the excursion area on subdomains of $\mathbb{T}^2$; as mentioned earlier, an extension of Theorem 3.1 to the torus has been given in [14]. It should be noted that arithmetic random waves can be viewed as an instance of random trigonometric polynomials, whose zeroes have been studied, among others, in [2, 4].

5.2. The Euclidean case: Berry's random waves

Spherical harmonics on the sphere $\mathbb{S}^2$ are known to exhibit a scaling limit; i.e., after a change of coordinates they converge locally to a Gaussian random process on $\mathbb{R}^2$ which is isotropic, zero mean, and has covariance function

$$\mathbb{E}\big[f(x)f(y)\big] = J_0\big(2\pi\|x-y\|\big), \quad x, y \in \mathbb{R}^2,\ J_0(z) := \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k}}{(k!)^2 2^{2k}};$$

here $J_0(\cdot)$ corresponds to the standard Bessel functions, for which the following scaling asymptotics hold:

$$P_\ell\Big(\cos\frac{\psi}{\ell}\Big) \to_{\ell\to\infty} J_0(\psi), \quad \psi \in \mathbb{R}.$$

The behaviour of nodal lines $\mathcal{L}_E(f) = \{x \in \mathbb{R}^2 : f(x) = 0, \|x\| < 2\pi\sqrt{E}\}$ can then be studied in the asymptotic regime $E \to \infty$; this is indeed the physical setting under which Berry first investigated cancellation phenomena in his pioneering paper [8]. The topology of nodal sets for Berry's random waves was studied in [17, 43, 55] and others. Concerning nodal lengths, a (quantitative) central limit theorem was established in [47], where intersections of independent random waves were also investigated; more recently, [60] proved a result analogous to Theorem 4.2, namely that, as $E \to \infty$,

$$\begin{aligned} &\mathcal{L}_E(f) - \mathbb{E}\big[\mathcal{L}_E(f)\big] \\ &\quad = -\frac{1}{4}\frac{2\pi}{4!}\sqrt{\frac{E}{2}} \int_{\|x\|<2\pi\sqrt{E}} H_4\big(f(x)\big)\,dx + o_p\Big(\sqrt{\mathrm{Var}\left\{\mathcal{L}_E(f)\right\}}\Big). \end{aligned} \tag{5.1}$$

We expect that results analogous to (4.1) and (5.1) will hold for more general Riemannian waves on 2-dimensional manifolds [64]; extensions to random waves in $\mathbb{R}^3$

have been studied, among others, in [18], but in these higher-dimensional settings it is no longer the case that nodal volumes are dominated by a single chaotic component.

5.3. Shrinking domains

As a final issue, we recall how some of the previous results can be extended to shrinking subdomains of the torus and of the sphere. In this respect, a surprising result was derived in [5] concerning the asymptotic behaviour of the nodal length on a suitably shrinking subdomain $B_n \subset \mathbb{T}^2$; indeed it was shown that, for density one subsequences in n,

$$\lim_{n\to\infty} \operatorname{Corr}\big(\operatorname{Len}(\mathbb{T}^2, f_n), \operatorname{Len}(\mathbb{T}^2 \cap B_n, f_n)\big) = 1,$$

entailing that the behaviour of the nodal length on the whole torus is fully determined by its behaviour on any shrinking disk B_n, provided the radius of this disk is not smaller than $n^{-1/2+\varepsilon}$, some $\varepsilon > 0$. Of course, the asymptotic variance and distributions of the nodal length in this shrinking domain are then immediately shown to be the same as those for the full torus, up to a normalizing factor. Interestingly, the same phenomenon does not occur on the sphere, where on the contrary it was shown in [59] that

$$\lim_{\ell\to\infty} \operatorname{Corr}\big(\operatorname{Len}(\mathbb{S}^2, f_\ell), \operatorname{Len}(\mathbb{S}^2 \cap B_\ell, f_\ell)\big) = 0,$$

so that the nodal length when evaluated on a shrinking subset B_ℓ of the 2-dimensional sphere is actually asymptotically independent from its global value; in the same paper, it is indeed shown that (4.1) generalizes to

$$\begin{aligned} &\operatorname{Len}(\mathbb{S}^2 \cap B_\ell, f_\ell) - \mathbb{E}\big[\operatorname{Len}(\mathbb{S}^2 \cap B_\ell, f_\ell)\big] \\ &\qquad = -\frac{1}{4}\sqrt{\frac{\lambda_\ell}{2}}\frac{1}{4!} h_{\ell;4}(B_\ell) + o_p\Big(\sqrt{\operatorname{Var}\{h_{\ell;4}(B_\ell)\}}\Big), \qquad (5.2) \\ &h_{\ell;4}(B_\ell) = \int_{B_\ell} H_4\big(f_\ell(x)\big)\,dx; \end{aligned}$$

from this characterization, a central limit theorem follows easily along the same lines that we discussed in Section 4; see [59] for more details and discussion.

Acknowledgments. I am grateful to Valentina Cammarota, Maurizia Rossi, Anna Paola Todino, and an anonymous referee for a number of comments and suggestions on an earlier draft.

Funding. Research supported in part by the MIUR Departments of Excellence Program *Math@Tov*.

References

[1] R. J. Adler and J. E. Taylor, *Random Fields and Geometry*. Springer Monogr. Math., Springer, New York, 2007 Zbl 1149.60003 MR 2319516

[2] J. Angst, V.-H. Pham, and G. Poly, Universality of the nodal length of bivariate random trigonometric polynomials. *Trans. Amer. Math. Soc.* **370** (2018), no. 12, 8331–8357 Zbl 1407.42001 MR 3864378

[3] J.-M. Azaïs and M. Wschebor, *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons, Hoboken, NJ, 2009 Zbl 1168.60002 MR 2478201

[4] V. Bally, L. Caramellino, and G. Poly, Non universality for the variance of the number of real roots of random trigonometric polynomials. *Probab. Theory Related Fields* **174** (2019), no. 3-4, 887–927 Zbl 07081459 MR 3980307

[5] J. Benatar, D. Marinucci, and I. Wigman, Planck-scale distribution of nodal length of arithmetic random waves. *J. Anal. Math.* **141** (2020), no. 2, 707–749 Zbl 1458.81020 MR 4179775

[6] P. Bérard, Volume des ensembles nodaux des fonctions propres du Laplacien. In *Bony-Sjöstrand-Meyer Seminar, 1984–1985*, p. Exp. No. 14, École Polytech., Palaiseau, 1985 Zbl 0589.58033 MR 819780

[7] M. V. Berry, Regular and irregular semiclassical wavefunctions. *J. Phys. A* **10** (1977), no. 12, 2083–2091 Zbl 0377.70014 MR 489542

[8] M. V. Berry, Statistics of nodal lines and points in chaotic quantum billiards: perimeter corrections, fluctuations, curvature. *J. Phys. A* **35** (2002), no. 13, 3025–3038 Zbl 1044.81047 MR 1913853

[9] J. Buckley and I. Wigman, On the number of nodal domains of toral eigenfunctions. *Ann. Henri Poincaré* **17** (2016), no. 11, 3027–3062 Zbl 1361.35051 MR 3556515

[10] V. Cammarota, Nodal area distribution for arithmetic random waves. *Trans. Amer. Math. Soc.* **372** (2019), no. 5, 3539–3564 Zbl 07089869 MR 3988618

[11] V. Cammarota and D. Marinucci, A quantitative central limit theorem for the Euler-Poincaré characteristic of random spherical eigenfunctions. *Ann. Probab.* **46** (2018), no. 6, 3188–3228 Zbl 1428.60067 MR 3857854

[12] V. Cammarota and D. Marinucci, A reduction principle for the critical values of random spherical harmonics. *Stochastic Process. Appl.* **130** (2020), no. 4, 2433–2470 Zbl 1457.60071 MR 4074706

[13] V. Cammarota and D. Marinucci, A reduction principle for the critical values of random spherical harmonics. *Stochastic Process. Appl.* **130** (2020), no. 4, 2433–2470 Zbl 1457.60071 MR 4074706

[14] V. Cammarota, D. Marinucci, and M. Rossi, Lipschitz-Killing curvatures for arithmetic random waves. 2020, arXiv:2010.14165

[15] V. Cammarota, D. Marinucci, and I. Wigman, On the distribution of the critical values of random spherical harmonics. *J. Geom. Anal.* **26** (2016), no. 4, 3252–3324 Zbl 1353.60020 MR 3544960

[16] V. Cammarota and I. Wigman, Fluctuations of the total number of critical points of random spherical harmonics. *Stochastic Process. Appl.* **127** (2017), no. 12, 3825–3869 Zbl 1377.60060 MR 3718098

[17] Y. Canzani and B. Hanin, Local universality for zeros and critical points of monochromatic random waves. *Comm. Math. Phys.* **378** (2020), no. 3, 1677–1712 Zbl 07250102 MR 4150887

[18] F. Dalmao, A. Estrade, and J. R. León, On 3-dimensional Berry's model. *ALEA Lat. Am. J. Probab. Math. Stat.* **18** (2021), no. 1, 379–399 Zbl 1465.60022 MR 4213863

[19] F. Dalmao, I. Nourdin, G. Peccati, and M. Rossi, Phase singularities in complex arithmetic random waves. *Electron. J. Probab.* **24** (2019), Paper No. 71 Zbl 1467.60034 MR 3978221

[20] H. Dehling and M. S. Taqqu, The empirical process of some long-range dependent sequences with an application to U-statistics. *Ann. Statist.* **17** (1989), no. 4, 1767–1783 Zbl 0696.60032 MR 1026312

[21] G. Dierickx, I. Nourdin, G. Peccati, and M. Rossi, Small scale CLTs for the nodal length of monochromatic waves. 2020, arXiv:2005.06577

[22] H. Donnelly and C. Fefferman, Nodal sets of eigenfunctions on Riemannian manifolds. *Invent. Math.* **93** (1988), no. 1, 161–183 Zbl 0659.58047 MR 943927

[23] A. Granville and I. Wigman, Planck-scale mass equidistribution of toral Laplace eigenfunctions. *Comm. Math. Phys.* **355** (2017), no. 2, 767–802 Zbl 1376.58012 MR 3681390

[24] S. Janson, *Gaussian Hilbert Spaces*. Cambridge Tracts in Math. 129, Cambridge University Press, Cambridge, 1997 Zbl 0887.60009 MR 1474726

[25] M. Kac, On the distribution of values of trigonometric sums with linearly independent frequencies. *Amer. J. Math.* **65** (1943), 609–615 Zbl 0061.13710 MR 9061

[26] M. Krishnapur, P. Kurlberg, and I. Wigman, Nodal length fluctuations for arithmetic random waves. *Ann. of Math. (2)* **177** (2013), no. 2, 699–737 Zbl 1314.60101 MR 3010810

[27] P. Kurlberg and I. Wigman, On probability measures arising from lattice points on circles. *Math. Ann.* **367** (2017), no. 3-4, 1057–1098 Zbl 1381.11055 MR 3623219

[28] P. Kurlberg, I. Wigman, and N. Yesha, The defect of toral Laplace eigenfunctions and arithmetic random waves. *Nonlinearity* **34** (2021), no. 9, 6651–6684 Zbl 1471.58012 MR 4304493

[29] A. Logunov, Nodal sets of Laplace eigenfunctions: polynomial upper estimates of the Hausdorff measure. *Ann. of Math. (2)* **187** (2018), no. 1, 221–239 Zbl 1384.58020 MR 3739231

[30] A. Logunov, Nodal sets of Laplace eigenfunctions: proof of Nadirashvili's conjecture and of the lower bound in Yau's conjecture. *Ann. of Math. (2)* **187** (2018), no. 1, 241–262 Zbl 1384.58021 MR 3739232

[31] A. Logunov and E. Malinnikova, Nodal sets of Laplace eigenfunctions: estimates of the Hausdorff measure in dimensions two and three. In *50 Years with Hardy Spaces*, pp. 333–344, Oper. Theory Adv. Appl. 261, Birkhäuser/Springer, Cham, 2018 Zbl 1414.31004 MR 3792104

[32] A. Logunov, E. Malinnikova, N. Nadirashvili, and F. Nazarov, The sharp upper bound for the area of the nodal sets of Dirichlet Laplace eigenfunctions. *Geom. Funct. Anal.* **31** (2021), no. 5, 1219–1244 MR 4356702

[33] R. W. Maffucci, Nodal intersections for arithmetic random waves against a surface. *Ann. Henri Poincaré* **20** (2019), no. 11, 3651–3691 Zbl 1441.58009 MR 4019200

[34] D. Marinucci, A central limit theorem and higher order results for the angular bispectrum. *Probab. Theory Related Fields* **141** (2008), no. 3-4, 389–409 Zbl 1141.60028 MR 2391159

[35] D. Marinucci and G. Peccati, *Random Fields on the Sphere. Representation, Limit Theorems and Cosmological Applications*. London Math. Soc. Lecture Note Ser. 389, Cambridge University Press, Cambridge, 2011 Zbl 1260.60004 MR 2840154

[36] D. Marinucci, G. Peccati, M. Rossi, and I. Wigman, Non-universality of nodal length distribution for arithmetic random waves. *Geom. Funct. Anal.* **26** (2016), no. 3, 926–960 Zbl 1347.60013 MR 3540457

[37] D. Marinucci and M. Rossi, Stein-Malliavin approximations for nonlinear functionals of random eigenfunctions on $\mathbb{S}^d$. *J. Funct. Anal.* **268** (2015), no. 8, 2379–2420 Zbl 1333.60033 MR 3318653

[38] D. Marinucci and M. Rossi, On the correlation between nodal and nonzero level sets for random spherical harmonics. *Ann. Henri Poincaré* **22** (2021), no. 1, 275–307 Zbl 1469.60165 MR 4201595

[39] D. Marinucci, M. Rossi, and I. Wigman, The asymptotic equivalence of the sample trispectrum and the nodal length for random spherical harmonics. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** (2020), no. 1, 374–390 Zbl 1465.60044 MR 4058991

[40] D. Marinucci and I. Wigman, The defect variance of random spherical harmonics. *J. Phys. A* **44** (2011), no. 35, 16 Zbl 1232.60039

[41] D. Marinucci and I. Wigman, On the area of excursion sets of spherical Gaussian eigenfunctions. *J. Math. Phys.* **52** (2011), no. 9, 093301, 21 Zbl 1272.82017 MR 2867816

[42] D. Marinucci and I. Wigman, On nonlinear functionals of random spherical eigenfunctions. *Comm. Math. Phys.* **327** (2014), no. 3, 849–872 Zbl 1322.60030 MR 3192051

[43] F. Nazarov and M. Sodin, On the number of nodal domains of random spherical harmonics. *Amer. J. Math.* **131** (2009), no. 5, 1337–1357 Zbl 1186.60022 MR 2555843

[44] M. Notarnicola, Fluctuations of nodal sets on the 3-torus and general cancellation phenomena. *ALEA Lat. Am. J. Probab. Math. Stat.* **18** (2021), no. 2, 1127–1194 Zbl 1468.60062 MR 4282185

[45] I. Nourdin and G. Peccati, Stein's method on Wiener chaos. *Probab. Theory Related Fields* **145** (2009), no. 1-2, 75–118 Zbl 1175.60053 MR 2520122

[46] I. Nourdin and G. Peccati, *Normal Approximations with Malliavin Calculus. From Stein's Method to Universality*. Cambridge Tracts in Math. 192, Cambridge University Press, Cambridge, 2012 Zbl 1266.60001 MR 2962301

[47] I. Nourdin, G. Peccati, and M. Rossi, Nodal statistics of planar random waves. *Comm. Math. Phys.* **369** (2019), no. 1, 99–151 Zbl 1431.60025 MR 3959555

[48] D. Nualart and G. Peccati, Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* **33** (2005), no. 1, 177–193 Zbl 1097.60007 MR 2118863

[49] S. O. Rice, Mathematical analysis of random noise. *Bell System Tech. J.* **24** (1945), 46–156 Zbl 0063.06487 MR 11918

[50] M. Rossi, The defect of random hyperspherical harmonics. *J. Theoret. Probab.* **32** (2019), no. 4, 2135–2165 Zbl 07120214 MR 4020703

[51] M. Rossi, Random nodal lengths and Wiener chaos. In *Probabilistic Methods in Geometry, Topology and Spectral Theory*, pp. 155–169, Contemp. Math. 739, Amer. Math. Soc., Providence, RI, 2019 Zbl 1458.60058 MR 4033918

[52] Z. Rudnick and I. Wigman, On the volume of nodal sets for eigenfunctions of the Laplacian on the torus. *Ann. Henri Poincaré* **9** (2008), no. 1, 109–130 Zbl 1142.60029 MR 2389892

[53] Z. Rudnick and I. Wigman, Nodal intersections for random eigenfunctions on the torus. *Amer. J. Math.* **138** (2016), no. 6, 1605–1644 Zbl 1373.58017 MR 3595496

[54] Z. Rudnick, I. Wigman, and N. Yesha, Nodal intersections for random waves on the 3-dimensional torus. *Ann. Inst. Fourier (Grenoble)* **66** (2016), no. 6, 2455–2484 Zbl 1360.60081 MR 3580177

[55] P. Sarnak and I. Wigman, Topologies of nodal sets of random band-limited functions. *Comm. Pure Appl. Math.* **72** (2019), no. 2, 275–342 Zbl 1414.58019 MR 3896022

[56] M. Stecconi, Kac-Rice formula for transverse intersections. *Anal. Math. Phys.* **12** (2022), no. 2, Paper No. 44 Zbl 07493058 MR 4386457

[57] G. Szegő, *Orthogonal Polynomials*. 4th edn., Amer. Math. Soc. Colloq. Publ. 23, Amer. Math. Soc., Providence, RI, 1975 Zbl 0305.42011 MR 0372517

[58] J. E. Taylor, A Gaussian kinematic formula. *Ann. Probab.* **34** (2006), no. 1, 122–158 Zbl 1094.60025 MR 2206344

[59] A. P. Todino, Nodal lengths in shrinking domains for random eigenfunctions on S^2. *Bernoulli* **26** (2020), no. 4, 3081–3110 Zbl 07256169 MR 4140538

[60] A. Vidotto, A note on the reduction principle for the nodal length of planar random waves. *Statist. Probab. Lett.* **174** (2021), Paper No. 109090 Zbl 07425242 MR 4237481

[61] I. Wigman, Fluctuations of the nodal length of random spherical harmonics. *Comm. Math. Phys.* **298** (2010), no. 3, 787–831 Zbl 1213.33019 MR 2670928

[62] I. Wigman, On the nodal lines of random and deterministic Laplace eigenfunctions. In *Spectral Geometry*, pp. 285–297, Proc. Sympos. Pure Math. 84, Amer. Math. Soc., Providence, RI, 2012 Zbl 1317.60013 MR 2985322

[63] S.-T. Yau, Open problems in geometry. In *Differential Geometry: Partial Differential Equations on Manifolds (Los Angeles, CA, 1990)*, pp. 1–28, Proc. Sympos. Pure Math. 54, Amer. Math. Soc., Providence, RI, 1993 Zbl 0801.53001 MR 1216573

[64] S. Zelditch, Real and complex zeros of Riemannian random waves. In *Spectral Analysis in Geometry and Number Theory*, pp. 321–342, Contemp. Math. 484, Amer. Math. Soc., Providence, RI, 2009 Zbl 1176.58021 MR 1500155

Domenico Marinucci

Department of Mathematics, University of Rome Tor Vergata, Via della Ricerca Scientifica 1,
00133 Rome, Italy; marinucc@mat.uniroma2.it

DOI 10.4171/8ECM/31

Looking at Euler flows through a contact mirror: Universality and undecidability

Robert Cardona, Eva Miranda, and Daniel Peralta-Salas

Abstract. The dynamics of an inviscid and incompressible fluid flow on a Riemannian manifold is governed by the Euler equations. In recent papers by Cardona, Miranda, and Peralta-Salas, several unknown facets of the Euler flows have been discovered, including universality properties of the stationary solutions to the Euler equations. The study of these universality features was suggested by Tao (2019) as a novel way to address the problem of global existence for Euler and Navier–Stokes. Universality of the Euler equations was proved by Cardona et al. (2019) for stationary solutions using a contact mirror which reflects a Beltrami flow as a Reeb vector field. This contact mirror permits the use of advanced geometric techniques in fluid dynamics. On the other hand, motivated by Tao's approach relating Turing machines to Navier–Stokes equations, a Turing complete stationary Euler solution on a Riemannian 3-dimensional sphere was constructed by Cardona et al. (2021). Since the Turing completeness of a vector field can be characterized in terms of the halting problem, which is known to be undecidable (as shown by Turing (1936)), a striking consequence of this fact is that a Turing complete Euler flow exhibits undecidable particle paths (as shown by Cardona et al. (2021)). In this article, we give a panoramic overview of this fascinating subject, and go one step further in investigating the undecidability of different dynamical properties of Turing complete flows. In particular, we show that variations of the work of Cardona et al. (2021) allow us to construct a stationary Euler flow of Beltrami type (and, via the contact mirror, a Reeb vector field) for which it is undecidable to determine whether its orbits through an explicit set of points are periodic.

1. Introduction

Back in 1936, Turing faced a fundamental question which had been driving the attention of many mathematicians since the 1920s: *Is there an answer for the decision problem for first-order logics?* A decision problem can be posed as a *yes/no* question depending on the input values. *Decidability* is the problem of the existence of an effective method, a test or automatic procedure to know whether certain premises entail certain conclusions. The halting problem is one of the first decision problems

2020 Mathematics Subject Classification. Primary 35Q31; Secondary 37J06, 03D78, 57R17.
Keywords. Euler equations, Reeb flows, Turing completeness, universality.

which was proved to be undecidable. Indeed, Alan Turing [32] proved that a general algorithm that solves the halting problem cannot exist (for all possible program-input pairs). In doing so, he, fortuitously, invented the basic model of modern digital computers, the so-called Turing machine.

The undecidability of the halting problem yields a cascade of related questions: *What kind of physics might be non-computational?* (Penrose [21]) *Is hydrodynamics capable of performing computations?* (Moore [19]). Given the Hamiltonian of a quantum many-body system, does there exist an algorithm to check whether it has a spectral gap? (this is known as *the spectral gap problem*, recently proved to be undecidable [10]). And last but not least, *can a mechanical system (including a fluid flow) simulate a universal Turing machine*? (Tao [27, 28, 30]).

Surprisingly, this last question is connected with the regularity of the Navier–Stokes equations [26], one of the unsolved problems in Clay's list of problems for the Millennium. In [29], Tao speculated on a relation between a potential blow-up of the Navier–Stokes equations, Turing completeness, and fluid computation. This is part of a more general program he launched in [26, 27, 29] to address the global existence problem for Euler and Navier–Stokes based on the concept of *universality*. Inspired by this proposal, in [8] we showed that the stationary Euler equations exhibit several universality features, in the sense that, any non-autonomous flow on a compact manifold can be extended to a smooth stationary solution of the Euler equations on a Riemannian manifold of possibly higher dimension. As a corollary, we established the Turing completeness of the steady Euler flows on a 17-dimensional sphere [8]. It is then natural to ask: Can this dimensional bound be improved?

We solved this problem affirmatively in [9] constructing stationary solutions of the Euler equations on a Riemannian 3-dimensional sphere that can simulate any Turing machine (i.e., they are Turing complete). In particular, these solutions exhibit undecidable paths in the sense that there are constructible points for which it is not possible to decide whether their associated trajectories will intersect a certain (explicit) open set or not. The type of flows that we considered are Beltrami fields, a particularly relevant class of stationary solutions. Our game plan combines the computational power of symbolic dynamics with techniques from contact topology. Contact topology enters into the scene because Beltrami fields correspond to Reeb flows under a contact mirror unveiled by Sullivan, Etnyre, and Ghrist more than two decades ago. The contact mirror thus reflects a problem in Fluid Dynamics as a problem in contact geometry and back.

The existence of Turing complete Euler flows gives rise to new questions concerning undecidability of different dynamical properties. One of the potential problems to consider is that of periodic orbits: ever, at least since the work of Poincaré [24], periodic orbits are known to be one of the major tools to understand the dynamics of Hamiltonian systems. Even though not every Hamiltonian system admits periodic orbits, the Weinstein conjecture asserts that under some topological (compact) and geometrical (contact) conditions on the manifold, Reeb vector fields admit at least one periodic orbit. The Weinstein conjecture is known to be true in dimension 3, so using our contact mirror we can conclude that the Turing complete Reeb flow we constructed in [9] has at least one periodic orbit (in fact, in our construction the Reeb vector field coincides with a Hopf field in the complement of a certain solid torus, so it has infinitely many periodic orbits). It is then natural to ask if for every point of the sphere it is possible to decide whether its corresponding orbit will be closed or not. We shall see in this article that such a decision problem has no answer. The undecidability of other dynamical properties of Reeb flows will be also discussed. In view of Gödel's incompleteness theorems, undecidability of such properties of dynamical systems seems to be an unsurmountable obstacle no matter what systems of axioms are considered.

Our goal in this article is to give an overview of this exciting area of research. Let us summarize the contents of this work. Next, in this introduction, we present the Euler equations and the Beltrami fields on Riemannian manifolds, in Section 1.1, and the connection between contact geometry and hydrodynamics (in particular, between Beltrami fields and Reeb flows), in Section 1.2. In Section 2, following [8], we introduce the theory of Reeb embeddings and their flexibility (in the form of a new h-principle), and apply it to prove several universality features of the stationary Euler flows in high dimensions. The construction of a Turing complete Reeb field on a 3-dimensional sphere [9] is presented in Section 3; as a novel feature, we show how variations of this result allow us to prove the existence of Reeb fields exhibiting different undecidable dynamical properties, including periodic orbits. Finally, in Section 4 we recall the main theorem of [7] establishing the existence of Turing complete time-dependent solutions to the Euler equations (on compact Riemannian manifolds of very high dimension), and discuss the implications of our results regarding computability with the Navier–Stokes equations.

1.1. The Euler equations on Riemannian manifolds

The Euler equations describe the dynamics of an incompressible fluid flow without viscosity. Even if they are classically considered on $\mathbb{R}^3$, they can be formulated on any n-dimensional Riemannian manifold (M, g), $n \geq 2$ (for an introduction to the geometric aspects of hydrodynamics see [2, 22]). The equations can be written as

$$\begin{cases} \frac{\partial}{\partial t} X + \nabla_X X = -\nabla p, \\ \operatorname{div} X = 0, \end{cases}$$

where p stands for the hydrodynamic pressure and X is the velocity field of the fluid (a non-autonomous vector field on M). Here $\nabla_X X$ denotes the covariant derivative of X along X. A solution to the Euler equations is called stationary whenever X does not depend on time, i.e., $\frac{\partial}{\partial t}X = 0$, and it models a fluid flow in equilibrium.

This extension of the Euler equations to high dimensional manifolds turns out to be very useful to show that the steady and time-dependent Euler flows exhibit remarkable dynamical [8] (see also [28, 30, 31]), computational [7] or topological [5] universality features. For non-specialists, we refer to [18] for an introduction to differential geometry.

A short comprehensive dictionary.

- A volume-preserving (autonomous) vector field X on M is Eulerisable [23] if there exists a Riemannian metric g on M compatible with the volume form, such that X satisfies the stationary Euler equations on (M, g):

$$\nabla_X X = -\nabla p, \quad \operatorname{div} X = 0 \tag{1.1}$$

  for some pressure function p.

- A divergence-free vector field X on an odd-dimensional manifold (M, g) of dimension $n = 2m + 1$ is Beltrami if

$$\operatorname{curl} X = fX,$$

  for some factor $f \in C^\infty(M)$. The curl of X is defined as the unique vector field $Y = \operatorname{curl} X$ that satisfies the equation

$$\iota_Y \mu = (dX^\flat)^m, \tag{1.2}$$

  where μ is the Riemannian volume form, the symbol $^\flat$ stands for the musical isomorphism associated to the metric g, and $\iota_Y \mu$ denotes the contraction of μ with Y. The classical Hopf fields on the round sphere $\mathbb{S}^{2m+1}$ and the ABC flows on the flat 3-torus $\mathbb{T}^3$ are examples of Beltrami fields.

1.2. Contact hydrodynamics

Let M^{2m+1} be an odd-dimensional manifold equipped with a hyperplane distribution ξ. Assume that there is a globally defined non-vanishing one-form $\alpha \in \Omega^1(M)$ with $\ker \alpha = \xi$ and satisfying $\alpha \wedge (d\alpha)^m > 0$ everywhere; i.e., it defines a volume form in M. Then we say that (M^{2m+1}, ξ) is a (cooriented) contact manifold.

The one-form α is called a contact form. Of course, the contact structure ξ does not depend on a particular choice of the defining contact one-form α, any other one-form $h \cdot \alpha$ with h a positive function in M is a contact form defining ξ as well. The contact condition $\alpha \wedge (d\alpha)^m > 0$ implies that $d\alpha$ induces a fiber-wise symplectic structure on the hyperplane distribution ξ (of even dimension $2m$). The unique Reeb vector field R associated to a given contact form α is uniquely determined by the equations

$$\iota_R\alpha = 1, \quad \iota_R d\alpha = 0. \tag{1.3}$$

These equations imply that the flow of R preserves the contact form, so, in particular, it preserves $\alpha \wedge d\alpha$ and hence R is a volume-preserving vector field. In contrast with the hyperplane distribution, the Reeb field can display drastically different dynamics depending on the particular choice of contact form.

We will now explain the connection between contact geometry and hydrodynamics. In order to understand this remarkable correspondence, it is convenient to rewrite the Euler equations in a *dual language*. Duality is given by contraction with the Riemannian metric g. With the one-form α defined as $\alpha := X^\flat$ and the Bernoulli function as $B := p + \frac{1}{2}g(X,X)$, the steady Euler equations can be equivalently formulated as

$$\begin{cases} \iota_X d\alpha = -dB, \\ d\iota_X\mu = 0, \end{cases}$$

where μ is the Riemannian volume form.

Observe that the following hold.

- The equation $\operatorname{curl} X = fX$, with $f \in C^\infty(M)$, satisfied by a Beltrami vector field on an odd-dimensional manifold, can be equivalently written as $(d\alpha)^m = f\iota_X\mu$. This follows from equation (1.2), that determines the curl of X, and the fact that $\alpha = X^\flat$. Assume that X is rotational, i.e., $f > 0$, then if X does not vanish on M we infer that
$$\alpha \wedge (d\alpha)^m = f\alpha \wedge \iota_X\mu > 0,$$
thus proving that α defines a contact structure on M.
- Obviously, X satisfies $\iota_X(d\alpha)^m = f\iota_X\iota_X\mu = 0$. Therefore, since $\alpha \wedge (d\alpha)^m > 0$, it is easy to conclude that $X \in \ker d\alpha$, and hence it is a reparametrization of the Reeb vector field R by the function $\alpha(X) = g(X,X)$. Indeed, the vector field $R = \frac{X}{\alpha(X)}$ satisfies equations (1.3).

These observations prove one of the implications of the following theorem, which is due to Etnyre and Ghrist [12].

Theorem 1.1. *Let M be a Riemannian odd-dimensional manifold. Any smooth, nonsingular rotational Beltrami field on M is a Reeb-like field for some contact form on M. Conversely, given a contact form α on M with Reeb field X, any nonzero rescaling of X is a smooth, nonsingular rotational Beltrami field for some Riemannian metric on M.*

Remark 1. The original proof by Etnyre and Ghrist is for three-dimensional manifolds. The fact that the correspondence holds on any odd-dimensional manifold was detailed in [8]. See also [6] for an extension of this result to b-manifolds.

2. Embedding dynamics into Reeb flows

In [8], we studied several universality features of the stationary Euler equations. In view of the correspondence established in Theorem 1.1, we can reformulate the question of embedding dynamics into steady Euler flows in terms of Reeb flows. Let us fix a nonvanishing vector field X on a compact manifold N and some compact contact manifold (M,ξ) of dimensions $n \leq m$, respectively. The question we answer in this section is the following: Can we give sufficient conditions for the existence of an embedding $e : N \hookrightarrow M$ and a contact form $\alpha \in \Omega^1(M)$ defining ξ such that the Reeb field R satisfies $e_*X = R|_{e(N)}$? In other words, can we find conditions which ensure the existence of a Reeb field, whose contact form defines ξ, such that $e(N)$ is an invariant submanifold of R and where the Reeb field coincides with X?

2.1. Flexibility of Reeb embeddings

We will address the question above using a classical framework for flexibility problems in contact geometry: the homotopy principle. The world of contact geometry exhibits a lot of flexibility which reduces geometrical problems to their associated purely homotopical algebraic problems. The pioneering work of Gromov [15] showed that this approach is extremely fruitful for symplectic and contact geometrical problems. Some of Gromov's results in contact geometry were generalized in [4] when the ambient manifold is closed and the contact structure is "overtwisted". We will not introduce this notion here, the only thing that we need in our discussion is that being "overtwisted" is a property that a given contact structure may satisfy.

A first observation concerning our motivating question of embedding dynamics on Reeb fields is that the vector field X cannot be arbitrary.

Definition 2.1. A vector field X on N is geodesible if there is some metric for which the orbits of X are geodesics.

When X is of unit length for such a metric, we say that X is geodesible of unit length. From now on, by geodesible we mean geodesible of unit length. A characterization of geodesible vector fields was given by Gluck in terms of differential forms: X is geodesible if and only if there is some one-form β such that $\beta(X) = 1$ and $\iota_X d\beta = 0$. In particular, if a Reeb vector field R defined by a form α on a contact manifold M has some invariant submanifold N, then R restricted to N is geodesible. Indeed, if X is the vector field R restricted on N and $i : N \hookrightarrow M$ is the inclusion of N into M, then $i^*\alpha$ satisfies

$$\begin{cases} i^*\alpha(X) = 1, \\ \iota_X d i^*\alpha = 0. \end{cases} \tag{2.1}$$

Note that $i^*\alpha$ is not necessarily a contact form, so that X is not necessarily a Reeb field (in general, it is not even volume-preserving). However, it is always geodesible according to Gluck's characterization.

Conversely, start with any geodesible (hence non-vanishing) vector field X on a compact manifold N.

Definition 2.2. An embedding $e : (N, X) \hookrightarrow (M, \xi)$ is called a *Reeb embedding* if there is a contact form α defining ξ such that the associated Reeb field satisfies $e_*X = R|_{e(N)}$.

The main theorem in [8] gives sufficient conditions in terms of the codimension of an arbitrary smooth embedding to be isotopic to a Reeb embedding.

Theorem 2.3 ([8]). *Let $e : (N, X) \hookrightarrow (M, \xi)$ be a smooth embedding of N into a contact manifold (M, ξ), where X is a geodesible vector field on N. Assume that* $\dim M \geq 3n + 2$. *Then e is isotopic to a C^0-close Reeb embedding $\tilde{e} : (N, X) \hookrightarrow (M, \xi)$.*

Remark 2. If we impose the additional assumption that (M, ξ) is an overtwisted contact manifold, then $\dim M \geq 3n$ is enough, although the Reeb embedding $\tilde{e}$ is not necessarily C^0 close to e if $\dim M < 3n + 2$. In [8], parametric versions of the previous statement are also discussed.

Example 2.4. The existence of a Reeb embedding of any pair (N, X) into some contact manifold is easy to establish, since there is a natural source of examples of such embeddings. Denote by β the one-form such that $\beta(X) = 1$ and $\iota_X d\beta = 0$. Gluck's characterization implies that there is a metric for which X is of unit-length and its orbits are geodesics which satisfies $g(X, \cdot) = \beta$. Recall that the cotangent bundle T^*N is equipped with the canonical Liouville one-form $\lambda_{\mathrm{std}} \in \Omega^1(T^*N)$. Such one-form is characterized by the property that, given any one-form γ on N, which can be understood as an embedding $\gamma : N \to T^*N$, we have $\gamma = \gamma^*\lambda_{\mathrm{std}}$. For a given metric one can define the unit tangent bundle STN defined fiberwise by $ST_pN = \{X \in T_pN \mid g_p(X, X) = 1\}$. A standard property (see e.g. [13, Section 1.5]) of λ_{std} is that given the metric g on N, it restricts on ST^*N (the unit cotangent bundle) as a contact form λ whose Reeb field is dual to the geodesic vector field on STN. In particular, the section β, seen as an embedding

$$\beta : N \to ST^*N,$$

satisfies $\beta^*\lambda = \beta$ and actually the Reeb field R defined by λ satisfies $\beta_*X = R$. Thus, it is a Reeb-embedding according to Definition 2.2. This further motivates a systematic examination of Reeb-embeddings from a contact topology point of view, a study that leads to Theorem 2.3.

Sketch of the proof of Theorem 2.3. The proof of Theorem 2.3 follows the usual procedure of h-principle type results. We first define a "formal" notion of Reeb embedding, which satisfies a property that is purely homotopic in terms of its differential. We then prove that, under certain conditions, any formal Reeb embedding is isotopic to a genuine Reeb embedding (i.e., they satisfy the h-principle). To conclude, we use obstruction theory to analyze the minimal codimension for which any smooth embedding is a formal Reeb embedding satisfying the conditions for the h-principle to apply. We will now sketch each of these steps of the proof, under the simplifying assumption that M is overtwisted.

Step 1: Iso-Reeb embeddings and extension lemma. Let X be a geodesible vector field on N, and denote by β a one-form such that $\beta(X) = 1$ and $\iota_X d\beta = 0$. We need to fix such a choice of one-form, and let $\eta := \ker \beta$ be the hyperplane distribution defined by the kernel of β (which in general will not be of contact type). Let (M, ξ) be an overtwisted contact manifold with some defining contact form α, i.e., $\ker \alpha = \xi$.

With a slight abuse of notation, given a monomorphism $F : TN \to TM$ we will denote $\alpha \circ F$ for $\alpha(F(\cdot))$ and $d\alpha \circ F$ for $d\alpha(F(\cdot), F(\cdot))$. This is also denoted by $F^*\alpha$ and $F^*d\alpha$ in the discussion of "generalized iso-contact immersions" in [11, Section 16.2].

Definition 2.5. An embedding $f : (N, X, \eta = \ker \beta) \to (M, \xi)$ is an *iso-Reeb* embedding if $f^*\xi = \eta$.

The corresponding formal notion is the following definition.

Definition 2.6. An embedding $f : (N, X, \eta) \to (M, \xi)$ is a *formal iso-Reeb* embedding if there exists a homotopy of monomorphisms

$$F_t : TN \to TM,$$

such that F_t covers[1] f, $F_0 = df$, $h_1\alpha \circ F_1 = \beta$, and $d\beta|_\eta = h_2 d\alpha \circ F_1|_\eta$ for some strictly positive functions h_1 and h_2 on N.

Any (genuine) iso-Reeb embedding is clearly a formal iso-Reeb embedding, with F_t constantly equal to df. Both conditions $h_1\alpha \circ F_1 = \beta$ and $d\beta|_\eta = h_2 d\alpha \circ F_1|_\eta$ have to be imposed, since F_1 does not commute with the exterior derivative in general (when F_1 is not holonomic). This formal notion of Reeb embedding is enough to obtain the main theorem for an overtwisted target contact manifold. For the most general case, an extra formal hypothesis needs to be imposed (confer [8]).

[1] We say that $F_t : TN \to TM$ covers $f : N \to M$ if the map between bases induced by F_t is constantly equal to f.

The following lemma by Inaba [16] (see also [8]) shows that the condition of being an iso-Reeb embedding is enough to answer positively our question: we can find a Reeb field in (M, ξ) extending the given geodesible vector field X.

Lemma 2.7. *Let N be a submanifold of (M, ξ), and denote by i the inclusion map of N into M. Let η be the restriction $i^*\xi$. A nonvanishing vector field X on N can be extended to a Reeb field on all M if and only if X is transverse to η and the flow of X preserves η.*

The vector field X is transverse to η and preserves it if and only if there is a one-form β such that $\beta(X) = 1$, $\iota_X d\beta = 0$, and $\ker \beta = \eta$. These are our hypotheses in the case of an iso-Reeb embedding, hence by the previous lemma there is a contact form whose Reeb field R satisfies $f_* X = R$. Observe that an iso-Reeb embedding f is, in particular, a Reeb embedding according to Definition 2.2, the only difference is that in the definition of iso-Reeb embedding the one-form β making X geodesible is fixed.

Step 2: An h-principle via isocontact embeddings. Our goal in this second step is to prove that any formal iso-Reeb embedding $e : (N, X, \eta) \to (M, \xi)$ into an overtwisted contact manifold is homotopic through formal iso-Reeb embeddings to a genuine iso-Reeb embedding. This is tantamount to saying that iso-Reeb embeddings satisfy an existence h-principle. Other versions of the h-principle (parametric, relative to the domain, etc.) are discussed in [8]. Recall that α is a defining contact form of ξ. The sketch of the argument is the following.

(1) The embedding e satisfies that $de(\eta) \subset TM|_N$, but $de(\eta)$ is not, in general, contained in $\ker \alpha = \xi$. We extend the homotopy F_t and use it inversely to deform ξ via a homotopy of symplectic vector bundles (ξ_t, ω_t) (defined over all M, but which is identically $(\xi, d\alpha)$ outside a neighborhood U of $e(N)$) such that $(\xi_0, \omega_0) = (\xi, d\alpha)$, (ξ_1, ω_1) satisfies $de(\eta) \subset \xi_1$ and $\omega_1|_\eta = d\beta$ along N. The last condition is guaranteed, up to a conformal transformation, by the formal iso-Reeb condition. The symplectic hyperplane bundle (ξ_1, ω_1) will no longer be a contact structure in general.

(2) Using partitions of unity, the fact that ω_1 is non-degenerate on ξ_1, and that $\omega_1|_\eta = d\beta$, it is now possible to make another deformation. We extend the homotopy (ξ_t, ω_t) to $t \in [1, 2]$ such that (ξ_2, ω_2) is a contact structure in a smaller neighborhood U' of $e(N)$ and still satisfies $de(\eta) \subset \xi_2$. In particular, we can achieve that $\omega_2 = d\gamma$ for some one-form γ such that γ satisfies $e^*\gamma = \beta$ (the form such that $\ker \beta = \eta$ and $\beta(X) = 1$). The pair (ξ_2, ω_2) will not be a contact structure globally, since this small neighborhood is a priori smaller than the neighborhood U, where (ξ_1, ω_1) was not anymore of contact type. Hence in some parts $U \setminus U'$, ξ_2 is not of contact type.

(3) We will now reduce to a formal isocontact embedding (confer [11, Section 12.3] for more details on such embeddings). We endow the neighborhood U' with the contact structure (ξ_2, ω_2). We use the previous deformations (ξ_t, ω_t), $t \in [0,2]$ defined on U' to endow the trivial embedding $\hat{e} : U' \to M$ (defined as a neighborhood extension of the embedding e) with a homotopy of monomorphisms $G_t : TU' \to TM$ such that $G_0 = d\hat{e}$, G_1 satisfies $\xi_2 = G_1^{-1}(\xi)$, and the map induces a conformally symplectic map.

(4) The map $\hat{e}$ is what is called a formal isocontact embedding of codimension 0 with open source manifold. The h-principle for such embeddings into overtwisted targets applies [4, Corollary 1.4]. We obtain an embedding $\tilde{e} : U' \to M$ (isotopic to $\hat{e}$ through formal isocontact embeddings) such that $d\tilde{e}$ satisfies $d\tilde{e}(\xi_2) = \xi$ and the map induces a conformally symplectic map. Since (ξ_2, ω_2) restricted to $N \subset U'$ corresponds to $(\eta, d\beta)$, we deduce that $\tilde{e}|_N$ satisfies $(\tilde{e}|_N)^*\xi = \eta$ and hence is a genuine iso-Reeb embedding isotopic to $e = \hat{e}|_N$.

Step 3: Obstruction theory. The final step of the proof consists in showing that for $\dim M \geq 3 \dim N$, any smooth embedding $e : N \to (M, \xi)$ is a formal iso-Reeb embedding for any choice of (X, β), where X is a non-vanishing geodesible field and β is a choice of one-form for which $\beta(X) = 1$ and $\iota_X d\beta = 0$. We will assume the following lemma; confer [8] for the details.

Lemma 2.8. *Let $e : (N, X, \eta) \to (M, \xi)$ be an embedding such that there is a homotopy of monomorphisms $F_t : TN \to TM$ covering e satisfying $F_0 = de$ and $F_1(\eta)$ is an isotropic subbundle of ξ. Then e is a formal iso-Reeb embedding.*

For $2m > \dim N$, standard obstruction theory shows that there is a family of monomorphisms $H_t : TN \to TM$ such that $F_1(X) \pitchfork \xi$, and furthermore $F_1(\eta) \subset \xi$. The previous lemma shows that a sufficient condition for being a formal iso-Reeb embedding is that $F_1(\eta)$ can be homotoped into an isotropic subbundle of ξ. Recall that n denotes the dimension of N, hence η has rank $n-1$. The manifold M is of dimension $2m+1$, hence ξ is of rank $2m$. Denote by $\mathrm{Gr} = \mathrm{Grass}(n-1, \mathbb{R}^{2m})$ the space of $(n-1)$-subspaces of $\mathbb{R}^m$. Similarly, denote by $\mathrm{Gr}_{\mathrm{is}} = \mathrm{Grass}_{\mathrm{is}}(n-1, \mathbb{R}^{2m})$ the space of isotropic subspaces of dimension $n-1$ in $\mathbb{R}^{2m}$ seen as $\mathbb{C}^m$. To find a path between η and an isotropic subspace of ξ over N, we need to find a global section of the bundle E over N whose fiber is

$$P = \mathrm{Path}\big(\mathrm{Grass}(n-1, \mathbb{R}^{2m}), \mathrm{Grass}_{\mathrm{is}}(n-1, \mathbb{R}^{2m})\big),$$

i.e., the space of paths between any $(n-1)$-subspace and any isotropic $(n-1)$-subspace of $\mathbb{R}^{2m}$. On the other hand, we know that the homotopy groups of such a path space depend on the relative homotopy groups

$$\pi_j(P) \cong \pi_{j+1}\big(\mathrm{Grass}(n-1,\mathbb{R}^{2m}), \mathrm{Grass}_{\mathrm{is}}(n-1,\mathbb{R}^{2m})\big).$$

We now use that

$$\begin{aligned} \mathrm{Gr} &\cong \frac{\mathrm{SO}(2m)}{\mathrm{SO}(n-1)\times \mathrm{SO}\big(2m-(n-1)\big)}, \\ \mathrm{Gr}_{is} &\cong \frac{\mathrm{U}(m)}{\mathrm{SO}(n-1)\times \mathrm{U}\big(m-(n-1)\big)}. \end{aligned}$$

Combining the exact sequence for relative pairs, the exact sequence for quotients, and using the stable range of the involved groups, we can show that

$$2m \geq 3n-1 \implies \pi_j(P) = 0 \quad \text{for all } j \leq n-1.$$

Hence, if $\dim M \geq 3\dim N$, we can find a global section along N. Using this section and the previous family of monomorphisms, we find a family of isomorphisms $G_t : TN \to TM$ covering the smooth embedding e such that $G_1(\eta)$ is an isotropic subbundle of ξ. Applying Lemma 2.8, we conclude that e is a formal iso-Reeb embedding.

Step 4: Conclusion. In Step 3, we showed that any smooth embedding is a formal iso-Reeb embedding for any pair (N, X) embedded into a contact manifold (M, ξ) such that $\dim M \geq 3\dim N$. Note that smooth embeddings in this context always exist by Whitney's embedding theorem. Under the assumption that M is overtwisted, we can apply the h-principle proved in Step 2 and deduce that there is an iso-Reeb embedding $\tilde{e}$ isotopic to e. Since an iso-Reeb embedding is, in particular, a Reeb embedding, we can find some contact form α defining ξ whose Reeb field R satisfies $\tilde{e}_* X = R|_{\tilde{e}(N)}$. This concludes the proof of the theorem. ■

The previous theorem "fixes" the target contact structure, which forces to take an embedding that is isotopic to the original smooth embedding $e : N \to (M, \xi)$. If we simply want to extend the vector field X to a Reeb vector field, without fixing the ambient contact structure, then we can fix the embedding.

Corollary 2.9. *Let X be a geodesible vector field on a compact manifold N. Let $e : N \to (M, \xi)$ be a smooth embedding into a contact manifold with $\dim M \geq 3\dim N + 2$. Then there is a contact form α on M whose Reeb field R satisfies $e_* X = R|_{e(N)}$. The contact form α defines a contact structure contactomorphic to ξ.*

Proof. It follows from Theorem 2.3 that there is a Reeb embedding $\tilde{e}$ (with respect to the contact structure ξ) isotopic to e. According to Definition 2.2, there is a contact one-form α' defining ξ such that the Reeb field R' of α' satisfies $\tilde{e}_* X = R'|_{\tilde{e}(N)}$. Let φ_t be an isotopy of M such that $\varphi_1 \circ \tilde{e} = e$. Then $\alpha := (\varphi_1^{-1})^* \alpha'$ is a contact one-form, defining a contact structure $(\varphi_1)_* \xi$, whose Reeb field $R = (\varphi_1)_* R'$ satisfies

$$e_* X = (\varphi_1)_* \circ \tilde{e}_* X = (\varphi_1)_* R' = R,$$

thus concluding the proof. ■

2.2. Applications to universality

We are now ready to give some applications of Theorem 2.3. The following concept is inspired by Tao's definition of Euler-extendibility in [30] (albeit it is different in the sense that it is adapted to the context of stationary solutions of the Euler equations).

Definition 2.10. A non-autonomous time-periodic vector field $u_0(\cdot, t)$ on a compact manifold N is *Euler-extendible* if there exists an embedding $e : N \times \mathbb{S}^1 \to \mathbb{S}^n$ for some dimension $n > \dim N + 1$ (that only depends on the dimension of N), and a Eulerisable flow u on $\mathbb{S}^n$, such that $e(N \times \mathbb{S}^1)$ is an invariant submanifold of u and $e_*(u_0(\cdot, \theta) + \partial_\theta) = u|_{e(N \times \mathbb{S}^1)}$, $\theta \in \mathbb{S}^1$. If the non-autonomous field $u_0(\cdot, t)$ is not time-periodic, we say that it is Euler-extendible if there exists a proper embedding $e : N \times \mathbb{R} \to \mathbb{R}^n$ for some dimension $n > \dim N + 1$ (that only depends on the dimension of N), and a Eulerisable flow u on $\mathbb{R}^n$, such that $e(N \times \mathbb{R})$ is an invariant submanifold of u and $e_*(u_0(\cdot, \theta) + \partial_\theta) = u|_{e(N \times \mathbb{R})}$, $\theta \in \mathbb{R}$. If any non-autonomous dynamics $u_0(\cdot, t)$ is Euler-extendible, we say that the stationary Euler flows are *universal*.

Roughly speaking, the extendibility of a non-autonomous dynamics implies that, in the appropriate local coordinates, u_0 describes the "horizontal" behavior of the integral curves of the extended vector field u. Observe that the original vector field u_0 is not assumed to be volume-preserving, although certainly u will be. We introduce another definition for embeddability of discrete dynamics.

Definition 2.11. We say that an (orientation-preserving) diffeomorphism $\phi : N \to N$ is *Euler-embeddable* if there exists a Eulerisable field u on $\mathbb{S}^n$ (for some n that only depends on the dimension of N) with an invariant submanifold exhibiting a cross-section diffeomorphic to N such that the first return map of u at this cross-section is conjugate to ϕ.

Two main corollaries of the previous construction can be expressed in terms of these two definitions.

Corollary 2.12 ([8]). *The stationary Euler flows are universal. Moreover, the dimension of the ambient manifold $\mathbb{S}^n$ or $\mathbb{R}^n$ is the smallest odd integer $n \in \{3 \dim N + 5, 3 \dim N + 6\}$. In the time-periodic case, the extended field u is a steady Euler flow* *with a metric $g = g_0 + \delta_P$, where g_0 is the canonical metric on $\mathbb{S}^n$ and δ_P is supported in a ball that contains the invariant submanifold $e(N \times \mathbb{S}^1)$.*

It is clear that the extension to a Euler flow u is not unique, since Theorem 2.3 shows that iso-Reeb embeddings exist in abundance. Corollary 2.9, via the correspondence theorem (Theorem 1.1), illustrates the flexibility of steady Euler flows in the sense that *any* fixed smooth embedding in high enough codimension can be realized as an invariant submanifold (with arbitrary induced geodesible dynamics) of a steady Euler flow. Our second corollary is expressed in terms of Definition 2.11.

Corollary 2.13 ([8]). *Let N be a compact manifold and ϕ an orientation-preserving diffeomorphism on N. Then ϕ is Euler-embeddable in $\mathbb{S}^n$, where n is the smallest odd integer $n \in \{3 \dim N + 5, 3 \dim N + 6\}$.*

As in Corollary 2.12, the metric can also be assumed to be the canonical one outside an embedding of the mapping torus of N by ϕ. This is ensured by applying Theorem 2.3 with a tight contact sphere as the target contact manifold. The dimensional bounds can be slightly improved if we use an overtwisted contact sphere as target manifold, as explained after the statement of Theorem 2.3. In the following section, we shall introduce the concept of “Turing complete” flows, which are flows that are universal in a computational sense. Using the fact that there are diffeomorphisms that simulate any Turing machine (see [27] for an example), and the fact that our construction via an h-principle is constructible (i.e., algorithmic), we obtain as a by-product that there is a Turing complete Euler flow on $\mathbb{S}^{17}$. In the next section, we will focus on this property and drastically improve the dimension of the ambient manifold.

3. A Turing complete steady Euler flow on $\mathbb{S}^3$

In this section, we review the construction of a Turing complete stationary Euler flow on a Riemannian three-sphere [9]. We end up by proving a new result (Corollary 3.7) on the existence of Reeb flows (and their Beltrami counterparts) with orbits whose periodicity is undecidable.

3.1. Turing machines and symbolic dynamics

A Turing machine is a mathematical model of a theoretical device manipulating a set of symbols on a tape following some specific rules. It receives, as input data, a sequence of symbols and, after a number of steps, it might return as output another string of symbols. More concretely, a Turing machine is defined via the following data:

- a finite set Q of “states” including an initial state q_0 and a halting state $q_{\rm halt}$;
- a finite set Σ which is the “alphabet” with cardinality at least two;
- a transition function $\delta : Q \times \Sigma \to Q \times \Sigma \times \{-1, 0, 1\}$.

We will denote by $q \in Q$ the current state, and by $t = (t_n)_{n\in\mathbb{Z}} \in \Sigma^{\mathbb{Z}}$ the current tape of the machine at a given step of the algorithm of the Turing machine. This gives a configuration (q, t) of the machine. In particular, the space of all possible *configurations* of a Turing machine is given by $\mathcal{P} := Q \times \Sigma^{\mathbb{Z}}$. The algorithm works as follows, for a given input tape $t \in \Sigma^{\mathbb{Z}}$.

(1) Set the current state q as the initial state and the current tape t as the input tape.

(2) If the current state is $q_{\rm halt}$, then halt the algorithm and return t as output. Otherwise, compute $\delta(q, t_0) = (q', t'_0, \varepsilon)$, with $\varepsilon \in \{-1, 0, 1\}$.

(3) Replace q with q' and t_0 with t'_0, obtaining a modified tape $\tilde{t} = (\cdots t_{-1} \cdot t'_0 t_1 \cdots)$.

(4) Shift $\tilde{t}$ by ε, obtaining a new tape t'. The resulting configuration is (q', t'). Return to step (2).

Our convention is that $\varepsilon = 1$ (resp. $\varepsilon = -1$) corresponds to the left shift (resp. the right shift). This algorithm (determined by the transition function δ) induces a global transition function in the space of configurations

$$\Delta : Q \setminus \{q_{\rm halt}\} \times \Sigma^{\mathbb{Z}} \to \mathcal{P},$$

which sends a non-halting configuration in $\mathcal{P}$ to the configuration obtained after one step of the algorithm.

Remark 3. Without loss of generality, one can assume that the configurations of the machine are those pairs $(q, t) \in Q \times \Sigma^{\mathbb{Z}}$ for which only a finite number of symbols in t are different from 0 (also called the “blank” symbol). We will not need this simplifying assumption in this section, although it is certainly useful in other constructions [7].

The halting problem. In computability theory, the halting problem is the problem of determining, from a description of an arbitrary computer program and an input, whether the program will finish running (halting state), or continue to run forever. Alan Turing proved in 1936 that a general algorithm to solve the halting problem for all possible program-input pairs cannot exist. A key part of the proof is the formulation of a mathematical definition of a computer and program, which is the previously introduced notion of Turing machine; the halting problem is undecidable for Turing machines. The halting problem is historically important as it was one of the first problems to be proved undecidable.

Turing machines and universality. A Eulerisable field on a manifold M is Turing complete if it can simulate any Turing machine. In fact, Turing machines can be simulated by dynamical systems in a large sense (a vector field, a diffeomorphism, a map, etc.). Following [27], we give a formal definition of such a "simulation".

Definition 3.1. Let X be a vector field on a manifold M. We say it is Turing complete if for any integer $k \geq 0$, given a Turing machine T, an input tape t, and a finite string $(t^*_{-k}, \dots, t^*_k)$ of symbols of the alphabet, there exist an explicitly constructible point $p \in M$ and an open set $U \subset M$ such that the trajectory of X through p intersects U if and only if T halts with an output tape whose positions $-k, \dots, k$ correspond to the symbols $t^*_{-k}, \dots, t^*_k$. A completely analogous definition holds for diffeomorphisms of M.

Remark 4. In the construction explained in this section, the point p depends on T, the input, and the finite string, while the open set U is always the same. In other constructions of Turing complete flows [6, 8, 27], the point p only depends on T and the input, and the open set U depends on the finite string of the output. In particular, for a fixed machine and input we construct a point p and we can "measure" a posteriori what is the output of the machine up to some precision by looking which open sets are intersected by the trajectory of the flow through p.

Remark 5. One might as well avoid fixing a finite string of the output $(t^*_{-k}, \dots, t^*_k)$ and just require that the machine halts if and only if the trajectory through p enters certain open set. As detailed in [9, Lemma 5.5], the computational power is the same with this simplification.

In 1991, Moore [19] introduced the notion of generalized shift to be able to *simulate any Turing machine*; a generalized shift is a map that acts on the space of infinite sequences on a given finite alphabet.

Let A be an alphabet and $S \in A^{\mathbb{Z}}$ an infinite sequence. A generalized shift $\phi : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ is specified by two maps F and G which depend on a finite number of specified positions of the sequence in $A^{\mathbb{Z}}$. Denote by $D_F = \{i, \dots, i + r - 1\}$ and $D_G = \{j, \dots, j + l - 1\}$ the sets of positions on which F and G depend, respectively. These functions take a finite number of different values since they depend on a finite number of positions. The function G modifies the sequence only at the positions indicated by D_G:

$$\begin{aligned} G : A^l &\to A^l \\ (s_j \cdots s_{j+l-1}) &\mapsto (s'_j \cdots s'_{j+l-1}). \end{aligned}$$

Here $s_j \cdots s_{j+l-1}$ are the symbols at the positions $j, \dots, j + l - 1$ of an infinite sequence $S \in A^{\mathbb{Z}}$.

On the other hand, the function F assigns to the finite subsequence of consecutive elements $(s_i, \dots, s_{i+r-1})$ of the infinite sequence $S \in A^{\mathbb{Z}}$ an integer

$$F : A^r \to \mathbb{Z}.$$

The generalized shift $\phi : A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ corresponding to F and G is defined as follows:

- compute $F(S)$ and $G(S)$;
- modify S changing the positions in D_G by the function $G(S)$, obtaining a new sequence S';
- shift S' by $F(S)$ positions. That is, we obtain a new sequence $s''_n = s'_{n+F(S)}$ for all $n \in \mathbb{Z}$.

The sequence S'' is then $\phi(S)$.

Given a Turing machine, there is a generalized shift ϕ conjugate to it. Conjugation means that there is an injective map $\varphi : \mathcal{P} \to A^{\mathbb{Z}}$ such that the global transition function of the Turing machine is given by $\Delta = \varphi^{-1}\phi\varphi$. In fact, if the Turing machine is reversible, it can be shown that the generalized shift is bijective.

Key observation. Generalized shifts are conjugate to maps of the *square Cantor set* $C^2 := C \times C \subset I^2$, where C is the (standard) *Cantor ternary set* in the unit interval $I = [0, 1]$.

Point assignment. Take $A = \{0, 1\}$ (this can be assumed without loss of generality). Given $s = (\cdots s_{-1} \cdot s_0 s_1 \cdots) \in A^{\mathbb{Z}}$, we can associate to it an *explicitly constructible point* in the square Cantor set. We just express the coordinates of the assigned point in base 3: the coordinate y corresponds to the *expansion* $(y_0, y_1, \dots)$, where $y_i = 0$ if $s_i = 0$ and $y_i = 2$ if $s_i = 1$. Analogously, the coordinate x corresponds to the *expansion* $(x_1, x_2, \dots)$ in base 3, where $x_i = 0$ if $s_{-i} = 0$ and $x_i = 2$ if $s_{-i} = 1$.

Moore proved that any generalized shift is conjugate to the restriction on the square Cantor set of a piecewise linear map defined on blocks of the Cantor set in I^2. This map consists of finitely many area-preserving linear components. If the generalized shift is bijective, then the image blocks are pairwise disjoint. An example is depicted in Figure 1. Each linear component is the composition of two linear maps: a *translation* and a positive (or negative) power of the *horseshoe map* (or the Baker's map).

3.2. Area-preserving maps and Turing complete Reeb flows

In [19], Moore proved that any bijective generalized shift, understood as a map of the square Cantor set onto itself, can be extended as a diffeomorphism of the disk isotopic to the identity. The construction suggests that this can be done by further imposing

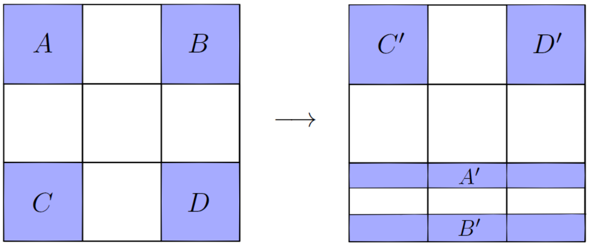


Figure 1. Example of a map by blocks of the square Cantor set.

the condition that this diffeomorphism is area-preserving. In [9], we formalized this proving that any bijective generalized shift can be extended to an area-preserving diffeomorphism of the disk which is the identity near the boundary. The proof of this result combines three ingredients: the aforementioned piecewise linear map defined on Cantor blocks, an explicit geometric construction using the homotopy extension property, and Moser's path method to ensure that the diffeomorphism that we obtain is area-preserving. The precise statement is the following:

Proposition 3.2. *For each bijective generalized shift and its associated map of the square Cantor set ϕ, there exists an area-preserving diffeomorphism of the disk $\varphi : D \to D$ which is the identity in a neighborhood of ∂D and whose restriction to the square Cantor set is conjugate to ϕ.*

Now the idea to construct a Turing complete Reeb flow is to take a Turing complete bijective generalized shift (which exists because there are universal Turing machines that are reversible as proved in the classical paper of Bennett [3]). Proposition 3.2 hence implies the existence of a Turing complete area-preserving diffeomorphism of the disk which is the identity on the boundary, as detailed in [9, Theorem 5.2]. Using a suspension construction in contact geometry, we can then show that any area-preserving diffeomorphism of the disk can be realized as the first-return map on a cross-section of a Reeb flow on any contact three-manifold. In particular, taking the aforementioned Turing complete diffeomorphism, we conclude the existence of Turing complete Reeb flows. More precisely, the following theorem holds.

Theorem 3.3. *Let (M,ξ) be a contact 3-manifold and $\varphi : D \to D$ an area-preserving diffeomorphism of the disk which is the identity (in a neighborhood of) the boundary. Then there exists a defining contact form α whose associated Reeb vector field R exhibits a Poincaré section with first return map conjugate to φ. In particular, there exists a Reeb field R on (M,ξ) which is Turing complete (in the sense of Definition 3.1).*

Combining Proposition 3.2, Theorem 3.3, and the correspondence theorem (Theorem 1.1) between Beltrami fields and Reeb flows, we obtain the desired result for stationary Euler flows.

Corollary 3.4. *There exists a Eulerisable field X on $\mathbb{S}^3$ that is Turing complete. The metric g that makes X a solution of the stationary Euler equations can be assumed to be the round metric in the complement of an embedded solid torus.*

The fact that the metric can be assumed to be the round one in the complement of an embedded solid torus needs some explanation. When applying Theorem 3.3, we take as ambient manifold the standard contact sphere $(\mathbb{S}^3, \xi_{\text{std}})$. Then, the contact form whose Reeb field realizes a given area-preserving diffeomorphism of the disk as a Poincaré map can be chosen to coincide with the standard contact form α_{std} outside a solid torus. To conclude, one can check that the metric associated to α via Theorem 1.1 can be taken to be the round one whenever α coincides with α_{std}.

Remark 6. The construction of a Turing complete Reeb flow in Theorem 3.3 is obtained by choosing a particular reversible universal Turing machine and realizing its associated generalized shift as the first return map of the flow restricted to a square Cantor set on a Poincaré section (see Proposition 3.2). Had we chosen another reversible Turing machine (not necessarily universal), its dynamics would have been induced in the square Cantor set via the first return map of a Reeb flow. We will use this observation in Corollary 3.8.

3.3. Undecidable dynamical properties in Reeb dynamics

In this subsection, we prove some new corollaries that follow from our construction in [9]. A straightforward implication of Theorem 3.3 is the existence of certain phenomena of contact dynamics that are undecidable. Specifically, there is no algorithm to assure that a Reeb trajectory will pass through a certain region of space in finite time. The precise formulation of this result is the following:

Corollary 3.5. *Let R be a Turing complete Reeb flow on (M, ξ). Then there exist an explicitly constructible compact set of points $K \subset M$ and an explicit open set $U \subset M$ such that it is an undecidable problem to determine if the (positive) integral curve of R through a point in K will intersect the set U or not.*

A variation of our construction also allows us to construct a Reeb field R for which there exist explicit points on M such that the problem of determining if the orbit of R through each of these points is closed is undecidable. The fact that generalized shifts have orbits whose periodicity is undecidable was proved by Moore in [19, Theorems 9 and 10]. In the lemma below, we give a complete formalization of an argument in [17, Theorem 8] that is similar to Moore's approach. This allows us to ensure that both properties required to prove Corollary 3.5 are satisfied: bijectivity of the associated generalized shift (i.e., reversibility of the Turing machine) and the equivalence between the halting of an input and the periodicity of the associated point in the disk.

Lemma 3.6. *There exists a Turing machine T' such that*

(1) *it is reversible;*

(2) *the image of the first component of the transition function δ does not contain q_0;*

(3) *it satisfies the "restart" property: if T' halts with input (q_0, t), then it halts with output (q_{halt}, t);*

(4) *T' is universal in the following sense: the halting of any Turing machine T and input c_0 is equivalent to the halting of T' for some explicit input (which depends on T and c_0).*

We are now ready to prove the undecidability of determining whether a trajectory is periodic or not.

Corollary 3.7. *Let (M, ξ) be a three-dimensional contact manifold. Then there is a contact form α defining ξ whose associated Reeb field R satisfies that there are explicit points on M for which determining whether the orbit through one of those points is periodic or not is an undecidable problem.*

Proof. Let $T = (Q, q_0, q_{\text{halt}}, \Sigma, \delta)$ be a universal Turing machine as in Lemma 3.6. We extend the transition function via $\delta(q_{\text{halt}}, t) = (q_0, t, 0)$, and construct a generalized shift ϕ conjugate to T by a map φ. Then given any input (q_0, t), the orbit of ϕ through $\varphi(q_0, t)$ is periodic if and only if T halts with input (q_0, t).

The map ϕ is bijective (since T is reversible), and by Proposition 3.2 we can find an area-preserving diffeomorphism of the disk $F : D \to D$ (which is the identity in a neighborhood of the boundary) whose restriction to the square Cantor set is conjugate to ϕ. Using Proposition 3.3, we construct a contact form α defining ξ whose Reeb flow has a cross-section with a first return map that is conjugate to F. It is then obvious that the orbit of the Reeb flow through a point representing an input of the Turing machine is periodic if and only if T halts with such an input. The result then follows from the undecidability of the halting problem. ■

Other special orbits can be constructed using the fact that the Turing machine is universal. For example, it is possible to construct an explicit point p such that the orbit of the Reeb flow through p is closed if and only if there is a counterexample to the Riemann hypothesis (using a discrete equivalent formulation [27]), and similarly with many other open problems in mathematics. This is achieved by constructing an initial input associated to a Turing machine which halts upon finding a counterexample.

Let us now give a proof of the auxiliary lemma (Lemma 3.6).

Proof of Lemma 3.6. As explained in [20, Section 6.1.2], we can find a reversible universal Turing machine $T = (Q, q_0, q_{\text{halt}}, \Sigma, \delta)$ which satisfies property (2): the initial state cannot be reached from any other state. Let us construct a universal Turing machine T' starting from T, which satisfies (1), (2), and (3).

This Turing machine is of the form $T' = (Q', q_0, q_{\text{halt}}, \Sigma, \delta')$. The space of states Q' is given by

$$Q' = \big(Q_0 \times \{-1, +1\}\big) \cup \{q_0, q_{\text{halt}}\},$$

where $Q_0 := Q \setminus \{q_0\}$. We basically take two copies of each state in Q except for q_0, and add q_0, q_{halt}. The sign in $\{-1, +1\}$ denotes the "direction" of the computation, a concept that will become clear in the construction. To simplify, for any state $q \in Q \setminus \{q_0, q_{\text{halt}}\}$, we denote $q_+ = q \times \{+1\} \in Q'$ and $q_- = q \times \{-1\} \in Q'$. The halting state of T' is q_{halt}, even if there are two additional states $q_{\text{halt}} \times \{1\}$ and $q_{\text{halt}} \times \{-1\}$ that we denote by q_{halt}^+ and q_{halt}^-.

For any input of T, given by (q_0, t), we associate the input (q_0, t) of T'. For any pair of the form (q_+, t) with $q \in Q \setminus \{q_0, q_{\text{halt}}\}$, we define the transition function of T' exactly as the transition function δ. To formalize this, we introduce the notation $(\tilde{q}, \tilde{t}_0, \varepsilon) = \delta(q, t_0)$. Then

$$\delta'(q_+, t_0) := (\tilde{q}_+, \tilde{t}_0, \varepsilon).$$

This is always well defined since $\tilde{q}$ is never equal to q_0. Similarly, for the initial state q_0 we also use the notation $(\tilde{q}, \tilde{t}_0, \varepsilon) = \delta(q_0, t_0)$ and we define

$$\delta'(q_0, t_0) := (\tilde{q}_+, \tilde{t}_0, \varepsilon).$$

When the machine reaches the state q_{halt}^+ (which happens when T halts with that input), we reverse the computation by defining

$$\delta'(q_{\text{halt}}^+, t_0) := (q_{\text{halt}}^-, t_0, 0). \tag{3.1}$$

The idea now is that instead of halting with the output of T, we swapped to a "reverse the computations" phase to undo the computations. For the states q_{halt}^- and q_- with $q \notin \{q_0, q_{\text{halt}}\}$, we define T' as the inverse Turing machine: a step of T' for a pair the form (q_-, t_0) is given by T^{-1}. See for instance [20, Section 5.1.4] for the construction of the inverse machine T^{-1}, which is also reversible. Denote by δ^{-1} the transition function of T^{-1}; notice that δ^{-1} is not defined on the state q_0 by property (2). Then, for $q \in Q \setminus \{q_0, q_{\text{halt}}\}$, if we set $\delta^{-1}(q, t_0) = (\tilde{q}, \tilde{t}_0, \varepsilon)$, we define

$$\delta'(q_-, t_0) := (\tilde{q}_-, \tilde{t}_0, \varepsilon) \quad \text{if } \tilde{q} \neq q_0.$$

If $\delta^{-1}(q, t_0) = (q_0, \tilde{t}_0, \varepsilon)$, it means that we have returned to the input configuration so we can define instead

$$\delta'(q_-, t_0) := (q_{\text{halt}}, \tilde{t}_0, \varepsilon). \tag{3.2}$$

Similarly, for q^-_{halt}, if $\delta^{-1}(q_{\text{halt}}, t_0) = (\tilde{q}, \tilde{t}_0, \varepsilon)$, we define

$$\delta'(q^-_{\text{halt}}, t_0) = (\tilde{q}_-, \tilde{t}_0, \varepsilon) \quad \text{if } \tilde{q} \neq q_0$$

and if $\tilde{q} = q_0$, then

$$\delta'(q^-_{\text{halt}}, t_0) = (q_{\text{halt}}, \tilde{t}_0, \varepsilon). \tag{3.3}$$

Notice that the image state $\tilde{q}$ via δ^{-1} cannot be q_{halt} because the transition function δ is not defined when $q = q_{\text{halt}}$.

The global transition function of T' on configurations with states q_0, q_+ coincides with the global transition function of T, where q^+_{halt} is identified with the halting state of T. Accordingly, it is injective there. Similarly, the global transition function on configurations with states q_- and q_{halt} coincides with that of T^{-1}, where q_{halt} is identified with the halting state of T' and q^-_{halt} is identified with the initial state of T'. So it is also injective there. Each configuration with state q^+_{halt} is sent to the same configuration with state q^-_{halt} in a trivial injective way. Summarizing, the global transition function of T' is injective everywhere so T' is reversible

The machine T' satisfies (2), since q_0 cannot be reached from δ, and in our construction we attain q_{halt} instead of q_0 when δ^{-1} is applied according to equation (3.3). The machine is universal since its halting is equivalent to the halting of T. Indeed, observe that the states of the form q_- in T' can only be reached if T halted, and q_{halt} can only be reached through negative states. This shows that if T does not halt with input (q_0, t), then T' does not halt. On the other hand, if T halts, T' will eventually reach a negative state, reverse the computation, and reach q_{halt}. In fact, T halts with input (q_0, t) if and only if T' halts with the same input. This shows that T' is universal.

Property (3) is also satisfied by construction. Whenever T' halts with input (q_0, t), it will reach a q^+_{halt}, then q^-_{halt} and reverse the computation to halt with configuration (q_{halt}, t). ■

Remark 7. Since any Turing machine can be simulated by a reversible Turing machine that satisfies property (2) (see e.g. [20, Section 6.1.2]), the construction presented in the proof of Lemma 3.6 allows one to start from any reversible Turing machine T, obtaining a reversible Turing machine T' which halts on the same inputs than T and has the "restart" property. In particular, any undecidable property associated to the inputs of T that halt is inherited by the inputs of T'.

Finally, we can mention a corollary which serves as a sample of dynamical properties of Reeb flows which simulate Turing machines that can be easily shown to be undecidable. Such undecidable properties are inherent to Turing machines and their associated generalized shifts [19, Theorem 10]. A key ingredient is Rice's theorem in computability theory, which in particular shows that non-trivial questions about the set of inputs for which the Turing machine halts are undecidable [25]. For example, Rice's theorem shows that there is no algorithm that can decide, for any given Turing machine, if there are at least k inputs that halt. From a logical point of view, this implies that there is at least one Turing machine for which determining if there are at least k inputs that halt is undecidable in the logical sense (i.e., the statement cannot be proved or disproved).

The following result is then a straightforward consequence of the previous discussion, Remark 7, Remark 6, and the existence of reversible Turing machines for which, respectively, determining if the set of inputs that halt has cardinality at least $k \geq 0$, is dense (in the set of all inputs) or has certain measure in the set of all inputs is undecidable in the logical sense.

Corollary 3.8. *Let (M, ξ) be a three-dimensional contact manifold. Then there is a contact form α defining ξ and an explicit set of points $K \subset M$ for which the following questions on the dynamics of R are undecidable (from the logical point of view, we remark that α depends on each question):*

- *Are there at least $k \geq 0$ points in K whose orbit is periodic?*
- *Is the set of points in K whose orbit is periodic dense in K?*
- *For a given $\mu > 0$, is the set of points in K whose orbit is periodic of measure greater than μ?*

In the previous corollary, the set K is simply the set of points associated to inputs of the Turing machine in the square Cantor set of the disk-like Poincaré section of the flow (these points lie on a finite union of blocks of the square Cantor set; see [9]).

Other dynamical properties of generalized shifts were proved to be undecidable by Moore, and could probably be adapted to establish analogous undecidability statements for Reeb flows. This includes convergence of orbits to a given point or the computability of Lyapunov exponents on a given invariant set (the orbits through the square Cantor set).

4. Time-dependent solutions of Euler and Navier–Stokes

In the previous sections, we have focused on stationary solutions to the Euler equations, first in high dimensions as a consequence of a new h-principle for Reeb embeddings, and then in dimension three using the power of symbolic dynamics. However, recall that the original motivation in [28–30] was to find a Turing complete time-dependent solution. The time-dependent Euler equations on a Riemannian manifold (M,g) define a dynamical system on the space of volume-preserving vector fields of the ambient manifold $\mathfrak{X}^\infty_{\mathrm{vol}}(M)$. The following definition of Turing completeness is adapted to this context by analogy with Definition 3.1.

Definition 4.1. Let (M,g) be a Riemannian manifold. The Euler equations on (M,g) are Turing complete if the following property is satisfied. For any integer $k \geq 0$, given a Turing machine T, an input tape t, and a finite string $(t^*_{-k},\dots,t^*_k)$ of symbols of the alphabet, there exist an explicitly constructible vector field $X_0 \in \mathfrak{X}^\infty_{\mathrm{vol}}(M)$ and a constructible open set $U \subset \mathfrak{X}^\infty_{\mathrm{vol}}(M)$ such that the solution to the Euler equations with initial datum X_0 is smooth for all time and intersects U if and only if T halts with an output tape whose positions $-k,\dots,k$ correspond to the symbols $t^*_{-k},\dots,t^*_k$.

In our recent article [7], we use a remarkable embedding theorem by Torres de Lizaur [31] (building on a previous embedding theorem into time-dependent Euler flows by Tao [28]) and the construction of Turing complete polynomial non-autonomous ODEs [14], to obtain Turing complete time-dependent solutions to the Euler equations:

Theorem 4.2 ([7]). *There exists a (constructible) compact Riemannian manifold (M,g) such that the Euler equations on (M,g) are Turing complete. In particular, the problem of determining whether a certain solution to the Euler equations with initial datum X_0 will reach a certain open set $U \subset \mathfrak{X}^\infty_{\mathrm{vol}}(M)$ is undecidable.*

This solves the question of the Turing universality of the time-dependent Euler equations in high dimensions with general Riemannian metrics.

We finish this article presenting an application of Corollary 3.4 in the context of the Navier–Stokes equations (following [8]). These equations describe the dynamics of an incompressible fluid flow with viscosity. On a Riemannian 3-manifold (M,g), they read as [1]

$$\begin{cases} \frac{\partial}{\partial t}X + \nabla_X X - \nu\Delta X = -\nabla p, \\ \operatorname{div} X = 0, \\ X(t=0) = X_0, \end{cases} \tag{4.1}$$

where $\nu > 0$ is the viscosity. In what follows, the differential operators are computed with respect to the metric g, and Δ stands for the Hodge Laplacian (whose action on a vector field is defined as $\Delta X := (\Delta X^\flat)^\sharp$).

Let us analyze what happens with the solution $X(t)$ when we take the Turing complete vector field X_0 constructed in Corollary 3.4 as initial condition (for the Navier–Stokes equations with the metric g that makes X_0 a stationary Euler flow). Specifically, using that $\operatorname{curl}_g(X_0) = X_0$, the solution to equation (4.1) with initial datum $X(t=0)=MX_0$, $M>0$ a real constant, is easily seen to be

$$\begin{cases} X(\cdot,t)=MX_0(\cdot)e^{-\nu t}, \\ p(\cdot,t)=c_0-\tfrac{1}{2}M^2e^{-2\nu t}g(X_0,X_0), \end{cases} \tag{4.2}$$

for any constant c_0. The integral curves (fluid particle paths) of the non-autonomous field X solve the ODE

$$\frac{dx(t)}{dt}=Me^{-\nu t}X_0\big(x(t)\big).$$

Accordingly, reparametrizing the time as

$$\tau(t):=\frac{M}{\nu}(1-e^{-\nu t}),$$

we show that the solution $x(t)$ can be written in terms of the solution $y(\tau)$ of the ODE

$$\frac{dy(\tau)}{d\tau}=X_0\big(y(\tau)\big),$$

as

$$x(t)=y\big(\tau(t)\big).$$

When $t\to\infty$, the new reparametrized "time" τ tends to $\frac{M}{\nu}$, and hence the integral curve $x(t)$ of the solution to the Navier–Stokes equations travels the orbit of X_0 just for the time interval $\tau\in[0,\frac{M}{\nu})$. In particular, the flow of the solution X only simulates a finite number of steps of a given Turing machine, so we cannot deduce the Turing completeness of the Navier–Stokes equations using the vector field MX_0 as initial condition. More number of steps of a Turing machine can be simulated if $\nu\to 0$ (the vanishing viscosity limit) or $M\to\infty$ (the L^2 norm of the initial datum blows up). For example, to obtain a universal Turing simulation we can take a family $\{M_kX_0\}_{k\in\mathbb{N}}$ of initial data for the Navier–Stokes equations, where $M_k\to\infty$ is a sequence of positive numbers. The energy (L^2 norm) of this family is not uniformly bounded, so it remains as a challenging open problem to know if there exists an initial datum of finite energy that gives rise to a Turing complete solution of the Navier–Stokes equations.

Funding. R. Cardona acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the María de Maeztu Programme for Units of Excellence in R&D (MDM-2014-0445) via an FPI grant. E. Miranda is supported by the Catalan Institution for Research and Advanced Studies via an ICREA Academia Prize 2016 and an ICREA Academia Prize 2021 and by the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (project CEX2020-001084-M). R. Cardona and E. Miranda are partially supported by the Spanish State Research Agency via the AEI grant PID2019-103849GB-I00 of MCIN/AEI/10.13039/501100011033. D. Peralta-Salas is supported by the grant MTM PID2019-106715GB-C21 (MICINN). This work was partially supported by the ICMAT-Severo Ochoa grant CEX2019-000904-S.

References

[1] M. Arnaudon and A. B. Cruzeiro, Lagrangian Navier–Stokes diffusions on manifolds: Variational principle and stability. *Bull. Sci. Math.* **136** (2012), no. 8, 857–881 Zbl 1254.35174 MR 2995006

[2] V. I. Arnold and B. A. Khesin, *Topological Methods in Hydrodynamics*. Appl. Math. Sci. 125, Springer, New York, 1998 Zbl 1475.76003 MR 1612569

[3] C. H. Bennett, Logical reversibility of computation. *IBM J. Res. Develop.* **17** (1973), 525–532 Zbl 0267.68024 MR 449020

[4] M. S. Borman, Y. Eliashberg, and E. Murphy, Existence and classification of overtwisted contact structures in all dimensions. *Acta Math.* **215** (2015), no. 2, 281–361 Zbl 1344.53060 MR 3455235

[5] R. Cardona, Steady Euler flows and Beltrami fields in high dimensions. *Ergodic Theory Dynam. Systems* **41** (2021), no. 12, 3610–3633 Zbl 07428300 MR 4336491

[6] R. Cardona, E. Miranda, and D. Peralta-Salas, Euler flows and singular geometric structures. *Philos. Trans. Roy. Soc. A* **377** (2019), no. 2158, 20190034, 15 Zbl 1462.76048 MR 4036383

[7] R. Cardona, E. Miranda, and D. Peralta-Salas, Turing universality of the incompressible Euler equations and a conjecture of Moore. *Int. Math. Res. Not. IMRN* **2022** (2022), no. 22, 18092–18109 Zbl 07635285 MR 4514463

[8] R. Cardona, E. Miranda, D. Peralta-Salas, and F. Presas, Universality of Euler flows and flexibility of Reeb embeddings. 2019, arXiv:1911.01963

[9] R. Cardona, E. Miranda, D. Peralta-Salas, and F. Presas, Constructing Turing complete Euler flows in dimension 3. *Proc. Natl. Acad. Sci. USA* **118** (2021), no. 19, Paper No. e2026818118 MR 4294081

[10] T. Cubitt, D. Perez-Garcia, and M. Wolf, Undecidability of the spectral gap. *Nature* **528** (2015), 207–211

[11] Y. Eliashberg and N. Mishachev, *Introduction to the h-Principle*. Grad. Stud. Math. 48, American Mathematical Society, Providence, RI, 2002 Zbl 1008.58001 MR 1909245

[12] J. Etnyre and R. Ghrist, Contact topology and hydrodynamics. I. Beltrami fields and the Seifert conjecture. *Nonlinearity* **13** (2000), no. 2, 441–458 Zbl 0982.76021 MR 1735969

[13] H. Geiges, *An Introduction to Contact Topology*. Cambridge Stud. Adv. Math. 109, Cambridge University Press, Cambridge, 2008 Zbl 1153.53002 MR 2397738

[14] D. S. Graça, M. L. Campagnolo, and J. Buescu, Computability with polynomial differential equations. *Adv. in Appl. Math.* **40** (2008), no. 3, 330–349 Zbl 1137.68025 MR 2402174

[15] M. Gromov, *Partial Differential Relations*. Ergeb. Math. Grenzgeb. (3) 9, Springer, Berlin, 1986 Zbl 0651.53001 MR 864505

[16] T. Inaba, Extending a vector field on a submanifold to a Reeb vector field on the whole contact manifold. 2019, http://www.math.s.chiba-u.ac.jp/∼inaba/ExtReeb0402.pdf

[17] J. Kari and N. Ollinger, Periodicity and immortality in reversible computing. In *Mathematical Foundations of Computer Science 2008*, pp. 419–430, Lecture Notes in Comput. Sci. 5162, Springer, Berlin, 2008 Zbl 1173.68521 MR 2539389

[18] J. M. Lee, *Introduction to Smooth Manifolds*. 2nd edn., Grad. Texts in Math. 218, Springer, New York, 2013 Zbl 1258.53002 MR 2954043

[19] C. Moore, Generalized shifts: unpredictability and undecidability in dynamical systems. *Nonlinearity* **4** (1991), no. 2, 199–230 Zbl 0725.58013 MR 1107005

[20] K. Morita, *Theory of Reversible Computing*. Monogr. Theoret. Comput. Sci. EATCS Ser., Springer, Tokyo, 2017 Zbl 1383.68002 MR 3822735

[21] R. Penrose, *The Emperor's New Mind. Concerning Computers, Minds, and the Laws of Physics. With a foreword by Martin Gardner*. Oxford University Press, New York, 1989 MR 1048125

[22] D. Peralta-Salas, Selected topics on the topology of ideal fluid flows. *Int. J. Geom. Methods Mod. Phys.* **13** (2016), no. suppl., 1630012, 23 Zbl 1469.76011 MR 3556103

[23] D. Peralta-Salas, A. Rechtman, and F. Torres de Lizaur, A characterization of 3D steady Euler flows using commuting zero-flux homologies. *Ergodic Theory Dynam. Systems* **41** (2021), no. 7, 2166–2181 Zbl 1468.35122 MR 4266368

[24] H. Poincaré, *Les méthodes nouvelles de la mécanique céleste. Tome III. Invariants intégraux*. Gauthier-Villars, Paris, 1899 Zbl 30.0834.08

[25] H. G. Rice, Classes of recursively enumerable sets and their decision problems. *Trans. Amer. Math. Soc.* **74** (1953), 358–366 Zbl 0053.00301 MR 53041

[26] T. Tao, Finite time blowup for an averaged three-dimensional Navier–Stokes equation. *J. Amer. Math. Soc.* **29** (2016), no. 3, 601–674 Zbl 1342.35227 MR 3486169

[27] T. Tao, On the universality of potential well dynamics. *Dyn. Partial Differ. Equ.* **14** (2017), no. 3, 219–238 Zbl 1383.37016 MR 3702540

[28] T. Tao, On the universality of the incompressible Euler equation on compact manifolds. *Discrete Contin. Dyn. Syst.* **38** (2018), no. 3, 1553–1565 Zbl 1397.35193 MR 3809006

[29] T. Tao, Searching for singularities in the Navier–Stokes equations. *Nat. Rev. Phys.* **1** (2019), 418–419

[30] T. Tao, On the universality of the incompressible Euler equation on compact manifolds, II. Non-rigidity of Euler flows. *Pure Appl. Funct. Anal.* **5** (2020), no. 6, 1425–1443 Zbl 1470.35273 MR 4196152

[31] F. Torres de Lizaur, Chaos in the incompressible Euler equation on manifolds of high dimension. *Invent. Math.* **228** (2022), no. 2, 687–715 Zbl 07514025 MR 4411730

[32] A. M. Turing, On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc. (2)* **42** (1936), no. 3, 230–265 Zbl 62.1059.03 MR 1577030

Robert Cardona
Laboratory of Geometry and Dynamical Systems, Department of Mathematics, Universitat Politècnica de Catalunya, Avinguda del Doctor Marañon 44-50, 08028 Barcelona, Spain; robert.cardona@upc.edu

Eva Miranda
Laboratory of Geometry and Dynamical Systems, Department of Mathematics, Universitat Politècnica de Catalunya, Avinguda del Doctor Marañon 44-50, 08028 Barcelona; Institut de Matemàtiques de la UPC-BarcelonaTech (IMTech), Pau Gargallo 14, 08028 Barcelona; and CRM Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra, Barcelona, Spain; eva.miranda@upc.edu

Daniel Peralta-Salas
Instituto de Ciencias Matemáticas (ICMAT), C/ Nicolás Cabrera, 13-15 Campus de Cantoblanco, Universidad Autónoma de Madrid, 28049 Madrid, Spain; dperalta@icmat.es

DOI 10.4171/8ECM/14

Lefschetz fibrations, open books, and symplectic fillings of contact 3-manifolds

Burak Ozbagci

Abstract. Ever since Donaldson showed that every closed symplectic 4-manifold admits a Lefschetz pencil and Giroux proved that every closed contact 3-manifold admits an adapted open book decomposition, Lefschetz fibrations and open books have been used fruitfully to obtain significant results about the topology of symplectic 4-manifolds and contact 3-manifolds. In this expository article, we present the highlights of our contribution to the subject at hand based on joint work with several coauthors during the past twenty years.

1. Introduction

At the turn of the century, two groundbreaking results have surfaced which had a long-lasting impact on the study of global topology of symplectic 4-manifolds and contact 3-manifolds. These results respectively are Donaldson's existence theorem [19] about Lefschetz pencils on closed symplectic 4-manifolds and Giroux's correspondence [30] between open books and contact structures on closed 3-manifolds.

In the first half of this short expository article, we briefly review the results of Donaldson and Giroux. In the last half, we first present an analogous result on Stein domains of complex dimension two, with an eye towards some applications to the study of the topology of symplectic fillings of contact 3-manifolds. Then we demonstrate how Lefschetz fibrations and open books interact with the classical theory of complex surface singularities as well as trisections of arbitrary smooth 4-manifolds, which were relatively recently discovered by Gay and Kirby [25].

2020 Mathematics Subject Classification. Primary 57K33; Secondary 57K43, 32Q28, 32S25, 32S30, 32S55, 14D05.
Keywords. Lefschetz fibration, open book decomposition, contact 3-manifold, symplectic 4-manifold, Stein surface, singularity link, trisection.

2. Topological characterization of symplectic 4-manifolds

Suppose that X and Σ are compact, oriented, and smooth manifolds of dimensions four and two, respectively, possibly with *nonempty* boundaries.

Definition 2.1. A *Lefschetz fibration* $\pi\colon X \to \Sigma$ is a submersion except for finitely many points $\{p_1, \dots, p_k\}$ in the interior of X, such that around each p_i and $\pi(p_i)$, there are orientation-preserving complex charts, on which π is of the form $\pi(z_1, z_2) = z_1^2 + z_2^2$.

The topology of Lefschetz fibrations is well understood with multiple points of view. We advise the reader to turn to the book [33] of Gompf and Stipsicz for an excellent introduction to the subject.

Lefschetz critical points can be viewed as complex analogs of Morse critical points, and they correspond to 2-handles. As a result, one obtains a handle decomposition of the 4-manifold X. Since a Lefschetz fibration is locally trivial in the complement of finitely many singular fibers, it can also be described combinatorially by means of its *monodromy*. Locally, the fiber of the map $(z_1, z_2) \to z_1^2 + z_2^2$ above $0 \neq t \in \mathbb{C}$ is smooth (topologically an annulus), while the fiber above the origin has a transverse double point (aka nodal singularity) and is obtained from the nearby fibers by collapsing an embedded simple closed curve called the *vanishing cycle*, as illustrated in Figure 1.

Let $\pi\colon X \to \Sigma$ be a Lefschetz fibration and let γ be a loop in Σ enclosing a single critical value, whose critical fiber has a single node. Then π restricts to surface fibration over γ, whose monodromy (a diffeomorphism of the fiber) is given by the right-handed Dehn twist about the vanishing cycle, as depicted in Figure 2.

For the purposes of this article, we assume that each singular fiber carries a *unique singularity* and there are *no homotopically trivial* vanishing cycles. Moreover, we restrict our attention to the following two cases.

First case, $\Sigma = S^2$, $\partial X = \emptyset$, and hence the fibers are closed surfaces. Suppose that $q_1, \dots, q_k \in D^2 \subset S^2$ are the critical values of a genus g Lefschetz fibration $\pi\colon X \to S^2$. Let $q_0 \in D^2$ be a regular value and for each $1 \leq i \leq k$, let $\gamma_i \subset D^2$ be a loop based at q_0 enclosing a single critical value q_i as shown in Figure 3. By the discussion above, the monodromy of the fibration over each γ_i is a positive Dehn twist along the corresponding vanishing cycle.

Since the fibration π is trivial over the complement $S^2 \setminus D^2$, the product of positive Dehn twists along the vanishing cycles is isotopic to the identity. The upshot is that a Lefschetz fibration $\pi\colon X \to S^2$ is characterized by a positive Dehn twist factorization of the identity element in Map_g, the mapping class group of an oriented closed surface of genus g.

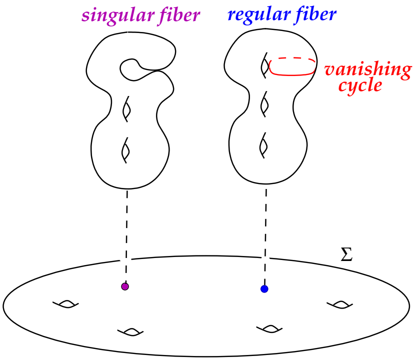


Figure 1. A nodal singularity.

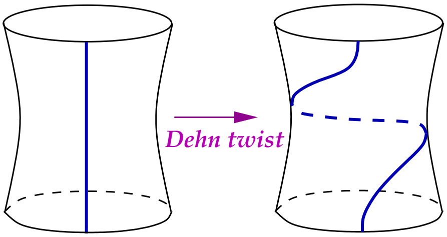


Figure 2. The right-handed (positive) Dehn twist.

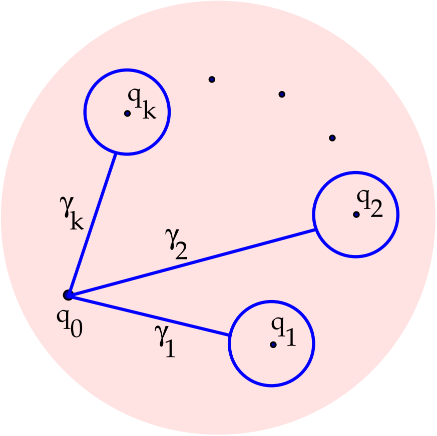


Figure 3. Loops in the base disk.

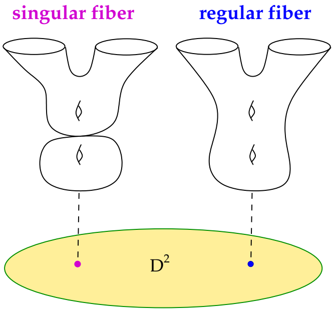


Figure 4. Fibers in a Lefschetz fibration.

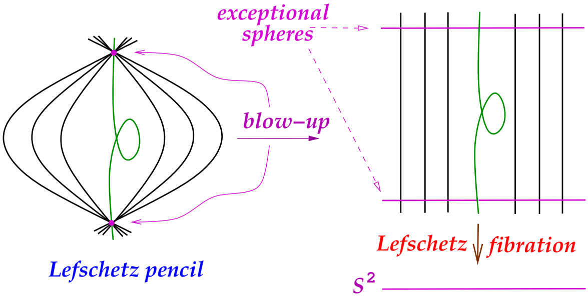


Figure 5. Blowing up the base-locus of a Lefschetz pencil.

Second case, $\Sigma = D^2$, the fibers have nonempty boundary and hence $\partial X \neq \emptyset$. In this case, the global monodromy over the boundary of the base disk D^2 is a product of positive Dehn twists in $\text{Map}_{g,r}$ (the mapping class group of an oriented genus g surface with $r > 0$ boundary components), *with no other constraints* (see Figure 4). Moreover, ∂X inherits a natural *open book decomposition*, which we will discuss in details later in Section 3.

Definition 2.2. A *Lefschetz pencil* on a closed and oriented 4-manifold X is a map $\pi\colon X - \{b_1, \dots, b_n\} \to S^2$, submersive except for a finite set $\{p_1, \dots, p_k\}$, conforming to local models

(i) $(z_1, z_2) \to z_1/z_2$ near each b_i and

(ii) $(z_1, z_2) \to z_1^2 + z_2^2$ near each p_j.

By blowing up X at the base-locus $\{b_1, \dots, b_n\}$, we obtain a Lefschetz fibration

$$X \# n \overline{\mathbb{CP}^2} \to S^2$$

with n disjoint sections, which are the exceptional spheres in the blow-up, as illustrated in Figure 5.

In the early twentieth century, Lefschetz showed that every *algebraic surface* (4-manifold arising as the zero-locus of a collection of homogeneous polynomials in $\mathbb{CP}^n$) admits "Lefschetz" pencils, which enabled him to study the topology of algebraic surfaces. This result was extended by Donaldson, to the case of the much larger class of symplectic 4-manifolds (i.e., those admitting closed non-degenerate 2-forms).

Theorem 2.3 (Donaldson [19]). *Any closed symplectic* 4*-manifold admits a Lefschetz pencil.*

For a sketch of the proof of Theorem 2.3 (other than Donaldson's original papers [18, 19]), the interested reader may consult the lecture notes [6] of Auroux and Smith, which is a wide-ranging survey, touching on the uses of Donaldon's theory of Lefschetz pencils and their relatives in 4-dimensional topology and mirror symmetry.

Conversely, generalizing a similar result of Thurston [58] on surface bundles over surfaces, Gompf [33] showed that if $\pi : X \to \Sigma$ is a Lefschetz fibration for which the fiber represents a non-torsion homology class,[1] then X admits a symplectic structure with symplectic fibers. As a corollary, he showed that any closed 4-manifold which admits a Lefschetz pencil, is symplectic.

Combining the results of Donaldson and Gompf, we obtain a *topological characterization* of symplectic 4-manifolds which has lead to a renewed interest in Lefschetz pencils/fibrations and hundreds of papers have been devoted to the study of various aspects and generalizations of Lefschetz fibrations, over the past twenty years. Here is one of the earlier results.

Theorem 2.4 (Ozbagci and Stipsicz [47]). *There are infinitely many pairwise non-homeomorphic closed* 4*-manifolds, each of which admits a genus two Lefschetz fibration over* S^2 *but does not carry complex structure with either orientation.*[2]

The examples in Theorem 2.4 are obtained by fiber sums of genus two Lefschetz fibrations $S^2 \times T^2 \# 4\,\overline{\mathbb{CP}^2} \to S^2$ of Matsumoto [39], which also shows that fiber sums of *holomorphic* Lefschetz fibrations are *not necessarily holomorphic*.

3. Topological characterization of contact 3-manifolds

Definition 3.1. An *open book decomposition* of a closed and oriented 3-manifold Y is a pair (B, π) consisting of an oriented link $B \subset Y$, and a locally trivial fibration $\pi\colon Y - B \to S^1$ such that B has a trivial tubular neighborhood $B \times D^2$ in which π is

[1] This hypothesis is automatically satisfied if the fiber genus is not equal to one.

[2] This result was independently observed by Ivan Smith.

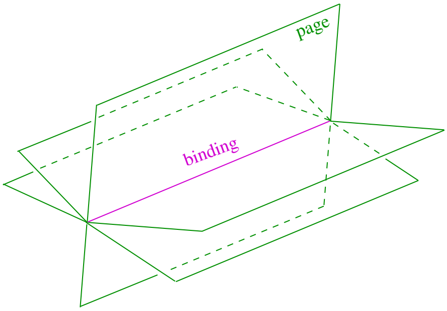


Figure 6. I am an open book!

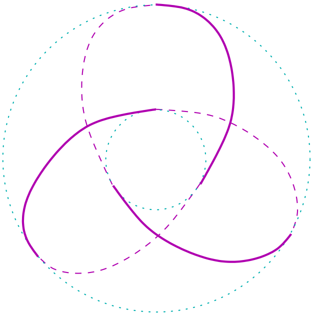


Figure 7. $(2, 3)$-torus knot (*the trefoil*).

given by the angular coordinate in the D^2-factor (see Figure 6). Here B is called the *binding* and the closure of each fiber of π, which is a Seifert surface for B, is called a *page*.

Example 3.2 (Milnor's fibration). Consider the polynomial $f\colon \mathbb{C}^2 \to \mathbb{C}$ given by $f(z_1, z_2) = z_1^p + z_2^q$, where $p, q \geq 2$ are relatively prime. Then $B = f^{-1}(0) \cap S^3$ is the (p, q)-torus knot in S^3 whose complement fibers over S^1:

$$\pi\colon S^3 - B \to S^1 := \frac{f(z_1, z_2)}{\left|f(z_1, z_2)\right|}.$$

Hence (B, π) is an open book for S^3 with connected binding. The torus knot for the case $p = 2$ and $q = 3$ is depicted in Figure 7.

For any given open book, one can choose a vector field which is transverse to the pages and meridional near the binding. Then the isotopy class of the first return map on a fixed page is called the *monodromy* of the open book. The topology of an open book is determined by the topology of its page and its monodromy.

Suppose that $\pi : X \to D^2$ is a Lefschetz fibration such that the regular fiber F has nonempty boundary ∂F. Then ∂X is the union of two pieces:

- the horizontal boundary, $\partial F \times D^2$ (see Figure 8) and
- the vertical boundary, $\pi^{-1}(\partial D^2)$ (see Figure 9),

glued together along the tori $\partial F \times \partial D^2$. It follows that ∂X inherits a natural open book, whose page is the fiber F and whose monodromy coincides with the monodromy of the Lefschetz fibration $\pi : X \to D^2$.

A differential 1-form α on a 3-manifold Y is called a *contact form* if $\alpha \wedge d\alpha$ is a volume form. A 2-dimensional distribution ξ in TY is called a contact structure if it can be given as the kernel of a contact form α. The pair (Y, ξ) is called a *contact 3-manifold*.

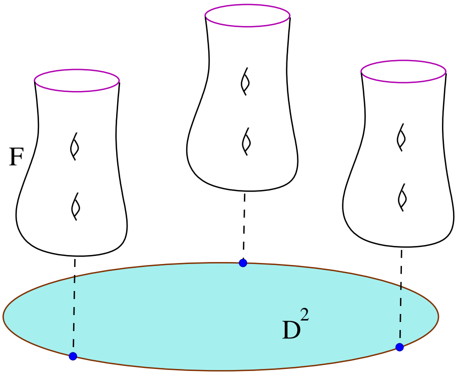


Figure 8. The vertical boundary: $\pi^{-1}(\partial D^2)$.

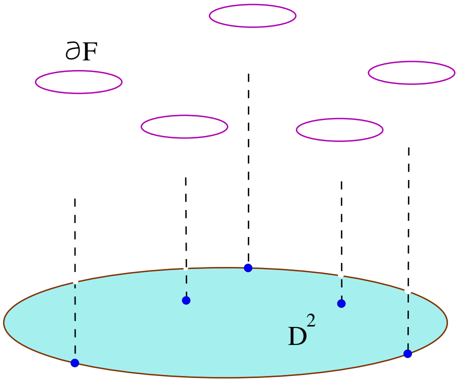


Figure 9. The horizontal boundary: $\partial F \times D^2$.

There are no local invariants of contact structures by *Darboux's theorem*, which says that any point in a contact 3-manifold has a neighborhood isomorphic to a neighborhood of the origin in the standard contact structure $\xi = \ker(dz + xdy)$ in $\mathbb{R}^3$, which is depicted in Figure 10.

We advise the reader to turn to the book [28] of Geiges, for a thorough introduction to contact topology in general dimensions and to the book [49] of Stipsicz and the author for a rapid course in dimension 3.

A classical theorem of Alexander [5] says that every closed oriented 3-manifold admits an open book decomposition and Martinet [38] showed that every closed oriented 3-manifold carries a contact structure. In 1975, Thurston and Winkelnkemper [59] presented an alternate proof of Martinet's theorem by constructing contact forms on closed 3-manifolds using open books.

Definition 3.3. A contact structure ξ on a 3-manifold Y is said to be supported by an open book (B, π) if ξ can be given by a contact form α such that $\alpha(B) > 0$ and $d\alpha > 0$ on every page.

In view of Definition 3.3, the result of Thurston and Winkelnkemper can be rephrased as follows: every open book on a closed oriented 3-manifold supports a contact structure.

The converse (i.e., every contact structure on a closed oriented 3-manifold is supported by an open book) was proven by Giroux. In fact, he proved the following theorem, which is known as *Giroux's correspondence*.

Theorem 3.4 (Giroux [30]). *On a closed oriented* 3*-manifold, there is a one-to-one correspondence between the set of isotopy classes of contact structures and open books up to positive stabilization.*

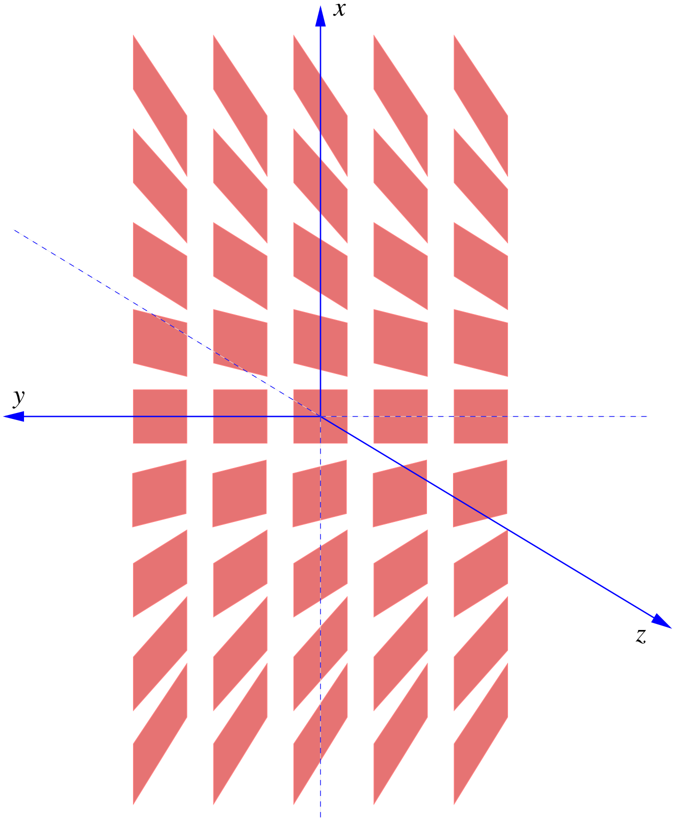


Figure 10. The standard contact structure $\xi = \ker(dz + x dy)$ in $\mathbb{R}^3$.

For a detailed sketch of the proof of Theorem 3.4, we refer to Etnyre's lecture notes [21].

4. Topological characterization of Stein domains of complex dimension two

Definition 4.1. A *Stein manifold* is an affine complex manifold, i.e., a complex manifold that admits a proper holomorphic embedding into some $\mathbb{C}^N$.

Suppose that $\phi\colon X \to \mathbb{R}$ is a smooth function on a complex manifold (X, J). Let ω_ϕ denote the 2-form $-d(d\phi \circ J)$. Then the map $\phi\colon X \to \mathbb{R}$ is called *J-convex* (aka *strictly plurisubharmonic*) if $\omega_\phi(u, Ju) > 0$ for all nonzero vectors $u \in TX$. It follows that ω_ϕ is an exact symplectic form on X.

Grauert's characterization. A complex manifold (X, J) is Stein if and only if it admits a proper J-convex function $\phi\colon X \to [0, \infty)$.

We advise the reader to turn to the book [17] of Eliashberg and Cieliebak, for a meticulous treatment of Stein (and Weinstein) manifolds. For the purposes of this article, we now restrict our attention to *Stein surfaces* (of complex dimension two), for which the reader may consult [32] for an elaborate discussion.

Suppose that (X, J) is a Stein surface. For any proper J-convex *Morse* function $\phi\colon X \to [0, \infty)$, each regular level set Y of ϕ is a contact 3-manifold, where the contact structure is given by the kernel of $\alpha_\phi = -d\phi \circ J$ or, equivalently, by the *complex tangencies* $TY \cap JTY$. For any regular value c of ϕ, the sublevel set $W = \phi^{-1}([0, c])$ is called a *Stein domain*. We also say that the compact 4-manifold (W, J) is a *Stein filling* of its contact boundary $(\partial W, \ker \alpha_\phi)$.

By the work of Eliashberg [20] and Gompf [32] a handle decomposition of a Stein domain (W, J) is well understood: it consists of a 0-handle, some 1-handles, and some 2-handles attached along Legendrian knots (those tangent to the contact planes) with framing -1 relative to the contact planes.

The following theorem, whose proof is based on the handle decomposition above, is somewhat analogous to Donaldson's theorem on the existence of Lefschetz pencils on closed symplectic manifolds.

Theorem 4.2 (Akbulut and Ozbagci [1] and Loi and Piergallini [36]). *A Stein domain admits an allowable*[3] *Lefschetz fibration over D^2 and, conversely, any allowable Lefschetz fibration over D^2 admits a Stein structure.*

Moreover, by modifying the proof of Akbulut and the author, Plamenevskaya [52] showed that the contact structure induced on the boundary of the Stein domain is supported by the open book inherited by the Lefschetz fibration. As a result we have the diagram

$$\begin{tikzcd}
\textit{Stein domain} \arrow[r, leftrightarrow] \arrow[d, "\partial"] & \begin{array}{c}\textit{allowable}\\ \textit{Lefschetz fibration}\end{array} \arrow[d, "\partial"] \\
\textit{contact structure} \arrow[r, leftrightarrow] & \textit{open book}
\end{tikzcd}$$

which gives a *criterion for Stein fillability: a contact 3-manifold is Stein fillable if and only if it admits a supporting open book whose monodromy can be factorized into positive Dehn twists.*[4]

[3] The vanishing cycles are homologically non-trivial.

[4] This was independently proved by Giroux.

Definition 4.3. A compact symplectic 4-manifold (X, ω) is a (strong) *symplectic filling* of a contact 3-manifold (Y, ξ) if $\partial X = Y$ (as oriented manifolds), ω is exact near the boundary, and its primitive α can be chosen so that $\ker(\alpha|_Y) = \xi$. A symplectic filling is called *minimal* if it does not contain any symplectically embedded sphere of self-intersection -1.

An active line of research in symplectic/contact topology is to classify *all Stein fillings* or more generally *all minimal symplectic fillings* of a given contact 3-manifold, up to diffeomorphism. It is clear by definition that every Stein filling is a minimal symplectic filling. The converse, however, is *not true* as shown by Ghiggini [29], using the celebrated Ozsváth–Szabó contact invariants [50].

The classification of Stein or more generally minimal symplectic fillings of a given contact 3-manifold is difficult in general. Nevertheless, this problem has been solved for many contact 3-manifolds, each of which has finitely many fillings. See the author's survey article [46] for the state of affairs until 2015.

The existence of a contact 3-manifold which admits infinitely many distinct Stein fillings was discovered by Stipsicz and the author. Let Y_g denote the closed 3-manifold, which is the total space of the open book whose page is a genus g surface with connected boundary and whose monodromy is the square of the boundary Dehn twist. Let ξ_g denote the contact structure on Y_g supported by this open book.

Theorem 4.4 (Ozbagci and Stipsicz [48]). *For each odd integer $g \geq 3$, the contact 3-manifold (Y_g, ξ_g) admits infinitely many pairwise non-homeomorphic Stein fillings.*

Outline of proof. A positive word in Map_g, for $g \geq 3$ (generalizing Matsumoto's genus two word [39]), was discovered independently by Cadavid [12] and Korkmaz [34]. For g odd, the word is $(c_0 c_1 c_2 \cdots c_g a^2 b^2)^2 = 1$, where, by an abuse of notation, each letter represents the right-handed Dehn twist along the curve decorated with the same letter, depicted in Figure 11. For each odd integer $g \geq 3$, there is a Lefschetz fibration over S^2, which corresponds to the aforementioned word. First we take (twisted) fiber sums of two copies of this Lefschetz fibration over S^2 and then remove a regular neighborhood of the union of a section and a regular fiber to get Stein fillings of the common contact boundary. The Stein fillings are distinguished by the torsion in their first homology groups, coming from the twistings in the fiber sums.

Remark 4.5. For a fixed odd integer $g \geq 3$, all the Stein fillings mentioned in Theorem 4.4 have the same Euler characteristic and the signature. In contrast, Baykur and Van Horn-Morris [8] showed that there are vast families of contact 3-manifolds each member of which admits infinitely many Stein fillings with arbitrarily large Euler characteristics and arbitrarily small signatures.

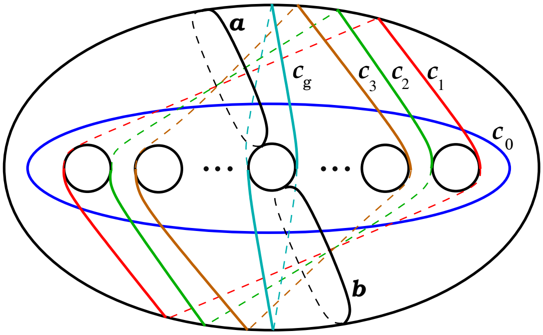


Figure 11. Curves on a genus g surface, for odd g.

5. Canonical contact structures on the links of isolated complex surface singularities

A fruitful source of Stein fillable contact 3-manifolds is given by the links of isolated complex surface singularities. Let $(X, 0) \subset (\mathbb{C}^N, 0)$ be an isolated complex surface singularity. Then for a sufficiently small sphere $S_\varepsilon^{2N-1} \subset \mathbb{C}^N$ centered at the origin, $Y = X \cap S_\varepsilon^{2N-1}$ is a closed, oriented, and smooth 3-dimensional manifold, which is called *the link of the singularity*.

If J denotes the complex structure on X, then the plane field given by the complex tangencies $\xi := TY \cap JTY$ is a contact structure on Y—called the canonical (aka *Milnor fillable*) contact structure on the singularity link. The contact 3-manifold (Y, ξ) is called the *contact singularity link*. Note that ξ is determined uniquely, up to isomorphism, by a theorem of Caubel, Némethi, and Popescu-Pampu [14].

We advise the reader to turn to the comprehensive lecture notes [54] of Popescu-Pampu for an introduction to complex singularity theory and its relation to contact topology.

The *minimal resolution* of an isolated complex surface singularity provides a Stein filling of its contact singularity link (Y, ξ), by the work of Bogomolov and de Oliveira [11]. Moreover, if the singularity is smoothable, the general fiber X of a smoothing is called a *Milnor fiber*, which is a compact smooth 4-manifold such that $\partial X = Y$. Furthermore, X has a natural Stein structure so that it provides a Stein (hence minimal symplectic) filling of (Y, ξ). Therefore, a natural question arises as follows (see, for example, [41]): Does there exist a contact singularity link which admits Stein (or minimal symplectic) fillings other than the Milnor fibers (and the minimal resolution)?

The answer is negative for simple and simple elliptic singularities as shown by Ohta and Ono [43–45]. The answer is negative for cyclic quotient singularities as shown by the culmination of the work of several people: McDuff [40], Christophersen

[16], Stevens [56], Lisca [35], and Némethi and Popescu-Pampu [42]. The answer is negative for non-cyclic quotient singularities as well by the work of Stevens [57], Bhupal and Ono [9], and H. Park, J. Park, Shin, and Urzúa [51].

The first examples where the answer is affirmative were discovered by Akhmedov and the author.

Theorem 5.1 (Akhmedov and Ozbagci [3]). *There exists an infinite family of Seifert fibered contact singularity links such that each member of this family admits infinitely many exotic*[5] *Stein fillings. Moreover, none of these Stein fillings are homeomorphic to Milnor fibers.*

The exotic fillings mentioned in Theorem 5.1 are not simply connected. The first examples of infinitely many exotic simply-connected Stein fillings were discovered by Akhmedov, Etnyre, Mark, and Smith [2].

Moreover, Plamenevskaya and Starkston [53] recently showed that many *rational singularities* admit simply-connected Stein fillings that are not diffeomorphic to any Milnor fibers.

Theorem 5.2 (Akhmedov and Ozbagci [4]). *For any finitely presented group G, there exists a contact singularity link which admits infinitely many exotic Stein fillings such that the fundamental group of each filling is G.*

Some key ingredients in the proofs of Theorem 5.1 and Theorem 5.2 are Luttinger surgery [37], symplectic sum [31], Fintushel–Stern knot surgery [24], and the Seiberg–Witten invariants [61].

We now turn our attention to Lefschetz fibrations on minimal symplectic fillings of lens spaces. Let ξ denote the canonical contact structure on the lens space $L(p,q)$, which is the link of a *cyclic quotient surface singularity*. The minimal symplectic fillings of $(L(p,q),\xi)$ have been classified by Lisca [35], generalizing the classification by McDuff [40] for $(L(p,1),\xi)$.

Theorem 5.3 (Bhupal and Ozbagci [10]). *There is an algorithm to describe any minimal symplectic filling of $(L(p,q),\xi)$ as an explicit genus-zero allowable Lefschetz fibration over D^2. Moreover, any minimal symplectic filling of $(L(p,q),\xi)$ is obtained by a sequence of rational blowdowns*[6] *starting from the minimal resolution of the corresponding cyclic quotient singularity.*

Theorem 5.3 was recently extended to the case of non-cyclic quotient singularities by H. Choi and J. Park [15].

[5] Homeomorphic but pairwise not diffeomorphic.

[6] Rational blow-down is a surgery operation discovered by Fintushel and Stern [23], where a negative definite linear plumbing submanifold is replaced by a rational 4-ball.

Remark 5.4. Since $(L(p,q),\xi)$ is known to be planar [55], i.e., it admits a planar open book that supports ξ, it also follows by a theorem of Wendl [60], that each minimal symplectic filling of $(L(p,q),\xi)$ is *deformation equivalent* to a genus-zero allowable Lefschetz fibration over D^2, although we have not relied on Wendl's theorem in our proof of Theorem 5.3.

6. Lefschetz fibrations and trisections

A handlebody is a compact manifold admitting a handle decomposition with a single 0-handle and some 1-handles. A *trisection* of a closed 4-manifold X is a decomposition of X into three 4D-handlebodies, whose pairwise intersections are 3D-handlebodies and whose triple intersection is a closed embedded surface.

A trisection of a 4-manifold is analogous to a Heegaard splitting of a closed 3-manifold, which is a decomposition into two 3D-handlebodies whose intersection is an embedded surface. Moreover, trisections can be presented by *trisection diagrams*, similar to the Heegaard diagrams. We refer to Gay's lecture notes [27] for a gentle introduction to trisections of 4-manifolds.

Theorem 6.1 (Gay and Kirby [25]). *Every closed oriented* 4*-manifold admits a trisection.*

Based on a splitting of an arbitrary closed 4-manifold into two *achiral*[7] Lefschetz fibrations over D^2 due to Etnyre and Fuller [22] and a gluing technique for *relative* trisections for 4-manifolds with boundary, Castro and the author [13] obtained an alternate proof of Theorem 6.1 using *Lefschetz fibrations* and *contact geometry*, instead of *Cerf theory* as utilized by Gay and Kirby. The following result is an application of this alternate proof.

Theorem 6.2 (Castro and Ozbagci [13]). *Suppose that X is a closed, oriented* 4*-manifold which admits a Lefschetz fibration over S^2 with a section of square* -1*. Then, an explicit trisection of X can be described by a corresponding trisection diagram, which is determined by the vanishing cycles of the Lefschetz fibration.*

We would like to point out that Gay [26] also constructed a trisection of any 4-manifold which admits a Lefschetz pencil, turning one type of decomposition into another, but without describing an *explicit* trisection diagram.

Remark 6.3. Baykur and Saeki [7] obtained yet another proof of Theorem 6.1, setting up a correspondence between *broken* Lefschetz fibrations and trisections, using

[7] Possibly including nodes with opposite orientation.

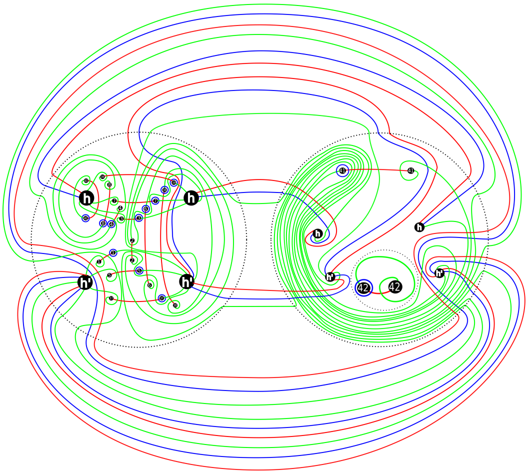


Figure 12. A trisection diagram for the Horikawa surface $H'(1)$.

a method which is very different from ours. They also proved a stronger version of Theorem 6.2.

Example 6.4 ([13]). The Horikawa surface $H'(1)$, a simply-connected complex surface of general type, admits a genus two Lefschetz fibration over S^2 with a section of square -1. The trisection diagram obtained by applying Theorem 6.2 is depicted in Figure 12. Notice that $H'(1)$ is an *exotic copy* of $5\,\mathbb{CP}^2\,\#\,29\,\overline{\mathbb{CP}^2}$.

References

[1] S. Akbulut and B. Ozbagci, Lefschetz fibrations on compact Stein surfaces. *Geom. Topol.* **5** (2001), 319–334 Zbl 1002.57062 MR 1825664

[2] A. Akhmedov, J. B. Etnyre, T. E. Mark, and I. Smith, A note on Stein fillings of contact manifolds. *Math. Res. Lett.* **15** (2008), no. 6, 1127–1132 Zbl 1156.57019 MR 2470389

[3] A. Akhmedov and B. Ozbagci, Singularity links with exotic Stein fillings. *J. Singul.* **8** (2014), 39–49 Zbl 1300.57025 MR 3213526

[4] A. Akhmedov and B. Ozbagci, Exotic Stein fillings with arbitrary fundamental group. *Geom. Dedicata* **195** (2018), 265–281 Zbl 1397.57045 MR 3820506

[5] J. W. Alexander, A lemma on systems of knotted curves. *Proc. Natl. Acad. Sci. USA* **9** (1923), 93–95 Zbl 49.0408.03

[6] D. Auroux and I. Smith, Lefschetz pencils, branched covers and symplectic invariants. In *Symplectic 4-Manifolds and Algebraic Surfaces*, pp. 1–53, Lecture Notes in Math. 1938, Springer, Berlin, 2008 Zbl 1142.14008 MR 2441411

[7] R. I. Baykur and O. Saeki, Simplified broken Lefschetz fibrations and trisections of 4-manifolds. *Proc. Natl. Acad. Sci. USA* **115** (2018), no. 43, 10894–10900 Zbl 1421.57026 MR 3871793

[8] R. I. Baykur and J. Van Horn-Morris, Families of contact 3-manifolds with arbitrarily large Stein fillings. With an appendix by Samuel Lisi and Chris Wendl. *J. Differential Geom.* **101** (2015), no. 3, 423–465 Zbl 1348.57036 MR 3415768

[9] M. Bhupal and K. Ono, Symplectic fillings of links of quotient surface singularities. *Nagoya Math. J.* **207** (2012), 1–45 Zbl 1258.53088 MR 2957141

[10] M. Bhupal and B. Ozbagci, Symplectic fillings of lens spaces as Lefschetz fibrations. *J. Eur. Math. Soc. (JEMS)* **18** (2016), no. 7, 1515–1535 Zbl 1348.57037 MR 3506606

[11] F. A. Bogomolov and B. de Oliveira, Stein small deformations of strictly pseudoconvex surfaces. In *Birational Algebraic Geometry (Baltimore, MD, 1996)*, pp. 25–41, Contemp. Math. 207, Amer. Math. Soc., Providence, RI, 1997 Zbl 0889.32021 MR 1462922

[12] C. A. Cadavid, *On a remarkable set of words in the mapping class group*. Ph.D. thesis, The University of Texas at Austin, 1998

[13] N. A. Castro and B. Ozbagci, Trisections of 4-manifolds via Lefschetz fibrations. *Math. Res. Lett.* **26** (2019), no. 2, 383–420 Zbl 1427.57013 MR 3999550

[14] C. Caubel, A. Némethi, and P. Popescu-Pampu, Milnor open books and Milnor fillable contact 3-manifolds. *Topology* **45** (2006), no. 3, 673–689 Zbl 1098.53064 MR 2218761

[15] H. Choi and J. Park, A Lefschetz fibration on minimal symplectic fillings of a quotient surface singularity. *Math. Z.* **295** (2020), no. 3-4, 1183–1204 Zbl 1446.57025 MR 4125685

[16] J. A. Christophersen, On the components and discriminant of the versal base space of cyclic quotient singularities. In *Singularity Theory and its Applications, Part I (Coventry, 1988/1989)*, pp. 81–92, Lecture Notes in Math. 1462, Springer, Berlin, 1991 Zbl 0735.14002 MR 1129026

[17] K. Cieliebak and Y. Eliashberg, *From Stein to Weinstein and Back. Symplectic Geometry of Affine Complex Manifolds*. Amer. Math. Soc. Colloq. Publ. 59, Amer. Math. Soc., Providence, RI, 2012 Zbl 1262.32026 MR 3012475

[18] S. K. Donaldson, Lefschetz fibrations in symplectic geometry. *Doc. Math.* **Extra Vol.** (1998), 309–314 Zbl 0909.53018 MR 1648081

[19] S. K. Donaldson, Lefschetz pencils on symplectic manifolds. *J. Differential Geom.* **53** (1999), no. 2, 205–236 Zbl 1040.53094 MR 1802722

[20] Y. Eliashberg, Topological characterization of Stein manifolds of dimension > 2. *Internat. J. Math.* **1** (1990), no. 1, 29–46 Zbl 0699.58002 MR 1044658

[21] J. B. Etnyre, Lectures on open book decompositions and contact structures. In *Floer Homology, Gauge Theory, and Low-Dimensional Topology*, pp. 103–141, Clay Math. Proc. 5, Amer. Math. Soc., Providence, RI, 2006 Zbl 1108.53050 MR 2249250

[22] J. B. Etnyre and T. Fuller, Realizing 4-manifolds as achiral Lefschetz fibrations. *Int. Math. Res. Not. IMRN* **2006** (2006), Art. ID 70272 Zbl 1118.57019 MR 2219214

[23] R. Fintushel and R. J. Stern, Rational blowdowns of smooth 4-manifolds. *J. Differential Geom.* **46** (1997), no. 2, 181–235 Zbl 0896.57022 MR 1484044

[24] R. Fintushel and R. J. Stern, Knots, links, and 4-manifolds. *Invent. Math.* **134** (1998), no. 2, 363–400 Zbl 0914.57015 MR 1650308

[25] D. Gay and R. Kirby, Trisecting 4-manifolds. *Geom. Topol.* **20** (2016), no. 6, 3097–3132 Zbl 1372.57033 MR 3590351

[26] D. T. Gay, Trisections of Lefschetz pencils. *Algebr. Geom. Topol.* **16** (2016), no. 6, 3523–3531 Zbl 1375.57023 MR 3584265

[27] D. T. Gay, From Heegaard splittings to trisections; porting 3-dimensional ideas to dimension 4. *Winter Braids Lect. Notes* **5** (2018), Exp. No. 4 Zbl 1444.57014 MR 4157116

[28] H. Geiges, *An Introduction to Contact Topology*. Cambridge Stud. Adv. Math. 109, Cambridge University Press, Cambridge, 2008 Zbl 1153.53002 MR 2397738

[29] P. Ghiggini, Strongly fillable contact 3-manifolds without Stein fillings. *Geom. Topol.* **9** (2005), 1677–1687 Zbl 1091.57018 MR 2175155

[30] E. Giroux, Géométrie de contact: de la dimension trois vers les dimensions supérieures. In *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)*, pp. 405–414, Higher Ed. Press, Beijing, 2002 Zbl 1015.53049 MR 1957051

[31] R. E. Gompf, A new construction of symplectic manifolds. *Ann. of Math. (2)* **142** (1995), no. 3, 527–595 Zbl 0849.53027 MR 1356781

[32] R. E. Gompf, Handlebody construction of Stein surfaces. *Ann. of Math. (2)* **148** (1998), no. 2, 619–693 Zbl 0919.57012 MR 1668563

[33] R. E. Gompf and A. I. Stipsicz, 4-*Manifolds and Kirby Calculus*. Grad. Stud. Math. 20, Amer. Math. Soc., Providence, RI, 1999 Zbl 0933.57020 MR 1707327

[34] M. Korkmaz, Noncomplex smooth 4-manifolds with Lefschetz fibrations. *Int. Math. Res. Not. IMRN* **2001** (2001), no. 3, 115–128 Zbl 0977.57020 MR 1810689

[35] P. Lisca, On symplectic fillings of lens spaces. *Trans. Amer. Math. Soc.* **360** (2008), no. 2, 765–799 Zbl 1137.57026 MR 2346471

[36] A. Loi and R. Piergallini, Compact Stein surfaces with boundary as branched covers of B^4. *Invent. Math.* **143** (2001), no. 2, 325–348 Zbl 0983.32027 MR 1835390

[37] K. M. Luttinger, Lagrangian tori in $\mathbf{R}^4$. *J. Differential Geom.* **42** (1995), no. 2, 220–228 Zbl 0861.53029 MR 1366546

[38] J. Martinet, Formes de contact sur les variétés de dimension 3. In *Proceedings of Liverpool Singularities Symposium, II (1969/1970)*, pp. 142–163, Lecture Notes in Math. 209, 1971 Zbl 0215.23003 MR 0350771

[39] Y. Matsumoto, Lefschetz fibrations of genus two—a topological approach. In *Topology and Teichmüller Spaces (Katinkulta, 1995)*, pp. 123–148, World Sci. Publ., River Edge, NJ, 1996 Zbl 0921.57006 MR 1659687

[40] D. McDuff, The structure of rational and ruled symplectic 4-manifolds. *J. Amer. Math. Soc.* **3** (1990), no. 3, 679–712 Zbl 0723.53019 MR 1049697

[41] A. Némethi, Some meeting points of singularity theory and low dimensional topology. In *Deformations of Surface Singularities*, pp. 109–162, Bolyai Soc. Math. Stud. 23, János Bolyai Math. Soc., Budapest, 2013 Zbl 1325.32001 MR 3203577

[42] A. Némethi and P. Popescu-Pampu, On the Milnor fibres of cyclic quotient singularities. *Proc. Lond. Math. Soc. (3)* **101** (2010), no. 2, 554–588 Zbl 1204.32020 MR 2679701

[43] H. Ohta and K. Ono, Simple singularities and topology of symplectically filling 4-manifold. *Comment. Math. Helv.* **74** (1999), no. 4, 575–590 Zbl 0957.57022 MR 1730658

[44] H. Ohta and K. Ono, Symplectic fillings of the link of simple elliptic singularities. *J. Reine Angew. Math.* **565** (2003), 183–205 Zbl 1044.57008 MR 2024651

[45] H. Ohta and K. Ono, Simple singularities and symplectic fillings. *J. Differential Geom.* **69** (2005), no. 1, 1–42 Zbl 1085.53079 MR 2169581

[46] B. Ozbagci, On the topology of fillings of contact 3-manifolds. In *Interactions Between Low-Dimensional Topology and Mapping Class Groups*, pp. 73–123, Geom. Topol. Monogr. 19, Geom. Topol. Publ., Coventry, 2015 Zbl 1332.57026 MR 3609904

[47] B. Ozbagci and A. I. Stipsicz, Noncomplex smooth 4-manifolds with genus-2 Lefschetz fibrations. *Proc. Amer. Math. Soc.* **128** (2000), no. 10, 3125–3128 Zbl 0951.57015 MR 1670411

[48] B. Ozbagci and A. I. Stipsicz, Contact 3-manifolds with infinitely many Stein fillings. *Proc. Amer. Math. Soc.* **132** (2004), no. 5, 1549–1558 Zbl 1045.57014 MR 2053364

[49] B. Ozbagci and A. I. Stipsicz, *Surgery on Contact* 3*-Manifolds and Stein Surfaces*. Bolyai Soc. Math. Stud. 13, Springer, Berlin, 2004 Zbl 1067.57024 MR 2114165

[50] P. Ozsváth and Z. Szabó, Heegaard Floer homology and contact structures. *Duke Math. J.* **129** (2005), no. 1, 39–61 Zbl 1083.57042 MR 2153455

[51] H. Park, J. Park, D. Shin, and G. Urzúa, Milnor fibers and symplectic fillings of quotient surface singularities. *Adv. Math.* **329** (2018), 1156–1230 Zbl 1390.14018 MR 3783436

[52] O. Plamenevskaya, Contact structures with distinct Heegaard Floer invariants. *Math. Res. Lett.* **11** (2004), no. 4, 547–561 Zbl 1064.57031 MR 2092907

[53] O. Plamenevskaya and L. Starkston, Unexpected Stein fillings, rational surface singularities, and plane curve arrangements. *Geom. Topol.* (to appear); arXiv:2006.06631

[54] P. Popescu-Pampu, Complex singularities and contact topology. *Winter Braids Lect. Notes* **3** (2016), Exp. No. 3 Zbl 1430.53002 MR 3707744

[55] S. Schönenberger, Determining symplectic fillings from planar open books. *J. Symplectic Geom.* **5** (2007), no. 1, 19–41 Zbl 1136.53062 MR 2371183

[56] J. Stevens, On the versal deformation of cyclic quotient singularities. In *Singularity Theory and its Applications, Part I (Coventry, 1988/1989)*, pp. 302–319, Lecture Notes in Math. 1462, Springer, Berlin, 1991 Zbl 0747.14002 MR 1129040

[57] J. Stevens, Partial resolutions of quotient singularities. *Manuscripta Math.* **79** (1993), no. 1, 7–11 Zbl 0791.14005 MR 1213355

[58] W. P. Thurston, Some simple examples of symplectic manifolds. *Proc. Amer. Math. Soc.* **55** (1976), no. 2, 467–468 Zbl 0324.53031 MR 402764

[59] W. P. Thurston and H. E. Winkelnkemper, On the existence of contact forms. *Proc. Amer. Math. Soc.* **52** (1975), 345–347 Zbl 0312.53028 MR 375366

[60] C. Wendl, Strongly fillable contact manifolds and J-holomorphic foliations. *Duke Math. J.* **151** (2010), no. 3, 337–384 Zbl 1207.32022 MR 2605865

[61] E. Witten, Monopoles and four-manifolds. *Math. Res. Lett.* **1** (1994), no. 6, 769–796 Zbl 0867.57029 MR 1306021

Burak Ozbagci
Department of Mathematics, Koç University, Rumelifeneri Yolu, 34450 Istanbul, Turkey;
bozbagci@ku.edu.tr

DOI 10.4171/8ECM/20

Finite groups of birational transformations

Yuri Prokhorov

Abstract. We survey new results on finite groups of birational transformations of algebraic varieties.

1. Introduction

We work over a field $\Bbbk$ of characteristic 0. Typically, unless otherwise mentioned, we assume that $\Bbbk$ is algebraically closed. The *Cremona group* $\operatorname{Cr}_n(\Bbbk)$ of rank n is the group of $\Bbbk$-automorphisms of the field $\Bbbk(x_1, \dots, x_n)$ of rational functions in n independent variables. Equivalently, $\operatorname{Cr}_n(\Bbbk)$ can be viewed as the group of birational transformations of the projective space $\mathbb{P}^n$. It is easy to show that for $n = 1$, the group $\operatorname{Cr}_n(\Bbbk)$ consists of linear projective transformations:

$$\operatorname{Cr}_1(\Bbbk) = \operatorname{PGL}_2(\Bbbk).$$

On the other hand, for $n \geq 2$, the group $\operatorname{Cr}_n(\Bbbk)$ has an extremely complicated structure. In particular, it contains linear algebraic subgroups of arbitrary dimension and has a lot of normal non-algebraic subgroups [18,24]. We refer to [3,22,23,38,48,95] for surveys, historical résumés, and introductions to the subject.

Examples. (i) Any matrix $A = \|a_{i,j}\| \in \operatorname{GL}_n(\mathbb{Z})$ defines an element $\varphi_A \in \operatorname{Cr}_n(\Bbbk)$ via the following action on $\Bbbk(x_1, \dots, x_n)$:

$$\varphi_A : x_i \mapsto x_1^{a_{1,i}} x_2^{a_{2,i}} \cdots x_n^{a_{n,i}}.$$

Such Cremona transformations are called *monomial*. For $n = 2$ and $A = -\mathrm{id}$, the transformation φ_A is known as the *standard quadratic involution*

$$(x_1, x_2) \mapsto (x_1^{-1}, x_2^{-1}).$$

2020 Mathematics Subject Classification. Primary 14E07; Secondary 14J50, 14J45, 14E30.
Keywords. Cremona group, birational transformation, Fano variety, minimal model program.

(ii) Let S be an algebraic variety admitting a generically finite rational map

$$\pi : S \dashrightarrow \mathbb{P}^{n-1}$$

of degree 2. In an affine piece and suitable coordinates, S can be given by the equation $y^2 = f(x_1, \dots, x_{n-1})$. One can associate with (S, π) an involution $\tau \in \mathrm{Cr}_n(\Bbbk)$ acting on $\Bbbk(x_1, \dots, x_{n-1}, y)$ via

$$\tau : (x_1, \dots, x_{n-1}, y) \mapsto \bigl(x_1, \dots, x_{n-1}, f(x_1, \dots, x_{n-1}) \cdot y^{-1}\bigr).$$

If $n = 2$ and S is a hyperelliptic curve, then τ is known as the *de Jonquières involution*.

The study of the Cremona group has a very long history. Basically, it was started in earlier works of A. Cayley and L. Cremona, and since then, this group has been the object of many studies. In these notes, we concentrate on the following particular problem.

Problem 1.1. Describe the structure of finite subgroups of $\mathrm{Cr}_n(\Bbbk)$.

Note, however, that the projective space is not an exceptional variety from the algebro-geometric point of view. So one can ask a similar question replacing $\mathrm{Cr}_n(\Bbbk)$ with the group of birational transformations $\mathrm{Bir}(X)$ of an arbitrary algebraic variety X. Hence it is natural to pose the following problem.

Problem 1.2. Describe the structure of finite subgroups of $\mathrm{Bir}(X)$, where X is an algebraic variety.

We deal with the most recent results related to these problems. Definitely, our survey is not exhaustive.

2. Equivariant minimal model program

In this section, we collect basic facts on the so-called G-minimal model program (abbreviated as G-MMP). This program is the main tool in the study of finite groups of birational transformations. For a detailed exposition, we refer to [89].

Let G be a *finite* group. Following Yu. Manin [68], we say that an algebraic variety X is a *G-variety* if it is equipped with a regular faithful action $G \curvearrowright X$, i.e., if there exists an injective homomorphism $\alpha : G \hookrightarrow \mathrm{Aut}(X)$. A morphism (resp. rational map) $f : X \to Y$ of G-varieties is a *G-morphism* (resp. *G-map*) if there exists a group automorphism $\varphi : G \to G$ such that, for any $g \in G$,

$$f \circ \alpha(g) = \beta\bigl(\varphi(g)\bigr) \circ f,$$

where $\alpha : G \hookrightarrow \mathrm{Aut}(X)$ and $\beta : G \hookrightarrow \mathrm{Aut}(Y)$ are the embeddings corresponding to the actions $G \curvearrowright X$ and $G \curvearrowright Y$, respectively.

For any G-variety X, the action $G \curvearrowright X$ induces an embedding $G \hookrightarrow \mathrm{Aut}_{\Bbbk}(\Bbbk(X))$ to the automorphism group of the field of rational functions $\Bbbk(X)$. Conversely, given any finitely generated extension $\mathbb{K}/\Bbbk$ and any finite subgroup $G \subset \mathrm{Aut}_{\Bbbk}(\mathbb{K})$, there exists a G-variety X and an isomorphism $\Bbbk(X) \simeq_{\Bbbk} \mathbb{K}$ inducing $G \subset \mathrm{Aut}_{\Bbbk}(\mathbb{K})$. Thus, we have the following fact.

Proposition 2.1. *Let $\mathbb{K}/\Bbbk$ be finitely generated field extension. Then there exists a* 1-1 *correspondence between finite subgroups $G \subset \mathrm{Aut}_{\Bbbk}(\mathbb{K})$ considered modulo conjugacy and G-varieties X such that $\Bbbk(X) \simeq_{\Bbbk} \mathbb{K}$ considered modulo G-birational equivalence.*

Recall that a variety X is said to be *rational* if it is birationally equivalent to the projective space $\mathbb{P}^n$ or, equivalently, if the field extension $\Bbbk(X)/\Bbbk$ is purely transcendental.

Corollary. *There exists a* 1-1 *correspondence between finite subgroups $G \subset \mathrm{Cr}_n(\Bbbk)$ considered modulo conjugacy and rational G-varieties X such that $\Bbbk(X) \simeq_{\Bbbk} \mathbb{K}$ considered modulo G-birational equivalence.*

Next, due to the equivariant resolution theorem (see e.g. [1]), it is possible to replace X with a smooth projective model.

Proposition 2.2 (see, e.g., [89, Lemma 14.1.1]). *For any G-variety X, there exists a* smooth projective *G-variety Y that is G-birationally equivalent to X.*

Thus the above considerations allow us to reduce the problem of classification of finite subgroups of $\mathrm{Bir}(X)$ to the study of subgroups in $\mathrm{Aut}(Y)$, where Y is a smooth projective variety. The main difficulty arising here is that this G-variety Y is not unique in its G-birational equivalence class. So, given G-birational equivalence class of algebraic G-varieties, we need to choose some good representative in it. This can be done by means of the G-MMP. The higher-dimensional MMP forces us to consider varieties with certain very mild, so-called terminal singularities.

Definition. A normal variety X has *terminal singularities* if some multiple mK_X of the canonical Weil divisor K_X is Cartier, and for any birational morphism $f : Y \to X$, one can write

$$mK_Y = f^*mK_X + \sum a_i E_i,$$

where E_i are all the exceptional divisors and $a_i > 0$ for all i. The smallest positive m such that mK_X is Cartier is called the *Gorenstein index* of X.

Definition. A G-variety X has *$G\mathbb{Q}$-factorial singularities* if a multiple of any G-invariant Weil divisor on X is Cartier.

It is important to note that terminal singularities lie in codimension ≥ 3. In particular, terminal surface singularities are smooth.

Example ([72, 93]). Let the cyclic group $\boldsymbol{\mu}_r$ act on $\mathbb{A}^4$ diagonally via

$$(x_1, x_2, x_3, x_4) \mapsto (\zeta x_1, \zeta^{-1} x_2, \zeta^a x_3, x_4), \quad \zeta = \zeta_r = \exp(2\pi\,\mathrm{i}/r), \quad \gcd(a, r) = 1.$$

Then for a polynomial $f(u, v)$, the singularity of the quotient

$$\{x_1 x_2 + f(x_3^r, x_4) = 0\}/\boldsymbol{\mu}_r$$

at 0 is terminal whenever it is isolated.

The aim of the G-MMP is to replace a G-variety with another one, which is "minimal" in some sense. As we mentioned above, running the G-MMP we have to consider singular varieties, and the class of terminal $G\mathbb{Q}$-factorial singularities is the smallest class that is closed under the G-MMP.

Definition (For simplicity, we assume that $\Bbbk$ is uncountable). A variety X is *uniruled* if for a general point $x \in X$, there exists a rational curve $C \subset X$ passing through x. A variety X is *rationally connected* if two general points $x_1, x_2 \in X$ can be connected by a rational curve.

Note that a rationally connected surface is rational, and an uniruled surface is birationally equivalent to $C \times \mathbb{P}^1$, where C is a curve.

Definition. Let Y be a G-variety with only terminal $G\mathbb{Q}$-factorial singularities and let $f : Y \to Z$ be a G-equivariant morphism with connected fibers to a lower-dimensional variety Z, where the action of G on Z is not necessarily faithful. Then f is called *G-Mori fiber space* (abbreviated as G-Mfs) if the anti-canonical class $-K_Y$ is f-ample and $\operatorname{rk}\operatorname{Pic}(Y/Z)^G = 1$. If Z is a point, then $-K_Y$ is ample, and Y is called *$G\mathbb{Q}$-Fano variety*. Two-dimensional $G\mathbb{Q}$-Fano varieties are traditionally called *G-del Pezzo surfaces*.

Definition. A G-variety Y is said to be a *G-minimal model* if it has only terminal $G\mathbb{Q}$-factorial singularities and the canonical class K_Y is numerically effective (nef).

It is not difficult to show that the concepts of G-minimal model and G-Mori fiber space are mutually exclusive. Moreover, if $f : Y \to Z$ is a G-Mfs, then its general fiber is rationally connected; hence Y is uniruled. On the other hand, a G-minimal model is never uniruled [70]. The following assertions are usually formulated for varieties without group actions. The corresponding equivariant versions can be easily deduced from non-equivariant ones (see [89]).

Theorem 2.3 ([14]). *Let X be an uniruled G-variety. Then there exists a birational G-map $X \dashrightarrow Y$, where Y has a structure of G-Mfs $f : Y \to Z$.*

Conjecture 2.4. *Let X be a non-uniruled G-variety. Then there exists a birational G-map $X \dashrightarrow Y$, where Y is a G-minimal model.*

The conjecture is known to be true in dimension ≤ 4 (see [73, 99]), as well as in the case where K_X is big [14], and in some other cases. In arbitrary dimension, a weaker notion of quasi-minimal models works quite satisfactory [82].

3. Cremona group of rank 2

The G-MMP for surfaces is much more simple than in higher dimensions. It was developed in the works of Yu. Manin and V. Iskovskikh (see [68]). In the two-dimensional case, the G-MMP works in the category of smooth G-surfaces, and all the birational transformations are contractions of disjoint unions of (-1)-curves. For a G-Mfs $f : Y \to Z$, there are two possibilities:

(i) Z is a point and then Y is a G-del Pezzo surface,

(ii) Z is a curve, any fiber of f is a reduced plane conic and $\operatorname{rk}\operatorname{Pic}(Y)^G = 2$. In this case, f is called *G-conic bundle*.

Thus to study finite subgroups of $\mathrm{Cr}_2(\Bbbk)$, one has to consider the above two classes of G-Mfs's in detail. The classification of del Pezzo surfaces is well known and very short. Hence, to study the case (i) one has to list all finite subgroups $G \subset \operatorname{Aut}(Y)$ satisfying the condition $\operatorname{rk}\operatorname{Pic}(Y)^G = 1$. The full list was obtained by Dolgachev and Iskovskikh [40]. In contrast, the class of conic bundles is huge and consists of an infinite number of families. In this case, a reasonable approach is to find an algorithm of enumerating conic bundles Y/Z together with subgroups $G \subset \operatorname{Aut}(Y/Z)$ satisfying $\operatorname{rk}\operatorname{Pic}(Y)^G = 2$. This also was done by Dolgachev and Iskovskikh [40] (see also [103]). However, even using this algorithm, it is very hard to get a complete list of corresponding groups.

As an example, we present a well-known classical result on the classification of subgroups of order 2 in $\mathrm{Cr}_2(\Bbbk)$. It was obtained by E. Bertini [12] in 1877; however, his arguments were incomplete from a modern point of view. A new rigorous proof was given by L. Bayle and A. Beauville [8].

Theorem 3.1. *Let $G = \{1, \tau\} \subset \mathrm{Cr}_2(\Bbbk)$ be a subgroup of order 2. Then the embedding $G \subset \mathrm{Cr}_2(\Bbbk)$ is induced by one of the following actions on a rational surface X:*

τ	X and τ
1^o Linear involution	$\mathbb{P}^2$
2^o de Jonquières involution of genus $g \geq 1$	$X = \{y_1 y_2 = p(x_1, x_2)\} \subset \mathbb{P}(1, 1, g+1, g+1)$ p is a homogeneous form of degree $2g+2$, τ is the deck involution of the projection $X \overset{2:1}{\dashrightarrow} \mathbb{P}(1, 1, g+1)$, $(x_1, x_2, y_1, y_2) \mapsto (x_1, x_2, y_1 + y_2)$
3^o Geiser involution	$X = \{y^2 = p(x_1, x_2, x_3)\} \subset \mathbb{P}(1, 1, 1, 2)$, p is a homogeneous form of degree 4, τ is the deck involution of the projection $X \xrightarrow{2:1} \mathbb{P}(1, 1, 1) = \mathbb{P}^2$
4^o Bertini involution	$X = \{z^2 = p(x_1, x_2, y)\} \subset \mathbb{P}(1, 1, 2, 3)$, p is a quasihomogeneous form of degree 6, τ is the deck involution of the projection $X \xrightarrow{2:1} \mathbb{P}(1, 1, 2)$

Here $\mathbb{P}(w_1, \dots, w_n)$ denotes the weighted projective space with corresponding weights.

In the cases 1^o, 3^o, and 4^o, the variety X is a del Pezzo surface of degree 9, 2, and 1, respectively. In the case 2^o, the projection $X \dashrightarrow \mathbb{P}(1, 1) = \mathbb{P}^1$ becomes a G-conic bundle after blowing up the indeterminacy points.

The G-MMP was successfully applied for the classification of various classes of finite subgroups in $\mathrm{Cr}_2(\Bbbk)$: groups of prime order [36], p-elementary groups [9], abelian groups [15, 16], and finally, arbitrary groups [40]. Here is another example of classification results.

Theorem 3.2 ([40]). *Let $G \subset \mathrm{Cr}_2(\mathbb{C})$ be a finite simple group. Then G is isomorphic to one of the following:*

$$\mathfrak{A}_5, \quad \mathfrak{A}_6, \quad \mathrm{PSL}_2(\mathbf{F}_7),$$

where $\mathfrak{A}_n$ is the alternating group of degree n and $\mathrm{PSL}_n(\mathbf{F}_q)$ is the projective special linear group over the finite field $\mathbf{F}_q$.

Moreover, if $G \not\simeq \mathfrak{A}_5$, then the embedding $G \subset \mathrm{Cr}_2(\Bbbk)$ is induced by one of the following actions on a del Pezzo surface X:

G	$\lvert G\rvert$	X
$\mathfrak{A}_6$	360	$\mathbb{P}^2$
$\mathrm{PSL}_2(\mathbf{F}_7)$	168	$\mathbb{P}^2$
$\mathrm{PSL}_2(\mathbf{F}_7)$	168	$\{y^2 = x_1^3 x_2 + x_2^3 x_3 + x_3^3 x_1\} \subset \mathbb{P}(1, 1, 1, 2)$

A complete classification of embeddings $\mathfrak{A}_5 \hookrightarrow \mathrm{Cr}_2(\Bbbk)$ can be found in [31].

4. Cremona group of rank 3

The MMP in dimension 3 is more complicated than the two-dimensional one, but still, it is developed very well. In particular, terminal threefold singularities are classified up to analytic equivalence [72,93]. The structure of all intermediate steps of the MMP and Mfs's is also studied relatively well (see [89] for a survey).

For a three-dimensional G-Mori fiber space $f : Y \to Z$, there are three possibilities:

(i) Z is a point, then Y is a (possibly singular) $G\mathbb{Q}$-Fano threefold,

(ii) Z is a curve, then f is called a $G\mathbb{Q}$-del Pezzo fibration,

(iii) Z is a surface, then f is a $G\mathbb{Q}$-conic bundle.

A $G\mathbb{Q}$-conic bundle can be birationally transformed into a *standard G-conic bundle*, i.e., $G\mathbb{Q}$-conic bundle such that both X and Z are smooth [6]. For $G\mathbb{Q}$-del Pezzo fibrations, there are only some partial results of this type (see [35, 66]). Nevertheless, the main difficulty in the application G-MMP to the classification of finite groups of birational transformations is the lack of a complete classification of Fano threefolds with terminal singularities. At the moment, only some very particular classes of $G\mathbb{Q}$-Fano threefolds are studied (see [4, 5, 52, 79, 80, 88] and references therein). Some roundabout methods work in the case of "large" in some sense (in particular, simple) finite groups.

Theorem 4.1 ([78]). *Let $G \subset \mathrm{Cr}_3(\mathbb{C})$ be a finite simple subgroup. Then G is isomorphic to one of the following:*

$$\mathfrak{A}_5, \quad \mathfrak{A}_6, \quad \mathfrak{A}_7, \quad \mathrm{PSL}_2(\mathbf{F}_7), \quad \mathrm{PSL}_2(\mathbf{F}_8), \quad \mathrm{PSp}_4(\mathbf{F}_3),$$

where $\mathrm{PSp}_4(\mathbf{F}_3)$ is the projective symplectic group over $\mathbf{F}_3$. All the possibilities occur.

This classification is a consequence of the following more general result.

Theorem 4.2 ([78]). *Let Y be a rationally connected threefold and let $G \subset \mathrm{Bir}(Y)$ be a finite simple group. If G is not embeddable to $\mathrm{Cr}_2(\mathbb{C})$, then Y is G-birationally equivalent to one of the following $G\mathbb{Q}$-Fano threefolds:*

	G	X	Rational?
1^o	$\mathfrak{A}_7$	$X_6' = \{\sigma_{1,7} = \sigma_{2,7} = \sigma_{3,7} = 0\} \subset \mathbb{P}^5 \subset \mathbb{P}^6$	no
2^o	$\mathfrak{A}_7$	$\mathbb{P}^3$	yes
3^o	$\mathrm{PSp}_4(\mathbf{F}_3)$	$\mathbb{P}^3$	yes
4^o	$\mathrm{PSp}_4(\mathbf{F}_3)$	Burkhardt quartic $X_4^{\mathrm{b}} = \{\sigma_{1,6} = \sigma_{4,6} = 0\} \subset \mathbb{P}^4 \subset \mathbb{P}^5$	yes
5^o	$\mathrm{PSL}_2(\mathbf{F}_8)$	Special Fano threefold $X_{12}^{\mathrm{m}} \subset \mathbb{P}^8$ of genus 7	yes
6^o	$\mathrm{PSL}_2(\mathbf{F}_{11})$	Klein cubic $X_3^{\mathrm{k}} = \{x_1x_2^2 + x_2x_3^2 + \cdots x_5x_1^2 = 0\} \subset \mathbb{P}^4$	no
7^o	$\mathrm{PSL}_2(\mathbf{F}_{11})$	Special Fano threefold $X_{14}^{\mathrm{a}} \subset \mathbb{P}^9$ of genus 8	no

Here $\sigma_{d,k} = \sigma_{d,k}(x_1, \dots, x_k)$ is the elementary symmetric polynomial of degree d in k variables.

Below we outline the proof of Theorem 4.2.

Assume that G is not embeddable to $\mathrm{Cr}_2(\Bbbk)$, i.e., it is not isomorphic to any of the groups listed in Theorem 3.2. First, Proposition 2.2 allows us to assume that the action of G is regularized on some smooth projective G-variety X. By running the equivariant MMP, we may assume that X has a structure of a G-Mfs $f : X \to Z$ (because X is rationally connected). Consider the case $\dim Z > 0$. Since G is a simple group, it must act faithfully on the base Z or on the general fiber F. Since the varieties F and Z are rational, this means that G is contained in the plane Cremona group $\mathrm{Cr}_2(\Bbbk)$. The contradiction proves Theorem 4.2 in the case $\dim Z > 0$.

Hence, we may further assume that Z is a point and X is a $G\mathbb{Q}$-Fano threefold. Consider the case where X is not Gorenstein, i.e., the canonical class K_X is not a Cartier divisor. It turns out that this case does not occur. Let $P_1, \dots, P_n \in X$ be all non-Gorenstein points and let $r_1, \dots, r_n$ be the corresponding Gorenstein indices. Arguments based on Bogomolov–Miyaoka inequality (see [55, 57] and [89, §12]) show that

$$\sum \left(r_i - \frac{1}{r_i} \right) < 24.$$

Hence, $n \le 15$. Then using the classification of transitive actions of simple groups [33] and analyzing the action of stabilizers of P_i, one obtains the only possibility:

- $n = 11$, $G \simeq \mathrm{PSL}_2(\mathbf{F}_{11})$, $r_1 = \cdots = r_n = 2$.

This case is excluded by a more detailed geometric consideration (see [78, §6]).

Thus, we may assume that K_X is a Cartier divisor. In this case, according to [74], the variety X has a smoothing, that is, there exists a one-parameter flat family $\mathfrak{X}/\mathfrak{B} \ni o$ such that the special fiber $\mathfrak{X}_o$ is isomorphic to X, and a general geometric fiber $\mathfrak{X}_t$ is a smooth Fano threefold. Hence some discrete invariants of X, such as the Picard lattice $\mathrm{Pic}(X)$ and the anticanonical degree $-K_X^3$, are the same as for smooth Fano threefolds, which are completely classified (see [52]). Recall that the Fano index $\iota(X)$ of X is the maximal integer that divides the canonical class K_X in the lattice $\mathrm{Pic}(X)$ [52]. By [80], we have $\operatorname{rk}\mathrm{Pic}(X) \le 4$. Since $\mathrm{Pic}(X)^G \simeq \mathbb{Z}$ and a simple group that is not isomorphic to $\mathfrak{A}_5$ cannot have a nontrivial integer representation of dimension ≤ 4, we have $\operatorname{rk}\mathrm{Pic}(X) = 1$. If $\iota(X) \ge 4$ (resp, $\iota(X) = 3$), then X is isomorphic to the projective space $\mathbb{P}^3$ (resp. a quadric in $\mathbb{P}^4$) [52]. Then from the classification of finite subgroups in $\mathrm{PSL}_4(\Bbbk)$ and $\mathrm{PSL}_5(\Bbbk)$, we get cases 2^o and 3^o. Three-dimensional Fano varieties with $\iota(X) = 2$ are called del Pezzo threefolds. G-Fano threefolds of this type were studied in [79]. As a consequence of these results, we get the case of the group $G = \mathrm{PSL}_2(\mathbf{F}_{11})$ acting on the Klein cubic (case 6^o).

Finally, let $\operatorname{Pic}(X) = \mathbb{Z} \cdot K_X$. Recall that in this case, the anticanonical degree is written in the form $-K_X^3 = 2\,\mathrm{g}(X) - 2$, where $\mathrm{g}(X) \in \{2, 3, \dots, 10, 12\}$ [52]. For $\mathrm{g}(X) \leq 5$, the variety X has a natural embedding to a (weighted) projective space as a complete intersection [52]. Using this and some facts from representation theory, we obtain for the group G two cases, 1^o and 4^o. The case $\mathrm{g}(X) = 6$ can be excluded using [37, Corollary 3.11]. For $\mathrm{g}(X) \geq 7$, the variety X must be smooth (see [78, Lemma 5.17] and [88]). Further, using some facts about automorphisms of smooth Fano threefolds [63], we obtain for the group G two possibilities, 5^o and 7^o. This completes our sketch of the proof of Theorem 4.2. ■

A similar technique was applied to the study of finite p-subgroups and quasi-simple subgroups in $\mathrm{Cr}_3(\Bbbk)$ (see [17, 64, 67, 77, 81, 86]).

Note that Theorem 4.2 does not describe *embeddings* of groups $\mathfrak{A}_5$, $\mathfrak{A}_6$, and $\mathrm{PSL}_2(\mathbf{F}_7)$ to the space Cremona group. It is obvious that such embeddings exist, but their full classification should be significantly more difficult. There are only some partial results in this direction (see e.g. [26–29, 62]).

5. Jordan property

The methods and results of [40] show that one cannot expect a reasonable classification of all finite subgroups of Cremona groups of higher rank. Thus it is natural to concentrate on the study of general properties of these subgroups. Recall the following two famous results by C. Jordan and H. Minkowski.

Theorem 5.1 ([53]). *There exists a function* $\mathrm{j}(n)$ *such that for any finite subgroup* $G \subset \mathrm{GL}_n(\mathbb{C})$*, there exists a normal abelian subgroup* $A \subset G$ *of index at most* $\mathrm{j}(n)$.

Theorem 5.2 ([69]). *There exists a function* $\mathrm{b}(n)$ *such that for every finite subgroup* $G \subset \mathrm{GL}_n(\mathbb{Q})$*, one has* $|G| \leq \mathrm{b}(n)$.

J.-P. Serre [94, 96] asked if these properties hold for Cremona groups. Complete answers to these questions were given in [82, 83] (see below). The following very convenient definitions were suggested by V. L. Popov [75].

Definition. • A group Γ is *Jordan* if there exists a constant $\mathrm{j}(\Gamma)$ such that any finite subgroup $G \subset \Gamma$ has a normal abelian subgroup A of index $[G : A] \leq \mathrm{j}(\Gamma)$.

- A group Γ is *bounded* (or satisfy *bfs* property) if there exists a constant $\mathrm{b}(\Gamma)$ such that for any finite subgroup $G \subset \Gamma$, one has $|G| \leq \mathrm{b}(\Gamma)$.

Rationally connected varieties

Theorem 5.3 ([13, 83]). *Let X be a rationally connected variety. Then* $\operatorname{Bir}(X)$ *is Jordan. Moreover,* $\operatorname{Bir}(X)$ *is* uniformly *Jordan; that is, the constant* $\mathrm{j}(\operatorname{Bir}(X))$ *depends only on* $\dim(X)$.

As a consequence, we obtain that the group $\mathrm{Cr}_n(\Bbbk)$ is Jordan.

Originally, Theorem 5.3 was proved modulo the so-called BAB conjecture (in a weak form), which is now settled by C. Birkar:

Theorem 5.4 ([13]). *Fix $d > 0$. The set of all Fano varieties X of dimension at most d with at worst terminal singularities form a bounded family; i.e., they are parameterized by a scheme of finite type.*

It follows from Theorem 5.3 that there is a constant $L = L(n)$ such that for any rationally connected variety X of dimension n and for any prime $p > L(n)$, every finite p-subgroup of $\mathrm{Bir}(X)$ is abelian and generated by at most n elements (see [83]). Recently this result was essentially improved by Jinsong Xu [104]; he showed that $L(n) = n + 1$. The proof is based on a result by O. Haution [47]. Thus we have the following theorem.

Theorem 5.5. *Let X be a rationally connected variety of dimension n and let $G \subset \mathrm{Bir}(X)$ be a finite p-subgroup. If $p > n + 1$, then G is abelian and is generated by at most n elements.*

The results of Theorems 5.3 and 5.5 were applied in the proof of Jordan property of local fundamental groups of log terminal singularities [20, 71].

Varieties over non-closed fields

Theorem 5.6 ([13, 82]). *Let X be a variety over a field $\Bbbk$ of characteristic 0, which is finitely generated over $\mathbb{Q}$. Then the group $\mathrm{Bir}(X)$ is bfs.*

Similar to Theorem 5.3, the proof of this result is based on the BAB conjecture (Theorem 5.4).

In the case $X = \mathbb{P}^2$, an explicit bound was obtained in [94] (see also [41]) in terms of cyclotomic invariants of the field $\Bbbk$. Theorem 5.6 can be reformulated in an algebraic form, which gives the positive answer to a question of J.-P. Serre [96].

Theorem 5.6a. *Let $\mathbb{K}$ be a finitely generated field over $\mathbb{Q}$. Then the group $\mathrm{Aut}(\mathbb{K})$ is bfs.*

Jordan constants. Define the *Jordan constant* of a group Γ as the number $\mathrm{j}(\Gamma)$ that appears in the definition of Jordan property. The *weak Jordan constant* $\bar{\mathrm{j}}(\Gamma)$ of Γ is the minimal j such that for any finite subgroup $G \subset \Gamma$, there exists an abelian (not necessarily normal) subgroup $A \subset G$ such that $[G : A] \le j$. Easy group-theoretic arguments show that

$$\bar{\mathrm{j}}(\Gamma) \le \mathrm{j}(\Gamma) \le \bar{\mathrm{j}}(\Gamma)^2.$$

The exact value of the Jordan constant is known only for the Cremona group of rank two: $\mathrm{j}(\mathrm{Cr}_2(\Bbbk)) = 7200$ (see [105]). On the other hand, weak Jordan constants are easier to compute. It was proved in [84] that

$$\bar{\mathrm{j}}(\mathrm{Cr}_2) = 288, \quad \bar{\mathrm{j}}(\mathrm{Cr}_3) = 10368.$$

Moreover, the inequality $\bar{\mathrm{j}}(\mathrm{Bir}(X)) \leq 10368$ holds for any rationally connected threefold X.

Jordan property of arbitrary varieties. It turns out that the group of birational transformations of an algebraic variety is not always Jordan. The first example was discovered by Yu. Zarhin.

Example ([106]). Let C be an elliptic curve and let $X = C \times \mathbb{P}^1$. Then the group $\mathrm{Bir}(X)$ is not Jordan.

On the other hand, the exceptions as above are very rare.

Theorem 5.7 (V. L. Popov [75]). *Let X be an algebraic surface. The group* $\mathrm{Bir}(X)$ *is not Jordan if and only if X is birationally equivalent to $\mathbb{P}^1 \times C$, where C is an elliptic curve.*

The proof of this theorem given in [75] essentially uses a result of I. Dolgachev, which in turn is based on the classification of algebraic surfaces. Later, Theorem 5.7 was generalized to higher dimensions with classification independent proofs.

Theorem 5.8 ([82]). *Let X be an algebraic variety. Then the following assertions hold.*

(i) *If X either is non-uniruled or has irregularity $q(X) = 0$, then* $\mathrm{Bir}(X)$ *is Jordan.*

(ii) *If X is non-uniruled and $q(X) = 0$, then* $\mathrm{Bir}(X)$ *is bfs.*

Similar to Theorems 5.6 and 5.3, the proof of Theorem 5.8(i) is based on the boundedness of terminal Fano varieties (Theorem 5.4).

In dimension three, there is the following much more precise result.

Theorem 5.9 ([85]). *Let X be a three-dimensional algebraic variety. Then* $\mathrm{Bir}(X)$ *is not Jordan if and only if either*

(i) *X is birationally equivalent to $C \times \mathbb{P}^2$, where C is an elliptic curve, or*

(ii) *X is birationally equivalent to $S \times \mathbb{P}^1$, where S is one of the following:*

- *a surface of Kodaira dimension $\varkappa(S) = 1$ such that the Jacobian fibration of the pluricanonical map $\phi\colon S \to B$ is locally trivial;*
- *S is either an abelian or bielliptic surface (and $\varkappa(S) = 0$).*

Below we explain the main idea of the proof of the necessity. So we assume that $\operatorname{Bir}(X)$ is not Jordan. By Theorems 5.3 and 5.8, the variety X is uniruled, but it is not rationally connected. Hence there exists a map $X \dashrightarrow Z$ with rationally connected fibers (so-called maximal rationally connected fibration) such that Z is not uniruled and $\dim(Z) = 1$ or 2 (see [56]). We have a natural exact sequence

$$1 \to \operatorname{Bir}(X_\eta) \to \operatorname{Bir}(X) \to \operatorname{Bir}(Z),$$

where X_η is the generic scheme-theoretic fiber. Since X_η is rationally connected and Z is not uniruled, the groups $\operatorname{Bir}(X_\eta)$ and $\operatorname{Bir}(Z)$ must be Jordan. Then group-theoretic arguments show that both groups $\operatorname{Bir}(X_\eta)$ and $\operatorname{Bir}(Z)$ are not bfs (see, e.g., [82, Lemma 2.8]). In the case where Z is a curve, this implies that Z is elliptic, and applying the following fact with $\mathbb{K} = \Bbbk(Z)$ and $S := X_\eta$, we obtain that X is birationally equivalent to $Z \times \mathbb{P}^2$.

Proposition 5.10 ([85]). *Let $\mathbb{K}$ be a field containing all roots of* 1 *and let S be a surface over $\mathbb{K}$ such that S is not $\mathbb{K}$-rational, S is $\bar{\mathbb{K}}$-rational, and $S(\mathbb{K}) \neq \varnothing$. Then the group* $\operatorname{Bir}(S)$ *is bfs.*

Note that the condition of the existence of a $\mathbb{K}$-point on S in the above statement is important. The groups of (birational) automorphisms of geometrically rational surfaces without rational points were studied in the series of papers [100–102].

Now assume that Z is a surface. According to the main result of [7], the threefold X is birationally equivalent to $Z \times \mathbb{P}^1$. By Theorem 5.8 we have $q(Z) > 0$. Thus in the case $\varkappa(Z) = 0$, the surface Z must be either abelian or bielliptic. Since the group $\operatorname{Bir}(Z)$ is not finite in our case, Z cannot be a surface of general type. Consider the case $\varkappa(Z) = 1$. Then the pluricanonical map $\phi : Z \to B$ is a $\operatorname{Bir}(Z)$-equivariant elliptic fibration. Let

$$\operatorname{Jac}(\phi) : E \to B$$

be the corresponding Jacobian fibration. The automorphism group $\operatorname{Aut}(Z_\eta)$ of the generic fiber Z_η over B is embedded to $\operatorname{Bir}(Z)$ as a normal subgroup. Analyzing singular fibers, one can conclude that $\operatorname{Aut}(Z_\eta)$ is of finite index in $\operatorname{Bir}(Z)$. In turn, $\operatorname{Aut}(Z_\eta)$ has a subgroup $\operatorname{Aut}'(Z_\eta)$ of index at most 6 isomorphic to the group of $\Bbbk(B)$-points of E_η. Assume that the fibration $\operatorname{Jac}(\phi)$ is not locally trivial. Then by the functional version of Mordell–Weil theorem, known as Lang–Néron theorem (see, e.g., [32]), the group of $\Bbbk(B)$-points of E_η is finitely generated, and in particular, the torsion subgroup of the group of points of E_η is finite. This implies that $\operatorname{Aut}'(Z_\eta)$ is finite. ■

6. Invariants and rigidity

The most important part of the classification of finite subgroups in $\mathrm{Bir}(X)$ is to distinguish conjugacy classes.

Problem 6.1. Let $G, G' \subset \mathrm{Bir}(X)$ be finite subgroups such that $G \simeq G'$. How can one conclude that G and G' are *not* conjugate?

This is equivalent to the following.

Problem 6.1a. Let X and X' be G-varieties. How can one conclude that X and X' are *not* G-birational?

Below we describe a few approaches to solve the above problems. Note, however, that there are no universal methods.

Fixed point locus. Let X be a smooth projective G-variety. By $\mathrm{Fix}(X, G)$, we denote the set of G-fixed points. It is not difficult to show (see [87]) that $\mathrm{Fix}(X, G)$ has at most one codimension one component that is not uniruled. Denote this component by $\mathrm{F}^{\mathrm{nu}}(X, G)$. This is a natural birational invariant in the category of smooth projective G-varieties.

Proposition 6.2 ([87]). *Let X and X' be smooth projective G-varieties. If X and X' are G-birational, then $\mathrm{F}^{\mathrm{nu}}(X, G_0)$ and $\mathrm{F}^{\mathrm{nu}}(X', G_0)$ are birational for any subgroup $G_0 \subset G$.*

If $G_0 \subset G$ is a normal subgroup, then the set $\mathrm{F}^{\mathrm{nu}}(X, G_0)$ (if it is not empty) has a structure of (G/G_0)-variety. Clearly, the birational type of this (G/G_0)-variety is also a birational invariant (cf. [16]).

Example. According to Theorem 3.1 for subgroups $G \subset \mathrm{Cr}_2(\Bbbk)$ of order 2, we have one of the following possibilities:

	Involution $\tau \in G$	$\mathrm{F}^{\mathrm{nu}}(X, G)$
1^o	Linear on $\mathbb{P}^2$	$\varnothing$
2^o	de Jonquières of genus $g \geq 1$	Hyperelliptic curve of genus g
3^o	Geiser	Non-hyperelliptic curve of genus 3
4^o	Bertini	Special non-hyperelliptic curve of genus 4

Thus the curve $\mathrm{F}^{\mathrm{nu}}(X, G)$ distinguishes conjugacy classes in this case. The same assertion is true for subgroups of prime order [36], but it fails in general [15].

Cohomological invariants. It is not difficult to see that for a smooth projective G-variety X, the cohomology group

$$H^1\big(G,\ \mathrm{Pic}(X)\big)$$

is a G-birational invariant (see [19]). More generally, we say that G-varieties X and X' are *stably* G-birationally equivalent if for some n and m the products $X \times \mathbb{P}^n$ and $X' \times \mathbb{P}^m$ are G-birationally equivalent, where the action of G on $\mathbb{P}^n$ and $\mathbb{P}^m$ is supposed to be trivial. Then we have the following theorem.

Theorem 6.3 ([19]). *Let X and X' be smooth projective G-varieties. If X and X' are stably G-birationally equivalent, then*

$$H^1\big(G, \mathrm{Pic}(X)\big) \simeq H^1\big(G,\ \mathrm{Pic}(X')\big).$$

Surprisingly, in some cases, the invariant $H^1(G, \mathrm{Pic}(X))$ can be computed in terms of G-fixed locus.

Theorem 6.4 ([19]). *Let G be a cyclic group of prime order p and let X be a smooth projective rational G-surface. Assume that* $\mathrm{F}^{\mathrm{nu}}(X, G)$ *is a curve of genus g. Then*

$$H^1\big(G, \mathrm{Pic}(X)\big) \simeq (\mathbb{Z}/p\mathbb{Z})^{2g}.$$

This theorem was slightly generalized with a more conceptual proof in [97]. Another cohomological invariant which is called *Amitsur group* was introduced in [17].

As a consequence of Theorem 6.4, one can see that involutions from different families in Theorem 3.1 are not stably conjugate in $\mathrm{Cr}_2(\Bbbk)$. Note, however, that $H^1(G, \mathrm{Pic}(X))$ is a discrete invariant. For example, stable conjugacy of involutions whose curves $\mathrm{F}^{\mathrm{nu}}(X, G)$ are non-isomorphic but have the same genus is not known.

A natural question that arises here is to find examples of subgroups in $\mathrm{Cr}_n(\Bbbk)$ that are stably conjugate but not conjugate. This question is similar to the birational Zariski problem [11].

Example. Let $G = \mathfrak{S}_3 \times \boldsymbol{\mu}_2$. There are two embeddings of this group into the Cremona group $\mathrm{Cr}_2(\Bbbk)$ induced by the following actions:

(i) action on $\mathbb{P}^2 = \{x_1 + x_2 + x_3 = 0\} \subset \mathbb{P}^3$ by permutation and reversing signs;

(ii) action on the sextic del Pezzo surface $\{y_1y_2y_3 = y_1'y_2'y_3'\} \subset \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ by permutation and taking inverses.

It was shown in [65] that these two subgroups in $\mathrm{Cr}_2(\Bbbk)$ are stably conjugate; in fact, they are conjugate in $\mathrm{Cr}_4(\Bbbk)$. On the other hand, they are not conjugate [51].

Here is another example of this kind, which was pointed out to us by Yuri Tschinkel.

Example ([92]). Let V and W be faithful linear representations of G with $\dim(V) = \dim(W) = n$. Assume that the images of G in $\mathrm{GL}(V)$ and $\mathrm{GL}(W)$ do not contain non-identity scalar matrices. Then by a variant of the no-name lemma [39], we have the following G-birational equivalences of G-varieties:

$$\mathbb{P}(V) \times \Bbbk^{n+1} \underset{\mathrm{bir}}{\sim} V \times W \underset{\mathrm{bir}}{\sim} \mathbb{P}(W) \times \Bbbk^{n+1},$$

where $\Bbbk^{n+1}$ is viewed as the trivial representation. Hence G-varieties V and W are stably G-birationally equivalent. On the other hand, it may happen that they are *not* G-birationally equivalent.

For example, Reichstein and Youssin [92] showed that the *determinant* of the action in the tangent space at a fixed point of a finite abelian group, up to sign, is a birational invariant of the action. This allowed them to produce nonbirational linear actions, e.g., of groups $\boldsymbol{\mu}_{p^n}$ on $\mathbb{P}^n$, with $p \geq 5$. Many new examples of nonbirational linear actions were given in [60, Section 10-11]; these are based on new invariants introduced in [61] (see also [46, 59]). These invariants take into account more refined information about the action on subvarieties with nontrivial abelian stabilizers.

A prime number p is said to be a *torsion prime* for the group $\mathrm{Bir}(X)$ if there is a finite abelian p-subgroup $G \subset \mathrm{Bir}(X)$ not contained in any algebraic torus of $\mathrm{Bir}(X)$ [76]. Note that if a group G is contained in an algebraic torus $T \subset \mathrm{Bir}(X)$, then for any smooth projective birational model Y of X on which T acts biregularly, we have $H^1(G, \mathrm{Pic}(Y)) = 0$. Then by Theorem 6.3, the inequality $H^1(G, \mathrm{Pic}(Y)) \neq 0$ for a finite p-subgroup $G \subset \mathrm{Aut}(Y)$ implies that a prime number p is a torsion prime for $\mathrm{Bir}(Y)$ and for $\mathrm{Bir}(Y \times \mathbb{P}^n)$ for any n. Using Theorem 6.4 and the classification [36], one can immediately see that the set of all torsion primes for $\mathrm{Cr}_2(\Bbbk)$ is equal to $\{2, 3, 5\}$, and the numbers 2, 3, and 5 are torsion primes for $\mathrm{Cr}_n(\Bbbk)$ for any $n \geq 2$. This fact was proved in [76] by using another argument. In the case $n \geq 3$, the collection of all torsion primes for $\mathrm{Cr}_n(\Bbbk)$ is unknown.

Maximal singularities method. The maximal singularities method is the most powerful tool to study birational maps between Mfs's. It goes back to the works of G. Fano and even earlier works of other Italian geometers. However, the first application of this technique with rigorous proofs appeared much later in the breakthrough paper of Manin and Iskovskikh [49]. For an introduction to the "standard," non-equivariant maximal singularities method, we refer to the book [90]. Below we outline very briefly an equivariant version of the method.

Definition ([40, Definition 7.10], [29, Definition 3.1.1]). A $G\mathbb{Q}$-Fano variety X is said to be *G-birationally rigid* if given birational G-map $\Phi : X \dashrightarrow X^\sharp$ to the total space of another G-Mfs $X^\sharp/Z^\sharp$, there exists a birational G-selfmap $\psi : X \dashrightarrow X$

such that the composition $\Phi \circ \psi : X \dashrightarrow X^\sharp$ is an isomorphism (in particular, $Z^\sharp$ is a point; i.e., $X^\sharp$ is also a $G\mathbb{Q}$-Fano variety).

A $G\mathbb{Q}$-Fano variety X is said to be *G-birationally superrigid* if any birational G-map $\Phi : X \dashrightarrow X^\sharp$ to the total space of another G-Mfs $X^\sharp/Z^\sharp$ is an isomorphism.

The maximal singularities method allows to check G-birational (super)rigidity using only internal geometry of the original variety, without considering all other G-Mfs's. We need the following technical definition which has become common nowadays.

Definition. Let X be a normal variety, let $\mathcal{M}$ be a linear system of Weil divisors on X without fixed components, and let λ be a rational number. We say that the pair $(X, \lambda\mathcal{M})$ is *canonical* if some multiple $m(K_X + \lambda M)$ is Cartier, where $M \in \mathcal{M}$, and for any birational morphism $f : Y \to X$, one can write

$$m(K_Y + \lambda\mathcal{M}_Y) = f^*m(K_X + \lambda\mathcal{M}) + \sum a_i E_i,$$

where $\mathcal{M}_Y$ is the birational transform of $\mathcal{M}$, E_i are prime exceptional divisors, and $a_i \geq 0$ for all i.

In the surface case, the canonical property is very easy to check: a pair $(X, \lambda\mathcal{M})$ is canonical if and only if

$$\mathrm{mult}_P(\mathcal{M}) \leq 1/\lambda$$

for any point $P \in X$.

Now, suppose that a $G\mathbb{Q}$-Fano variety X is not G-birationally superrigid. Then the Noether–Fano inequality [34, Theorem 4.2] implies the existence of a G-invariant linear system $\mathcal{M}$ on X without fixed components such that the pair $(X, \lambda\mathcal{M})$ is not canonical, where $\lambda \in \mathbb{Q}$ is taken so that $K_X + \lambda\mathcal{M}$ is numerically trivial. Moreover, any $\mathcal{M}$ as above defines a birational G-map $X \dashrightarrow X^\sharp$ to the total space of a G-Mfs $X^\sharp/Z^\sharp$. To show the existence or non-existence of such $\mathcal{M}$, one needs to analyze the geometry of the variety X carefully.

Example. Let X be a del Pezzo surface of degree 1. Assume that X is a G-del Pezzo with respect to some group $G \subset \mathrm{Aut}(X)$. This means that G acts on X so that $\mathrm{rk}\,\mathrm{Pic}(X)^G = 1$. For example, this holds for any subgroup $G \subset \mathrm{Aut}(X)$ containing the Bertini involution. Let $\mathcal{M}$ be a G-invariant linear subsystem without fixed components. Since $\mathrm{Pic}(X)^G = \mathbb{Z} \cdot K_X$, we have $\mathcal{M} \subset |-nK_X|$ for some $n > 0$. Suppose that the pair $(X, \frac{1}{n}\mathcal{M})$ is not canonical. Then $\mathrm{mult}_P(\mathcal{M}) > n$. Since $\mathcal{M}$ has no fixed components,

$$n^2 = (-nK_X)^2 = \mathcal{M}^2 \geq \big(\mathrm{mult}_P(\mathcal{M})\big)^2 > n^2.$$

The contradiction shows that X is G-birationally superrigid.

Similar arguments show that any G-del Pezzo surface X of degree ≤ 3 is G-birationally rigid. Moreover, it is G-birationally superrigid if and only if G has no orbits of length $\leq K_X^2 - 2$ on X. In particular, $\mathrm{PSL}_2(\mathbf{F}_7)$-del Pezzo surface from Theorem 3.2 is G-birationally superrigid.

Example. All the $G\mathbb{Q}$-Fano threefolds from Theorem 4.2 are G-birationally superrigid [17, 28, 30]. In particular, different embeddings of $\mathrm{PSp}_4(\mathbf{F}_3)$ and $\mathrm{PSL}_2(\mathbf{F}_{11})$ are not conjugate in $\mathrm{Cr}_3(\Bbbk)$.

There is another relevant and very important notion called G-solidity [25]. For Fano varieties without group action, this notion has been introduced earlier by Shokurov [98] (who called solid Fano varieties primitive) and by Ahmadinezhad and Okada [2].

Definition ([25]). A G-Fano variety X is G-*solid* if X is not G-birational to a G-Mfs with a positive dimensional base.

For example, a G-del Pezzo surface X of degree 4 is G-solid if and only if G has no fixed points on X [40, §8].

A part of the maximal singularities method is the so-called Sarkisov program [34, 45]. It allows us to decompose any birational map between Mfs's into a composition of elementary ones. Refer to [50] for an explicit description of this program in dimension two and to [31] for examples and applications.

7. Application: Essential dimension

The notion of the essential dimension of a finite group G, denoted by $\mathrm{ed}(G)$, was introduced by Buhler and Reichstein [21]. Informally, $\mathrm{ed}(G)$ is the minimal number of algebraic parameters needed to describe a faithful representation. More precisely, given a faithful linear representation V of G viewed as a G-variety, the *essential dimension* $\mathrm{ed}(G, V)$ is the minimal value of $\dim(X)$, where X is taken from the set of all G-varieties admitting dominant rational G-equivariant map $V \dashrightarrow X$. It can be shown that $\mathrm{ed}(G, V)$ does not depend on V, so we can omit V in the notation. It is easy to see that $\mathrm{ed}(G) = 1$ if and only if G is cyclic or dihedral of order $2n$ where n is odd. Finite groups of essential dimension ≤ 2 have been classified [43].

The essential dimension of symmetric groups $\mathfrak{S}_n$ is important because it is equal to the minimal number of parameters needed to describe the general polynomial of degree n modulo Tschirnhaus transformations [21]. The values of $\mathrm{ed}(\mathfrak{S}_n)$, as well as of $\mathrm{ed}(\mathfrak{A}_n)$, are known for $n \leq 7$, and bounds exist for any n as follows.

Theorem 7.1 ([21,42]). *If $n \geq 6$, then*

$$n-3 \geq \operatorname{ed}(\mathfrak{S}_n) \geq \lfloor n/2 \rfloor,$$

$$\operatorname{ed}(\mathfrak{S}_n) \geq \operatorname{ed}(\mathfrak{A}_n) \geq \begin{cases} \frac{n}{2} & \text{if } n \text{ is even,} \\ 2\lfloor \frac{n+2}{4} \rfloor & \text{if } n \text{ is odd.} \end{cases}$$

In many cases, the computations of $\operatorname{ed}(G)$ use the machinery of G-varieties. As an example, following Serre [95], we show that $\operatorname{ed}(\mathfrak{A}_6) = 3$. Let V be the standard six-dimensional permutation representation of $\mathfrak{A}_6$. There exists an equivariant open embedding $V \subset (\mathbb{P}^1)^6$. On the other hand, the group $\operatorname{PSL}_2(\Bbbk)$ also acts on $(\mathbb{P}^1)^6$ so that the two actions commute. Hence we have a dominant rational $\mathfrak{A}_6$-map

$$V \hookrightarrow (\mathbb{P}^1)^6 \to (\mathbb{P}^1)^6 / \operatorname{PSL}_2(\Bbbk),$$

where $(\mathbb{P}^1)^6 / \operatorname{PSL}_2(\Bbbk)$ is a birational quotient. Since $\dim((\mathbb{P}^1)^6 / \operatorname{PSL}_2(\Bbbk)) = 3$, we have $\operatorname{ed}(\mathfrak{A}_6) \leq 3$. Thus it is sufficient to show that $\operatorname{ed}(\mathfrak{A}_6)$ is not equal to 2. If so, there exists a dominant rational G-map $V \dashrightarrow X$ to a surface which must be rational. According to Theorem 3.2, we may assume that $X = \mathbb{P}^2$. But in this case, a Sylow 3-subgroup $S \subset \mathfrak{A}_6$ is abelian and acts without fixed points on $\mathbb{P}^2$. On the other hand, S has a fixed point on V, and the same should be true for the image of any rational S-map to a projective variety [58]. Therefore, $\operatorname{ed}(\mathfrak{A}_6) = 3$ as claimed.

Using similar arguments and the classification of embeddings of $\mathfrak{A}_7$ to groups of birational transformations of rationally connected threefolds (Theorem 4.2), A. Duncan proved that $\operatorname{ed}(\mathfrak{A}_7) = \operatorname{ed}(\mathfrak{S}_7) = 4$ [42].

Denote by $\operatorname{rdim}(G)$ (resp. $\operatorname{cdim}(G)$) the minimal dimension of faithful representations of G (resp. the smallest n such that G is embeddable to $\operatorname{Cr}_n(\Bbbk)$). It immediately follows from the definition that

$$\operatorname{ed}(G) \leq \operatorname{rdim}(G).$$

If G is a p-group, then the equality holds $\operatorname{ed}(G) = \operatorname{rdim}(G)$ [54]. In general, this equality fails, but there is a bound in terms of Jordan constants.

Theorem 7.2 ([91]). $\operatorname{rdim}(G) \leq \operatorname{ed}(G) \cdot \operatorname{j}(\operatorname{ed}(G))$, *where* $\operatorname{j}(n)$ *is the Jordan constant.*

I. Dolgachev conjectured that $\operatorname{ed}(G) \geq \operatorname{cdim}(G)$ (see [44]). It would be interesting to test this conjecture for the group $G = \operatorname{PSL}_2(\mathbf{F}_{11})$. In fact, we have

$$3 \leq \operatorname{ed}\big(\operatorname{PSL}_2(\mathbf{F}_{11})\big) \leq 4$$

by Theorem 3.2 and because the group $\operatorname{PSL}_2(\mathbf{F}_{11})$ is simple and has a faithful five-dimensional representation. Assuming Dolgachev's conjecture, by Theorem 4.2 we would have $\mathrm{ed}(\mathrm{PSL}_2(\mathbf{F}_{11})) = 4$. But this is unknown. See [44] for interesting discussions. The computation of the essential dimension of $\mathrm{PSL}_2(\mathbf{F}_{11})$ should complete Beauville's classification of finite simple groups of essential dimension ≤ 3 [10].

Acknowledgments. The author would like to thank Alexander Duncan, Constantin Shramov, Yuri Tschinkel, Mikhail Zaidenberg, and the referee for helpful comments on the original version of this paper.

Funding. This work was performed at the Steklov International Mathematical Center and supported by the Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-15-2022-265).

References

[1] D. Abramovich and J. Wang, Equivariant resolution of singularities in characteristic 0. *Math. Res. Lett.* **4** (1997), no. 2-3, 427–433 Zbl 0906.14005 MR 1453072

[2] H. Ahmadinezhad and T. Okada, Birationally rigid Pfaffian Fano 3-folds. *Algebr. Geom.* **5** (2018), no. 2, 160–199 Zbl 1407.14038 MR 3769891

[3] M. Alberich-Carramiñana, *Geometry of the Plane Cremona Maps*. Lecture Notes in Math. 1769, Springer, Berlin, 2002 Zbl 0991.14008 MR 1874328

[4] A. Avilov, Automorphisms of threefolds that can be represented as an intersection of two quadrics. *Mat. Sb.* **207** (2016), no. 3, 3–18 Zbl 1370.14036 MR 3507481

[5] A. Avilov, Automorphisms of singular three-dimensional cubic hypersurfaces. *Eur. J. Math.* **4** (2018), no. 3, 761–777 Zbl 1423.14096 MR 3851116

[6] A. A. Avilov, Existence of standard models of conic bundles over algebraically nonclosed fields. *Mat. Sb.* **205** (2014), no. 12, 3–16 Zbl 1317.14091 MR 3309386

[7] T. Bandman and Y. G. Zarhin, Jordan groups, conic bundles and abelian varieties. *Algebr. Geom.* **4** (2017), no. 2, 229–246 Zbl 1388.14047 MR 3620637

[8] L. Bayle and A. Beauville, Birational involutions of $\mathbf{P}^2$. *Asian J. Math.* **4** (2000), no. 1, 11–17; Kodaira's issue Zbl 1055.14012 MR 1802909

[9] A. Beauville, p-elementary subgroups of the Cremona group. *J. Algebra* **314** (2007), no. 2, 553–564 Zbl 1126.14017 MR 2344578

[10] A. Beauville, Finite simple groups of small essential dimension. In *Trends in Contemporary Mathematics*, pp. 221–228, Springer INdAM Ser. 8, Springer, Cham, 2014 Zbl 1386.14173 MR 3586401

[11] A. Beauville, J.-L. Colliot-Thélène, J.-J. Sansuc, and P. Swinnerton-Dyer, Variétés stablement rationnelles non rationnelles. *Ann. of Math. (2)* **121** (1985), no. 2, 283–318 Zbl 0589.14042 MR 786350

[12] E. Bertini, Ricerche sulle trasformazioni univoche involutorie nel piano. *Annali di Mat. Pura Appl.* **8** (1877), 254–287 Zbl 09.0578.02

[13] C. Birkar, Singularities of linear systems and boundedness of Fano varieties. *Ann. of Math. (2)* **193** (2021), no. 2, 347–405 Zbl 1469.14085 MR 4224714

[14] C. Birkar, P. Cascini, C. D. Hacon, and J. McKernan, Existence of minimal models for varieties of log general type. *J. Amer. Math. Soc.* **23** (2010), no. 2, 405–468 Zbl 1210.14019 MR 2601039

[15] J. Blanc, Linearisation of finite abelian subgroups of the Cremona group of the plane. *Groups Geom. Dyn.* **3** (2009), no. 2, 215–266 Zbl 1170.14009 MR 2486798

[16] J. Blanc, Elements and cyclic subgroups of finite order of the Cremona group. *Comment. Math. Helv.* **86** (2011), no. 2, 469–497 Zbl 1213.14029 MR 2775137

[17] J. Blanc, I. Cheltsov, A. Duncan, and Y. Prokhorov, Finite quasisimple groups acting on rationally connected threefolds. 2018, arXiv:1809.09226

[18] J. Blanc, S. Lamy, and S. Zimmermann, Quotients of higher-dimensional Cremona groups. *Acta Math.* **226** (2021), no. 2, 211–318 Zbl 07378146 MR 4281381

[19] F. Bogomolov and Y. Prokhorov, On stable conjugacy of finite subgroups of the plane Cremona group, I. *Cent. Eur. J. Math.* **11** (2013), no. 12, 2099–2105 Zbl 1286.14016 MR 3111709

[20] L. Braun, S. Filipazzi, J. Moraga, and R. Svaldi, The Jordan property for local fundamental groups. *Geom. Topol.* **26** (2022), no. 1, 283–319 Zbl 07525902 MR 4404879

[21] J. Buhler and Z. Reichstein, On the essential dimension of a finite group. *Compositio Math.* **106** (1997), no. 2, 159–179 Zbl 0905.12003 MR 1457337

[22] S. Cantat, The Cremona group in two variables. In *European Congress of Mathematics*, pp. 211–225, Eur. Math. Soc., Zürich, 2013 Zbl 1364.14009 MR 3469123

[23] S. Cantat, The Cremona group. In *Algebraic Geometry: Salt Lake City 2015*, pp. 101–142, Proc. Sympos. Pure Math. 97, Amer. Math. Soc., Providence, RI, 2018 Zbl 1451.14037 MR 3821147

[24] S. Cantat and S. Lamy, Normal subgroups in the Cremona group. *Acta Math.* **210** (2013), no. 1, 31–94 Zbl 1278.14017 MR 3037611

[25] I. Cheltsov, A. Dubouloz, and T. Kishimoto, Toric G-solid Fano threefolds. 2020, arXiv:2007.14197

[26] I. Cheltsov, V. Przyjalkowski, and C. Shramov, Burkhardt quartic, Barth sextic, and the icosahedron. *Int. Math. Res. Not. IMRN* **2019** (2019), no. 12, 3683–3703 Zbl 1454.14036 MR 3973105

[27] I. Cheltsov and C. Shramov, Three embeddings of the Klein simple group into the Cremona group of rank three. *Transform. Groups* **17** (2012), no. 2, 303–350 Zbl 1272.14013 MR 2921069

[28] I. Cheltsov and C. Shramov, Five embeddings of one simple group. *Trans. Amer. Math. Soc.* **366** (2014), no. 3, 1289–1331 Zbl 1291.14060 MR 3145732

[29] I. Cheltsov and C. Shramov, *Cremona Groups and the Icosahedron*. Monogr. Res. Notes Math., CRC Press, Boca Raton, FL, 2016 Zbl 1328.14003 MR 3444095

[30] I. Cheltsov and C. Shramov, Finite collineation groups and birational rigidity. *Selecta Math. (N.S.)* **25** (2019), no. 5, Paper No. 71 Zbl 1440.14061 MR 4036497

[31] I. A. Cheltsov, Two local inequalities. *Izv. Ross. Akad. Nauk Ser. Mat.* **78** (2014), no. 2, 167–224 Zbl 1329.14021 MR 3234821

[32] B. Conrad, Chow's K/k-image and K/k-trace, and the Lang-Néron theorem. *Enseign. Math. (2)* **52** (2006), no. 1-2, 37–108 Zbl 1133.14028 MR 2255529

[33] J. H. Conway, R. T. Curtis, S. P. Norton, R. A. Parker, and R. A. Wilson, $\mathbb{ATLAS}$ *of Finite Groups. Maximal Subgroups and Ordinary Characters for Simple Groups. With computational assistance from J. G. Thackray*. Oxford University Press, Eynsham, 1985 Zbl 0568.20001 MR 827219

[34] A. Corti, Factoring birational maps of threefolds after Sarkisov. *J. Algebraic Geom.* **4** (1995), no. 2, 223–254 Zbl 0866.14007 MR 1311348

[35] A. Corti, Del Pezzo surfaces over Dedekind schemes. *Ann. of Math. (2)* **144** (1996), no. 3, 641–683 Zbl 0902.14026 MR 1426888

[36] T. de Fernex, On planar Cremona maps of prime order. *Nagoya Math. J.* **174** (2004), 1–28 Zbl 1062.14019 MR 2066103

[37] O. Debarre and A. Kuznetsov, Gushel–Mukai varieties: classification and birationalities. *Algebr. Geom.* **5** (2018), no. 1, 15–76 Zbl 1408.14053 MR 3734109

[38] J. Déserti, *Some Properties of the Cremona Group*. Ensaios Mat. 21, Sociedade Brasileira de Matemática, Rio de Janeiro, 2012 Zbl 1276.14020 MR 2934616

[39] I. V. Dolgachev, Rationality of fields of invariants. In *Algebraic Geometry, Bowdoin, 1985 (Brunswick, Maine, 1985)*, pp. 3–16, Proc. Sympos. Pure Math. 46, Amer. Math. Soc., Providence, RI, 1987 Zbl 0659.14009 MR 927970

[40] I. V. Dolgachev and V. A. Iskovskikh, Finite subgroups of the plane Cremona group. In *Algebra, Arithmetic, and Geometry: in Honor of Yu. I. Manin. Vol. I*, pp. 443–548, Progr. Math. 269, Birkhäuser, Boston, MA, 2009 Zbl 1219.14015 MR 2641179

[41] I. V. Dolgachev and V. A. Iskovskikh, On elements of prime order in the plane Cremona group over a perfect field. *Int. Math. Res. Not. IMRN* **2019** (2009), no. 18, 3467–3485 Zbl 1188.14007 MR 2535007

[42] A. Duncan, Essential dimensions of A_7 and S_7. *Math. Res. Lett.* **17** (2010), no. 2, 263–266 Zbl 1262.14057 MR 2644373

[43] A. Duncan, Finite groups of essential dimension 2. *Comment. Math. Helv.* **88** (2013), no. 3, 555–585 Zbl 1300.14044 MR 3093503

[44] A. Duncan and Z. Reichstein, Versality of algebraic group actions and rational points on twisted varieties. *J. Algebraic Geom.* **24** (2015), no. 3, 499–530 Zbl 1327.14210 MR 3344763

[45] C. D. Hacon and J. McKernan, The Sarkisov program. *J. Algebraic Geom.* **22** (2013), no. 2, 389–405 Zbl 1267.14024 MR 3019454

[46] B. Hassett, A. Kresch, and Y. Tschinkel, Symbols and equivariant birational geometry in small dimensions. In *Rationality of Varieties*, pp. 201–236, Progr. Math. 342, Birkhäuser, Cham, 2021 MR 4383699

[47] O. Haution, Fixed point theorems involving numerical invariants. *Compos. Math.* **155** (2019), no. 2, 260–288 Zbl 1441.14156 MR 3905117

[48] H. P. Hudson, *Cremona Transformations in Plane and Space*. Cambridge University Press, 1927

[49] V. A. Iskovskih and J. I. Manin, Three-dimensional quartics and counterexamples to the Lüroth problem. *Mat. Sb. (N.S.)* **86(128)** (1971), 140–166 MR 0291172

[50] V. A. Iskovskikh, Factorization of birational mappings of rational surfaces from the point of view of Mori theory. *Uspekhi Mat. Nauk* **51** (1996), no. 4(310), 3–72
Zbl 0914.14005 MR 1422227

[51] V. A. Iskovskikh, Two nonconjugate embeddings of the group $S_3 \times Z_2$ into the Cremona group. *Tr. Mat. Inst. Steklova* **241** (2003), no. Teor. Chisel, Algebra i Algebr. Geom., 105–109 Zbl 1078.14015 MR 2024046

[52] V. A. Iskovskikh and Y. G. Prokhorov, Fano varieties. In *Algebraic Geometry, V*, pp. 1–247, Encyclopaedia Math. Sci. 47, Springer, Berlin, 1999 Zbl 0912.14013
MR 1668579

[53] M. C. Jordan, Mémoire sur les équations différentielles linéaires à intégrale algébrique. *J. Reine Angew. Math.* **84** (1878), 89–215 Zbl 09.0096.01 MR 1581645

[54] N. A. Karpenko and A. S. Merkurjev, Essential dimension of finite p-groups. *Invent. Math.* **172** (2008), no. 3, 491–508 Zbl 1200.12002 MR 2393078

[55] Y. Kawamata, Boundedness of $\mathbf{Q}$-Fano threefolds. In *Proceedings of the International Conference on Algebra, Part 3 (Novosibirsk, 1989)*, pp. 439–445, Contemp. Math. 131, Amer. Math. Soc., Providence, RI, 1992 Zbl 0785.14024 MR 1175897

[56] J. Kollár, Y. Miyaoka, and S. Mori, Rationally connected varieties. *J. Algebraic Geom.* **1** (1992), no. 3, 429–448 Zbl 0780.14026 MR 1158625

[57] J. Kollár, Y. Miyaoka, S. Mori, and H. Takagi, Boundedness of canonical $\mathbf{Q}$-Fano 3-folds. *Proc. Japan Acad. Ser. A Math. Sci.* **76** (2000), no. 5, 73–77 Zbl 0981.14016
MR 1771144

[58] J. Kollár and E. Szabó, Fixed points of group actions and rational maps. Appendix to “Essential dimensions of algebraic groups and a resolution theorem for G-varieties” by Z. Reichstein and B. Youssin. *Canad. J. Math.* **52** (2000), no. 5, 1054–1056
Zbl 1044.14023 MR 1782331

[59] M. Kontsevich, V. Pestun, and Y. Tschinkel, Equivariant birational geometry and modular symbols. 2019, arXiv:1902.09894

[60] A. Kresch and Y. Tschinkel, Equivariant Burnside groups and representation theory. 2021, arXiv:2108.00518

[61] A. Kresch and Y. Tschinkel, Equivariant birational types and Burnside volume. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **23** (2022), no. 2, 1013–1052

[62] I. Krylov, Families of embeddings of the alternating group of rank 5 into the Cremona group. 2020, arXiv:2005.07354

[63] A. G. Kuznetsov, Y. G. Prokhorov, and C. A. Shramov, Hilbert schemes of lines and conics and automorphism groups of Fano threefolds. *Jpn. J. Math.* **13** (2018), no. 1, 109–185 Zbl 1406.14031 MR 3776469

[64] A. A. Kuznetsova, Finite 3-Subgroups in the Cremona Group of Rank 3. *Mat. Zametki* **108** (2020), no. 5, 725–749 Zbl 1469.14028 MR 4169699

[65] N. Lemire, V. L. Popov, and Z. Reichstein, Cayley groups. *J. Amer. Math. Soc.* **19** (2006), no. 4, 921–967 Zbl 1103.14026 MR 2219306

[66] K. Loginov, Standard models of degree 1 del Pezzo fibrations. *Mosc. Math. J.* **18** (2018), no. 4, 721–737 Zbl 1420.14037 MR 3914112

[67] K. Loginov, A note on 3-subgroups in the space Cremona group. *Comm. Algebra* **50** (2022), no. 9, 3704–3714 MR 4442466

[68] J. I. Manin, Rational surfaces over perfect fields. II. *Mat. Sb. (N.S.)* **72 (114)** (1967), 161–192 Zbl 0182.23701 MR 0225781

[69] H. Minkowski, Zur Theorie der positiven quadratischen Formen. *J. Reine Angew. Math.* **101** (1887), 196–202 Zbl 19.0189.01 MR 1580123

[70] Y. Miyaoka and S. Mori, A numerical criterion for uniruledness. *Ann. of Math. (2)* **124** (1986), no. 1, 65–69 Zbl 0606.14030 MR 847952

[71] J. Moraga, On a toroidalization for klt singularities. 2021, arXiv:2106.15019

[72] S. Mori, On 3-dimensional terminal singularities. *Nagoya Math. J.* **98** (1985), 43–66 Zbl 0589.14005 MR 792770

[73] S. Mori, Flip theorem and the existence of minimal models for 3-folds. *J. Amer. Math. Soc.* **1** (1988), no. 1, 117–253 Zbl 0649.14023 MR 924704

[74] Y. Namikawa, Smoothing Fano 3-folds. *J. Algebraic Geom.* **6** (1997), no. 2, 307–324 Zbl 0906.14019 MR 1489117

[75] V. L. Popov, On the Makar–Limanov, Derksen invariants, and finite automorphism groups of algebraic varieties. In *Affine Algebraic Geometry*, pp. 289–311, CRM Proc. Lecture Notes 54, Amer. Math. Soc., Providence, RI, 2011 Zbl 1242.14044 MR 2768646

[76] V. L. Popov, Tori in the Cremona groups. *Izv. Ross. Akad. Nauk Ser. Mat.* **77** (2013), no. 4, 103–134 Zbl 1278.14065 MR 3135700

[77] Y. Prokhorov, p-elementary subgroups of the Cremona group of rank 3. In *Classification of Algebraic Varieties*, pp. 327–338, EMS Ser. Congr. Rep., Eur. Math. Soc., Zürich, 2011 MR 2779480

[78] Y. Prokhorov, Simple finite subgroups of the Cremona group of rank 3. *J. Algebraic Geom.* **21** (2012), no. 3, 563–600 Zbl 1257.14011 MR 2914804

[79] Y. Prokhorov, G-Fano threefolds, I. *Adv. Geom.* **13** (2013), no. 3, 389–418 Zbl 1291.14024 MR 3100917

[80] Y. Prokhorov, G-Fano threefolds, II. *Adv. Geom.* **13** (2013), no. 3, 419–434 Zbl 1291.14025 MR 3100918

[81] Y. Prokhorov, 2-elementary subgroups of the space Cremona group. In *Automorphisms in Birational and Affine Geometry*, pp. 215–229, Springer Proc. Math. Stat. 79, Springer, Cham, 2014 Zbl 1327.14070 MR 3229353

[82] Y. Prokhorov and C. Shramov, Jordan property for groups of birational selfmaps. *Compos. Math.* **150** (2014), no. 12, 2054–2072 Zbl 1314.14022 MR 3292293

[83] Y. Prokhorov and C. Shramov, Jordan property for Cremona groups. *Amer. J. Math.* **138** (2016), no. 2, 403–418 Zbl 1343.14010 MR 3483470

[84] Y. Prokhorov and C. Shramov, Jordan constant for Cremona group of rank 3. *Mosc. Math. J.* **17** (2017), no. 3, 457–509 Zbl 1411.14018 MR 3711004

[85] Y. Prokhorov and C. Shramov, Finite groups of birational selfmaps of threefolds. *Math. Res. Lett.* **25** (2018), no. 3, 957–972 Zbl 1423.14094 MR 3847342

[86] Y. Prokhorov and C. Shramov, p-subgroups in the space Cremona group. *Math. Nachr.* **291** (2018), no. 8-9, 1374–1389 Zbl 1423.14099 MR 3817323

[87] Y. G. Prokhorov, On birational involutions of $\mathbb{P}^3$. *Izv. Ross. Akad. Nauk Ser. Mat.* **77** (2013), no. 3, 199–222 Zbl 1282.14025 MR 3098794

[88] Y. G. Prokhorov, Singular Fano manifolds of genus 12. *Mat. Sb.* **207** (2016), no. 7, 101–130 Zbl 1372.14032 MR 3535377

[89] Y. G. Prokhorov, Equivariant minimal model program. *Uspekhi Mat. Nauk* **76** (2021), no. 3(459), 93–182 Zbl 07402603 MR 4265398

[90] A. Pukhlikov, *Birationally Rigid Varieties*. Math. Surveys Monogr. 190, American Mathematical Society, Providence, RI, 2013 Zbl 1297.14001 MR 3060242

[91] Z. Reichstein, The Jordan property of Cremona groups and essential dimension. *Arch. Math. (Basel)* **111** (2018), no. 5, 449–455 Zbl 06951230 MR 3859426

[92] Z. Reichstein and B. Youssin, A birational invariant for algebraic group actions. *Pacific J. Math.* **204** (2002), no. 1, 223–246 Zbl 1054.14062 MR 1905199

[93] M. Reid, Young person's guide to canonical singularities. In *Algebraic Geometry, Bowdoin, 1985 (Brunswick, Maine, 1985)*, pp. 345–414, Proc. Sympos. Pure Math. 46, Amer. Math. Soc., Providence, RI, 1987 Zbl 0634.14003 MR 927963

[94] J.-P. Serre, A Minkowski-style bound for the orders of the finite subgroups of the Cremona group of rank 2 over an arbitrary field. *Mosc. Math. J.* **9** (2009), no. 1, 193–208 Zbl 1203.14017 MR 2567402

[95] J.-P. Serre, Le groupe de Cremona et ses sous-groupes finis. Séminaire Bourbaki. Volume 2008/2009. Exposés 997–1011. *Astérisque* **332** (2010), 75–100 Zbl 1257.14012 MR 2648675

[96] J.-P. Serre, Problems for the Edinburgh workshop on Cremona groups. 2010

[97] E. Shinder, The Bogomolov–Prokhorov invariant of surfaces as equivariant cohomology. *Bull. Korean Math. Soc.* **54** (2017), no. 5, 1725–1741 Zbl 1398.14042 MR 3708807

[98] V. V. Shokurov, Problems about Fano varieties. In *Birational Geometry of Algebraic Varieties, Open Problems*, pp. 30–32, Katata, 1988

[99] V. V. Shokurov, Prelimiting flips. *Tr. Mat. Inst. Steklova* **240** (2003), 82–219 Zbl 1082.14019 MR 1993750

[100] A. Shramov, Birational automorphisms of Severy–Brauer surfaces. *Mat. Sb.* **211** (2020), no. 3, 169–184 Zbl 1445.14025 MR 4070054

[101] C. Shramov, Automorphisms of cubic surfaces without points. *Internat. J. Math.* **31** (2020), no. 11, Article No. 2050083 Zbl 1461.14057 MR 4163640

[102] C. Shramov, Finite groups acting on Severi–Brauer surfaces. *Eur. J. Math.* **7** (2021), no. 2, 591–612 Zbl 1473.14025 MR 4256964

[103] V. I. Tsygankov, Equations of G-minimal conic bundles. *Mat. Sb.* **202** (2011), no. 11, 103–160 Zbl 1261.14006 MR 2907201

[104] J. Xu, A remark on the rank of finite p-groups of birational automorphisms. *C. R. Math. Acad. Sci. Paris* **358** (2020), no. 7, 827–829 Zbl 1454.14037 MR 4174816

[105] E. Yasinsky, The Jordan constant for Cremona group of rank 2. *Bull. Korean Math. Soc.* **54** (2017), no. 5, 1859–1871 Zbl 1428.14024 MR 3708815

[106] Y. G. Zarhin, Theta groups and products of abelian and rational varieties. *Proc. Edinb. Math. Soc. (2)* **57** (2014), no. 1, 299–304 Zbl 1311.14018 MR 3165026

Yuri Prokhorov
Steklov Mathematical Institute, 8 Gubkina street, 119991 Moscow; and AG Laboratory, HSE,
6 Usacheva str., 119048 Moscow, Russia; prokhoro@mi-ras.ru

DOI 10.4171/8ECM/03

Propositional proof complexity

Alexander A. Razborov

Abstract. Propositional proof complexity studies efficient provability of those statements that can be expressed in propositional logic, in various proof systems, and under various notions of "efficiency." Proof systems and statements of interest come from a variety of sources that, besides logic and combinatorics, include many other areas like combinatorial optimization and practical SAT solving. This article is an expanded version of the ECM talk in which we will attempt to convey some basic ideas underlying this vibrant area.

1. General overview

Like with many other areas in theoretical computer science, the framework of propositional proof complexity can be easily explained to a mathematically advanced high school student. In fact, its core definitions are so easy to give that we prefer to interlace them with the discussion rather than to separate the two.

Definition 1.1 (preliminaries). We fix a set of *Boolean* (that is, 0-1 valued, where 0 stands for FALSE and 1 stands for TRUE) variables. A *literal* is either a variable x or its negation that will be denoted by $\bar{x}$. The alternate notation $\neg x$ is also used in the literature, and sometimes we will use the uniform notation x^a, $a \in \{0,1\}$, where $x^1 \overset{\text{def}}{=} x$ and $x^0 \overset{\text{def}}{=} \bar{x}$. A *clause* C is a disjunction of literals: $C = x_{i_1}^{a_1} \vee \cdots \vee x_{i_w}^{a_w}$ in which no variable appears twice. A *conjunctive normal form* (CNF in what follows) is a conjunction of clauses $\tau = C_1 \wedge \cdots \wedge C_m$, often identified with the set $\{C_1, \ldots, C_m\}$ of which it is comprised. Whenever n appears as a subscript in τ_n, it always stands for the number of variables.

One very important complexity measure for this article is width. The *width* of a clause is the number of literals w in it. The *width* of a CNF is the maximal width of a clause in it. A k*-CNF* is a CNF of width $\leq k$. An *assignment* (sometimes called *truth assignment*) is a mapping $\alpha : V \to \{0,1\}$. It is naturally extended to literals, clauses, and CNFs. For example, for the assignment α given by $\alpha(x_1) = 1, \alpha(x_2) = 0,$

2020 Mathematics Subject Classification. Primary 03F20; Secondary 68Q17.
Keywords. Proof complexity, lower bounds, resolution.

$\alpha(x_3) = 1$, and $\alpha(x_4) = 0$ we have $\alpha(\bar{x}_2) = 1, \alpha(\bar{x}_1 \vee x_2 \vee x_4) = 0, \alpha(x_2 \vee x_3) = 1$, and $\alpha((\bar{x}_1 \vee x_2 \vee x_4) \wedge (x_2 \vee x_3)) = 0$. A CNF τ is *satisfiable* if there exists at least one truth assignment α such that $\alpha(\tau) = 1$; α itself is called then a *satisfying assignment*. Otherwise, τ is *unsatisfiable*.

The algorithmic problem SATISFIABILITY of determining whether a given CNF τ is satisfiable or not is NP-complete. In fact, it is the most fundamental NP-complete problem, as well as historically the first [4, Chapter 2.4]. It is central to the field of computational complexity.

In proof complexity, accents are slightly shifted. Instead of *deciding* whether τ is satisfiable or not, we want a *proof* of the answer, and we are interested in the resources necessary to *represent* this proof, in most cases abstracting away from the complexity of *finding* it.

If we want to certify the *satisfiability* of τ, then the task becomes trivial: a proof consists of a satisfying assignment α itself. Let us note in passing, however, that this immediately changes once we impose additional restrictions on the verification process. Significantly oversimplifying, any proof can be written in a special "holographic" form such that, once submitted, its validity can be checked by verifying a small number of "lemmas" in it, selected randomly. This leads to one of the most beautiful and difficult topics in the computational complexity theory called *probabilistically checkable proofs* (PCPs). Unfortunately, this topic is way beyond the scope of our article, so we refer the reader to [4, Chapter 11].

The main question of interest in the propositional proof complexity is how to prove efficiently that a CNF τ is *unsatisfiable*.

Remark 1.2. If we view τ itself as representing a mathematical statement, then what we call a "proof" is actually its *refutation*. The reason why this change of direction is very convenient will become clear below. For now, let us just warn the reader that the terminology is unfortunately rather inconsistent. Say, an unsatisfiable CNF may be called in the literature "a contradiction" or even "a tautology." In what follows we also may at times be sloppy about this.

Remark 1.3. We have restricted ourselves to CNFs mostly because this class is sufficiently broad to easily encompass virtually all statements we will be interested in. It will also be a must when we discuss so-called weak proof systems. But sometimes people do consider more complicated Boolean (and not only Boolean in fact) expressions to be proved/refuted.

Once we have determined that our goal is to study efficient provability of (the unsatisfiability of) CNFs, the next task is to define what we mean by a "proof system." In the most abstract form this definition was given in the seminal paper [25] by Cook and Reckhow.

Definition 1.4 (proof systems). Let UNSAT be the set of all unsatisfiable CNFs. A *propositional proof system* is a surjective polynomial-time computable function $P : \{0,1\}^* \twoheadrightarrow \text{UNSAT}$, where $\{0,1\}^*$ is the set of all finite binary strings.

The intuition is that proofs are encoded by binary strings w and the function P first checks whether w is a legitimate proof (and outputs something trivial like $x \wedge \bar{x}$ if it is not). Then $P(w)$ is the theorem that the proof w proves, and the surjectivity of P is the property of a proof system called *completeness*: every unsatisfiable CNF possesses at least one proof (that is, refutation).

In this abstract form, the definition has turned out very useful for general "structural" studies in proof complexity; see e.g. [40, 55]. But the main focus of our article is on *concrete* fixed proof systems that are interesting for some external reasons.

Before branching into specifics, we still can give a few crucial definitions at this level of generality.

Definition 1.5 (size complexity). For a propositional proof system P and $\tau \in \text{UNSAT}$, let $S_P(\tau \vdash 0)$ be the *size complexity* of τ defined as the minimal possible bit length $|w|$ of $w \in \{0,1\}^*$ such that $P(w) = \tau$. The proof system is *p-bounded* if $S_P(\tau \vdash 0)$ is bounded by a polynomial in the bit length $|\tau|$ of τ itself.

Whether p-bounded proof systems P exist is the main motivating question of proof complexity. It is not hard to see, however, that in this generality (that is without any other restrictions on P) this is equivalent to a major question in the computational complexity.

Theorem 1.6 ([25]). *A p-bounded proof system P exists if and only if* $\mathsf{NP} = co - \mathsf{NP}$.

The following will allow us to compare different proof systems according to their strength and arrange them into a hierarchy.

Definition 1.7 (simulation and equivalence). A proof system P *p-simulates* another proof system Q if there is a polynomial-time computable function s such that the following diagram commutes:

$$\begin{tikzcd}
\{0,1\}^* \arrow[rr, "s"] \arrow[dr, "Q"'] & & \{0,1\}^* \arrow[dl, "P"] \\
& \text{UNSAT} &
\end{tikzcd}$$

Informally, any Q-proof w can be efficiently converted into a P-proof $s(w)$ of the same theorem; note that the poly-time computability of s automatically implies that $|s(w)|$ is bounded by a polynomial in $|w|$. Two proof systems are *p-equivalent* if they p-simulate each other.

We now move on to consider concrete proof systems.

2. Strong proof systems

The classification of proof systems into “weak” and “strong” is loosely defined and it is not universally agreed upon. Roughly speaking, a proof system P is considered strong if we cannot rule out that it is p-bounded, and it is sufficiently widely believed that this inability is in a sense inherent. We will see below at least one proof system in the “gray area.”

Strong proof systems are usually associated with original motivations for the propositional proof complexity coming from mathematical logic, more exactly from the study of weak theories of bounded arithmetic. On this subject I will be very brief (as I was in my ECM presentation); the reader willing to learn more about these fascinating connections with classical proof theory is referred to the monographs [21, 24, 38, 39]. As before, we precede the discussion with a few definitions.

Definition 2.1 (Frege, informal). Take any textbook in the mathematical logic. It will most likely begin with a description of propositional calculus given as a Hilbert-style proof system. That is, it will contain finitely many *axiom schemes* like $A \Rightarrow (A \vee B)$ or $A \vee \neg A$ and *inference rules* like

$$\frac{A \qquad A \Longrightarrow B}{B} \quad \text{(modus ponens)}.$$

Here $A, B, C, \ldots$ are placeholders for which one can substitute an arbitrary Boolean formula. This is a *Frege proof system.*

Remark 2.2. One very important distinction in propositional proof complexity is whether we consider proofs in the tree-like form or allow arbitrary directed acyclic graphs (DAGs). In other words, do we allow intermediate “lemmas” to be used more than once or not? This is of little significance in the classical proof theory since any DAG can be expanded into a tree (if you need to use a lemma more than once, just repeat its inference). But this may result in an exponential increase in the size of the proof and, as a result, for weak proof systems we should strictly distinguish between the two possibilities. It is a non-trivial fact that for the Frege proof system these two versions are actually p-equivalent [37].

Textbooks in the mathematical logic seldom use the same finite sets of axioms and inference rules, and in many cases they use even different sets of Boolean connectives (e.g., we have just seen the implication $\Rightarrow$ that was not in our original de Morgan language $\{\neg, \wedge, \vee\}$). But it turns out that modulo polynomial equivalence all these choices are immaterial.

Theorem 2.3 ([60]). *Any two Frege proof systems, understood as Hilbert-style complete proof systems based on a finite number of axiom schemes and inference rules, are p-equivalent.*

Remark 2.2, along with Theorem 2.3, strongly suggests that the concept of Frege proof system is very robust and hence natural. This system is denoted by F; thus, the function $S_{\mathrm{F}}(\tau \vdash 0)$ is well defined up to a polynomial.

Definition 2.4 (extended Frege, informal). An *extended Frege proof system*, denoted by EF, is the Frege proof system augmented with the following *extension rule*. This rule allows to introduce at any moment a *fresh new* propositional variable x_A as an abbreviation for a formula A. The proof then may proceed using also the *extension axioms* $x_A \equiv A$, and this can happen recursively.

All that has been said about the robustness of Frege proof systems fully applies to EF as well. That is, $S_{\mathrm{EF}}(\tau \vdash 0)$ does not depend on whether it is DAG-like or tree-like or on the choice of the underlying Frege proof system.

Returning to the connections with weak arithmetic, these theories capture various complexity classes in the sense that, roughly speaking, all functions provably total in such a theory T are precisely the functions from that class. Total provability of a function $f(x)$ means that it is representable by a formula $A(x, y)$ such that T proves[1] $(\exists! y \leq t) A(x, y)$ and $A(n, f(n))$ is true for any n. It involves the bounded existential quantifier $(\exists y \leq t)$ in front. It turns out that if we are interested in the provability, in the same theory T, of "almost" quantifier-free formulas (for experts, Δ_0^b formulas), then such formulas can be translated into an increasing sequence $\{\tau_n\}$ of propositional formulas. Then the provability of the original statement in T becomes "essentially equivalent" to the *efficient* provability of its propositional translation in a proof system P_T naturally associated with T. In most cases, it simply means that $S_{P_T}(\tau_n \vdash 0)$ is bounded by a polynomial in n, and F and EF happen to correspond to the most central systems of weak arithmetic. For more details see the monographs [21, 24, 38, 39] already cited above.

Showing that F or EF are not p-bounded is widely believed to be out of reach of the current methods and in general even more difficult than solving notorious open problems in the computational complexity like $\mathsf{NC}^1 \neq \mathsf{P}$ or $\mathsf{P} \neq \mathsf{NP}$. They are paradigmatic strong systems in our informal classification. A good explanation, both philosophical and heuristical, predicates that the most important feature of a proof system P is the *expressive* (in the computational sense of the word) *power of its lines*, that is what computational power is afforded to concepts underlying auxiliary statements appearing in the proof. For a Frege proof system lines are just arbitrary Boolean expressions, and they correspond to the complexity class NC^1. For the extended Frege we get arbitrary Boolean circuits, and those correspond to the class P. It appears to be even more difficult, and usually way more difficult, to analyze what one can *prove* using concepts definable by a complexity class than what we can *compute* within this

[1] The exclamation mark stands for "unique."

class. I am not aware of any good explanation of this fact, this is just what has been happening in the area so far.

The final observation I would like to offer about F and EF strongly differentiates the propositional proof complexity from its sister discipline, circuit complexity. Let me remind the reader that in the latter field we know that almost all Boolean functions are hard; this is the famous *Shannon effect* (see e.g. [36, Chapter 1.4]). Moreover, we strongly believe that a variety of very natural Boolean functions corresponding to NP-complete problems are hard. That is, we have a host of *natural* and *explicit* candidates for hardness; we simply do not know yet how to prove that they are actually hard.

Nothing like that happens in proof complexity, and potential candidates are few and far between. In [16], Bonet, Buss, and Pitassi set off for a slightly modified task to find good tautologies *separating* F and EF, that is hard for F, easy for EF. Their own conclusion, to which I fully concur, was that "no particularly good or convincing examples are known." If we relax the requirement and simply ask for tautologies that would be good candidates to show that the Frege proof system is not p-bounded, I believe there are only two principles that have passed the test of time even by loose standards, and both are equally plausible to be hard for EF.

The first is random k-CNFs. Pick up sufficiently many clauses of width k at random. Then the resulting CNF will be in UNSAT w.h.p. but there does not appear to be even a good starting point for F or EF (or, for that matter, any other conceivable proof system) to certify the unsatisfiability in particular instances.

The second kind of examples is made by CNFs expressing facts like "NP does not have small size circuits." For an extensive discussion of these statements and their relations to other topics in proof and computational complexities I refer the reader to [59, Section 1].

All proof systems in the remainder of this article will be weak ("potentially" weak in one case).

3. Benchmarks

In computer science, a "benchmark" usually stands for a "good" standardized test, or a family of tests, used to run competing pieces of software or hardware to compare these pieces to each other. In the propositional proof complexity, it also turns out that there is a handful of combinatorial principles, expressible as unsatisfiable CNFs, that wander from one framework to another and appear in papers over and over again. This uniformity turns out indispensable for understanding the general picture and trying out new methods for proving both lower and upper bounds that can be then applied to many other tautologies.

For now, let us define two such principles that, arguably, are the most prominent and popular ones (we will see a few more later in the text).

Definition 3.1 (pigeonhole principle). Let $m > n$ be integers; introduce propositional variables x_{ij} ($i \in [m]$, $j \in [n]$). The *pigeonhole principle* (sometimes also called the *Dirichlet principle*, particularly in the Russian literature) is the unsatisfiable CNF PHP_n^m made of the following clauses:

- $x_{i1} \vee \cdots \vee x_{in}$, for all "pigeons" $i \in [m]$ ("every pigeon flies to a hole");
- $\bar{x}_{ij} \vee \bar{x}_{i'j}$, for all pairs of different "pigeons" $i \neq i' \in [m]$ and all "holes" $j \in [n]$ ("no two pigeons fly to the same hole").

This is the so-called "basic" pigeonhole principle. One can also add to it dual axioms, the *functionality axioms* $\bar{x}_{ij} \vee \bar{x}_{ij'}$ or the *surjectivity axioms* $x_{1j} \vee \cdots x_{mj}$. Varying the parameter $m = m(n)$ as well, we obtain a large family of pigeonhole principles and, somewhat surprisingly, they may display very different behavior with respect to the same proof system. I refer the reader to the survey [57] *entirely* devoted to the pigeonhole principle, with the warning that several important results have been obtained since its release.

Our second principle was introduced in [62] that, arguably, was the earliest paper in the propositional proof complexity.

Definition 3.2 (Tseitin tautologies). Let $G = (V, E)$ be a simple graph with odd number of vertices. Introduce propositional variables x_e, one variable per edge $e \in E$. The *Tseitin tautology* $\mathrm{Tseitin}(G)$ is the following system of linear equations over $\mathbb{F}_2$:

$$\bigoplus_{e \ni v} x_e = 1 \quad (v \in V)$$

($\oplus$ is the *parity function*, addition mod 2). This principle says that in any spanning sub-graph of G (determined by the values $(x_e \mid e \in E)$) there exists a vertex of even degree.

Remark 3.3. The attentive reader may have observed that, as stated, $\mathrm{Tseitin}(G)$ is not a CNF. It is usually converted into a CNF by straightforwardly expanding all parities into a family of clauses. For example, $x \oplus y \oplus z$ is the same as $(x \vee y \vee z) \wedge (\bar{x} \vee \bar{y} \vee z) \wedge (\bar{x} \vee y \vee \bar{z}) \wedge (x \vee \bar{y} \vee \bar{z})$. This expansion incurs an increase in the size of the contradiction by a factor of $2^{\Delta-1}$, where Δ is the maximal vertex degree of G. This is often unacceptable when Δ is large so in most applications Tseitin tautologies are considered only for constant-degree graphs that are also sometimes assumed to be regular (all vertices have the same degree).

It turns out that Tseitin tautologies work best when G is a good expander. There are several standard definitions of graph expansion, very much equivalent in the bounded-degree case. Here we only recall that of *edge* expansion, as the most convenient for our purposes.

Definition 3.4 (edge expansion). For a graph $G = (V, E)$ and $S \subseteq V$, let $E(S, \overline{S})$ be the set of all cross-edges between S and $\overline{S} \overset{\mathrm{def}}{=} V \setminus S$. The *(edge) expansion* $c(G)$ of G is defined as

$$c(G) \overset{\mathrm{def}}{=} \min \left\{ \frac{|E(S, \overline{S})|}{|S|} \mid S \subseteq V,\ 1 \leq |S| \leq |V|/2 \right\}.$$

4. Bounded-depth Frege

In this section, we will discuss several restrictions of the Frege proof system to which Theorem 2.3 no longer applies. On the other hand, the remark from Section 2 (that a proof system is largely determined by the expressive power of its lines) applies in full, and a bounded-depth Frege proof system is determined by the bound on depth and the set of propositional connectives (the *basis*) it employs.

Let us start with the standard *de Morgan basis* $\{\neg, \vee, \wedge\}$. The first useful observation is that using de Morgan rules $\neg(A \vee B) \equiv (\neg A \wedge \neg B)$, $\neg(A \wedge B) \equiv (\neg A \vee \neg B)$, any formula can be converted into a formula with *tight negations*, that is a formula in which negations occur only at the variables.

Definition 4.1 (bounded-depth Frege). The *logical depth* of a $\{\neg, \vee, \wedge\}$-formula with tight negations is the maximum number of *alternations* $\vee \vee \cdots \wedge \wedge \wedge \cdots \vee$ of $\vee$ and $\wedge$, where the maximum is taken over all paths from the root of the formula to its leaves (i.e., literals). Alternatively, we can allow disjunctions and conjunctions with an arbitrary number of arguments, and then logical depth becomes the ordinary depth (= height) of the tree representing the formula.

The *depth-d* Frege proof system F_d is the fragment of a Frege proof system over $\{\neg, \wedge, \vee\}$ in which all lines are required to have logical depth $\leq d$.

As in Definition 2.1, we do not specify axiom schemes and inference rules since all "reasonable" choices lead to p-equivalent systems. For most of this section, we view the depth d as arbitrarily large but fixed constant; this is what we mean by "bounded depth."

The corresponding circuit class, made of sequences of Boolean functions that can be computed by circuits of polynomial size and bounded depth, is well known in circuit complexity. It is denoted by AC^0 and by now it is relatively well understood, beginning with exponential size lower bounds for bounded-depth circuits proved in the celebrated series of papers [1, 34, 63].

While lower bounds for F_d were established with the same general method (so-called *restrictions*), this required to overcome a great deal of additional difficulties as compared to the case of circuits. But before we start discussing concrete results I find it prudent to make the following disclaimer.

This short article is not intended to be a comprehensive survey in the propositional proof complexity or its sub-areas; for more extended account, see e.g. the monograph [39] and historical remarks made therein. Its purpose is limited to giving the first impression about the area to non-specialists, and my choice of illustrating examples is necessarily incomplete and subjective.

That said, the first lower bounds for bounded-depth Frege were proved for the pigeonhole principle.

Theorem 4.2 ([41, 48]). $S_{\mathrm{F}_d}(\mathrm{PHP}_n^{n+1} \vdash 0) \geq \exp(\Omega(n^{1/5^d}))$.

Here, and in what follows, "Ω" is the notation dual to "big-O": $f \geq \Omega(g)$ means that there exists an absolute constant $\varepsilon > 0$ such that $f \geq \varepsilon g$ for all values of the parameters appearing in f, g.

Corollary 4.3. *For any fixed $d > 0$, F_d is not p-bounded.*

To illustrate one point made in Section 3, let us note that once we increase the number of pigeons to $2n$, the situation changes dramatically.

Theorem 4.4 ([5, 42]). $S_{\mathrm{F}_d}(\mathrm{PHP}_n^{2n} \vdash 0) \leq n^{(\log n)^{O(1/d)}}$. *For $d = 2$, this refines as* $S_{\mathrm{F}_2}(\mathrm{PHP}_n^{2n} \vdash 0) \leq n^{O(\log n)}$.

Whether this can be improved to polynomial, perhaps at the expense of using more pigeons, is open despite decades of research.

Problem 4.5. Does there exist a fixed $d > 0$ such that $S_{\mathrm{F}_d}(\mathrm{PHP}_n^{\infty} \vdash 0) \leq n^{O(1)}$?

As we noted above, once a method to analyze a proof system (in particular, to prove lower bounds for it) is established, it usually can be extended to other contradictions as well. As an illustration, the following was proved by a *direct* (albeit, very clever) reduction from Theorem 4.2.

Theorem 4.6 ([13]). *Let $\{G_n\}$ be a sequence of bounded-degree graphs with $c(G_n) \geq \Omega(1)$. Then for any fixed $d > 0$,* $S_{\mathrm{F}_d}(\mathrm{Tseitin}(G_n) \vdash 0) \geq \exp(\Omega(n^{1/5^d}))$.

But sometimes the next improvement/generalization requires a very serious enhancement of known techniques. Let us for example reverse the gears and instead of asking about size lower bounds in any *fixed* depth, ask what is the *largest* depth, as a function of the number of variables n, for which the bound still holds.

The bounds in Theorems 4.2 and 4.6 work up to $d = \varepsilon \log \log n$. It was recently improved to $d = o(\sqrt{\log n})$ in [50]. While this is still the same basic method of restrictions the previous work was based upon, this improvement literally had to take it to a new level of sophistication.

Theorem 4.7 ([50]). *For $\{G_n\}$ as in Theorem 4.6,*

$$S_{\mathrm{F}_d}\big(\mathrm{Tseitin}(G_n) \vdash 0\big) \geq n^{\Omega((\log n)/d^2)}.$$

Note that unlike Theorem 4.6, this bound is only quasi-polynomial. But it is good enough to prove that $\mathrm{F}_{d(n)}$ is not p-bounded when $d(n) = o(\sqrt{\log n})$.

In conclusion of this section, let us briefly discuss one extension.

Definition 4.8 (bounded-depth Frege with modular gates, informal). Let $m > 0$ be a fixed integer and $\mathrm{MOD}_m(x_1, \ldots, x_n)$ the propositional connective with the intended meaning $\mathrm{MOD}_m(x_1, \ldots, x_n) = 1$ iff $m | x_1 + \cdots + x_n$. Let $\mathrm{F}(\mathrm{MOD}_m)$ be a Frege system (p-equivalent to F) in the language $\{\neg, \wedge, \vee, \mathrm{MOD}_m\}$. The proof system $\mathrm{F}_d(\mathrm{MOD}_m)$ is its fragment in which the logical depth of all formulas is restricted to d, where axioms schemes and inference rules are chosen in any reasonable way (in particular, they should describe basic properties of the new connectives).

For some inspiration of what might be expected from this extension, we have to look again into the circuit complexity. The corresponding complexity class is denoted by $\mathsf{ACC}^0[m]$, and it turns out that the story crucially depends on m.

When m is a prime power, exponential lower bounds for this class of circuits have been known since [54, 61]. In all other cases (say, when $m = 6$) this is one of the most major and challenging open problems in circuit complexity: for all we know, $\mathsf{ACC}^0[6]$ may contain all of NP or, for that matter, $\mathsf{EXPTIME}$. For details, see e.g. [36, Chapter 12].

Accordingly, when m has at least two different prime divisors, $\mathrm{F}_d(\mathrm{MOD}_m)$ should definitely be classified as "strong." Somewhat embarrassingly, we have not been able to adapt the proofs from [54, 61] (based on the so-called *method of approximations*) to our context so far. The following is one of the main open problems in the area.

Problem 4.9. Prove that for any fixed $d > 0$ and any fixed prime $m > 0$ the system $\mathrm{F}_d(\mathrm{MOD}_m)$ is not p-bounded.

The only known partial results towards this problem pertain to its much weaker subsystems; we will now briefly mention one of them and another will appear in Section 6.1.

Definition 4.10 (counting principles). Let $m \nmid n$, and introduce propositional variables x_e, where $e \in \binom{[n]}{m}$, the family of all m-element subsets of $[n] \overset{\mathrm{def}}{=} \{1, 2, \ldots, n\}$. The *counting principle* Count^n_m is the unsatisfiable CNF consisting of the following clauses:

- $\bar{x}_e \vee \bar{x}_f$, for all $e \neq f$ such that $e \cap f \neq \emptyset$;
- $\bigvee_{e \ni i} x_e$, for all $i \in [n]$.

Intuitively, these clauses state that $(x_e \mid e \in \binom{[n]}{m})$ defines a partition of $[n]$ into sets of size m which may not exist since we assumed $m \nmid n$. The proof system $\mathrm{F}_d + \mathrm{Count}_m$ is obtained from F_d by adding to it all substitutional (de Morgan!) instances of Count_m^n, for arbitrary n, that are of logical depth $\leq d$.

The principle Count_m^n is easily provable in $\mathrm{F}_d(\mathrm{MOD}_m)$, hence $\mathrm{F}_d + \mathrm{Count}_m$ is indeed intermediate between F_d and $\mathrm{F}_d(\mathrm{MOD}_m)$ in the sense of Definition 1.7.

Theorem 4.11 ([11, 18]). *Let m, d, ℓ be fixed integers and assume that ℓ has a prime factor which is not a prime factor of m. Then $S_{\mathrm{F}_d + \mathrm{Count}_m}(\mathrm{Count}_\ell^n \vdash 0) \geq \exp(n^{\Omega(1)})$.*

Note, however, that this result holds for all m including, say, $m = 6$. This might be not so good sign for attempts to adapt these methods for solving Problem 4.9.

5. Resolution

In our notation, resolution is simply F_1. It obviously does not make much sense to consider *terms* $x_{i_1}^{a_1} \wedge \cdots \wedge x_{i_w}^{a_w}$ as lines in a proof, they can be always split into w lines consisting of single literals. Hence resolution uses clauses only and, given the importance of this proof system (that we will try to explain below), we prefer to break up with our own tradition and formulate its inference rules (there are no default axioms) very explicitly.

Definition 5.1 (resolution). Resoluton is the proof system operating with clauses, denoted by R. It has the inference rules

$$\frac{C}{C \vee D} \text{ (weakening)} \quad \frac{C \vee x \qquad D \vee \bar{x}}{C \vee D} \text{ (resolution rule)}.$$

A resolution proof is *regular* if, on any path in this proof, no variable x is resolved more than once. We will denote this subsystem of resolution by RR.

Remark 5.2. Resolution, as well as most systems we will see in the rest of this article, is too weak to speak of CNFs directly. It is therefore paramount (cf. Remark 1.2) that from now on we strictly adopt the "refutational" perspective: all "proofs" will be actually contradictions derived from a set of clauses.

Remark 5.3. The weakening rule is cosmetic and its removal does not change the complexity $S_R(\tau \vdash 0)$. Having this rule, however, is very convenient in many situations.

Resolution, as well as other proof systems that we will see below, is very relevant to various scenarios with practical flavor. The paradigm is somewhat similar in all these cases; let us spell it out for resolution in a few more details. Much more information on the topic, as well as all definitions missing in our description below, can be found e.g. in the very recent survey [19].

There is a large community of practice-oriented researchers working on finding feasible algorithms (which in this context means "actually implemented and delivering concrete results") for solving "interesting" instances of SATISFIABILITY. These programs are called *SAT solvers*. Now, what will happen if we feed a CNF τ to a SAT solver, it runs successfully and produces the correct answer?

When τ is satisfiable, in most cases the solver will be able to justify its answer by producing an actual satisfying assignment. But this case is not very inspiring for our purposes.

More interesting is the case when τ is unsatisfiable because if we understand the code and believe in its correctness, then we also must accept the *transcript of the solver's run* as a *proof* of unsatisfiability of τ. In mathematical terms, any practical scalable algorithm for solving SATISFIABILITY defines a propositional proof system in terms of Definition 1.4.

It turns out that in many scenarios the proof systems automatically associated in this way to algorithms are also mathematically elegant, and it is particularly visible in the case of SAT solvers. Namely, the algorithmic technique that has been dominating in that community for quite a while is called *conflict-driven clause learning* (CDCL). Then, a transcript of a run of a CDCL solver can be *identified* with a resolution proof, modulo a differing terminology. This connection is in fact so strong that it would not be too much of an exaggeration to describe the operation of CDCL solvers in this way: they search, in very ingenuous and specific ways, for resolution refutations of a CNF τ and declare it satisfiable if the search fails.

Thus, any *lower* bounds for the resolution proof system imply *inherent* limitations on CDCL solvers that cannot be overcome by any amount of clever engineering. They can also be used as a rough guidance of what to expect and what to avoid when building CDCL solvers.

An extremely interesting question is whether there is a connection in the opposite direction; that is, what algorithmic applications does the mere *existence* of a short resolution proof entail?

When the word "algorithmic" is understood in its most theoretical sense (that is, poly-time computable), this question is captured by the concept of "automatizability" (or "automation"), and we have recently seen a major progress in this direction [8] followed up in several other papers. Very loosely speaking, if $\mathsf{P} \neq \mathsf{NP}$, then no efficient algorithm will be able to find small resolution refutations in all cases when they exist, ever.

Another meaningful interpretation is to consider only algorithms based on the CDCL-architecture but allow them a limited amount of non-determinism in the choices they make. It turns out that this question is very sensitive to the choice of the model and, in my view, it is far from being answered conclusively. Some partial work in that direction is reported e.g. in [7, 12, 44]; once again, much more information can be found in [19].

Let us now return to mathematics, and we begin with several early prominent results.

Theorem 5.4 ([62]). *$S_{\mathrm{RR}}(\mathrm{Tseitin}(\mathrm{Grid}_{n,n}) \vdash 0) \geq \exp(\Omega(n))$, where $\mathrm{Grid}_{n,n}$ is the $n \times n$ grid graph.*

Theorem 5.5 ([33]). *$S_{\mathrm{R}}(\mathrm{PHP}_n^{n+1} \vdash 0) \geq \exp(\Omega(n))$.*

Theorem 5.6 ([22]). *Let τ_n be a random 3-CNF with $O(n)$ clauses. Then with probability $1 - o(1)$ we have $S_{\mathrm{R}}(\tau_n \vdash 0) \geq \exp(\Omega(n))$.*

As we already mentioned several times, it is highly desirable to have reasonably general methods for analyzing proof complexity, as opposed to those that are tailored to individual benchmarks. In that respect, the following prominent *width-size relation* clearly stands out.

Given a resolution refutation, its *width* is defined as the maximum width of its clauses, and let $w(\tau_n \vdash 0)$ be the minimum possible width of a resolution refutation of τ_n. In other words, we are trying to refute τ_n using only narrow clauses as our "lemmas," disregarding the question of how many of them we use. Then the width-size relation due to Ben–Sasson and Wigderson has the following neat and general form.

Theorem 5.7 ([15]). *For* any *sequence $\{\tau_n\}$ of unsatisfiable CNFs,*

$$w(\tau_n \geq 0) \leq O\left(\sqrt{n \cdot \log S_R(\tau_n \vdash 0)} + w_0\right),$$

where w_0 is the width of τ_n itself.

Parsing this expression, when w_0 is small (say, a constant) and $w(\tau_n \vdash 0) \geq \Omega(n)$, we get $S_R(\tau_n \vdash 0) \geq \exp(\Omega(n))$. In words, *linear* lower bounds on width imply *exponential* lower bounds on the resolution size.

And it turns out that width lower bounds are often much easier to prove. For example, we have the following (cf. Definition 3.4).

Theorem 5.8 ([15]). *For any sequence of bounded-degree graphs $\{G_n\}$ with $c(G_n) \geq \Omega(1)$, $w(\mathrm{Tseitin}(G_n) \vdash 0) \geq \Omega(n)$ (and hence by Theorem 5.7, $S_R(\mathrm{Tseitin}(G_n) \vdash 0) \geq \exp(\Omega(n))$).*

This recovers a stronger version of Theorem 4.6 for $d = 1$ but, again, the main strength of Theorem 5.7 lies in its generality. Two more important points highlighted by the width-size relation that have turned out very influential in proof complexity (we will see some examples below) are as follows.

(1) Diversity is good. Proof complexity measures more elaborated than the one stipulated by Definition 1.5 are inspiring *even* if one is primarily interested in size.

(2) Expansion is good as well. If a graph property imply hardness in the proof complexity, the odds are that expansion will also do the job.

The width-size relation can be successfully applied to an impressive array of various contradictions τ_n, often after some massaging. But, as is the case with any good method, it has its limitations. One notable principle it completely fails at is the pigeonhole principle with many (say, infinitely many) pigeons, which is the special case of Problem 4.5 for $d = 1$. For that, another technique of *pseudo-width* was developed in [49, 53, 58]. Unfortunately, this concept is a bit too technical to meaningfully address here, so let us simply state the end result for PHP[2].

Theorem 5.9. $S_{\mathrm{R}}(\mathrm{PHP}_n^\infty \vdash 0) \geq \exp(\Omega(n^{1/3}))$.

Remark 5.10. Surprisingly, the best known *upper* bound here is not the trivial $\exp(O(n))$ but $\exp(O(n^{1/2}))$ [20]. That would be nice to close the gap, particularly since most likely this will require developing new methods.

Problem 5.11. Determine the smallest $\alpha \in [1/3, 1/2]$ for which $S_{\mathrm{R}}(\mathrm{PHP}_n^\infty \vdash 0) \leq \exp(n^{\alpha+o(1)})$.

Among other things, Theorem 5.9 implies resolution lower bounds for the statement “NP does not have small size circuits” mentioned at the end of Section 2; see again [59], as well as [52], for more details and the context. The former paper also extends this to the proof system $\mathrm{Res}(O(1))$ operating with $O(1)$-CNFs but the proof is very indirect and complicated. On the other hand, Problem 4.5 remains wide open even for the system (say) Res(2) intermediate between F_1 and F_2. Moreover, now the upper bound of Theorem 4.4 no longer applies and we can state this conjecture in the stronger form.

Problem 5.12. Prove (or disprove) that $S_{\mathrm{Res}(2)}(\mathrm{PHP}_n^\infty \vdash 0) \geq \exp(n^{\Omega(1)})$.

More applications of the pseudo-width method can be found in the recent paper [28].

Are there prominent unsatisfiable CNFs that (in terms of their resolution complexity) resist analysis by both the width-size and pseudo-width methods? Let me conclude this section with my favorite example, the *small clique problem*.

[2] The last paper in the series [58] generalized the method to a much wider class of general *perfect matching principles* including, among others, the counting principles from Definition 4.10.

Definition 5.13. Let G be a *k-partite graph*, that is its vertices can be partitioned into k blocks, $V(G) = V_1 \dot\cup \cdots \dot\cup V_k$ (let us also assume that $|V_1| = \cdots = |V_k|$) such that there is no edge within each block. The CNF $\mathrm{Clique}_{\mathrm{Block}}(G, k)$ is defined as the following set of clauses in the variables $(x_v \mid v \in V(G))$:

- $\bigvee_{v \in V_i} x_v$ $(1 \leq i \leq k)$;
- $\bar{x}_v \wedge \bar{x}_w$ $((v, w)$ is *not* an edge of $G)$.

This CNF says that $(x_v \mid v \in V(G))$ encodes a k-clique in G and when the clique number $\omega(G)$ is at most $(k-1)$, this is a contradiction.

The obvious brute-force resolution refutation has size at most n^k, and the question is whether we can do any better. Motivated by the framework of parameterized (computational) complexity [29] and some research in circuit complexity, it is natural to ask about the existence of resolution refutations of size $f(k) \cdot n^{O(1)}$, where $f(k)$ is *any* function. Assuming that k is a fixed constant, the first term disappears and the question is whether $S_{\mathrm{R}}(\mathrm{Clique}_{\mathrm{Block}}(G, k) \vdash 0) \leq n^{O(1)}$, where the degree of the polynomial in the right-hand side must not depend on k.

The small clique problem is usually considered when G is the *Erdös–Renyi random graph*, that is when every potential edge between $v \in V_i$ and $w \in V_j$ is included i.i.d. with probability $p_{kn} > 0$, $n \overset{\mathrm{def}}{=} |V_i|$. Let us fix for definiteness

$$p_{kn} \overset{\mathrm{def}}{=} n^{-C/(k-1)}, \tag{5.1}$$

where $C > 2$ is an arbitrary constant, and let $\mathbf{G_{k,n}}$ be the corresponding Erdös–Renyi graph. The value (5.1) is a (weak) *threshold value*; it guarantees that the probability of the event $\omega(\mathbf{G_{k,n}}) = k$ is bounded away from both 0 and 1.

Theorem 5.14 ([6]). *For $k \leq n^{1/4} - \Omega(1)$, with probability $1 - o(1)$ we have*

$$S_{\mathrm{RR}}\big(\mathrm{Clique}_{\mathrm{Block}}(\mathbf{G_{k,n}}) \vdash 0\big) \geq n^{\Omega(k)}.$$

Problem 5.15. Prove that for any fixed $k > 0$, $S_{\mathrm{R}}(\mathrm{Clique}_{\mathrm{Block}}(\mathbf{G_{k,n}}) \vdash 0) \geq n^2$.

6. Algebraic and semi-algebraic proof systems

When you say 0 and 1, it is only mathematical logicians and computer scientists whose first association would be FALSE and TRUE. For anyone else, these are distinguished elements of a ring with particular algebraic (or semi-algebraic if the ring is ordered) properties. In this section, we will review, very briefly, a prominent family of proof systems heavily adopting this latter point of view and entirely abstracting from the logical interpretation of the statements they are proving. Besides [39, Chapter 16], the foundational material for this section, as well as a taxonomy of these proof systems, can be found in the early paper [32].

The first thing to decide is how exactly we are going to translate logic to algebra/geometry, and we should start with encoding clauses. There are essentially two different ways of doing it, and this choice largely determines what kind of proof systems we are aiming at.

The first possibility is to encode clauses by polynomial *equations* over a ground field $\mathbb{F}$. This is done in a very straightforward way; for example, the clause $C = x_1 \vee \bar{x}_2 \vee x_3$ is encoded as the equation $(1 - x_1)x_2(1 - x_3) = 0$.

For the second option we must assume that our ground field $\mathbb{F}$ is ordered, say $\mathbb{F} = \mathbb{Q}$ or $\mathbb{F} = \mathbb{R}$. In that case, we can encode clauses by linear *inequalities*. For example, $C = x_1 \vee \bar{x}_2 \vee x_3$ will be translated as $x_1 + (1 - x_2) + x_3 \geq 1$ that can be further simplified to $x_1 + x_3 \geq x_2$, if desired.

In either case, the original CNF is unsatisfiable if and only if the algebraic/semi-algebraic set defined by the corresponding system of polynomial equations/inequalities over $\mathbb{F}$ does not have 0-1 solutions. This reformulation allows us to employ tools from algebra/geometry, and we now treat the two cases separately.

6.1. Algebraic models

If we are allowed to use non-linear polynomials, the assumption that we are interested only in 0-1 solutions can be hardwired into the framework by introducing the default axioms $x_i^2 - x_i = 0$. It turns out to be very handy, albeit not strictly necessary, to factor out these relations at once and work in the $\mathbb{F}$-algebra

$$\Lambda_n \stackrel{\text{def}}{=} \mathbb{F}[x_1, \dots, x_n]/(x_i^2 - x_i \mid 1 \leq i \leq n).$$

This algebra was introduced to complexity theory (apparently) in [54, 61]; it consists of all *multi-linear* polynomials and hence has linear dimension 2^n. On the other hand, it is isomorphic to the algebra of all functions $\{0, 1\}^n \to \mathbb{F}$; $\mathrm{Hom}(\Lambda_n, \mathbb{F})$ is the set of all Boolean assignments to the variables $x_1, \dots, x_n$ etc.

Hilbert's Nullstellensatz tells us that a polynomial system $f_1(x_1, \dots, x_n) = \cdots = f_m(x_1, \dots, x_n) = 0$ ($f_i \in \Lambda_n$) does not have 0-1 solutions if and only if there exist $Q_1, Q_2, \dots, Q_m \in \Lambda_n$ such that

$$f_1 Q_1 + f_2 Q_2 + \cdots + f_m Q_m = 1. \tag{6.1}$$

Every such system of polynomials $(Q_1, \dots, Q_m) \in \Lambda_n$ can thus be considered as a *proof* of the statement that the algebraic set $(f_1 = 0, \dots, f_m = 0)$ does not contain 0-1 points. This proof system is called the *Nullstellensatz* proof system (over the field $\mathbb{F}$).

Remark 6.1. The question whether this system formally fits Definition 1.4 is slightly non-trivial. It may depend on the way the polynomials are represented, on their coefficients etc. We prefer not to dwell into these details as it has become much more customary (and it is way more clean mathematically, too) to measure the complexity of the proof $(Q_1, \dots, Q_m)$ by its *degree* defined as $\max_{1\leq i\leq m}(\deg(Q_i) + \deg(f_i))$.

By now, the Nullstellensatz proof system is fairly well understood. But since most results proved for it have been eventually generalized (and sometimes strengthened) to a stronger system that we will consider next, let us confine ourselves to just one prominent example.

Theorem 6.2 ([10]). *Every Nullstellensatz refutation of* PHP_n^∞ *must have degree* $\Omega(\sqrt{n})$.

Remark 6.3. Both for this result and those below, Definition 2.1 should be slightly adjusted. Namely, to avoid polynomials of prohibitively high degree, the pigeon axioms $x_{i1} \vee \cdots \vee x_{in}$ should be translated as $x_{i1} + \cdots + x_{in} - 1 = 0$ (note that this also implies that the funcionality axioms $\bar{x}_{ij_1} \vee \bar{x}_{ij_2}$ $(j_1 \neq j_2 \in [n])$ are also implicitly included).

The polynomial calculus (PC) is a *dynamic* version of this system in which we attempt to prove that 1 is in the ideal $(f_1, \dots, f_m) \subseteq \Lambda_n$ by generating its elements one by one instead of writing down a single expression like (6.1).

Definition 6.4. *Polynomial calculus* (over a ground field $\mathbb{F}$) is the algebraic proof system whose lines are elements of Λ_n. It has the following inference rules:

$$\frac{f = 0 \qquad g = 0}{\alpha f + \beta g = 0};\ \alpha, \beta \in \mathbb{F} \text{ (addition rule)}, \qquad \frac{f = 0}{fg = 0} \text{ (multiplication rule)}.$$

The *degree* of a PC proof is the maximum degree of its lines.

Remark 6.5. The main source of non-triviality of this system stems from the fact that at every step we completely expand the result as a sum of terms. When doing this, cancelations may (and typically do) lead to a substantial decrease in degree. On the other hand, there is a degree-size relation for the PC perfectly analogous to Theorem 5.7 (and actually proved earlier in [23]).

Remark 6.6. It is not very hard to see that every PC proof over $\mathbb{F}_p$ can be p-simulated by $F_2(\mathrm{MOD}_p)$. Thus, polynomial calculus over a finite field can be reasonably viewed as an "algebraic" component of $F_2(\mathrm{MOD}_p)$ while F_2 is its logical part. The main reason why Problem 4.9 appears to be so difficult is that the existing methods for understanding these two parts seem to be totally disjoint from each other.

There has been a fair amount of work attempting to build actual SAT solvers based upon algebraic principles, primarily the Gröbner basis algorithm. These solvers relate to the PC in precisely the same way CDCL-based solvers are related to resolution, cf. our discussion in Section 5. It would be fair to say that so far they have not been competitive with CDCL solvers but there does not seem to exist any good theoretical explanation of this fact. So perhaps the true potential of algebraic SAT solvers is yet to be revealed; we refer the reader to [19, Section 7.5.7] for more details.

As usual, we conclude with a few sample results. Historically the first lower bound for the PC generalized and strengthened Theorem 6.2.

Theorem 6.7 ([56]). *Every polynomial calculus refutation of* PHP_n^∞ *must have degree* $\Omega(n)$.

This also implies PC degree lower bounds for the statement "NP does not have small size circuits" we already mentioned several times before.

The proof method of Theorem 6.7 is rather ad hoc, it is based on the so-called "pigeon dance" specifically designed for the purpose. The next paper [17] introduced a very nice and remarkably simple method of analyzing PC refutations from *binomial*[3] axioms. Here is one concrete application that strengthens Theorem 5.8.

Theorem 6.8 ([17]). *For any sequence of bounded-degree graphs* $\{G_n\}$ *with* $c(G_n) \geq \Omega(1)$, *every polynomial calculus refutation of* $\mathrm{Tseitin}(G_n)$ *over any field of* odd or zero *characteristic must have degree* $\Omega(n)$.

The extension to random 3-CNFs, with the same restriction on the ground field $\mathbb{F}$, is not very difficult [14]. But the binomial method completely breaks down for $\mathbb{F} = \mathbb{F}_2$ which is one of the most interesting cases. Another method for proving PC degree lower bounds over an arbitrary field based on a general hardness criterion was proposed in [2]; see also [43] and the literature cited therein for more recent developments.

Theorem 6.9 ([2, 14]). *Let* τ_n *be a random* 3*-CNF with* $O(n)$ *clauses. Then any polynomial calculus refutation of* τ_n *over an* arbitrary *field* $\mathbb{F}$ *must have degree* $\Omega(n)$.

Let us finally note that the degree-size relation mentioned above immediately implies exponential *size* lower bounds for PC refutations in Theorems 6.8 and 6.9.

6.2. Semi-algebraic case

There are many prominent semi-algebraic proof systems: Sum-of-Squares, Cutting Planes, Lovász–Schrijver, Sherali–Adams to name a few. We will only touch, very briefly, on the first two; for a nicely organized exposition see [32]. Throughout this section we assume that $\mathbb{F} = \mathbb{Q}$ or $\mathbb{F} = \mathbb{R}$.

The Sum-of-Squares is also known under the name Positivestellensatz and is closely related to the so-called Lassierre hierarchy. There are several slight variations in its definition, we only present here (as many other authors do) the simplest version in which the original axioms are given as polynomial equations, like in Section 6.1.

[3] In the Rademacher $\{\pm 1\}$ framework.

Definition 6.10. An *SOS* (or *Positivestellensatz*) refutation of a polynomial system $(f_1 = \cdots = f_m = 0)$ $(f_i \in \Lambda_n)$ is a family of polynomials $(Q_1, \ldots, Q_m, g_1, \ldots, g_t)$ in Λ_n such that

$$f_1 Q_1 + \cdots + f_m Q_m + \sum_{j=1}^{t} g_j^2 = -1. \tag{6.2}$$

Its *degree* is defined as $\max(\max_{1 \le i \le m} \deg(f_i) + \deg(Q_i),\ 2\max_{1 \le j \le t} \deg(g_j))$.

The corresponding algorithmic technique has in recent years become extremely important in combinatorial optimization and approximation algorithms, largely due to the fact that it has turned out unexpectedly powerful. We refer the reader to the expository paper [9] although a great deal of important work has been done since that. The relation between combinatorial optimization and proof complexity follows the familiar pattern, and in fact in this case it is even more transparent. But one important difference is that unlike SAT solvers, algorithms in combinatorial optimization seldom output the exact answer but only an optimistic approximation to it which in most cases means relaxing the integrality constraints $x_i \in \{0, 1\}$ to $x_i \in [0, 1]$. In any case, the computation implies that one cannot beat the value of the goal function delivered by this relaxation, and then after a straightforward application of the PSD duality, it becomes an SOS proof in the sense of Definition 6.10. See again [9] for more details.

As for degree lower bounds, SOS is also relatively well understood although some important problems still remain open. The first lower bound had been proven by Grigoriev [31] and largely forgotten until the realization of the algorithmic significance of the SOS method came. This is the same binomial method we saw in Section 6.1, wisely put to a different use.

Theorem 6.11 ([31]). *Every SOS refutation of* Tseitin(G_n)*, where* $\{G_n\}$ *is a sequence of bounded-degree graphs with* $c(G_n) \ge \Omega(1)$*, must have degree* $\Omega(n)$.

More modern methods of handling SOS proofs are based upon the concept of a *pseudoexpectation* which is essentially an object dual to the expression (6.2) (therefore, it exists if and only if the system (6.2) is *not* solvable in Q_i, g_j of given degree).

The last system we discuss is Cutting Planes.

Definition 6.12. *Cutting Planes* is the proof system operating with affine inequalities, denoted by PC. It has default axioms $x \ge 0$ and $x \le 1$ for all variables x, as well as the following inference rules:

$$\frac{f \ge 0 \qquad g \ge 0}{\alpha f + \beta g \ge 0}; \quad \alpha, \beta \ge 0 \text{ (convex closure)},$$

$$\frac{f \ge a}{f \ge \lceil a \rceil}; \quad f \in \mathbb{Z}[x_1, \ldots, x_n] \text{ (cut rule)}.$$

To explain this terminology, it is convenient to adapt the dual, more geometric point of view. Namely, if we allow to apply all possible convex closure rules at once, then the set of constraints inferrable in this way will form a (convex) polyhedra. Its dual will be a polytope P that is actually a sub-polytope of $[0,1]^n$ (due to default axioms). The task is to show that $P \cap \{0,1\}^n = \emptyset$, that is that P does not contain any *integer* points. In this language, applying the cut rule means cutting off a small piece from this polytope (whence the name) guaranteed to not contain integer points.

From the algorithmic perspective, cutting planes correspond to the "geometric" part of very prominent method in combinatorial optimization called *Branch and Cut*. The full power of this method is captured by the proof system that, in the geometric language above, operates with finite unions of polytopes. This system is currently out of reach of the current methods although I would hesitate to classify it as "strong." A major development has been very recently reported on its subsystem $\mathrm{Res}(\mathrm{lin}_{\mathbb{R}})$ in which all polytopes are confined to the form $H \cap [0,1]^n$, H a hyperplane [47].

One proof complexity measure for cutting planes that has been extensively considered in the literature is their (Chvátal) *rank* (or *depth*). It is defined as follows: we allow to apply in parallel not only all possible convex closure rules but cut rules as well. Then the rank is simply the number of rounds that are necessary to arrive at the empty polytope. This complexity measure is rather well understood due to a very powerful technique called "protection lemmas," see [36, Chapter 19] for an excellent exposition.

As far as the *size* of cutting planes refutation is concerned, the situation is way more intriguing and dynamic. The first lower bounds were proved by Pudlák [51] using a prominent *feasible interpolation* method (or rather property). In the next theorem, Clique-Coloring(n,k) is the principle that says that a graph on n vertices may not simultaneously have a clique on k vertices and be k-colorable.

Theorem 6.13 ([51]). $S_{\mathrm{CP}}(\text{Clique-Coloring}(n, \sqrt{n})) \geq \exp(n^{\Omega(1)})$.

Remarkably, the method of feasible interpolation is not combinatorial or direct, instead it reduces a difficult problem in proof complexity to a difficult problem in circuit complexity (lower bounds for monotone circuits) that we fortunately know how to solve. As a by-side remark, let me mention that this kind of reductions is very important and welcome for the proof complexity. Still, it is also natural to wonder whether there are any "direct" methods (all other results in this article certainly qualify) to handle cutting planes. On this frontier we have seen recent exciting developments that defy several pieces of "common wisdom."

Firstly, it somehow makes sense to assume that random $O(\log n)$-CNFs and $O(1)$-CNFs should be "morally similar." Nonetheless, the proof method of the following theorem (a very clever use of feasible interpolation) seems to completely break apart for $O(1)$-CNFs.

Theorem 6.14 ([30, 35]). *With probability* $1 - o(1)$, *for a random* $\Theta(\log n)$*-CNF* τ_n *with* $n^{O(1)}$ *clauses we have* $S_{\mathrm{CP}}(\tau_n \vdash 0) \geq \exp(n^{\Omega(1)})$.

Even more striking and unexpected is the following recent result. In all our previous scenarios, random $O(1)$-CNFs and Tseitin tautologies for expanders went hand in hand, and it was a general feeling that morally they should be sort of the same (well, unless the characteristics of the field is 2). Given this feeling, the following *upper* bound, very surprising in itself, also does not seem to generalize to random $O(1)$-CNFs.

Theorem 6.15 ([26]). *For any sequence* $\{G_n\}$ *of bounded-degree graphs,*

$$S_{\mathrm{CP}}\big(\mathrm{Tseitin}(G_n) \vdash 0\big) \leq n^{O(\log n)}.$$

All these developments make the following problem particularly exciting.

Problem 6.16. Is it true that for random $O(1)$-CNFs τ_n with $O(n)$ clauses,

$$S_{\mathrm{CP}}(\tau_n \vdash 0) \geq \exp(n^{\Omega(1)}) \text{ w.h.p.?}$$

It is expected that solving this in the affirmative would require development of long-sought *direct* techniques, combinatorial or geometric, for analyzing the size complexity of cutting planes. But then it also had been expected from the principles featuring in the last two theorems.

7. In lieu of conclusion

There are several important topics in the modern proof complexity that, due to time and space constraints, we have either skipped entirely or given them much less attention than they deserve. Let me conclude with a list of such topics, saying (literally) a few words about each of them and providing some pointers to the literature.

Space complexity. Size complexity measures roughly correspond to the framework in which a complete proof is written as a single piece made ready for submission or verification. Space complexity deals with more dynamic, "classroom" scenario when the proof is presented on a blackboard and lemmas that are no longer needed can be erased to save space. See [45] for a nice exposition.

Feasible interpolation and automatizability. These were already mentioned in Sections 5 and 6.1. The book [39] treats the subject extensively in Chapters 17 and 18.

Relations between various proof systems and complexity measures. We have already seen some of those but, with the exception for Theorem 5.7, they were somewhat straightforward. There are, however, many other realtions, particularly involving space complexity measures, that are rather intricate and unexpected. The paper [46] aims at providing a general picture using an appropriate notion of reduction.

Pseudo-random generators in proof complexity. This is an ongoing effort to adjust to the needs of proof complexity the concept that is omnipresent in computational complexity. It is largely motivated by studying (efficient) provability of the principle "NP does not possess small circuits" we already mentioned several times. See (again) [59, Section 1] or [39, Chapter 19.4] for more details.

Lifting techniques. This is a very recent general approach to lower bounds in circuit complexity, communication complexity, and proof complexity remarkably uniting the three themes. I am not aware of an expository source (this is very much work in progress!) so let me instead refer to one of the latest papers in this direction [27].

Ideal proof system. This is an intriguing and bold attempt to stretch the Cook–Reckhow framework (Definition 1.4) and bring it closer to the concept of PCPs discussed earlier in Section 1. The paper [3] is one of the latest texts on the subject.

References

[1] M. Ajtai, Σ_1^1-formulae on finite structures. *Ann. Pure Appl. Logic* **24** (1983), no. 1, 1–48 Zbl 0519.03021 MR 706289

[2] M. V. Alekhnovich and A. A. Razborov, Lower bounds for polynomial calculus: the non-binomial ideal case. *Tr. Mat. Inst. Steklova* **242** (2003), 23–43 Zbl 1079.03047 MR 2054483

[3] Y. Alekseev, D. Grigoriev, E. A. Hirsch, and I. Tzameret, Semi-algebraic proofs, IPS lower bounds, and the τ-conjecture: can a natural number be negative? In *STOC '20—Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 54–67, ACM, New York, 2020 Zbl 07298230 MR 4141742

[4] S. Arora and B. Barak, *Computational Complexity. A Modern Approach*. Cambridge University Press, Cambridge, 2009 Zbl 1193.68112 MR 2500087

[5] A. Atserias, Improved bounds on the weak pigeonhole principle and infinitely many primes from weaker axioms. In *Mathematical Foundations of Computer Science, 2001 (Mariánské Lázně)*, pp. 148–158, Lecture Notes in Comput. Sci. 2136, Springer, Berlin, 2001 Zbl 0999.03052 MR 1907008

[6] A. Atserias, I. Bonacina, S. F. De Rezende, M. Lauria, J. Nordström, and A. A. Razborov, Clique is hard on average for regular resolution. *J. ACM* **68** (2021), no. 4, Art. 23 Zbl 1427.68102 MR 4313110

[7] A. Atserias, J. K. Fichte, and M. Thurley, Clause-learning algorithms with many restarts and bounded-width resolution. *J. Artificial Intelligence Res.* **40** (2011), 353–373 Zbl 1214.68340 MR 2805245

[8] A. Atserias and M. Müller, Automating resolution is NP-hard. *J. ACM* **67** (2020), no. 5, Art. 31 Zbl 07273098 MR 4152445

[9] B. Barak and D. Steurer, Sum-of-squares proofs and the quest toward optimal algorithms. In *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*, pp. 509–533, Kyung Moon Sa, Seoul, 2014 Zbl 1373.68253 MR 3727623

[10] P. Beame, S. Cook, J. Edmonds, R. Impagliazzo, and T. Pitassi, The relative complexity of NP search problems. In *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing (STOC'95)*, pp. 303–314, ACM, New York, 1995 Zbl 0978.68526

[11] P. Beame, R. Impagliazzo, J. Krajíček, T. Pitassi, and P. Pudlák, Lower bounds on Hilbert's Nullstellensatz and propositional proofs. *Proc. London Math. Soc. (3)* **73** (1996), no. 1, 1–26 Zbl 0853.03017 MR 1387081

[12] P. Beame, H. Kautz, and A. Sabharwal, Towards understanding and harnessing the potential of clause learning. *J. Artificial Intelligence Res.* **22** (2004), 319–351 Zbl 1080.68651 MR 2129471

[13] E. Ben-Sasson, Hard examples for bounded depth Frege. In *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*, pp. 563–572, ACM, New York, 2002 Zbl 1192.03041 MR 2121182

[14] E. Ben-Sasson and R. Impagliazzo, Random CNF's are hard for the polynomial calculus. *Comput. Complexity* **19** (2010), no. 4, 501–519 Zbl 1216.03064 MR 2746277

[15] E. Ben-Sasson and A. Wigderson, Short proofs are narrow—resolution made simple. *J. ACM* **48** (2001), no. 2, 149–169 Zbl 1089.03507 MR 1868713

[16] M. L. Bonet, S. R. Buss, and T. Pitassi, Are there hard examples for Frege systems? In *Feasible Mathematics, II (Ithaca, NY, 1992)*, pp. 30–56, Progr. Comput. Sci. Appl. Logic 13, Birkhäuser, Boston, MA, 1995 Zbl 0834.03021 MR 1322273

[17] S. Buss, D. Grigoriev, R. Impagliazzo, and T. Pitassi, Linear gaps between degrees for the polynomial calculus modulo distinct primes. *J. Comput. System Sci.* **62** (2001), no. 2, 267–289 Zbl 1007.03052 MR 1820593

[18] S. Buss, R. Impagliazzo, J. Krajíček, P. Pudlák, A. A. Razborov, and J. Sgall, Proof complexity in algebraic systems and bounded depth Frege systems with modular counting. *Comput. Complexity* **6** (1996/97), no. 3, 256–298 Zbl 0890.03030 MR 1486929

[19] S. Buss and J. Nordström, Proof complexity and SAT solving. In *Handbook of Satisfiability*, chap. 7, pp. 233–350, Frontiers in Artificial Intelligence and Applications 336, IOS Press, Amsterdam, 2021 Zbl 1456.68001

[20] S. Buss and T. Pitassi, Resolution and the weak pigeonhole principle. In *Computer Science Logic (Aarhus, 1997)*, pp. 149–156, Lecture Notes in Comput. Sci. 1414, Springer, Berlin, 1998 Zbl 0910.03036 MR 1727809

[21] S. R. Buss, *Bounded Arithmetic*. Studies in Proof Theory. Lecture Notes 3, Bibliopolis, Naples, 1986 Zbl 0649.03042 MR 880863

[22] V. Chvátal and E. Szemerédi, Many hard examples for resolution. *J. Assoc. Comput. Mach.* **35** (1988), no. 4, 759–768 Zbl 0712.03008 MR 1072398

[23] M. Clegg, J. Edmonds, and R. Impagliazzo, Using the Groebner basis algorithm to find proofs of unsatisfiability. In *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, pp. 174–183, ACM, New York, 1996 Zbl 0938.68825 MR 1427512

[24] S. Cook and P. Nguyen, *Logical Foundations of Proof Complexity*. Perspect. Log., Cambridge University Press, Cambridge, 2014 Zbl 1284.03001 MR 3241240

[25] S. A. Cook and R. A. Reckhow, The relative efficiency of propositional proof systems. *J. Symbolic Logic* **44** (1979), no. 1, 36–50 Zbl 0408.03044 MR 523487

[26] D. Dadush and S. Tiwari, On the complexity of branching proofs. In *35th Computational Complexity Conference*, p. Art. No. 34, LIPIcs. Leibniz Int. Proc. Inform. 169, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2020 MR 4129296

[27] S. de Rezende, O. Meir, J. Nordström, T. Pitassi, R. Robere, and M. Vinyals, Lifting with simple gadgets and applications to circuit and proof complexity. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pp. 24–30, IEEE Computer Soc., Los Alamitos, CA, 2020 MR 4232019

[28] S. F. de Rezende, J. Nordström, K. Risse, and D. Sokolov, Exponential resolution lower bounds for weak pigeonhole principle and perfect matching formulas over sparse graphs. In *35th Computational Complexity Conference*, p. Art. No. 28, LIPIcs. Leibniz Int. Proc. Inform. 169, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2020 MR 4129290

[29] R. G. Downey and M. R. Fellows, *Parameterized Complexity*. Monographs in Computer Science, Springer, New York, 1999 MR 1656112

[30] N. Fleming, D. Pankratov, T. Pitassi, and R. Robere, Random $\Theta(\log n)$-CNFs are hard for cutting planes. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pp. 109–120, IEEE Computer Soc., Los Alamitos, CA, 2017 MR 3734222

[31] D. Grigoriev, Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. *Theoret. Comput. Sci.* **259** (2001), no. 1-2, 613–622 Zbl 0974.68192 MR 1832812

[32] D. Grigoriev, E. A. Hirsch, and D. V. Pasechnik, Complexity of semialgebraic proofs. *Mosc. Math. J.* **2** (2002), no. 4, 647–679 Zbl 1027.03044 MR 1986085

[33] A. Haken, The intractability of resolution. *Theoret. Comput. Sci.* **39** (1985), no. 2-3, 297–308 Zbl 0586.03010 MR 821207

[34] J. Håstad, *Computational limitations on Small Depth Circuits*. Ph.D. thesis, Massachusetts Institute of Technology, 1986

[35] P. Hrubeš and P. Pudlák, Random formulas, monotone circuits, and interpolation. In *58th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2017*, pp. 121–131, IEEE Computer Soc., Los Alamitos, CA, 2017 MR 3734223

[36] S. Jukna, *Boolean Function Complexity. Advances and Frontiers*. Algorithms Combin. 27, Springer, Heidelberg, 2012 Zbl 1235.94005 MR 2895965

[37] J. Krajíček, Lower bounds to the size of constant-depth propositional proofs. *J. Symbolic Logic* **59** (1994), no. 1, 73–86 Zbl 0798.03056 MR 1264964

[38] J. Krajíček, *Bounded Arithmetic, Propositional Logic, and Complexity Theory*. Encyclopedia Math. Appl. 60, Cambridge University Press, Cambridge, 1995 Zbl 0835.03025 MR 1366417

[39] J. Krajíček, *Proof Complexity*. Encyclopedia Math. Appl. 170, Cambridge University Press, Cambridge, 2019 Zbl 07044161 MR 3929744

[40] J. Krajíček and P. Pudlák, Propositional proof systems, the consistency of first order theories and the complexity of computations. *J. Symbolic Logic* **54** (1989), no. 3, 1063–1079 Zbl 0696.03029 MR 1011192

[41] J. Krajíček, P. Pudlák, and A. Woods, An exponential lower bound to the size of bounded depth Frege proofs of the pigeonhole principle. *Random Structures Algorithms* **7** (1995), no. 1, 15–39 Zbl 0843.03032 MR 1346282

[42] A. Maciel, T. Pitassi, and A. R. Woods, A new proof of the weak pigeonhole principle. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pp. 368–377, ACM, New York, 2000 Zbl 1296.03033 MR 2114552

[43] M. Mikša and J. Nordström, A generalized method for proving polynomial calculus degree lower bounds. In *30th Conference on Computational Complexity*, pp. 467–487, LIPIcs. Leibniz Int. Proc. Inform. 33, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2015 Zbl 1434.03134 MR 3441818

[44] N. Mull, S. Pang, and A. A. Razborov, On CDCL-based proof systems with the ordered decision strategy. In *Theory and Applications of Satisfiability Testing—SAT 2020*, pp. 149–165, Lecture Notes in Comput. Sci. 12178, Springer, Cham, 2020 Zbl 07331019 MR 4140309

[45] J. Nordström, Pebble games, proof complexity and time-space trade-offs. *Log. Methods Comput. Sci.* **9** (2013), no. 3, 3:15, 63 Zbl 1285.03070 MR 3109599

[46] T. Papamakarios and A. A. Razborov, Space characterizations of complexity measures and size-space trade-offs in propositional proof systems. Tech. Rep. TR21-074, Electronic Colloquium on Computational Complexity, 2021

[47] F. Part and I. Tzameret, Resolution with counting: dag-like lower bounds and different moduli. *Comput. Complexity* **30** (2021), no. 1, Paper No. 2 Zbl 07355181 MR 4199854

[48] T. Pitassi, P. Beame, and R. Impagliazzo, Exponential lower bounds for the pigeonhole principle. *Comput. Complexity* **3** (1993), no. 2, 97–140 Zbl 0784.03034 MR 1233662

[49] T. Pitassi and R. Raz, Regular resolution lower bounds for the weak pigeonhole principle. *Combinatorica* **24** (2004), no. 3, 503–524 Zbl 1063.03044 MR 2085370

[50] T. Pitassi, B. Rossman, R. A. Servedio, and L.-Y. Tan, Poly-logarithmic Frege depth lower bounds via an expander switching lemma. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 644–657, ACM, New York, 2016 Zbl 1373.03125 MR 3536603

[51] P. Pudlák, Lower bounds for resolution and cutting plane proofs and monotone computations. *J. Symbolic Logic* **62** (1997), no. 3, 981–998 Zbl 0945.03086 MR 1472134

[52] R. Raz, P $\neq$ NP, propositional proof complexity, and resolution lower bounds for the weak pigeonhole principle. In *Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002)*, pp. 685–693, Higher Ed. Press, Beijing, 2002 Zbl 1012.68081 MR 1957570

[53] R. Raz, Resolution lower bounds for the weak pigeonhole principle. *J. ACM* **51** (2004), no. 2, 115–138 Zbl 1317.03036 MR 2145651

[54] A. A. Razborov, Lower bounds on the size of bounded-depth networks over a complete basis with logical addition. *Mat. Zametki* **41** (1987), no. 4, 598–607; English translation *Math. Notes* **41** (1987), 333–338 MR 897705

[55] A. A. Razborov, On provably disjoint **NP**-pairs. Tech. Rep. RS-94-36, Basic Research in Computer Science Center, Aarhus, Denmark, 1994, https://www.brics.dk/RS/94/36/BRICS-RS-94-36.pdf

[56] A. A. Razborov, Lower bounds for the polynomial calculus. *Comput. Complexity* **7** (1998), no. 4, 291–324 Zbl 1026.03043 MR 1691494

[57] A. A. Razborov, Proof complexity of pigeonhole principles. In *Developments in Language Theory (Vienna, 2001)*, pp. 110–116, Lecture Notes in Comput. Sci. 2295, Springer, Berlin, 2002 Zbl 1073.03540 MR 1964164

[58] A. A. Razborov, Resolution lower bounds for perfect matching principles. *J. Comput. System Sci.* **69** (2004), no. 1, 3–27 Zbl 1106.03049 MR 2070797

[59] A. A. Razborov, Pseudorandom generators hard for k-DNF resolution and polynomial calculus resolution. *Ann. of Math. (2)* **181** (2015), no. 2, 415–472 Zbl 1376.03055 MR 3275844

[60] R. A. Reckhow, On the lengths of proofs in the propositional calculus. Tech. Rep. 87, University of Toronto, 1976

[61] R. Smolensky, Algebraic methods in the theory of lower bounds for Boolean circuit complexity. In *Proceedings of the 19th ACM Symposium on Theory of Computing*, pp. 77–82, 1987

[62] G. S. Tseitin, On the complexity of derivations in propositional calculus. In *Studies in Constructive Mathematics and Mathematical Logic, Part II*, Consultants Bureau, New York, 1968

[63] A. Yao, Separating the polynomial-time hierarchy by oracles. In *Proceedings of the 26th IEEE FOCS*, pp. 1–10, 1985

Alexander A. Razborov
Departments of Mathematics and Computer Science, University of Chicago, Eckhart Hall,
5734 S University Ave, Chicago, IL 60637, USA; and Steklov Mathematical Institute of
the Russian Academy of Sciences, 8 Gubkina St., Moscow 119991, Russia;
razborov@uchicago.edu, razborov@mi-ras.ru

DOI 10.4171/8ECM/21

Covering and growth for group subsets and representations

Aner Shalev

Abstract. Deep results on products of subsets of finite groups, and of finite simple groups in particular, were obtained this century. Gowers' theory of quasi-random groups, further developed and applied by Nikolov and Pyber, focuses on covering results, while the theory of approximate subgroups and the product theorem, developed by Helfgott, Hrushovski, Breuillard, Green and Tao, and Pyber and Szabó, focus on growth results.

In recent joint works with Larsen and Tiep, following works with Liebeck and Tiep, we explore analogous problems in representation theory. We replace subsets of a group by its characters, and subset products by products of characters. We also study covering and growth for normal subsets of finite simple groups and derive various applications. In particular, we prove that every element of a sufficiently large finite simple transitive permutation group is a product of two derangements.

The product theorem establishes 3-step growth of the form $|A^3| \geq |A|^{1+\varepsilon}$ for (certain) subsets A of finite simple groups of Lie type of bounded rank. Surprisingly, stronger results hold for characters. We obtain 2-step growth for characters of finite simple groups of Lie type, including those of unbounded rank. For a character χ of G, we set $|\chi| = \sum_i \chi_i(1)^2$, where χ_i are the (distinct) irreducible constituents of χ. For a finite simple group G of Lie type, we show that for every $\delta > 0$ there exists $\varepsilon > 0$ such that if χ is an irreducible character of G satisfying $|\chi| \leq |G|^{1-\delta}$, then $|\chi^2| \geq |\chi|^{1+\varepsilon}$. In addition, we obtain results for reducible characters and establish faster growth of the form $|\chi^2| \geq |\chi|^{2-\varepsilon}$ if $|\chi| \leq |G|^{\delta}$.

Following a recent work of Sellke, we also study covering phenomena in representation theory, proving that if $|\chi_1| \cdots |\chi_m|$ is a sufficiently large power of $|G|$, then every irreducible character of G is a constituent of $\chi_1 \cdots \chi_m$. Finally, we obtain related results for characters of compact semisimple Lie groups.

1. Subset products and covering

In the past two decades, there has been considerable interest in the products of subsets of finite groups, especially (nonabelian) finite simple groups. The so-called Gowers' trick, which is part of the theory of quasi-random groups (see [1, 17, 57]), establishes

2020 Mathematics Subject Classification. Primary 20C33; Secondary 20D06.
Keywords. Finite simple groups, growth, covering, characters, approximate subgroups.

a useful 3-step covering result. Let $m(G)$ denote the minimal degree of a non-trivial irreducible character of a finite group G. The *density* of a subset $A \subseteq G$ is defined as $|A|/|G|$. Let A, B, and C be subsets of G satisfying $|A||B||C| \geq |G|^3/m(G)$. Then, Gowers' trick shows that $ABC = G$. In particular, if the density of A is at least $m(G)^{-1/3}$, then $A^3 = G$.

A family F of finite groups is said to be *quasi-random* if $m(G) \to \infty$ as G ranges over the groups in F. It follows that if F is a quasi-random family of finite groups, $\varepsilon > 0$, $G \in F$, and the density of $A, B, C \subseteq G$ is at least ε, then $ABC = G$ provided that $|G|$ is sufficiently large.

Much less is known about the products of two subsets, which is the main topic of this section. It is easy to see that the above assertion fails to hold for length 2 products AB. Moreover, for every positive integer k, there exist infinitely many finite groups G and subsets $A, B \subset G$ such that $|A|, |B| \geq |G|/(k+1)$ and $|AB| \leq |G| - k$. To see this, fix $k \geq 1$, and let G be any finite group of order divisible by $k+1$. Let $A, S \subset G$ be subsets satisfying $|A| = |G|/(k+1)$ and $|S| = k$. Define $B := G \setminus A^{-1}S$ (where $A^{-1} := \{a^{-1} : a \in A\}$). Note that $|B| = |G| - |A^{-1}S| \geq |G| - |S||A| = |G| - k|G|/(k+1) = |G|/(k+1)$. Clearly, $AB \cap S = \emptyset$ (if $ab = s$ for some $a \in A$, $b \in B$, and $s \in S$, then $b = a^{-1}s$, so $B \cap A^{-1}S \neq \emptyset$, a contradiction). Thus, $AB \subseteq G \setminus S$ and $|AB| \leq |G| - k$, proving the claim.

Can we still obtain 2-step covering results under suitable stronger assumptions?

A trivial observation (which is still useful) is that if subsets $A, B \subseteq G$ have densities α and β, respectively, and $\alpha + \beta > 1$, then $AB = G$. In particular, if $A, B \subseteq G$ have size greater than $|G|/2$, then $AB = G$. This observation will play a role in the complicated proof of the main result of Section 2 (see Theorems 2.4 and 2.8).

Next, let us assume that F is the family of all finite simple groups G. It is well known that F is quasi-random (see [29] for detailed information on $m(G)$). Let $S, T \subseteq G$ be *normal* subsets of G; this means that S, T are closed under conjugation by elements of G, and so they are unions of conjugacy classes of G. What can we say about the product ST and about related distributions?

Products of two (or more) normal subsets of finite simple groups have been extensively studied. This includes the challenging case of products of two conjugacy classes. A major motivation is a longstanding conjecture of Thompson, which asserts that every finite simple group G has a conjugacy class C such that $C^2 = G$. In spite of considerable progress (see Ellers and Gordeev [8] and the references therein) and the proof of the related Ore conjecture (see [42]), Thompson's conjecture is still open for various infinite families of groups of Lie type over fields with $q \leq 8$ elements. A weaker result, that all finite simple groups G of order exceeding 2^{630} have conjugacy classes C_1, C_2 such that $C_1C_2 \supseteq G \setminus \{e\}$, is obtained in [36]; this was improved, using computational group theory and other tools, by Guralnick and Malle in [18], where the same conclusion is established for *all* finite simple groups.

See also [65], where a probabilistic approximation to Thompson's conjecture is obtained. It is shown there that, for finite simple groups G and random (not necessarily independent) elements $x, y \in G$, the sizes of $x^G y^G$ and of $(x^G)^2$ are $(1 - o(1))|G|$. Thus, the square of the conjugacy class x^G of a random element $x \in G$ almost covers G as $|G| \to \infty$. Very recently, Larsen and Tiep [39] have proved Thompson's conjecture for additional infinite families of finite simple groups.

For normal subsets S (not equal to $\emptyset$, $\{e\}$) of arbitrary finite simple groups G, the minimal $k > 0$ such that $S^k = G$ is determined by Liebeck and me in [45] up to an absolute multiplicative constant. Indeed, we show there that $\log |G| / \log |S| \leq k \leq c \log |G| / \log |S|$ and derive various applications. Note that the lower bound on k is trivial and that the upper bound on k is also an upper bound on the diameter of the Cayley graph $\Gamma(G, S)$.

A beautiful improvement of this result in the case $G = \mathrm{PSL}_n(q)$ was obtained by Rodgers and Saxl [59]. They show that if $C_1, \ldots, C_k$ are conjugacy classes of G satisfying $|C_1| \cdots |C_k| > |G|^{12}$, then $C_1 \cdots C_k = G$.

Very recently, Maróti and Pyber [54] have obtained an impressive common extension of [45, 59], proving the following covering result.

Theorem 1.1 (Maróti and Pyber, 2021). *There exists an absolute constant c such that if G is any finite simple group, $k \in \mathbb{N}$, and $T_1, \ldots, T_k \subseteq G$ are normal subsets of G satisfying $|T_1| \cdots |T_k| \geq |G|^c$, then $T_1 \cdots T_k = G$.*

In [41], Liebeck, Nikolov, and I conjectured that there is an absolute constant c such that if G is any finite simple group and $A \subseteq G$ is any subset of size at least two, then there is $k \leq c \log |G| / \log |A|$ and elements $g_1, \ldots, g_k \in G$ such that $A^{g_1} \cdots A^{g_k} = G$ (where $A^g = g^{-1} A g$). This conjecture is still open in general, but Gill, Pyber, Short, and Szabó [15] confirmed it for finite simple groups of Lie type of bounded rank.

The following stronger covering conjecture, which implies Theorem 1.1, was stated by Gill, Pyber, and Szabó in [16] and proved there for finite simple groups of Lie type of bounded rank.

Conjecture 1.2. *There is an absolute constant c such that if G is any finite simple group, $k \in \mathbb{N}$, and $A_1, \ldots, A_k \subseteq G$ are subsets of G satisfying $|A_1| \cdots |A_k| \geq |G|^c$, then there exist elements $g_1, \ldots, g_k \in G$ such that $A_1^{g_1} \cdots A_k^{g_k} = G$.*

Some covering results have deep applications, also to the theory of expander graphs. It is now known that all finite simple groups can be made expanders uniformly with respect to bounded generating sets. Remarkable pioneering work by Kassabov established this for alternating groups [28] and then for special linear groups of unbounded rank. A key step in proving expansion for the remaining finite simple groups was to present the simple groups of Lie type of bounded rank (except the

Suzuki groups, shown to be expanders in [6]) as a bounded product of subgroups of the types $\mathrm{SL}_2(q)$ and $\mathrm{PSL}_2(q)$. This is done effectively in [40] with explicit bounds and ineffectively (using a model-theoretic approach based on work by Hrushovski and Pillay) by Lubotzky in [51].

In the work [43], Liebeck, Schul, and I show that the product of two small normal subsets of finite simple groups has size close to the product of their sizes (see Section 3 for more details).

An interesting context in which the products of normal subsets of finite simple groups play a role is the Waring problem for finite simple groups; see, for instance, [21,22,32,33,36,37,42,55,60,66], the references therein, and Segal's monograph [61] on word width and the affirmative solution by Nikolov and Segal of Serre's problem, whether every finite index subgroup of a (topologically) finitely generated profinite group is open.

By a *word*, we mean an element $w = w(x_1, \dots, x_d)$ of the free group F_d freely generated by $x_1, \dots, x_d$. A word w and a group G give rise to a word map $w : G^d \to G$ induced by substituting group elements $g_1, \dots, g_d$ in $x_1, \dots, x_d$, respectively; its image, denoted by $w(G)$, is a normal subset of G.

The classical Waring problem in number theory, solved by Hilbert and subsequently by Hardy and Littlewood using the circle method, deals with sums of nth powers of natural numbers (see [56]). In [55, 60], the analogous problem for finite simple groups G is studied; it is shown there that for every integer $n > 1$ there is a number $f(n)$ such that if G is a finite simple group not satisfying the identity $x^n = 1$, then every element of G is a product of $f(n)$ nth powers. In other words, if $w = w_n := x^n$, then $w(G)^{f(n)} = G$ for such groups G.

This result is improved in [66] for sufficiently large finite simple groups in several ways: the inexplicit function f (depending on w) is replaced by the fixed small number 3; the equality $w(G)^3 = G$ holds for all non-trivial words w provided that $|G| \geq N_w$; moreover, it is shown in [66] that, for all non-trivial words $w_1, w_2, w_3 \in F_d$ and all sufficiently large finite simple groups G, we have $w_1(G)w_2(G)w_3(G) = G$. This is improved by Larsen, Tiep, and me in [36] for length 2 products; i.e., for non-trivial words $w_1, w_2 \in F_d$ and all sufficiently large finite simple groups G, we have

$$w_1(G)w_2(G) = G. \tag{1.1}$$

The tools we apply in proving this and other results on word maps include representation theory and the Deligne–Lustig theory of characters, as well as algebraic geometry and some model theory; see Hrushovski's work on the elementary theory of Frobenius automorphism [27] and Varshavsky's strengthening of Fujiwara's proof of Deligne's conjecture [68].

There are various asymptotic results showing that word maps associated with words $w \neq 1$ on finite simple groups G have large image; see [31–33, 57]. In particular, it is shown by Larsen in [31] that $|w(G)| \geq |G|^{1-\varepsilon}$ for any $\varepsilon > 0$ provided that $|G| \gg 0$, and that for G of Lie type and bounded rank there exists $\varepsilon > 0$ (depending only on the rank of G) such that for all words $w \neq 1$ we have $|w(G)| \geq \varepsilon|G|$. In the recent preprint [35], we attempt to understand to what extent (1.1) can be extended to products of arbitrary large normal subsets of finite simple groups.

Let $\varepsilon > 0$ be a constant. Let G be a finite simple group and S and T normal subsets of G such that $|S|, |T| > \varepsilon|G|$. We are particularly interested in the following questions.

Question 1.3. Does every element in $G \smallsetminus \{e\}$ lie in ST if $|G|$ is sufficiently large?

Question 1.4. Does the ratio between the number of representations of each $g \in G \smallsetminus \{e\}$ and $\frac{|S||T|}{|G|}$ tend uniformly to 1 as $|G| \to \infty$?

Question 1.5. What happens in the special case $S = T$?

An affirmative answer to Question 1.4 implies an affirmative answer to Question 1.3 (and the same holds in the special case $S = T$).

We exclude the identity element e of G in Questions 1.3 and 1.4 because every conjugacy class C in a non-trivial finite group G satisfies $|C| = \frac{|G|}{n}$ for some $n \geq 2$, so each such group has a normal subset S with $\frac{|G|}{3} \leq |S| \leq \frac{2|G|}{3}$. Setting $T = G \smallsetminus S^{-1}$, we have $|T| \geq \frac{|G|}{3}$ and $e \notin ST$.

If G is non-trivial and we do not assume that $S, T \subseteq G$ are normal subsets, then we may choose $S, T \subseteq G$ of size at least $\lfloor \frac{|G|}{2} \rfloor$ such that $ST \not\supseteq G \smallsetminus \{e\}$; indeed, fix $g \in G \smallsetminus \{e\}$, choose S of the specified size, and let $T = G \smallsetminus S^{-1}g$.

Our answers to these questions are listed below.

Theorem 1.6. (1) *The answers to Questions 1.3 and 1.4 are negative if G ranges over all finite simple groups, or even just over the alternating groups, or just over all projective special linear groups.*

(2) *In the $S = T$ case, the answer to Question 1.4 is still negative for alternating groups.*

(3) *In the $S = T$ case, the answer to Question 1.3 is positive for alternating groups.*

(4) *If G is a group of Lie type of bounded rank, then the answers to Questions 1.3 and 1.4 are both positive.*

In view of this, we may say that the simple groups of Lie type of bounded rank are the most well behaved in this context, and that the alternating groups are mildly well behaved.

Let us now outline the proof of Theorem 1.6, starting with the case of alternating groups A_n. Part (3) in this case follows from the more detailed result below.

Proposition 1.7. *For every $s, t \geq 0$ with $s + t \leq 1$, there are normal subsets $S_n, T_n \subset A_n$ such that $|S_n|/|A_n| \to s$, $|T_n|/|A_n| \to t$, and $S_n T_n$ contains no 3-cycle.*

It follows that, for normal subsets $S, T \subset A_n$, even the inequalities $|S|, |T| \geq (1/2 - o(1))|A_n|$ do not imply $ST \supseteq A_n \setminus \{e\}$.

As for part (3) of Theorem 1.6, namely, the positive result for A_n in the case $S = T$, the following more detailed proposition shows that we obtain a covering result even when $\varepsilon \to 0$ rather fast.

Proposition 1.8. *For every $0 < \alpha < 1/4$, there exists $N > 0$ such that if $n \geq N$ and $T \subseteq A_n$ is a normal subset satisfying $|T| \geq \exp(-n^{\alpha}) \cdot |A_n|$, then $T^2 = A_n$.*

The main tool in the proof of Proposition 1.8 are strong character bounds for symmetric groups obtained in [32]. Roughly speaking, we show that for each $\sigma \in S_n$ there is a well-defined $E(\sigma) \in [0, 1]$ such that

$$\left|\chi(\sigma)\right| \leq \chi(1)^{E(\sigma)+o(1)} \quad \text{for all } \chi \in \operatorname{Irr}(S_n).$$

Applying these character bounds and other tools, we deduce that $E(\sigma) < 1/4$ implies $(\sigma^{S_n})^2 = A_n$ for all $n \gg 0$. It is also shown in [32] that, for every subset $T \subseteq A_n$ satisfying $|T| \geq \exp(-n^{\alpha}) \cdot |A_n|$ with $\alpha < 1/4$, a random $\sigma \in T$ satisfies $E(\sigma) < 1/4$ almost surely. We therefore deduce that there is $\sigma \in T$ such that $(\sigma^{S_n})^2 = A_n$ if $n \gg 0$. Finally, replacing σ^{S_n} with σ^{A_n} and using Erdős–Turán's statistical group theory (see, for instance, [9]), we show that $T^2 = A_n$ for $n \gg 0$.

We now turn to projective special linear groups $\mathrm{PSL}_n(q)$. We show the following.

Proposition 1.9. *Let q be a fixed prime power. Then, there exists $\varepsilon > 0$ such that, for every $n \geq 2$ which is relatively prime to $q - 1$, there are normal subsets $S_n, T_n \subset \mathrm{SL}_n(q) \cong \mathrm{PSL}_n(q)$ of density at least ε such that $S_n T_n$ does not contain any transvection.*

This result completes the proof of part (1) of Theorem 1.6.

Our proof of part (4) of Theorem 1.6, dealing with Lie-type groups of bounded rank, depends on a new result in algebraic geometry, which may be of independent interest; it may be regarded as a refinement of the classical Lang–Weil estimate [30] (see also Varshavsky [68]), which concerns the number of points in a finite product set inside a product variety which lies on a subvariety of the product variety. Another major ingredient of the proof is character theory. To explain the connection, we need some notation. For normal subsets $R_1, \ldots, R_k$ of a finite group G and $g \in G$, let $P_{R_1,\ldots,R_k}(g)$ denote the probability that $x_1 \cdots x_k = g$, where $x_i \in R_i$ are randomly chosen. Using this notation, we formulate and establish the following result, which is equivalent to part (4) of Theorem 1.6.

Theorem 1.10. *Let $G = X_r(q)$, a finite simple group of Lie type of rank r over F_q. Suppose that r is bounded and $q \to \infty$. Fix $\varepsilon > 0$ and let $S, T \subseteq G$ be normal subsets of size $\geq \varepsilon|G|$. Then, for every $g \in G \setminus \{e\}$ we have*

$$P_{S,T}(g) = \bigl(1 + o(1)\bigr)|G|^{-1}.$$

The relevance of character theory and character bounds to the proof of Theorem 1.10 stems from a classical result of Frobenius: let $C_1, \ldots, C_k \subset G$ be conjugacy classes, and $g \in G$. Then,

$$P_{C_1,\ldots,C_k}(g) = |G|^{-1} \sum_{\chi \in \mathrm{Irr}(G)} \frac{\chi(C_1) \cdots \chi(C_k)\overline{\chi(g)}}{\chi(1)^{k-1}}.$$

Frobenius' formula above is also useful for classical groups of unbounded rank. In this case, part (1) of Theorem 1.6 and the more detailed Proposition 1.8 provide counterexamples to 2-step covering by large normal subsets. It turns out that 3-step covering is achieved. More specifically, Question 1.3 for products of three normal subsets has a positive answer with a tiny $\varepsilon = |G|^{-\delta}$, which tends to zero as $|G| \to \infty$.

Theorem 1.11. *There exists a fixed $\delta > 0$ such that if G is a finite simple classical group and $R, S, T \subseteq G$ are normal subsets of size $\geq |G|^{1-\delta}$, then $RST = G$.*

Note that this result does not follow from Gowers' trick. Indeed, for G of rank $r \to \infty$, $|G|^{-\delta} \sim q^{-ar^2}$ is much smaller than $m(G)^{-1/3} \sim q^{-br}$.

The proof of Theorem 1.11 relies heavily on recent developments in representation theory and, more specifically, on the theory of *exponential character bounds* for finite simple groups G; these are bounds of the form

$$\bigl|\chi(g)\bigr| \leq \chi(1)^{\alpha(g)},$$

for various $g \in G$, where $\alpha(g) \in [0, 1]$ is an explicit function of g.

For symmetric and alternating groups, such bounds were first obtained by Fomin and Lulov [10] in 1997 for the so-called homogeneous permutations. Bounds for almost homogeneous permutations were subsequently obtained in [46] (see also [48]) with various applications to Fuchsian groups, Higman's conjecture, subgroup growth, and representation varieties. In [32], Larsen and I obtain essentially best-possible exponential character bounds for most permutations in S_n, with applications to word maps and other topics.

Exponential character bounds for finite simple groups of Lie type were recently obtained by Bezrukavnikov, Liebeck, Tiep, and me in [2], by Guralnick, Larsen, and Tiep in [19, 20], and by Taylor and Tiep in [67].

The proof of Theorem 1.11 relies mainly on the level theory of characters developed by Guralnick, Larsen, and Tiep in [19, 20], combined with earlier results on the Witten zeta function

$$\zeta^G(s) = \sum_{\chi \in \mathrm{Irr}(G)} \chi(1)^{-s}$$

and its abscissa of convergence obtained by Liebeck and me in [47].

More specifically, we apply a theorem from [20] according to which there exists an absolute constant $\gamma > 0$ such that if G is a finite simple classical group and $g \in G$ satisfies $|C_G(g)| \leq |G|^\gamma$, then $|\chi(g)| \leq \chi(1)^{1/4}$ for all $\chi \in \mathrm{Irr}(G)$.

We then apply Frobenius' formula and [47, Theorem 1.2], stating that, for any fixed $s > 0$ and r sufficiently large (in terms of s), $\zeta^G(s)$ converges and tends to 1 as $r \to \infty$. In fact, the case $s = 1/4$ suffices.

We now turn to applications of Theorem 1.6. We start with a direct (yet highly non-trivial) application to word maps. A major application, the proof of which is considerably harder, will be discussed in the next section (see Theorem 2.8).

For a non-trivial word $w \in F_d$ and a finite group G, consider the word map $w : G^d \to G$, and define $P_{w,G}(g) := |w^{-1}(g)|/|G|^d$. Thus, $P_{w,G}(g)$ is the probability that $w(g_1, \ldots, g_d) = g$ as $g_1, \ldots, g_d \in G$ are chosen uniformly and independently.

In [37, Theorem 4], we show that for every $\ell \geq 1$ there exists $N = N(\ell) := 2 \cdot 10^{18} \cdot \ell^4$ such that if $1 \neq w_1, \ldots, w_N \in F_d$ are pairwise disjoint words of length $\leq \ell$, G is a finite simple group, and U_G is the uniform distribution on G, then

$$\|P_{w_1 \cdots w_N, G} - U_G\|_\infty \to 0 \quad \text{as } |G| \to \infty;$$

namely, $P_{w_1 \cdots w_N, G}$ is almost uniform with respect to the L^∞ norm.

Surprisingly, changing the probabilistic model and using Theorem 1.10, we obtain an almost uniform distribution in L^∞ much more rapidly.

Corollary 1.12. *Let $1 \neq w_1, w_2 \in F_d$ and let G be a finite simple group of Lie type of bounded rank. Then,*

$$\|P_{w_1(G), w_2(G)} - U_G\|_\infty \to 0 \quad \textit{as } |G| \to \infty.$$

A version for classical groups of unbounded rank was previously implicitly obtained by Nikolov and Pyber in [57] using Gowers' theory of quasi-random groups; it shows that

$$\|P_{w_1(G), w_2(G), w_3(G)} - U_G\|_\infty \to 0 \quad \text{as } |G| \to \infty.$$

Note that Theorem 1.11 extends this result, since $w_i(G)$ are normal subsets of size at least $|G|^{1-\delta}$ by [31].

2. Permutation groups and derangements

A major application of our results from Section 1 on products of normal subsets concerns permutation groups and fixed-point-free permutations, also called *derangements*.

The study of derangements goes back three centuries.

In 1708, Monmort proved that the proportion of derangements in the symmetric group S_n (in its natural action) tends to $1/e$ as $n \to \infty$. Passing to general permutation groups $G \leq S_n$, it is easy to see that if G is intransitive it need not contain derangements (e.g., all permutations in G may have a common fixed point).

In the 1870s, Jordan showed that if $G \leq S_n$ is transitive and $2 \leq n < \infty$, then there exists a derangement $g \in G$ (this result fails to hold for infinite transitive permutation groups).

What can be said about the proportion of derangements in finite transitive permutation groups?

In 1990, Cameron and Cohen [7] proved that the proportion of derangements in transitive permutation groups of degree n is at least $1/n$ and that this lower bound is sharp (as shown by the example of Frobenius groups). Subsequently, it was conjectured that a much better lower bound holds for finite *simple* transitive permutation groups.

Conjecture 2.1 (Boston–Shalev, 1990s). *The proportion of derangements in any finite simple transitive permutation group is at least ε for some fixed $\varepsilon > 0$.*

Let $G \leq S_n$ be a transitive permutation group. Let $D(G)$ denote the set of derangements in G. Clearly, $D(G) = D(G)^{-1}$ and $D(G)$ is a normal subset of G. Let H be a point stabilizer in G. The set of derangements in G in this case is also denoted by $D(G, H)$. Clearly,

$$D(G, H) = G \setminus \bigcup_{g \in G} H^g.$$

Conjecture 2.1 states that if G is simple, then

$$\big|D(G)\big| \geq \varepsilon |G|,$$

for some absolute positive constant ε.

Impressive work on Conjecture 2.1 was carried out in 2002–2018 by Fulman and Guralnick (see, e.g., [11–13]), culminating in the following result.

Theorem 2.2 (Fulman–Guralnick, 2018). *The Boston–Shalev conjecture holds. Moreover, if G is sufficiently large we may take $\varepsilon = 0.016$.*

It would be nice to find an explicit number N such that if the finite simple transitive permutation group G has order at least N, then $|D(G)| \geq 0.016|G|$ or to find an explicit (possibly smaller) $\varepsilon > 0$ such that $|D(G)| \geq \varepsilon|G|$ without exceptions.

Since finite simple groups are quasi-random, Theorem 2.2, combined with Gowers' trick, yields the following.

Corollary 2.3. *For all sufficiently large finite simple transitive permutation groups G, every permutation in G is a product of three derangements, namely, $D(G)^3 = G$.*

Can we replace three by two? Note that the proof of Corollary 2.3 does not use the fact that $D(G)$ is a normal subset of G. Using the normality of $D(G)$, Theorem 1.6 becomes highly relevant. Applying parts (3) and (4) of it, noting that $e \in D(G)^2$, we immediately obtain the following.

Theorem 2.4. *Let G be a finite simple transitive permutation group which is alternating or of Lie type of bounded rank. If $|G| \gg 0$, then $D(G)^2 = G$; namely, every element of G is a product of two derangements.*

Indeed, we proved for the groups above that $T^2 = G$ for every normal subset $T \subseteq G$ of size $\geq \varepsilon|G|$, so take $T := D(G)$.

In order to extend Theorem 2.4 to all types of finite simple groups, it remains to deal with classical groups G of unbounded rank (since the sporadic groups have bounded order). We may assume that G is primitive; i.e., a point stabilizer $H < G$ is a maximal subgroup. Indeed, if H is not maximal, it is contained in a maximal subgroup M of G, and

$$D(G,H) = G \setminus \bigcup_{g \in G} H^g \supseteq G \setminus \bigcup_{g \in G} M^g = D(G,M),$$

so $D(G,M)^2 = G$ implies that $D(G,H)^2 = G$.

We need some additional tools in order to deal with the remaining case of classical groups of unbounded rank.

In 1993, Łuczak and Pyber [52] proved a conjecture of Cameron that as $n \to \infty$, almost all permutations in S_n do not lie in a proper transitive subgroup (not containing A_n). In the same paper, they pose a similar problem for $\mathrm{GL}_n(p)$, where p is a fixed prime and $n \to \infty$. In 1998, this problem was solved in [64].

Theorem 2.5. *Let q be a fixed prime power. Then, as $n \to \infty$, almost all matrices in $\mathrm{GL}_n(q)$ do not lie in a proper irreducible subgroup (not containing $\mathrm{SL}_n(q)$).*

In 2018, Fulman and Guralnick [13] proved a stronger result for all classical groups in dimension $n \to \infty$, where the size of the underlying field need not be fixed. In the case $G = \mathrm{Sp}_n(2^k)$, they exclude (apart from irreducible subgroups) the subgroups $O_n^{\pm}(2^k)$. We apply this to obtain the following.

Corollary 2.6. *Let G be a finite simple classical group in dimension $n \gg 0$. Let $H < G$ be a maximal subgroup. Suppose that H is irreducible and not $O_n^{\pm}(2^k)$ when $G = \mathrm{Sp}_n(2^k)$. Then, $D(G,H)^2 = G$.*

To prove this, let $X(G)$ denote the union of the above maximal subgroups $H < G$. Then, $|X(G)|/|G| \to 0$ as $n \to \infty$. Therefore, $|X(G)| < |G|/2$ for $n \gg 0$. Fixing one such subgroup H (noting that $X(G)$ is closed under conjugation), we see that $\bigcup_{g \in G} H^g \subseteq X(G)$ implies that $|D(G, H)| \geq |G| - |X(G)| > |G|/2$. Applying an observation from Section 1, it follows that $D(G, H)^2 = G$.

Hence we may assume that H is reducible (namely, a parabolic subgroup) or $G = \mathrm{Sp}_n(2^k)$ and $H = O_n^{\pm}(2^k)$.

Our next result settles the problem in additional cases.

Proposition 2.7. *There are absolute constants c_1, c_2 such that the following holds. Let $G \in \mathrm{Cl}_n(q)$ be a finite simple classical primitive permutation group with point stabilizer H. If q is even, assume $(G, H) \neq (\mathrm{Sp}_n(\mathbb{F}_q), O_n^{\pm}(\mathbb{F}_q))$. Suppose that $n \geq c_1$ and the action is not a subspace action on subspaces of dimension $k \leq c_2$. Then, $D(G)^2 = G$.*

To show this, we may assume that H is reducible; namely, G acts in subspace action, say on subspaces (non-degenerate or totally singular for $G \neq \mathrm{PSL}_n(q)$) of dimension k, with $1 \leq k \leq n/2$. Theorems 6.4, 9.4, 9.10, 9.17, and 9.30 of [12] show that, as $k \to \infty$, the proportion of derangements in G is $1 - O(k^{-0.005})$, which tends to 1. The result follows as before.

We are left with very concrete cases, of subspace action on subspaces of bounded dimension and of $\mathrm{Sp}_n(\mathbb{F}_q)$ for q even acting on cosets of $GO_n^{\pm}(\mathbb{F}_q)$. These cases are handled using character methods and the theory of symbols. Roughly speaking, we apply the method of [53] and its extension in [36] and use weakly orthogonal tori T_1, T_2 and regular semisimple elements $t_i \in T_i$ such that only few (unipotent) characters $\chi \in \mathrm{Irr}(G)$ satisfy $\chi(t_1)\chi(t_2) \neq 0$. This helps show that $t_1^G t_2^G \supseteq G \setminus \{e\}$. This rather long case-by-case study completes the proof of the main result of this section (see [34]).

Theorem 2.8. *Let G be a finite simple transitive permutation group. If G is sufficiently large, then every element of G is a product of two derangements.*

We conjecture that the assumption that G is sufficiently large is not needed; namely:

Conjecture 2.9. *Let G be a finite simple transitive permutation group. Then, every element of G is a product of two derangements.*

Computations carried out by Eamonn O'Brien provide strong evidence in favor of this conjecture, but proving it seems to require new methods.

3. Subset growth

The celebrated product theorem of [5,58], which is part of the deep theory of approximate subgroups [4] following the pioneeing work of Helfgott on $\mathrm{SL}_2(p)$ [24] (see also [25]) and Hrushovski's model-theoretic approach [26], shows that for finite simple groups G of Lie type and bounded rank there exists $\varepsilon > 0$ such that, for every subset $A \subseteq G$ which generates G, either $|A^3| \geq |A|^{1+\varepsilon}$ or $A^3 = G$.

Can we extend this result to general finite simple groups? The answer is known to be negative, as shown by counterexamples for classical groups of unbounded rank and alternating groups of unbounded degree.

However, the situation changes dramatically if we replace arbitrary subsets by normal subsets. A first result in this direction was obtained in [66] before the product theorem was fully established. Indeed, Theorem 2.7 there states the following.

Theorem 3.1. *For any $\delta > 0$, there exists $\varepsilon > 0$ depending only on δ such that if G is a finite simple group and C is a conjugacy class of G of size at most $|G|^{1-\delta}$, then $|C^3| \geq |C|^{1+\varepsilon}$.*

Note that an upper bound on the size of C of the type above is necessary for the conclusion to hold. The proof of Theorem 3.1 uses tools from character theory, properties of the Witten zeta function obtained by Liebeck and me in [47], as well as [24, Lemma 2.2] of Helfgott and its proof.

Can we extend this 3-step growth result to 2-step growth results, replacing C^3 by C^2? It turns out that the answer is positive if G is a finite simple group of Lie type of bounded rank. Indeed, we have the following (see [66, Proposition 10.4]).

Theorem 3.2. *If C is a conjugacy class of a finite simple group G of Lie type, then $|C^2| \geq |C|^{1+\varepsilon}$, where $\varepsilon > 0$ depends only on the rank of G.*

The above result was extended by Gill, Pyber, Short, and Szabó in [15, Theorem 1.5], where conjugacy classes C are replaced by arbitrary normal subsets T, and G is an arbitrary finite simple group.

Theorem 3.3. *There are absolute constants $b \in \mathbb{N}$ and $\varepsilon > 0$ such that, for any normal subset T of a finite simple group G, either $T^b = G$ or $|T^2| \geq |T|^{1+\varepsilon}$.*

Subsequently, Liebeck, Schul, and I obtained stronger expansion results for normal subsets in [43].

Theorem 3.4. *Given any $\varepsilon > 0$, there exists $b \in \mathbb{N}$ such that, for any normal subset T of any finite simple group G, either $T^b = G$ or $|T^2| \geq |T|^{2-\varepsilon}$.*

Theorem 3.4 is deduced from the following result.

Theorem 3.5. *Given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that if T is a normal subset of a finite simple group G satisfying $|T| \leq |G|^{\delta}$, then $|T^2| \geq |T|^{2-\varepsilon}$.*

Obviously $|T^2| \leq |T|^2$, so Theorem 3.5 shows that small normal subsets of finite simple groups expand almost as quickly as possible.

Note that some upper bound on the size of T is needed in order for the conclusion to hold.

To deduce Theorem 3.4, fix $\varepsilon > 0$ and choose $\delta = \delta(\varepsilon) > 0$ as above. Recall that, by the main result of [45], there exists an absolute constant c and a positive integer $k \leq c \log |G| / \log |T|$ such that $T^k = G$ for every (non-trivial) normal subset T of a finite simple group G. Hence, if $|T| \geq |G|^{\delta}$, then $T^k = G$ for some $k \leq c\delta$. Thus, Theorem 3.4 holds with $b = \lfloor c\delta \rfloor$.

Theorem 3.5 holds vacuously for simple groups of bounded order or of bounded rank, since for these we may choose δ so small that $|T| > |G|^{\delta}$ for all non-trivial normal subsets T; in particular, it holds for the sporadic groups and the exceptional groups of Lie type. It therefore remains to prove the theorem for classical groups of large rank and alternating groups of large degree.

We deduce Theorem 3.5 from the following more general result.

Theorem 3.6. *Given any $\varepsilon > 0$, there exists $\delta > 0$ such that if T_1, T_2 are normal subsets of a finite simple group G satisfying $|T_1|, |T_2| \leq |G|^{\delta}$, then $|T_1 T_2| \geq (|T_1|\,|T_2|)^{1-\varepsilon}$.*

The proof of Theorem 3.6 in [43] is based on results from [44, 45, 47], together with some new results of independent interest on the size of the conjugacy classes in classical groups and in symmetric groups in terms of the *support* of their elements.

The support of a permutation $x \in S_n$ is the number of points moved by x. Let $C \subset S_n$ be a non-trivial conjugacy class and let s be the support of its elements (obviously all the elements of C have the same support), which may be regarded as the support of C. Then, $2 \leq s \leq n$. For our purpose, it is essential to obtain good estimates on the size of C in terms of its support s. We show that

$$|C| \leq \frac{n!}{(n-s)!s}$$

for all s and that

$$|C| \geq \frac{n!}{(n-s)!2^{s/2}\lfloor s/2 \rfloor!}$$

for all $s \neq 3$.

Note that the above lower bound on $|C|$ is best possible, since it is attained in the case where the permutations in C decompose into $s/2$ 2-cycles (s even). The upper bound on $|C|$ is also sharp; it is attained when C consists of s-cycles. Finally, if $s = 3$, then the lower bound does not hold, but it does hold for all s if we replace $\lfloor s/2 \rfloor$ by $\lceil s/2 \rceil$.

Next, let G be one of the classical groups $L_n^{\pm}(q)$, $PSp_n(q)$ or $PO_n^{\pm}(q)$, and let $V = V_n(q^u)$ be the natural module for G with n large, where $u = 2$ if G is unitary and $u = 1$ otherwise. Let $\overline{\mathbb{F}}$ be the algebraic closure of $\mathbb{F}_q$, and let $\overline{V} = V \otimes \overline{\mathbb{F}}$. Let $x \in G$, and let $\hat{x}$ be a preimage of x in $\mathrm{GL}(V)$. Define

$$\nu(x) = \nu_{V,\overline{\mathbb{F}}}(x) = \min\left\{\dim[\overline{V}, \lambda\hat{x}] : \lambda \in \overline{\mathbb{F}}^*\right\},$$

where $[\overline{V}, \lambda\hat{x}]$ denotes the subspace $\overline{V}(\lambda\hat{x} - Id_{\overline{V}})$. We shall refer to $\nu(x)$ as the *support* of x.

Define

$$a(G) = \begin{cases} 1, & \text{if } G = L_n^{\pm}(q), \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

The inequalities we state below, which are an extension of [44, Lemma 3.4], show that $\nu(x)$ is closely related to the size of the conjugacy class $C = x^G$. Suppose that $\nu(x) = s < \frac{n}{2}$, and let $a = a(G)$. We prove that

$$c_1 q^{2as(n-s-1)} \leq |x^G| \leq c_2 q^{as(2n-s+1)}$$

for some absolute constants $c_1, c_2 > 0$.

In fact, under the assumptions of Theorem 3.6, we establish a stronger conclusion: there exists a single conjugacy class $C \subseteq T_1T_2$ such that $|C| \geq (|T_1|\,|T_2|)^{1-\varepsilon}$. The notion of the support of elements of G plays a key role in our argument.

A similar result for k subsets follows inductively from Theorem 3.6.

Corollary 3.7. *Given $\varepsilon > 0$ and $k \in \mathbb{N}$, there exists $\delta > 0$ such that if $T_1, \ldots, T_k \subseteq G$ are normal subsets of a finite simple group G with $|T_i| \leq |G|^\delta$ for $i = 1, \ldots, k$, then $|T_1 \cdots T_k| \geq (|T_1| \cdots |T_k|)^{1-\varepsilon}$. In particular, $|T^k| \geq |T|^{k-\varepsilon}$ for every normal subset T of G satisfying $|T| \leq |G|^\delta$, where δ depends on ε and k.*

Finally, we prove a result analogous to Theorem 3.6 for algebraic groups over algebraically closed fields.

Theorem 3.8. *For any $\varepsilon > 0$, there exists $\delta > 0$ such that if C_1, C_2 are conjugacy classes in a simple algebraic group G satisfying $\dim C_i \leq \delta \dim G$ for $i = 1, 2$, then the product C_1C_2 contains a conjugacy class of dimension at least $(1 - \varepsilon)(\dim C_1 + \dim C_2)$.*

4. Character growth and covering

The main goal of this section, based mainly on the recent preprint [38] with Larsen and Tiep, is to study covering and growth phenomena in representation theory, with emphasis on (complex) representations of the finite simple groups G of Lie type. Here, products of subsets of G are replaced by tensor products of representations (or equivalently, products of characters). Our results on tensor product growth are somewhat stronger than the product theorem in two senses: instead of 3-step growth, we establish 2-step growth, as well as uniform growth when the rank of G tends to infinity.

In some cases, the results of this section are character-theoretic analogues of results from the previous section, dealing with product growth of conjugacy classes (corresponding to irreducible characters) and normal subsets (corresponding to arbitrary characters). An irreducible constituent of an arbitrary character may be regarded as an analogue of a conjugacy class contained in a normal subset.

Covering results by products of characters of finite simple groups were obtained by Liebeck, Tiep, and me in the recent papers [49,50]. These papers study the McKay graphs of finite simple groups, with emphasis on their diameter.

We need some background and notation. For a finite group G and a complex character α of G, the *McKay graph* $MC(G,\alpha)$ is defined to be the directed graph with vertex set $\mathrm{Irr}(G)$, and with an edge from χ_1 to χ_2 if and only if χ_2 is a constituent of $\alpha\chi_1$.

A classical result of Burnside and Brauer [3] shows that $MC(G,\alpha)$ is connected if and only if α is faithful; furthermore, in this case an upper bound for the diameter $\mathsf{diam}\, MC(G,\alpha)$ is given by $N-1$, where N is the number of distinct values of α. This means that $\sum_{j=0}^{N-1} \alpha^j$ contains every irreducible character of G.

An obvious lower bound for $\mathsf{diam}\, MC(G,\alpha)$ (when $\alpha(1) > 1$) is given by $\frac{\log \mathsf{b}(G)}{\log \alpha(1)}$, where $\mathsf{b}(G)$ is the largest degree of an irreducible character of G. This lower bound (which can be slightly improved) is in general far from tight. However, finite simple groups often behave better than arbitrary groups, and for them we stated the following conjecture in [50].

Conjecture 4.1. *There is an absolute constant c such that, for any finite non-abelian simple group G and any non-trivial irreducible character α of G, we have*

$$\mathsf{diam}\, MC(G,\alpha) \le c\frac{\log |G|}{\log \alpha(1)}.$$

This conjecture may be regarded as a representation-theoretic analogue of [45, Theorem 1.1] on the diameter of the Cayley graph $\Gamma(G,S)$ of a finite simple group G with respect to a (non-trivial) normal subset S.

Various results supporting this conjecture were obtained in [49, 50], where it is proved for several families of groups of Lie type and for alternating groups following Sellke's paper [62] proving it for symmetric groups. In [49], we also obtain some results showing that, under suitable assumptions, products $\chi_1 \cdots \chi_m$ of possibly different characters cover $\mathrm{Irr}(G)$ (namely, every irreducible character is a constituent of the above product). In [39], Larsen and Tiep have completed the proof of Conjecture 4.1.

A more recent covering result of Sellke [63] is a character-theoretic analogue of Gowers' trick and the theory of quasi-random groups discussed in Section 1, which focuses on 3-step covering. We need some notation.

Let G be a finite group. We say that an arbitrary complex character ψ of G *covers* $\operatorname{Irr}(G)$ if every irreducible character of G is a constituent of ψ. If $X = \{\chi_1, \ldots, \chi_k\}$ is a set of (pairwise distinct) irreducible characters of G, we define

$$|X| = \sum_{i=1}^{k} \chi_i(1)^2.$$

This is the Plancherel measure, normalized so that $|\operatorname{Irr}(G)| = |G|$. If χ is any character of G, we define $|\chi| = |X_\chi|$, where X_χ denotes the set of distinct irreducible constituents of χ. We show in [38] that the function sending χ to $|\chi|$ has convenient properties: it is sub-multiplicative and satisfies the triangle inequality in the sense that

$$|\chi_1\chi_2| \leq |\chi_1| \cdot |\chi_2| \quad \text{and} \quad |\chi_1 + \chi_2| \leq |\chi_1| + |\chi_2|. \tag{4.1}$$

We can now state the covering result mentioned above, which is the main part of [63, Theorem 1.3]. For a finite group G, let $c(G)$ denote the minimal size of a conjugacy class $\neq \{e\}$ in G. Let us say that a collection F of finite groups is *conjugacy-random* if $c(G) \to \infty$ as G ranges over the groups in F.

Theorem 4.2 (Sellke, 2021). *Let F be a conjugacy-random set of finite groups. Fix $\varepsilon > 0$. Let $G \in F$ and let χ_1, χ_2, χ_3 be (not necessarily irreducible) characters of G with the property that $|\chi_1|, |\chi_2|, |\chi_3| \geq \varepsilon|G|$. Then, $\chi_1\chi_2\chi_3$ covers* $\operatorname{Irr}(G)$ *provided $|G|$ is sufficiently large.*

While most of our results below establish rapid tensor product growth in various situations, some of them, i.e., Theorems 4.8 and 4.9, are covering results, while Theorem 4.7 establishes a growth-or-covering phenomenon.

Recall that G is *quasisimple* if $G = [G, G]$ and $G/\mathbf{Z}(G)$ is simple.

Our first growth results are as follows.

Theorem 4.3. *For all $\delta > 0$, there exists $\varepsilon > 0$ such that if G is a finite quasisimple group of Lie type and χ is an irreducible character of G with $|\chi| \leq |G|^{1-\delta}$, then $|\chi^2| \geq |\chi|^{1+\varepsilon}$ and $|\chi\overline{\chi}| \geq |\chi|^{1+\varepsilon}$.*

We also have a version of this result for general characters in groups of high rank.

Theorem 4.4. *For all $\delta > 0$, there exist $\varepsilon > 0$ and $R > 0$ such that if G is a finite quasisimple group of Lie type and rank $\geq R$ and χ is any (not necessarily irreducible) character of G with $|\chi| \leq |G|^{1-\delta}$, then $|\chi^2| \geq |\chi|^{1+\varepsilon}$ and $|\chi\overline{\chi}| \geq |\chi|^{1+\varepsilon}$.*

An essential tool in the proofs of most of the results in this section is a new uniform character bound obtained by Larsen and Tiep [39, Theorem A]. The proofs of Theorems 4.3 and 4.4 present ε as an explicit function of δ, e.g., $\varepsilon = \frac{c\delta}{4+2c(1-\delta)}$ in Theorem 4.3, where $c > 0$ is the absolute constant in [39, Theorem A]. Moreover, if G is sufficiently large but of bounded rank r, and χ is irreducible, then $\varepsilon = \frac{\delta}{2-2\delta}$ will do; for example, any irreducible character χ of G with $|\chi| \leq |G|^{1/2}$ satisfies $|\chi^2| \geq |\chi|^{3/2}$. Can we establish faster growth when $|\chi|$ is smaller?

Our next result provides an affirmative answer and may be regarded as a character-theoretic analogue of the main result of Section 3, namely, Theorem 3.6 (which obviously implies Theorem 3.4).

Theorem 4.5. *For any $\varepsilon > 0$, there exists an explicit $\delta > 0$ such that the following statement holds. If G is a finite quasisimple group of Lie type and χ_1, χ_2 are any (not necessarily irreducible) characters of G with $|\chi_1|, |\chi_2| \leq |G|^\delta$, then*

$$|\chi_1\chi_2| \geq \big(|\chi_1| \cdot |\chi_2|\big)^{1-\varepsilon}.$$

In particular, if χ is a character of G satisfying $|\chi| \leq |G|^\delta$, then $|\chi^2| \geq |\chi|^{2-2\varepsilon}$.

The inequality $|\chi_1\chi_2| \leq |\chi_1| \cdot |\chi_2|$ mentioned in (4.1) shows that the growth established in Theorem 4.5 is almost best possible.

Theorem 4.5 is easily extended to products of arbitrary length, in the spirit of Corollary 3.7 for k normal subsets.

Corollary 4.6. *For any $\varepsilon > 0$ and any integer $k \geq 1$, there exists an explicit $\gamma = \gamma(\varepsilon, k) > 0$ such that the following statement holds. If G is a finite quasisimple group of Lie type and $\chi_1, \chi_2, \dots, \chi_k$ are any (not necessarily irreducible) characters of G with $|\chi_1|, |\chi_2|, \dots, |\chi_k| \leq |G|^\gamma$, then*

$$|\chi_1\chi_2\cdots\chi_k| \geq \big(|\chi_1| \cdot |\chi_2| \cdots |\chi_k|\big)^{1-\varepsilon}.$$

In particular, if χ is a character of G satisfying $|\chi| \leq |G|^\gamma$, then $|\chi^k| \geq |\chi|^{k-k\varepsilon}$.

To show this, we prove by induction on $k \geq 2$ the following equivalent statement.

For any $\varepsilon > 0$ and any $k \geq 2$, there exists an explicit $\gamma > 0$ (depending on both ε and k) such that if G is a finite quasisimple group of Lie type and $\chi_1, \chi_2, \dots, \chi_k$ are any characters of G with $|\chi_1|, |\chi_2|, \dots, |\chi_k| \leq |G|^\gamma$, then

$$|\chi_1\chi_2\cdots\chi_k| \geq \big(|\chi_1| \cdot |\chi_2| \cdots |\chi_k|\big)^{1-k\varepsilon}.$$

We will show that this statement holds with $\gamma := \delta/(k-1)$, where δ is the constant in Theorem 4.5. The case $k = 2$ is already established in Theorem 4.5. For the inductive step, note that (4.1) and the induction hypothesis yield

$$\big(|\chi_2|\cdots|\chi_k|\big)^{1-(k-1)\varepsilon} \le |\chi_2\cdots\chi_k| \le \prod_{i=2}^{k}|\chi_i| \le |G|^{\gamma(k-1)} \le |G|^{\delta}.$$

Since $|\chi_1| \le |G|^{\delta}$, by Theorem 4.5 we have

$$\begin{aligned}|\chi_1\chi_2\cdots\chi_k| &\ge \big(|\chi_1|\cdot|\chi_2\cdots\chi_k|\big)^{1-\varepsilon}\\ &\ge \Big(|\chi_1|\cdot\big(|\chi_2|\cdots|\chi_k|\big)^{1-(k-1)\varepsilon}\Big)^{1-\varepsilon}\\ &\ge \big(|\chi_1|\cdot|\chi_2|\cdots|\chi_k|\big)^{1-k\varepsilon}.\end{aligned}$$

The above result shows that, for any $\varepsilon > 0$ and any integer $k \ge 2$, there exists an explicit $\delta = \delta(\varepsilon, k) > 0$ such that, for G as above and any (not necessarily irreducible) character χ of G satisfying $|\chi| \le |G|^{\delta}$, we have $|\chi^k| \ge |\chi|^{k-\varepsilon}$; indeed, define $\delta(\varepsilon, k) = \gamma(\varepsilon/k, k)$.

Applying Theorem 4.5, we deduce the following result, which is a character-theoretic analogue of Theorem 3.4.

Theorem 4.7. *For all $\varepsilon > 0$, there exists an explicit positive integer b such that if G is a finite simple group of Lie type and χ is any (not necessarily irreducible) character of G, then either χ^b contains every irreducible character of G or $|\chi^2| \ge |\chi|^{2-\varepsilon}$.*

In view of Gowers' theorem, it is natural to ask whether $b = 3$ suffices in Theorem 4.7 when $|\chi|$ is sufficiently large. Sellke's theorem (Theorem 4.2) shows that the answer to this question is affirmative for large G provided that $|\chi| \ge |G|^{\delta}$ for some fixed $\delta > 0$. We therefore ask the following.

Question. If G is a finite simple group of Lie type and χ is an arbitrary character of G such that $|\chi| \ge |G|^{\delta}$ for some fixed $\delta > 0$, is it true that $|\chi^3| = |G|$ provided $|G| \gg 0$?

We remark that the stronger equality $|\chi^2| = |G|$ does not always hold, as shown by the example of $\mathrm{PSU}_{2n+1}(q)$ (see [23, Theorem 1.2]). On the other hand, for certain simple groups of Lie type, we can bring b down to 6 or 7.

Theorem 4.8. *If $G = \mathrm{PSL}_n(q)$ and q is sufficiently large in terms of n, then $|\chi| \ge |G|^{11/12}$ implies that $|\chi^6| = |G|$. If $G = \mathrm{PSU}_n(q)$ and q is sufficiently large in terms of n, then $|\chi| \ge |G|^{11/12}$ implies that $|\chi^7| = |G|$.*

Our next result is an analogue of Theorem 1.1 by Maróti and Pyber (following [45] and Rodgers–Saxl [59]), where the normal subsets are replaced by characters of G. In the case where the characters are irreducible, this analogue was conjectured by Gill in [14] and proved by Larsen and Tiep in [39, Theorem 8.5]. A more general version, for arbitrary characters, is given below.

Theorem 4.9. *There exists an explicit constant $c > 0$ such that the following statement holds. If G is a finite simple group of Lie type, $m \geq 1$ any integer, and $\chi_1, \chi_2, \dots, \chi_m$ are any (not necessarily irreducible) characters of G with $\prod_{i=1}^m |\chi_i| \geq |G|^c$, then $|\chi_1 \chi_2 \cdots \chi_m| = |G|$ and thus $\chi_1 \chi_2 \cdots \chi_m$ contains every irreducible character of G.*

Finally, we prove an analogue of Theorem 4.3 for compact semisimple Lie groups.

Theorem 4.10. *Let G be a compact semisimple Lie group. Then, there exists $\varepsilon > 0$ such that, for each irreducible character χ of G, we have $|\chi^2| \geq |\chi|^{1+\varepsilon}$.*

Funding. The author was supported in part by BSF grant 2016072, ISF grant 686/17, and the Vinik Chair of Mathematics which he holds.

References

[1] L. Babai, N. Nikolov, and L. Pyber, Product growth and mixing in finite groups. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 248–257, ACM, New York, 2008 Zbl 1192.60016 MR 2485310

[2] R. Bezrukavnikov, M. W. Liebeck, A. Shalev, and P. H. Tiep, Character bounds for finite groups of Lie type. *Acta Math.* **221** (2018), no. 1, 1–57 Zbl 06983623 MR 3877017

[3] R. Brauer, A note on theorems of Burnside and Blichfeldt. *Proc. Amer. Math. Soc.* **15** (1964), 31–34 Zbl 0122.27503 MR 158004

[4] E. Breuillard, Lectures on approximate groups and Hilbert's 5th problem. In *Recent Trends in Combinatorics*, pp. 369–404, IMA Vol. Math. Appl. 159, Springer, Cham, 2016 Zbl 1407.11019 MR 3526417

[5] E. Breuillard, B. Green, and T. Tao, Approximate subgroups of linear groups. *Geom. Funct. Anal.* **21** (2011), no. 4, 774–819 Zbl 1229.20045 MR 2827010

[6] E. Breuillard, B. Green, and T. Tao, Suzuki groups as expanders. *Groups Geom. Dyn.* **5** (2011), no. 2, 281–299 Zbl 1247.20017 MR 2782174

[7] P. J. Cameron and A. M. Cohen, On the number of fixed point free elements in a permutation group. A collection of contributions in honour of Jack van Lint. *Discrete Math.* **106/107** (1992), 135–138 Zbl 0813.20001 MR 1181907

[8] E. W. Ellers and N. Gordeev, On the conjectures of J. Thompson and O. Ore. *Trans. Amer. Math. Soc.* **350** (1998), no. 9, 3657–3671 Zbl 0910.20007 MR 1422600

[9] P. Erdős and P. Turán, On some problems of a statistical group-theory. II. *Acta math. Acad. Sci. Hungar.* **18** (1967), 151–163 Zbl 0189.31302 MR 0207810

[10] S. V. Fomin and N. Lulov, On the number of rim hook tableaux. *J. Math. Sci. (New York)* **87** (1997), 4118–4123 Zbl 0909.05046

[11] J. Fulman and R. Guralnick, Bounds on the number and sizes of conjugacy classes in finite Chevalley groups with applications to derangements. *Trans. Amer. Math. Soc.* **364** (2012), no. 6, 3023–3070 Zbl 1256.20048 MR 2888238

[12] J. Fulman and R. Guralnick, Derangements in subspace actions of finite classical groups. *Trans. Amer. Math. Soc.* **369** (2017), no. 4, 2521–2572 Zbl 1431.20033 MR 3592520

[13] J. Fulman and R. Guralnick, Derangements in finite classical groups for actions related to extension field and imprimitive subgroups and the solution of the Boston–Shalev conjecture. *Trans. Amer. Math. Soc.* **370** (2018), no. 7, 4601–4622 Zbl 06862789 MR 3812089

[14] N. Gill, A Rodgers–Saxl type conjecture for characters. https://nickpgill.github.io/a-rodgers-saxl-conjecture-for-characters

[15] N. Gill, L. Pyber, I. Short, and E. Szabó, On the product decomposition conjecture for finite simple groups. *Groups Geom. Dyn.* **7** (2013), no. 4, 867–882 Zbl 1355.20013 MR 3134028

[16] N. Gill, L. Pyber, and E. Szabó, A generalization of a theorem of Rodgers and Saxl for simple groups of bounded rank. *Bull. Lond. Math. Soc.* **52** (2020), no. 3, 464–471 Zbl 07224936 MR 4171380

[17] W. T. Gowers, Quasirandom groups. *Combin. Probab. Comput.* **17** (2008), no. 3, 363–387 Zbl 1191.20016 MR 2410393

[18] R. Guralnick and G. Malle, Products of conjugacy classes and fixed point spaces. *J. Amer. Math. Soc.* **25** (2012), no. 1, 77–121 Zbl 1286.20007 MR 2833479

[19] R. M. Guralnick, M. Larsen, and P. H. Tiep, Character levels and character bounds. II. 2019, arXiv:1904.08070v1

[20] R. M. Guralnick, M. Larsen, and P. H. Tiep, Character levels and character bounds. *Forum Math. Pi* **8** (2020), e2 Zbl 07158138 MR 4061963

[21] R. M. Guralnick, M. W. Liebeck, E. A. O'Brien, A. Shalev, and P. H. Tiep, Surjective word maps and Burnside's $p^a q^b$ theorem. *Invent. Math.* **213** (2018), no. 2, 589–695 Zbl 1397.20037 MR 3827208

[22] R. M. Guralnick and P. H. Tiep, Effective results on the Waring problem for finite simple groups. *Amer. J. Math.* **137** (2015), no. 5, 1401–1430 Zbl 1338.20009 MR 3405871

[23] G. Heide, J. Saxl, P. H. Tiep, and A. E. Zalesski, Conjugacy action, induced representations and the Steinberg square for simple groups of Lie type. *Proc. Lond. Math. Soc. (3)* **106** (2013), no. 4, 908–930 Zbl 1372.20017 MR 3056296

[24] H. A. Helfgott, Growth and generation in $\mathrm{SL}_2(\mathbb{Z}/p\mathbb{Z})$. *Ann. of Math. (2)* **167** (2008), no. 2, 601–623 Zbl 1213.20045 MR 2415382

[25] H. A. Helfgott, Growth in $\mathrm{SL}_3(\mathbb{Z}/p\mathbb{Z})$. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 3, 761–851 Zbl 1235.20047 MR 2781932

[26] E. Hrushovski, Stable group theory and approximate subgroups. *J. Amer. Math. Soc.* **25** (2012), no. 1, 189–243 Zbl 1259.03049 MR 2833482

[27] E. Hrushovski, The elementary theory of the Frobenius automorphisms. 2022, arXiv: math/0406514v2

[28] M. Kassabov, Symmetric groups and expander graphs. *Invent. Math.* **170** (2007), no. 2, 327–354 Zbl 1191.20002 MR 2342639

[29] V. Landazuri and G. M. Seitz, On the minimal degrees of projective representations of the finite Chevalley groups. *J. Algebra* **32** (1974), 418–443 Zbl 0325.20008 MR 360852

[30] S. Lang and A. Weil, Number of points of varieties in finite fields. *Amer. J. Math.* **76** (1954), 819–827 Zbl 0058.27202 MR 65218

[31] M. Larsen, Word maps have large image. *Israel J. Math.* **139** (2004), 149–156 Zbl 1130.20310 MR 2041227

[32] M. Larsen and A. Shalev, Characters of symmetric groups: sharp bounds and applications. *Invent. Math.* **174** (2008), no. 3, 645–687 Zbl 1166.20009 MR 2453603

[33] M. Larsen and A. Shalev, Word maps and Waring type problems. *J. Amer. Math. Soc.* **22** (2009), no. 2, 437–466 Zbl 1206.20014 MR 2476780

[34] M. Larsen, A. Shalev, and P. H. Tiep, Products of derangements in simple permutation groups. *Int. Math. Res. Not. IMRN*, to appear

[35] M. Larsen, A. Shalev, and P. H. Tiep, Products of normal subsets. Preprint

[36] M. Larsen, A. Shalev, and P. H. Tiep, The Waring problem for finite simple groups. *Ann. of Math. (2)* **174** (2011), no. 3, 1885–1950 Zbl 1283.20008 MR 2846493

[37] M. Larsen, A. Shalev, and P. H. Tiep, Probabilistic Waring problems for finite simple groups. *Ann. of Math. (2)* **190** (2019), no. 2, 561–608 Zbl 1448.20063 MR 3997129

[38] M. Larsen, A. Shalev, and P. H. Tiep, Representations and tensor product growth. 2021, arXiv:2104.11716

[39] M. Larsen and P. H. Tiep, Uniform character bounds for finite classical groups. Preprint

[40] M. W. Liebeck, N. Nikolov, and A. Shalev, Groups of Lie type as products of SL_2 subgroups. *J. Algebra* **326** (2011), 201–207 Zbl 1225.20016 MR 2746060

[41] M. W. Liebeck, N. Nikolov, and A. Shalev, Product decompositions in finite simple groups. *Bull. Lond. Math. Soc.* **44** (2012), no. 3, 469–472 Zbl 1250.20018 MR 2966992

[42] M. W. Liebeck, E. A. O'Brien, A. Shalev, and P. H. Tiep, The Ore conjecture. *J. Eur. Math. Soc. (JEMS)* **12** (2010), no. 4, 939–1008 Zbl 1205.20011 MR 2654085

[43] M. W. Liebeck, G. Schul, and A. Shalev, Rapid growth in finite simple groups. *Trans. Amer. Math. Soc.* **369** (2017), no. 12, 8765–8779 Zbl 06790363 MR 3710643

[44] M. W. Liebeck and A. Shalev, Simple groups, permutation groups, and probability. *J. Amer. Math. Soc.* **12** (1999), no. 2, 497–520 Zbl 0916.20003 MR 1639620

[45] M. W. Liebeck and A. Shalev, Diameters of finite simple groups: sharp bounds and applications. *Ann. of Math. (2)* **154** (2001), no. 2, 383–406 Zbl 1003.20014 MR 1865975

[46] M. W. Liebeck and A. Shalev, Fuchsian groups, coverings of Riemann surfaces, subgroup growth, random quotients and random walks. *J. Algebra* **276** (2004), no. 2, 552–601 Zbl 1068.20052 MR 2058457

[47] M. W. Liebeck and A. Shalev, Character degrees and random walks in finite groups of Lie type. *Proc. London Math. Soc. (3)* **90** (2005), no. 1, 61–86 Zbl 1077.20020 MR 2107038

[48] M. W. Liebeck and A. Shalev, Fuchsian groups, finite simple groups and representation varieties. *Invent. Math.* **159** (2005), no. 2, 317–367 Zbl 1134.20059 MR 2116277

[49] M. W. Liebeck, A. Shalev, and P. H. Tiep, McKay graphs for alternating and classical groups. *Trans. Amer. Math. Soc.* **374** (2021), no. 8, 5651–5676 Zbl 07377376 MR 4293783

[50] M. W. Liebeck, A. Shalev, and P. H. Tiep, On the diameters of McKay graphs for finite simple groups. *Israel J. Math.* **241** (2021), no. 1, 449–464 Zbl 1475.20016 MR 4242157

[51] A. Lubotzky, Finite simple groups of Lie type as expanders. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 5, 1331–1341 Zbl 1257.20016 MR 2825166

[52] T. Łuczak and L. Pyber, On random generation of the symmetric group. *Combin. Probab. Comput.* **2** (1993), no. 4, 505–512 Zbl 0817.20002 MR 1264722

[53] G. Malle, J. Saxl, and T. Weigel, Generation of classical groups. *Geom. Dedicata* **49** (1994), no. 1, 85–116 Zbl 0832.20029 MR 1261575

[54] A. Maróti and L. Pyber, A generalization of the diameter bound of Liebeck and Shalev for finite simple groups. *Acta Math. Hungar.* **164** (2021), no. 2, 350–359 Zbl 07377593 MR 4279340

[55] C. Martinez and E. Zelmanov, Products of powers in finite simple groups. *Israel J. Math.* **96** (1996), no. part B, 469–479 Zbl 0890.20013 MR 1433702

[56] M. B. Nathanson, *Additive Number Theory. The Classical Bases*. Grad. Texts in Math. 164, Springer, New York, 1996 Zbl 0859.11002 MR 1395371

[57] N. Nikolov and L. Pyber, Product decompositions of quasirandom groups and a Jordan type theorem. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 4, 1063–1077 Zbl 1228.20020 MR 2800484

[58] L. Pyber and E. Szabó, Growth in finite simple groups of Lie type. *J. Amer. Math. Soc.* **29** (2016), no. 1, 95–146 Zbl 1371.20010 MR 3402696

[59] D. M. Rodgers and J. Saxl, Products of conjugacy classes in the special linear groups. *Comm. Algebra* **31** (2003), no. 9, 4623–4638 Zbl 1032.20031 MR 1995555

[60] J. Saxl and J. S. Wilson, A note on powers in simple groups. *Math. Proc. Cambridge Philos. Soc.* **122** (1997), no. 1, 91–94 Zbl 0890.20014 MR 1443588

[61] D. Segal, *Words: Notes on Verbal Width in Groups*. London Math. Soc. Lecture Note Ser. 361, Cambridge University Press, Cambridge, 2009 Zbl 1198.20001 MR 2547644

[62] M. Sellke, Covering Irrep(S_n) with tensor products and powers. 2020, arXiv: 2004.05283v3

[63] M. Sellke, Tensor quasi-random groups. *Proc. Amer. Math. Soc. Ser. B* **9** (2022), 12–21 MR 4377265

[64] A. Shalev, A theorem on random matrices and some applications. *J. Algebra* **199** (1998), no. 1, 124–141 Zbl 0910.20031 MR 1489358

[65] A. Shalev, Mixing and generation in simple groups. *J. Algebra* **319** (2008), no. 7, 3075–3086 Zbl 1146.20057 MR 2397424

[66] A. Shalev, Word maps, conjugacy classes, and a noncommutative Waring-type theorem. *Ann. of Math. (2)* **170** (2009), no. 3, 1383–1416 Zbl 1203.20013 MR 2600876

[67] J. Taylor and P. H. Tiep, Lusztig induction, unipotent supports, and character bounds. *Trans. Amer. Math. Soc.* **373** (2020), no. 12, 8637–8676 Zbl 07301836 MR 4177271

[68] Y. Varshavsky, Lefschetz–Verdier trace formula and a generalization of a theorem of Fujiwara. *Geom. Funct. Anal.* **17** (2007), no. 1, 271–319 Zbl 1131.14019 MR 2306659

Aner Shalev
Einstein Institute of Mathematics, Hebrew University, Jerusalem 9190401, Israel;
aner.shalev@mail.huji.ac.il

DOI 10.4171/8ECM/22

HMS symmetries and hypergeometric systems

Špela Špenko

Abstract. The derived category of an algebraic variety might be a source of a myriad of new (categorical) symmetries. Some are predicted by homological mirror symmetry, to be obtained from the fundamental group of the space of complex structures of its mirror partner. These finally lead to differential equations. We expositorily unravel a part of this conjectural master plan for a class of toric varieties.

nasvidenje, Marjan, nekoč … nekje …

1. Overview

Hilbert's 21st problem asks about the existence of Fuchsian linear differential equations on the Riemann sphere with prescribed singular points and monodromy representation of the fundamental group of the complement of the singular points [25]. The first (slightly erroneous) solution was proposed by the Slovenian mathematician Plemelj [38]. A suitably adapted version of this problem was solved and generalised, depending on the context, by Deligne [16], Kashiwara [33], Mebkhout [37], Beĭlinson–Bernstein [5], and others. The solution is now known as the Riemann–Hilbert correspondence.

Homological mirror symmetry (HMS) predicts the existence of an action of the fundamental group of the "stringy Kähler moduli space (SKMS)" on the derived category of an algebraic variety. The prediction was established by Halpern-Leistner and Sam for certain toric varieties [24]. A decategorification of this action yields a representation of the fundamental group of the SKMS, and our joint work with Michel Van den Bergh shows that it corresponds under the Riemann–Hilbert correspondence to a hypergeometric system of differential equations [42].

In this expository note, we aim to explain the above terms and finally present the mentioned results.

2020 Mathematics Subject Classification. Primary 13A50; Secondary 53D37, 32S45, 16S38, 14F06.
Keywords. Riemann–Hilbert correspondence, homological mirror symmetry, geometric invariant theory.

2. Hilbert's 21st problem

We begin with a classical problem, namely Hilbert's 21st problem. It is a part of the list of 23 problems [25, 26], published by Hilbert in 1900, which has been influential for the future mathematical development. The 21st one had the following formulation:

> *To show that there always exists a linear differential equation of the Fuchsian class, with given singular points and monodromic group.*

We shall first decipher the problem a little bit.

2.1. Fuchsian type

A system of linear differential equations

$$\begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = A(z) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \tag{2.1}$$

is of *Fuchsian* type if $A(z)$ is holomorphic on $\overline{\mathbb{C}} \setminus \{a_1, \dots, a_N\}$ with a pole of order 1 at a_j, $1 \le j \le N$, where we denote $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

In particular,[1] $\sum_{i=0}^{n} q_i(z) y^{(n-i)} = 0$, $q_n(z) = 1$, is Fuchsian if and only if the familiar Fuchsian condition is satisfied, i.e., $q_i(z)(z-a)^i$ is holomorphic at $z = a$ for $a \in \mathbb{C}$ and $q_i(z)z^i$ is holomorphic at $z = \infty$, for $0 \le i \le n$.[2]

2.2. Monodromy

Assume that we have a system of linear differential equations (2.1) with singularities at finitely many points $\{a_1, \dots, a_N\}$. Let γ be a closed path (so $\gamma(0) = \gamma(1)$) in $\overline{\mathbb{C}} \setminus \{a_1, \dots, a_N\}$.

Let $y_1, \dots, y_n$ be a basis of solutions of the system on an open set around $\gamma(0)$ (they exist by the local existence theorem for differential equations). These solutions are guaranteed to exist a priori only locally. However, we can analytically continue them along γ. Let us denote by $\tilde{y}_1, \dots, \tilde{y}_n$ analytic continuations of $y_1, \dots, y_n$ along γ.

Because both $y_1, \dots, y_n$ and $\tilde{y}_1, \dots, \tilde{y}_n$ form a basis of solutions around $\gamma(0)$, they should be related via an invertible linear map. We denote it by ρ_γ. It turns out that ρ_γ only depends on the homotopy class of γ. Therefore, we obtain a group homo-

[1] We may take $y_i = y^{(i)}$, where $0 \le i \le n$ and for A an $((n+1) \times (n+1)$-)matrix with nonzero entries only on the first upper diagonal where they are equal to 1 and in the last row.

[2] This follows by taking the $n \times n$-matrix with $a_{i,i+1} = -1$, $a_{n,i} = q_{n-i+1}$, and $a_{ij} = 0$ otherwise.

morphism

$$\pi_1\big(\overline{\mathbb{C}} \setminus \{a_1, \dots, a_N\}\big) \to \mathrm{GL}_n(\mathbb{C}), \quad [\gamma] \mapsto \rho_\gamma.$$

This is what is called a *monodromy* representation.

To close this discussion, we look at a concrete example of a differential equation.

Example 2.1. We take the differential equation $zy' - \alpha y = 0$. First, note that it has singularities at 0 and at ∞. (It is of Fuchsian type.) We take a loop γ around 0. A local solution is equal to $y = z^\alpha$ and its analytic continuation along γ equals $\tilde{y} = e^{2\pi i\alpha} z^\alpha$. To construct a monodromy representation, we first notice that the fundamental group of $\overline{\mathbb{C}} \setminus \{0, \infty\} = \mathbb{C} \setminus \{0\}$ is isomorphic to $\mathbb{Z}$, and we can identify the generator 1 with the homotopy class of γ. The monodromy representation is then given by

$$\rho : \pi_1\big(\overline{\mathbb{C}} \setminus \{0, 1\}\big) \cong \mathbb{Z} \to \mathrm{GL}_1(\mathbb{C}) \cong \mathbb{C}^*, \quad k \mapsto e^{2\pi i k\alpha}.$$

2.3. Formulation

Let us now restate the problem. As input we have

- a finite set of points $\{a_1, \dots, a_N\}$, and
- a representation ρ of $\pi_1(\overline{\mathbb{C}} \setminus \{a_1, \dots, a_N\})$.

Then Hilbert's 21st problem reads as follows: *Does there exist a system of linear differential equations of Fuchsian type with singular points $\{a_1, \dots, a_N\}$ and the monodromy representation equal to ρ?*

2.4. Progress

Already in 1908, Plemelj proposed a complete solution [38]. Unfortunately, it turned out that Plemelj's solution was not entirely correct. (Nevertheless, Plemelj's proof shows that one can find a system of linear differential equations which is Fuchsian at all but one point, where it is regular, see Section 3.3.) In 1988, Bolibrukh found a counterexample for $N = 4$ and a ρ of degree 3 [8].

The problem then transformed into classifying the input data that correspond to systems of differential equations of Fuchsian type.

Among algebraic geometers, the focus was however directed towards higher dimensions with a suitably rendered condition. Instead of Fuchsian type, one requires regularity, a weaker condition.

3. Riemann–Hilbert correspondence

There are plentiful variants of the Riemann–Hilbert correspondence. We first present one in line with the previous discussion, and then its powerful generalisation to the context of D-modules. We mostly follow [28]. We also mention [34] for a very nice review of Deligne's work on Hilbert's 21st problem.

3.1. Integrable connections

We first need to make sense of differential equations on general manifolds where we have no global coordinates at our disposal.

Let X be a complex manifold. Let $\mathcal{T}_X$ be the tangent sheaf on X (i.e., the sheaf of vector fields).[3]

Definition 3.1. An integrable connection on X is a pair (M, ∇), where M is a finite dimensional vector bundle on X and a linear map $\nabla : \mathcal{T}_X \otimes \widetilde{M} \to \widetilde{M}$, where $\widetilde{M}$ is the sheaf of sections of M such that[4]

- $\nabla_{f\theta}(m) = f\nabla_\theta(m)$ for $f \in \mathcal{O}_X, \theta \in \mathcal{T}_X, m \in \widetilde{M}$,
- $\nabla_\theta(fm) = \theta(f)m + f\nabla_\theta(m)$ for $f \in \mathcal{O}_X, \theta \in \mathcal{T}_X, m \in \widetilde{M}$,
- $\nabla_{[\theta_1,\theta_2]}(m) = [\nabla_{\theta_1}, \nabla_{\theta_2}](m)$ for $\theta_1, \theta_2 \in \mathcal{T}_X, m \in \widetilde{M}$.

With the natural definition of morphisms, we obtain an abelian category of connections on X which we denote by $\operatorname{Conn}(X)$.

Remark 3.2. For a system of differential equations (2.1) on $X = \mathbb{C} \setminus \{0\}$ (i.e., 0 is the only singularity different from ∞), M is the trivial vector bundle of rank n, and ∇ is given by $\nabla_{\partial/\partial z}(y) = y' - A(z)y$ for $y \in \widetilde{M} = (\mathcal{O}_X)^n$.

Conversely, if (M, ∇) is an integrable connection on $X = \mathbb{C} \setminus \{0\}$, then M is a trivial vector bundle, say of rank n. We choose an $\mathcal{O}_X$-basis $(e_i)_i$ of $\widetilde{M} = \mathcal{O}_X^n$. Define $a_{ij}(z)$, $1 \leq i, j \leq n$, by $\nabla_{\partial/\partial z}(e_j) = -\sum_{i=1}^n a_{ij}(z)e_i$. Then $\nabla_{\partial/\partial z}(y) = \nabla_{\partial/\partial z}(\sum_i y_i e_i) = \sum_i y_i' e_i + \sum_i y_i \nabla_{\partial/\partial z}(e_i) = y' - A(z)y$ for $y \in \widetilde{M}$.

The solutions of an integrable connection are defined as $\{m \in \widetilde{M} \mid \nabla_\theta(m) = 0 \text{ for all } \theta \in \mathcal{T}_X\}$ and are called *horizontal sections*.

3.2. Meromorphic connections

We now extend the concept of integrable connections to allow poles as well. Let $D \subset X$ a divisor. Let $\mathcal{O}_X[D]$ be a sheaf of meromorphic functions on X, holomorphic on $X \setminus D$ with poles along D.

Definition 3.3. A coherent $\mathcal{O}_X[D]$-module M[5] is a *meromorphic connection* if there exists a map $\nabla : M \to \Omega_X^1 \otimes_{\mathcal{O}_X} M$ such that

- $\nabla(fs) = df \otimes s + f\nabla s$,
- $[\nabla_\theta, \nabla_{\theta'}] = \nabla_{[\theta,\theta']}$ for $\theta, \theta' \in \mathcal{T}_X$ (where $\nabla_\theta : M \to M$ is ∇'_θ for $\nabla' : \mathcal{T}_X \otimes M \to M$ obtained from ∇).

[3] Note that $\mathcal{T}_X$ may also be identified with derivations in $\mathcal{E}nd_{\mathbb{C}_X}(\mathcal{O}_X)$.

[4] We use standard notation $\nabla_\theta(m) := \nabla(\theta \otimes m)$.

[5] We note that the definition implies that the restriction $M_{X\setminus D}$ of a meromorphic connection M to $X \setminus D$ is a locally free $\mathcal{O}_{X\setminus D}$-module.

With the natural definition of morphisms between meromorphic connections, we obtain an abelian category $\mathrm{Conn}(X; D)$ of meromorphic connections.

Remark 3.4. This remark is an analogue of Remark 3.2. We obtain a natural one-to-one correspondence between linear differential equations on $\mathbb{C}$ with possible poles at 0 and meromorphic connections in $\mathrm{Conn}(X; D)$.

3.3. Regular singularities

Here we define regular singularities of differential equations, which are a generalisation of the differential equations of Fuchsian type.

Definition 3.5. In complex dimension 1, a system of differential equations has *regular singularities* if every solution y on a punctured angular sector around a singular point in $\{a_1, \dots, a_N\}$ has moderate growth, i.e.,

- a_j finite: $|y(z)| = O(|z - a_j|^{-m})$ for some $m \geq 0$ as $z \to a_j$,
- $a_j = \infty$: $|y(z)| = O(|z|^m)$ for some $m \geq 0$ as $z \to \infty$.

This also has an algebraic interpretation which can be moreover generalised to higher dimensions and all manifolds.

Definition 3.6. A meromorphic connection (M, ∇) in $\mathrm{Conn}(X; D)$ is *regular* if $(i^* M)_0$ is regular for every $i : B \to X$ such that $i^{-1} D = \{0\}$.

We also mention that with the natural definition of morphisms between regular meromorphic connections on (X, D) we obtain an abelian category $\mathrm{Conn}^{\mathrm{reg}}(X; D)$.

3.4. Deligne's Riemann–Hilbert correspondence

Theorem 3.7 ([16]). *Let X be a complex manifold and let D be a divisor in X. Then the restriction functor induces an equivalence* $\mathrm{Conn}^{\mathrm{reg}}(X; D) \xrightarrow{\sim} \mathrm{Conn}(X \setminus D)$.

Deligne's theorem constitutes the essential part of the correspondence between systems of differential equations on X with regular singularities along D and representations of the fundamental group of $X \setminus D$.

Corollary 3.8. *There is an equivalence of categories between* $\mathrm{Conn}^{\mathrm{reg}}(X; D)$ *and* $\mathrm{rep}(\pi_1(X \setminus D))$.

This equivalence factors as

$$\begin{tikzcd}
\mathrm{Conn}^{\mathrm{reg}}(X; D) \arrow[r, "\sim"] \arrow[d, "\wr"'] & \mathrm{rep}\left(\pi_1(X \setminus D)\right) \\
\mathrm{Conn}(X \setminus D) \arrow[r, "\sim"] & \mathrm{Loc}(X \setminus D), \arrow[u, "\wr"']
\end{tikzcd} \tag{3.1}$$

where $\mathrm{Loc}(X \setminus D)$ is the category of local systems, i.e., locally constant sheaves of finite dimensional $\mathbb{C}$-vector spaces. The first (vertical) equivalence is the restriction from Theorem 3.7, the second is obtained by taking the horizontal sections ("solutions of the system"), and the last (vertical) arrow is a well-known equivalence (see e.g. [1]) which sends a local system L to the representation of $\pi_1(X \setminus D)$ on L_{x_0} that associates to every path an isomorphism of L_{x_0} along itself (which exists as L is locally constant).

The statement holds also in the context of smooth algebraic varieties which was Deligne's original motivation.

In short, we could say that topology, here measured by the fundamental group, is somewhat determined by analysis or algebra, here represented by differential equations with regular singularities.

3.5. D-modules

We continue towards a generalisation of Deligne's correspondence to other systems of linear differential equations.

For this we move on the left-hand side of the above diagram a bit more towards the algebra direction, and replace the differential equations with modules over the ring of differential operators. We enter the framework of so-called D-modules. We follow [28, Introduction].

Let X be an open submanifold in $\mathbb{C}^n$ and let $\mathcal{O}(X)$ be holomorphic functions globally defined on X. With D we denote the set of partial differential operators with coefficients in $\mathcal{O}(X)$. Namely,

$$D = \left\{ \sum_{i_1,\dots,i_n} f_{i_1\cdots i_n} \left(\frac{\partial}{\partial x_1}\right)^{i_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{i_n} \mid f_{i_1\cdots i_n} \in \mathcal{O}(X) \right\},$$

where x_i are coordinate functions on $\mathbb{C}^n$. Note that D also has a ring structure. For example, D contains the n-th Weyl algebra for $X = \mathbb{C}^n$ (we take only polynomial coefficients).

Now take P in D. Then P corresponds to a differential equation.[6] We can represent the holomorphic (global) solutions as follows:

$$\{u \in \mathcal{O}(X) \mid Pu = 0\} \cong \mathrm{Hom}_D\left(D/DP, \mathcal{O}(X)\right), \quad u \mapsto (d \mapsto du).$$

We can proceed similarly if we have a collection of $P_{ij} \in D$, $1 \le i \le k$, $1 \le j \le l$, corresponding to a system of differential equations. Then the solution (u_j) of the system given by the matrix (P_{ij}) can be identified with

$$\{(u_j) \mid (P_{ij})(u_j) = 0\} \cong \mathrm{Hom}_D\left(M, \mathcal{O}(X)\right),$$

[6] For example, $x\frac{\partial}{\partial x} - \alpha$ corresponds to the equation $xy' - \alpha y = 0$.

where M is defined by the short exact sequence

$$D^k \xrightarrow{(P_{ij})} D^l \to M \to 0.$$

In sum, we have found a way to turn systems of differential equations into finitely presented D-modules, and have described their (global) solutions purely algebraically using homomorphisms.

However, solutions may not exist globally, so therefore we should use a tool that takes into account also local solutions. From modules, we should pass to sheaves, as we have already done in the beginning of this section. Now $\mathcal{O}$ denotes the sheaf of holomorphic functions. Similarly, we replace D by $\mathcal{D}$ ($\mathcal{D}(U)$ consists of partial differential operators with coefficients in $\mathcal{O}(U)$). Then we can look at the sheaf $\mathcal{H}om_{\mathcal{D}}(\mathcal{M}, \mathcal{O})$ ($U \mapsto \mathrm{Hom}_{\mathcal{D}(U)}(\mathcal{M}(U), \mathcal{O}(U))$).

There is another caveat to consider. We may be interested in relating different systems of differential equations; i.e., from solutions of two systems deduce something about solutions of the system that is formed as the union of the two systems. The problem that we encounter here is that the functor $\mathcal{H}om_{\mathcal{D}}(-, \mathcal{O})$ is not exact. So we should also consider "higher solutions", namely the extension modules $\mathcal{E}xt^i_{\mathcal{D}}(\mathcal{M}, \mathcal{O})$.

It will turn out that higher solutions give us almost all the topological data that we need. Perhaps it is then a good point to ask what kind of sheaves these higher solutions are. We know they are sheaves of $\mathbb{C}$-vector spaces. Is there any other property that distinguishes them?

Recall from (3.1) (applied with $D = \emptyset$) that if $\mathcal{M}$ is associated to a connection, then we obtain a local system, i.e., a locally constant sheaf of finite dimensional $\mathbb{C}$-vector spaces. It turns out that this correspondence generalises if we restrict to holonomic modules[7], they are those that roughly speaking give finite dimensional (higher) solution spaces. With this assumption, all the higher solution sheaves $\mathcal{E}xt^i_{\mathcal{D}}(\mathcal{M}, \mathcal{O})$ are *constructible*, which means that they are built from local systems. More precisely, there exists a stratification of $X = \sqcup_\alpha X_\alpha$ into locally closed sets such that $F_i|_{X_\alpha}$ is a local system for all i.

This is a prelude to a correspondence between holonomic $\mathcal{D}$-modules on the algebraic side and constructible sheaves on the topological side. Note that on the topological side we obtain an entire sequence of constructible sheaves, and to compute those we should also know something about the projective resolution of the modules, again on the algebraic side. A convenient machinery to process all this data at once and without losing too much information is the derived category.

[7] A coherent $\mathcal{D}_X$-module M is holonomic if $\dim \mathrm{Ch}(M) = \dim X$. Here $\mathrm{Ch}(M)$ denotes the characteristic variety of M, i.e., the support of the associated graded module $\mathrm{gr}\, M$ (for a "good" filtration) on the cotangent bundle of X.

3.6. Derived categories

Let $\mathcal{A}$ be an abelian category, for example the category mod($\mathcal{D}_X$) of $\mathcal{D}_X$-modules on X, or the category mod($\mathbb{C}_X$) of sheaves of finite dimensional vector spaces on X, the categories that we have just seen.

Let $C(\mathcal{A})$ be the category of complexes on $\mathcal{A}$. We say that a map $f : X^\bullet \to Y^\bullet$ between two complexes is a quasi-isomorphism if it induces isomorphisms on cohomology, i.e., $H^i(f) : H^i(X^\bullet) \xrightarrow{\sim} H^i(Y^\bullet)$ for all i.

We want that the derived category does not distinguish between two complexes which are connected via a quasi-isomorphism. So we formally invert quasi-isomorphisms (see e.g. [43, 04VB] for localisation in categories) and define the derived category as

$$D(\mathcal{A}) = C(\mathcal{A})[\text{qis}^{-1}].$$

Furthermore, if a covariant, resp. contravariant, functor $F : \mathcal{A} \to \mathcal{B}$ between two abelian categories (with $\mathcal{A}$ having enough injectives, resp. projectives) is left-exact, then there exists a corresponding functor $RF : D^+(\mathcal{A}) \to D^+(\mathcal{B})$, resp. $RF : D^-(\mathcal{A}) \to D^+(\mathcal{B})$, between the derived categories (of bounded-below, resp. above/below, complexes).

Let us zoom this in on our example.

Example 3.9. We take for F the solution functor $F = \mathcal{H}om_{\mathcal{D}}(-, \mathcal{O})$. Then the derived functor $RF : D^-(\mathcal{D}_X)^o \to D^+(\mathbb{C}_X)$ is such that its cohomology sheaves are exactly the higher solutions; i.e., $H^i(RF) = \mathcal{E}xt^i(-, \mathcal{O}_X)$. So the derived solution functor carries the information about all higher solutions. (Note that here and later we for brevity omit writing mod.)

3.7. Riemann–Hilbert correspondence

We are ready to state the Riemann–Hilbert correspondence in its full power and complexity, to connect all the module data with the data of solutions and higher solutions.

We need to restrict to a subclass of complexes of $\mathcal{D}_X$-modules that have regular[8] and holonomic cohomology. Roughly these conditions guarantee that the solution spaces are finite dimensional and have moderate growth. We denote the derived category of bounded complexes of $\mathcal{D}_X$ modules with regular and holonomic cohomology by $D^b_{rh}(\mathcal{D}_X)$. On the topological side, we look at those bounded complexes of sheaves of $\mathbb{C}$-vector spaces on X that have constructible cohomology, and we denote the corresponding derived category by $D^b_c(\mathbb{C}_X)$.

Under these restrictions, the derived solution functor gives the celebrated anti-equivalence of categories.

[8] For the definition of regularity for $\mathcal{D}_X$-modules on a complex manifold X, we refer to [28, Definition 6.1.8].

Theorem 3.10 ([5, 32, 33, 37][9]). *There is an anti-equivalence of (triangulated) categories*

$$R\,\mathcal{H}om_{\mathcal{D}_X}(-,\mathcal{O}_X) : D^b_{rh}(\mathcal{D}_X)^o \xrightarrow{\sim} D^b_c(\mathbb{C}_X).$$

First we remark that we really need to pass to the derived level contrary to Deligne's Riemann–Hilbert correspondence. Indeed, as mentioned earlier, the solution functor is not exact so it cannot induce an equivalence of abelian categories. This theorem is from an algebraic point of view a real advancement, and a vast generalisation of Deligne's Riemann Hilbert correspondence, since we can, in particular, to every (regular holonomic) $\mathcal{D}_X$-module associate a topological object, a complex of sheaves of $\mathbb{C}$-vector spaces on X (with constructible cohomology), and vice versa.

These associated complexes are also rather special, they form an abelian category, and they are called *perverse sheaves*, i.e.,

$$\mathrm{Perv}(X) := R\,\mathcal{H}om_{\mathcal{D}_X}(-,\mathcal{O}_X)(\mathrm{mod}_{rh}\,\mathcal{D}^o_X)[\dim X].$$

4. Homological mirror symmetry symmetries

We divert the story to mirror symmetry. There we will encounter representations of some fundamental groups and our aim will be to realise them as monodromy representations of differential equations.

4.1. Mirror symmetry

Let us first very briefly say a few words on mirror symmetry, a theory that has its origins in physics, more precisely in string theory. Typically, the spaces that appear in this context have both a complex and a symplectic structure. Moreover, the spaces come in mirror pairs X and X^o, with the complex and symplectic structures interlaced. The complex geometry of X mirrors the symplectic geometry of its mirror X^o, and vice versa. The picture is still highly speculative. We refer to [15, Introduction] for a survey of its origins and multiple predictions that mirror symmetry provides to algebraic geometry.

4.2. HMS categorical symmetries

Mirror symmetry has been enhanced to a homological statement about the equivalence of certain categories (the derived category and the Fukaya category) that reflect complex and symplectic geometry, respectively. The correspondence has been conjectured by Kontsevich [36] and nowadays goes under the name of *homological mirror symmetry*.

[9] Beĭlinson and Bernstein proved the theorem in the algebraic setting.

We discuss here one of the consequences of HMS. For a more precise explanation of heuristics, see [24, §1.1]. Assume that we regard X as a complex manifold. Then the symplectic structure of the mirror X^o is fixed, but there is still room for different complex structures. Denote by $\mathcal{K}_X$ the space of complex structures of X^0.[10]

Then HMS predicts the following.

Conjecture 4.1. *There exists an action*[11]

$$\pi_1(\mathcal{K}_X) \curvearrowright D^b(X).$$

As an immediate corollary of this, we would get the following result about the Grothendieck group of X.

Corollary 4.2. *There exists an action*

$$\pi_1(\mathcal{K}_X) \curvearrowright K_0(X)_{\mathbb{C}}.$$

It is this action about which we will wonder which system of differential equations it corresponds to.

4.3. Example

We look at the conifold, $Y = \operatorname{Spec}(\mathbb{C}[x, y, z, u]/(xu - yz))$.[12] We define $X = \operatorname{Bl}_{(x,y)} Y$, a small resolution of Y. (In the framework of toric geometry, we might represent Y as a cone in $\mathbb{R}^3$ over the unit square in $\mathbb{R}^2 \times \{1\}$. To obtain X we should add a diagonal hyperplane.)

There is another viewpoint that will be more useful for us. Let $\mathbb{C}^*$ act on $\mathbb{C}^4$ as $t \cdot (v_1, v_2, v_3, v_4) = (t^{-1}v_1, t^{-1}v_2, tv_3, tv_4)$. Then we may view Y as the (categorical) quotient $\mathbb{C}^4 /\!\!/ \mathbb{C}^*$ $(= \operatorname{Spec}\mathbb{C}[x_1, x_2, x_3, x_4]^{\mathbb{C}^*}, x_i = v_i^*)$.[13] We obtain X as the geometric invariant theory (GIT) quotient $(\mathbb{C}^4 \setminus V(x_1, x_2)) /\!\!/ \mathbb{C}^*$.[14]

Heuristics from physics [4] yield that $\mathcal{K}_X = \mathbb{P}^1 \setminus \{0, 1, \infty\}$.

[10] $\mathcal{K}_X$ is also called the "stringy Kähler moduli space" (SKMS) of X (i.e., the space of Kähler structures on X coming from symplectic geometry of X). The tangent space to the SKMS is $H^2(X, \mathbb{C})$ (the space of complexified symplectic forms). However, there is no global definition; $\mathcal{K}_X$ has only been explicitly defined in very few examples, the difficulty being the determination of the mirror pair.

[11] We might think of $D^b(X)$ as bounded complexes of vector bundles on X.

[12] One can describe the conifold also as a cone over $\mathbb{P}^1 \times \mathbb{P}^1$.

[13] The homomorphism $\mathbb{C}[x, y, z, u]/(xu - yz) \to \mathbb{C}[x_1, x_2, x_3, x_4]^{\mathbb{C}^*}$, $x \mapsto x_1x_3$, $y \mapsto x_1x_4$, $z \mapsto x_2x_3$, $u \mapsto x_2x_4$, is an isomorphism.

[14] Let us assume that $t \cdot v = tv$, and take $s = v^*$, assume that $\deg s = 1$, $\deg x_i = 0$, $1 \le i \le 4$. Then the GIT quotient $(\mathbb{C}^4 \setminus V(x_1, x_2)) /\!\!/ \mathbb{C}^*$ is defined as $\operatorname{Proj}(\mathbb{C}[x_1, x_2, x_3, x_4, s]^{\mathbb{C}^*})$.

To construct a representation of $\pi_1(\mathcal{K}_X)$ on $D^b(X)$, we first view $D^b(X)$ as the (full thick) subcategory of $D^b([\mathbb{C}^4/\mathbb{C}^*])$[15], generated by $\mathcal{O}_{\mathbb{C}^4}$, $\mathcal{O}_{\mathbb{C}^4} \otimes V(1)$, where $V(n)$ denotes the irreducible (1-dimensional) representation of $\mathbb{C}^*$ with character n, i.e., $t \cdot v = t^n v$ for $v \in V(n)$; see [44, Theorem 8.6].

Then it turns out that in the basis $\{\mathcal{O}_{\mathbb{C}^4} \otimes V(1), \mathcal{O}_{\mathbb{C}^4}\}$ the action of the three generating loops $\gamma_0, \gamma_1, \gamma_\infty \in \pi_1(\mathcal{K}_X)$ is given by

$$\gamma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \gamma_0 = \begin{pmatrix} 2 & 1 \\ -1 & 0 \end{pmatrix}, \quad \gamma_\infty = \begin{pmatrix} 0 & -1 \\ 1 & 2 \end{pmatrix}.$$

See e.g. [17, 24, 41].

5. HMS symmetries: toric varieties

We will approach the conjecture in the setting of toric varieties.

5.1. Setting

We assume that $W = \mathbb{C}^d$ is a $T := (\mathbb{C}^*)^n$-representation which is unimodular (i.e., the sum of weights is equal to 0).

We describe how to obtain an analogue of the variety X in the case of the conifold; cf. Section 4.3. We should remove some undesirable locus of W and then take the GIT quotient. The variety (or stack) X that we obtain in this way is a (crepant) resolution of singularities of $W /\!\!/ T$ $(= \operatorname{Spec} \mathbb{C}[W]^T)$.

Let $X(T)$ be the character group of T and $Y(T)$ the group of 1-parameter subgroups of T. We take a generic $\chi \in X(T)_{\mathbb{R}}$. Let $W^{\chi,u}$ be the χ-unstable locus, i.e., the set of points $w \in W$ such that if $\lim_{t\to 0} \lambda(t)w$ for $\lambda \in Y(T)$ exists, then $\chi(\lambda) \geq 0$. Then we take

$$X = \big[(W \setminus W^{\chi,u})/T\big].$$

This is a priori a Deligne–Mumford quotient stack, a quotient stack whose points have finite stabilizers. In the case that all stabilizers are trivial, the corresponding GIT quotient variety can replace the stack (i.e., in this case the quotient stack and the quotient variety are isomorphic). The GIT quotient is defined in the analogy with Footnote 14.[16]

[15] Here $[\mathbb{C}^4/\mathbb{C}^*]$ denotes the quotient stack. The category $\operatorname{mod}([\mathbb{C}^4/\mathbb{C}^*])$ consists of $\mathbb{C}^*$-equivariant $\mathbb{C}[x_1, x_2, x_3, x_4]$-modules and the category $\operatorname{coh}([\mathbb{C}^4/\mathbb{C}^*])$ of $\mathbb{C}^*$-equivariant coherent sheaves on $\mathbb{C}^4$. It follows that $D^b([\mathbb{C}^4/\mathbb{C}^*]) = D^b(\operatorname{mod}([\mathbb{C}^4/\mathbb{C}^*])) = D^b(\operatorname{coh}([\mathbb{C}^4/\mathbb{C}^*]))$.

[16] We assume that $V = \mathbb{C}v$ is the 1-dimensional T-representation with character χ; i.e., $t \cdot v = \chi(t)v$. Let w_i be a basis of W such that $t \cdot w_i = \beta_i(t) w_i$ for $\beta_i \in X(T)$. Set $x_i = w_i^*$, $1 \leq i \leq d$, $d = v^*$. We assume that $\deg x_i = 0$ and $\deg s = 1$. Then $(W \setminus W^{\chi,u}) /\!\!/ T := \operatorname{Proj}(\mathbb{C}[x_1, \dots, x_d, s]^T)$.

Remark 5.1. The varieties above are exactly affine normal Gorenstein toric varieties whose class group is a torus (i.e., it has no finite group part).

5.2. Space of complex structures on X^o

In the case of toric varieties physics heuristics are rather reliable. In [18, §4.1] there is an explicit recipe for $\mathcal{K}_X$ that refers for evidence to [13].[17]

Set $d = \dim W$. Let $(\beta_i)_{i=1}^d$ be T-characters of W. Note that $X(T) \cong \mathbb{Z}^n$ and set $B = (\beta_i)_{i=1}^d \in M_{n\times d}(\mathbb{Z})$. We define A (up to an automorphism of $\mathbb{Z}^{d-n}$) by the exact sequence

$$0 \to \mathbb{Z}^{d-n} \xrightarrow{A} \mathbb{Z}^d \xrightarrow{B} \mathbb{Z}^n \to 0. \tag{5.1}$$

Then $\mathcal{K}_X$ is the complement of a hypersurface $V(E_A) \subset T$, where E_A is the *principal A-determinant*. We refer to [20, §10.1.A] for the definition.[18] Alternatively, see [18, 35].

In a sufficiently symmetric case, $V(E_A)$ is much simpler.

Theorem 5.2 ([35]). *If W is quasi-symmetric*[19]*, then $\mathcal{K}_X$ is the complement of a hyperplane arrangement (in logarithmic coordinates) in $T = (\mathbb{C}^*)^n$.*

The hyperplane arrangement in $1/(2\pi i)\log T = X(T)_{\mathbb{C}}$ can be explicitly described. Let Δ be the Minkowski sum of $[0, (1/2)\beta_i]$. Let $(H_i)_i$ be the supporting (affine) hyperplanes of Δ. Then the hyperplane arrangement is the complexification of the real hyperplane arrangement $\bigcup_i (-H_i) + X(T)$ (up to a suitable translation). This is an infinite, but locally finite, hyperplane arrangement.

This hyperplane arrangement was prior to the result of Kite heuristically predicted to coincide with $\mathcal{K}_X$ in [24].

Example 5.3. We make a quick sanity check in the case of the conifold; cf. Section 4.3. Then $\mathcal{K}_X = \mathbb{P}^1 \setminus \{0, 1, \infty\}$. Applying $1/(2\pi i)\log$ to $\mathbb{P}^1 \setminus \{0, 1, \infty\} = \mathbb{C} \setminus \{0, 1\}$, we obtain $\mathbb{C} \setminus \mathbb{Z}$. On the other hand, by the above recipe, $\Delta = [-1, 1]$ (as $(\beta_i)_{i=1}^4 = (-1, -1, 1, 1)$) and the hyperplane arrangement is given by $\mathbb{Z}$. Thus, the two descriptions are consistent.

5.3. HMS symmetries: quasi-symmetric case

Assume that $\mathbb{C}^d$ is a quasi-symmetric representation of $(\mathbb{C}^*)^n$. In this case, Halpern-Leistner and Sam [24] confirmed Conjecture 4.1.

[17] The heuristics are derived from the speculations that a mirror is given by a family of Landau–Ginzburg models [27]. See also [12, 29].

[18] In loc. cit. E_A stands for A', where $A = (A', 1)$ which we may assume since $\sum_i \beta_i = 0$.

[19] W is quasi-symmetric if for all lines $0 \in \ell \in X(T)_{\mathbb{R}}$, $\sum_{\beta_i \in \ell} \beta_i = 0$.

Theorem 5.4 ([24]). *There exists an action of $\pi_1(\mathcal{K}_X)$ on $D^b(X)$.*

As in Section 4.3, $D^b(X)$ is identified with the (full thick) subcategory D of $D^b([W/T])$ generated by $\{\mathcal{O}_W \otimes V(\mu) \mid \mu \in (\nu + \Delta) \cap X(T)\}$, where $V(\mu)$ is the irreducible T-representation with character μ, and $\nu \in X(T)_{\mathbb{R}}$ is generic [23, 40][20].

Then this action can be explicitly described, especially relying on the concrete description of the fundamental group of the complement of a complexified hyperplane arrangement [39]. See Section 7.1.2.

Remark 5.5. The statement can be generalised to some reductive groups, i.e., those groups G for which $X(G) \neq 0$, if some genericity assumptions are satisfied.[21] See [24].

6. HMS differential equations: quasi-symmetric case

In this section, we assume that we are in the setting of Section 5.1. Moreover, we assume that W is quasi-symmetric. Having Theorem 5.4, providing evidence for Conjecture 4.1, at our disposal, we also obtain Corollary 4.2. Hence, $\pi_1(\mathcal{K}_X)$ acts on $K_0(X)_{\mathbb{C}}$. We want to determine which (regular) system of differential equations on $(\mathbb{C}^*)^n$ this action corresponds to.

6.1. Example

We first want to understand the monodromy representation in the case of the conifold; cf. Section 4.3.

We look at the *Gauss hypergeometric equation*

$$z(1-z)y'' + \big(c - (a+b+1)z\big)y' - aby = 0.$$

The monodromy is given by, see e.g. [7],

$$\gamma_1 = \begin{pmatrix} 1 & -e^{2\pi \mathrm{i}(c-b)} - e^{2\pi \mathrm{i}(c-a)} + e^{2\pi \mathrm{i} c} + 1 \\ 0 & e^{2\pi \mathrm{i}(c-a-b)} \end{pmatrix},$$

$$\gamma_0 = \begin{pmatrix} 1 + e^{-2\pi \mathrm{i} c} & 1 \\ -e^{-2\pi \mathrm{i} c} & 0 \end{pmatrix},$$

$$\gamma_\infty = \begin{pmatrix} 0 & -e^{2\pi \mathrm{i}(a+b)} \\ 1 & e^{2\pi \mathrm{i} a} + e^{2\pi \mathrm{i} b} \end{pmatrix}.$$

[20] ν is not parallel to any face of Δ.

[21] The condition $\sum_i \mathbb{R}\beta_i = X(T)$ should be satisfied and there should exist $\chi \in X(G)$ which is not parallel to any face of Δ.

Setting $a = b = c = 0$, we obtain matrices that we have already encountered in Section 4.3. From this, one may deduce that the action of $\pi_1(\mathcal{K}_X)$ on $K_0(X)_{\mathbb{C}}$ from Theorem 5.4 in the case of the conifold corresponds to $z(1-1)y'' - zy' = 0$, i.e., the Gauss differential equations with parameters $a = b = c = 0$ (which is regular on $\mathbb{P}^1$ with singularities at $0, 1, \infty$).

6.2. Example with parameters

We change the focus a bit and ask whether we can find an action of $\pi_1(\mathcal{K}_X)$ on $K_0(X)_{\mathbb{C}}$ that would give the Gauss hypergeometric equation also for other parameters. We obtained the original action from an action of $\pi_1(\mathcal{K}_X)$ on $D^b(X)$. We would want to tweak this action a little bit to open the route to other parameters.

For this, first observe that $(\mathbb{C}^*)^4$ acts on $\mathbb{C}^4$ coordinate-wise. The initial $\mathbb{C}^*$ embeds in it via the map $t \mapsto (t^{-1}, t^{-1}, t, t)$ determined by the action of $\mathbb{C}^*$ on $\mathbb{C}^4$; cf. Section 4.3 and (5.1). This inclusion splits, and the complement is $(\mathbb{C}^*)^3$. We seem to be well on the way, the dimension of the complement torus coincides with the number of parameters in the Gauss hypergeometric equation.

Now a slightly more technical part follows. To get an action for other a, b, c, we need to replace $D^b(X)$ by a bigger category $\widetilde{D}$ such that $X((\mathbb{C}^*)^3)$ acts on it.

We define $\widetilde{D}$ as the (full thick) subcategory of $D^b([\mathbb{C}^4/(\mathbb{C}^*)^4])$ generated by

$$\mathcal{O}_{\mathbb{C}^4} \otimes V(\mu), \quad \mu \in X\big((\mathbb{C}^*)^4\big)$$

such that $B\mu \in \{0, 1\}$ (see (5.1) for B).

It turns out that $\pi_1(\mathcal{K}_X)$ still acts on $\widetilde{D}$. However, $K_0(\widetilde{D})_{\mathbb{C}}$ is a (free rank 2) module over $\mathbb{C}\{X((\mathbb{C}^*)^3)\} \cong \mathbb{C}[(\mathbb{C}^*)^3]$[22]. Specialising at (sufficiently generic[23]) $h \in (\mathbb{C}^*)^3$, we obtain an action of $\pi_1(\mathcal{K}_X)$ on a 2-dimensional $\mathbb{C}$-vector space. This action corresponds to the Gauss hypergeometric equation with parameters $-1/(2\pi i)\log h$.

6.3. GKZ hypergeometric systems

The GKZ hypergeometric systems are systems of differential equations that generalise the Gauss hypergeometric differential equation, as well as Appell, Lauricella, Horn, etc. They were introduced and studied by Gelfand, Kapranov, and Zelevinsky [19, 21, 22]. Allegedly, they were introduced as a unified approach to the multidimensional generalisations of the Gauss hypergeometric functions. In some sense, the construction of the GKZ hypergeometric system is dictated by the desired set of solutions, which should be hypergeometric power series. See Remark 6.3.

[22] Here $\mathbb{C}\{X((\mathbb{C}^*)^3)\}$ is the group algebra of $X((\mathbb{C}^*)^3)$, while $\mathbb{C}[(\mathbb{C}^*)^3]$ is the coordinate ring of $(\mathbb{C}^*)^3$.

[23] This is, in particular, satisfied if $a, b, a-c, b-c$ are all non-integers. However, one might check that $a = b = c = 0$ as in Section 6.1 also work.

Let $\alpha \in \mathbb{C}^{d-n}$. Recall the exact sequence (5.1). Let $B^* : \mathbb{Z}^n \to \mathbb{Z}^d$ be the dual of B. Then the hypergeometric GKZ system with parameter α is defined by the differential operators

- homogeneity relations: $\sum_{j=1}^{d} a_{ij} x_j \partial_j - \alpha_i$, $1 \le i \le d-n$,
- box relations: $\Box_l = \prod_{l_i>0} \partial_i^{l_i} - \prod_{l_i<0} \partial_i^{-l_i}$, $l \in B^*\mathbb{Z}^n$.

Note that this is a system of differential equations on $(\mathbb{C}^*)^d$. However, the homogeneity relations allow to descend these differential equations to $(\mathbb{C}^*)^n$.[24]

This descent also allows us to recover the Gauss hypergeometric equation from the GKZ hypergeometric system corresponding to the conifold, i.e., for the example $B = (-1,-1,1,1)$.

Example 6.1. In the case of the conifold, we may take

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Then a solution Φ of the GKZ hypergeometric system satisfies

$$\begin{aligned} (-x_1\partial_1 + x_2\partial_2)\Phi &= \alpha_1\Phi, \\ (x_1\partial_1 + x_3\partial_3)\Phi &= \alpha_2\Phi, \\ (x_1\partial_1 + x_4\partial_4)\Phi &= \alpha_3\Phi, \\ \partial_1\partial_2 - \partial_3\partial_4 &= 0. \end{aligned}$$

Setting $\alpha = (c-1,-a,-b)$, a simple manipulation yields

$$\big(x_3^{-1}x_4^{-1}\big(x_1\partial_1^2 - (1+a+b)x_1\partial_1 - ab\big) - x_2^{-1}(x_1\partial_1^2 - c\partial_1)\big)\Phi = 0.$$

Then $F(x) := \Phi(x,1,1,1)$ is a solution of the Gauss hypergeometric equation. Moreover, by homogeneity relations, F determines Φ.

We denote the corresponding $\mathcal{D}_{(\mathbb{C}^*)^d}$-module, cf. Section 3.5, by $\mathcal{P}(\alpha)$, and its restriction to $(\mathbb{C}^*)^n$ by $P(\alpha)$. The next proposition reveals that they are well behaved, as required for the Riemann–Hilbert correspondence.

Proposition 6.2 ([2]). *The $\mathcal{D}_{(\mathbb{C}^*)^d}$-module $\mathcal{P}(\alpha)$ is holonomic with regular singularities. The same holds for the $\mathcal{D}_{(\mathbb{C}^*)^n}$-module $P(\alpha)$.*

[24] The corresponding D-module on $(\mathbb{C}^*)^d$ is weakly equivariant for the action of $(\mathbb{C}^*)^{d-n}$, hence it descends to $(\mathbb{C}^*)^n$; see e.g. [42, Corollary A.11].

Remark 6.3. To follow on the introduction to this subsection we record here that the multidimensional hypergeometric (formal) series[25]

$$\Phi_\gamma(x_1,\dots,x_d)=\sum_{l\in\mathbb{Z}^n}\prod_{i=1}^{d}\frac{x_i^{B^*l+\gamma_i}}{\Gamma(B^*l+\gamma_i+1)},$$

where $\gamma\in(\mathbb{C})^d$ is such that $A\gamma=\alpha$, is a (formal) solution of the GKZ hypergeometric system [22].[26] Moreover, also Euler integrals generalise [19] to give solutions to the GKZ hypergeometric system. Another handy class of solutions is given by Mellin–Barnes integrals [6] that we crucially employ in the proof of Theorem 6.4 below.

6.4. Decategorification of HMS symmetries

We want to determine the system of differential equations whose monodromy representation coincides with the representation of $\pi_1(\mathcal{K}_X)$ on $K_0(X)_\mathbb{C}$ obtained from Theorem 5.4. However, we cannot quite do that. Instead, we tweak the action a bit, as in Section 6.2.

Analogously to Section 6.2, we note that $(\mathbb{C}^*)^d$ acts on $\mathbb{C}^d$ coordinate-wise, and we have the inclusion $T=(\mathbb{C}^*)^n\hookrightarrow(\mathbb{C}^*)^d$ which splits. The complement is $(\mathbb{C}^*)^{d-n}$.

We replace $D^b(X)$ by a bigger category $\widetilde{D}$, the (full thick) subcategory of $D^b([\mathbb{C}^d/(\mathbb{C}^*)^d])$ generated by

$$\{\mathcal{O}_{\mathbb{C}^d}\otimes V(\mu)\mid B\mu\in(\nu+\Delta)\cap X(T)\};$$

cf. the paragraph below Theorem 5.4. Then $X((\mathbb{C}^*)^{d-n})$ acts on $\widetilde{D}$ and $K_0(\widetilde{D})_\mathbb{C}$ is a $\mathbb{C}\{X((\mathbb{C}^*)^{d-n})\}\cong\mathbb{C}[(\mathbb{C}^*)^{d-n}]$-module.

Theorem 6.4 ([42]). *Assume that $\alpha\in\mathbb{C}^{d-n}$ is generic.*[27] *The monodromy representation of the GKZ system of differential equations with parameter α restricted to $\mathcal{K}_X$ is isomorphic to the representation of $\pi_1(\mathcal{K}_X)$ on $K_0(\widetilde{\mathcal{D}})_\mathbb{C}$ specialised at $e^{-2\pi i\alpha}$*[28].

As a corollary, we obtain, in particular, a description of the full monodromy of such "quasi-symmetric" GKZ hypergeometric systems. In [6], Beukers describes the "local" monodromy.

[25] We abuse the notation and denote by B^* also the "complexified" $B^*:\mathbb{C}^n\to\mathbb{C}^d$.

[26] By appropriately varying γ, one can achieve that such power series are a basis of solutions that converge on an open set.

[27] We require that α is non-resonant, i.e., α does not belong to the hyperplane arrangement consisting of $\mathbb{Z}^{d-n}$-translates of the supporting hyperplanes of the cone $\mathbb{R}_+A$.

[28] More precisely, $K_0(\widetilde{\mathcal{D}})_\mathbb{C}\otimes_{\mathbb{C}[(\mathbb{C}^*)^{d-n}]}\mathbb{C}$ for $\mathbb{C}[(\mathbb{C}^*)^{d-n}]\to\mathbb{C}$, $p\mapsto p(e^{-2\pi i\alpha})$.

Remark 6.5. There are various other results where an interesting system of differential equations is obtained from actions on derived categories, often also inspired by mirror symmetry. We mention here [3, 10, 11].

7. Liftings

Theorem 5.4 (and accordingly Theorem 6.4) extend a bit further, in analogy with D-modules introduced in Section 3.5 and the associated perverse sheaves, defined as the image of the abelian category of D-modules by the derived solution functor Section 3.7.

7.1. Perverse schobers

Recall that a representation of $\pi_1(\mathcal{K}_X)$ corresponds to a local system on $\mathcal{K}_X$; cf. Section 3.4. If $\pi_1(\mathcal{K}_X)$ acts instead on a category, we might say that it corresponds to a local system of categories on $\mathcal{K}_X$. In the quasi-symmetric setting, $\mathcal{K}_X$ is a complement of a hyperplane arrangement in $(\mathbb{C}^*)^n$ (in logarithmic coordinates); cf. Theorem 5.2. We may extend a local system on $\mathcal{K}_X$ to a perverse sheaf on $(\mathbb{C}^*)^n$. This extension for the particular action of Theorem 5.4 also lifts on the level of derived categories, and we get what we might call a perverse sheaf of categories on $(\mathbb{C}^*)^n$ [41]. It also goes under the name of a *perverse schober*, which was coined by Kapranov and Schechtman [30] for a categorification of a perverse sheaf.

The rest of this subsection builds on this extension and, in return, also illuminates the proof of Theorem 5.4. Unfortunately, it is rather technical.

7.1.1. Perverse sheaves over real hyperplane arrangements. While in general the abelian category of perverse sheaves might be difficult to describe, in the case of complements of complexified real hyperplane arrangements there exists a concrete combinatorial description [31], which is apt for categorification.

Let $\mathcal{H}$ be an affine hyperplane arrangement in a finite dimensional real vector space $V = \mathbb{R}^n$. Then $\mathcal{H}$ stratifies V into a set $\mathcal{C}$ of locally closed subsets.[29] We partially order $\mathcal{C}$ by $C' \leq C$ iff $C' \subset \overline{C}$. A triple of faces (C_1, C_2, C_3) is *collinear* if there exists $C' \leq C_1, C_2, C_3$ and there exist $c_i \in C_i$ such that $c_2 \in [c_1, c_3]$.

We denote by $\mathrm{vec}(\mathbb{C})$ the category of finite dimensional $\mathbb{C}$-vector spaces.

Theorem 7.1 ([31]). *The category of perverse sheaves on $V_{\mathbb{C}}$ with respect to the stratification induced by $\mathcal{H}_{\mathbb{C}}$ is equivalent to the category of diagrams consisting of*

- *finite dimensional vector spaces E_C, $C \in \mathcal{C}$, and*
- *linear maps $\gamma_{C'C} : E_{C'} \to E_C$, $\delta_{CC'} : E_C \to E_{C'}$ for $C' \leq C$*

[29] The elements of $\mathcal{C}$ are level sets for $(\operatorname{sign} f_H)_{H \in \mathcal{H}}$, where f_H is the affine map defining H.

such that $(E_C, (\gamma_{C'C})_{CC'})$*, resp.* $(E_C, (\delta_{CC'})_{CC'})$*, is a representation of* $(\mathscr{C}, \leq)$*, resp.* $(\mathscr{C}, \geq)$*, in* $\mathrm{vec}(\mathbb{C})$*, and the following conditions are satisfied.*

- $\gamma_{C'C}\delta_{CC'} = \mathrm{id}_{E_C}$ *for* $C' \leq C$*. In particular,* $\phi_{C_1C_2} := \gamma_{C'C_2}\delta_{C_1C'}$ *for* $C' \leq C_1, C_2$ *is well defined.*
- $\phi_{C_1C_2}$ *is an isomorphism for all* C_1, C_2*,* $C_1 \neq C_2$*, of the same dimension* ℓ*, which lie in the same* ℓ*-dimensional affine space and share a facet.*
- $\phi_{C_1C_3} = \phi_{C_2C_3}\phi_{C_1C_2}$ *for collinear triples of faces* (C_1, C_2, C_3).

7.1.2. Perverse schobers over real hyperplane arrangements. To define perverse schobers over real hyperplane arrangements, we may word for word translate the description of perverse sheaves from Theorem 7.1 to the setting of triangulated categories. When we apply $K_0(-)_{\mathbb{C}}$, we get back the data defining a perverse sheaf.

Definition 7.2 ([9]). A perverse schober on $V_{\mathbb{C}}$ with respect to the stratification induced by $\mathscr{H}_{\mathbb{C}}$[30] is given by

- triangulated categories $\mathscr{E}_C$, $C \in \mathscr{C}$, and
- adjoint exact functors $(\delta_{CC'} : \mathscr{E}_C \to \mathscr{E}_{C'}, \gamma_{C'C} : \mathscr{E}_{C'} \to \mathscr{E}_C)$ for $C' \leq C$

such that $(\mathscr{E}_C, (\delta_{C'C})_{C'C})$ defines a pseudo-functor from $(\mathscr{C}, \geq)$ to the 2-category of triangulated categories satisfying the following conditions.

- The unit of the adjunction $(\delta_{CC'}, \gamma_{C'C})$ defines a natural isomorphism
$$\mathrm{id}_{\mathscr{E}_C} \xrightarrow{\cong} \gamma_{C'C}\delta_{CC'}$$
for $C' \leq C$, and thus $\phi_{C_1C_2} := \gamma_{C'C_2}\delta_{C_1C'}$ for $C' \leq C_1, C_2$ is well defined up to canonical natural isomorphism.
- $\phi_{C_1C_2}$ is an equivalence for all C_1, C_2, $C_1 \neq C_2$, of the same dimension ℓ, which lie in the same ℓ-dimensional affine space and share a facet.
- The counit of the adjunction $(\delta_{C_0C_2}, \gamma_{C_2C_0})$ defines a natural isomorphism
$$\phi_{C_2C_3}\phi_{C_1C_2} \xrightarrow{\cong} \phi_{C_1C_3}$$
for collinear triples of faces (C_1, C_2, C_3).

This definition also sheds some light on the proof of Theorem 5.4 (cf. the paragraph following it) and allows its extension.

Theorem 7.3 ([41]). *The local system on* $\mathcal{K}_X$ *from Theorem 5.4 extends to a perverse schober on* $(\mathbb{C}^*)^n$*.*[31]

[30] A perverse schober in this context is also called an $\mathscr{H}$-schober.

[31] We identify $(\mathbb{C}^*)^n$ with $\mathbb{C}^n/\mathbb{Z}^n$, and in order to use Definition 7.2 we should also impose an action of $\mathbb{Z}^n$ on a perverse schober, which consists of isomorphisms $\mathcal{S}_C \to \mathcal{S}_{gC}$ for $g \in \mathbb{Z}^n$ satisfying some compatibility conditions; see e.g. [41, §3.3].

Remark 7.4. By a suitable tweak as in Theorem 6.4, we obtain perverse schobers whose decategorifications are the perverse sheaves obtained as solution complexes of GKZ hypergeometric D-modules [42].

7.2. HMS predictions

GKZ hypergeometric systems appeared here rather ad hoc, and not really motivated. In fact, it is HMS that indicates that they should be there [10, 14, 29].

While we only combinatorially match the two perverse sheaves, one would desire to construct a canonical correspondence via the following sequence of maps (GM denotes Gauss–Manin):

$$\begin{aligned}
&\big(K_0\big(D^b([W/T])\big) \supset \big)K_0(D) \xrightarrow{\sim} K_0(X) \xrightarrow{\sim}\\
&\quad H^*(X) \text{ (for. quantum conn.)} \xrightarrow{\text{mirror map}} \{\text{rel. tw. DR-coh. at } \infty\}\text{(for. GM conn.)}\\
&\qquad\qquad\qquad\qquad\qquad\qquad \xrightarrow{\text{anal. cont.}} \{\text{solutions to GKZ system}\}.
\end{aligned}$$

However, the heuristics of why this action would lift to an action on the derived category of X are still somewhat mysterious.[32]

Acknowledgements. We are foremost grateful to Michel Van den Bergh for a journey to kaleidoscopic areas that would otherwise remain inaccessible to us. Moreover, we thank Geoffrey Janssens, Urban Jezernik, Igor Klep, and the referee for a generous assortment of comments and suggestions.

Funding. The work on this survey was partially supported by the FNRS project MIS/BEJ F.4545.21 and the ARC project “Algebra”.

References

[1] P. Achar, Local systems and constructible sheaves. 2007, applications to homological algebra: Introduction to perverse sheaves, https://www.math.lsu.edu/~pramod/

[2] A. Adolphson, Hypergeometric functions and rings generated by monomials. *Duke Math. J.* **73** (1994), no. 2, 269–290 Zbl 0804.33013 MR 1262208

[3] R. Anno, R. Bezrukavnikov, and I. Mirković, Stability conditions for Slodowy slices and real variations of stability. *Mosc. Math. J.* **15** (2015), no. 2, 187–203, 403 Zbl 1342.14034 MR 3427420

[32] The RHS of the mirror map corresponds to the B-side of the LG-model, which would in turn lead to a GKZ system on the Fukaya category of X, rather than on the derived category.

[4] P. S. Aspinwall, A point's point of view of stringy geometry. *J. High Energy Phys.* (2003), no. 1, 002, 15 Zbl 1225.14033 MR 1970097

[5] A. Beĭlinson and J. Bernstein, Localisation de g-modules. *C. R. Acad. Sci. Paris Sér. I Math.* **292** (1981), no. 1, 15–18 Zbl 0476.14019 MR 610137

[6] F. Beukers, Monodromy of A-hypergeometric functions. *J. Reine Angew. Math.* **718** (2016), 183–206 Zbl 1355.33017 MR 3545882

[7] F. Beukers and G. Heckman, Monodromy for the hypergeometric function ${}_nF_{n-1}$. *Invent. Math.* **95** (1989), no. 2, 325–354 Zbl 0663.30044 MR 974906

[8] A. A. Bolibrukh, The Riemann–Hilbert problem on the complex projective line. *Mat. Zametki* **46** (1989), no. 3, 118–120 Zbl 0687.30004 MR 1032917

[9] A. Bondal, M. Kapranov, and V. Schechtman, Perverse schobers and birational geometry. *Selecta Math. (N.S.)* **24** (2018), no. 1, 85–143 Zbl 1436.14037 MR 3769727

[10] L. A. Borisov and R. P. Horja, Mellin–Barnes integrals as Fourier–Mukai transforms. *Adv. Math.* **207** (2006), no. 2, 876–927 Zbl 1137.14314 MR 2271990

[11] T. Bridgeland, Y. Qiu, and T. Sutherland, Stability conditions and the A_2 quiver. *Adv. Math.* **365** (2020), 107049, 33 Zbl 07184853 MR 4064773

[12] E. Clader and Y. Ruan, Mirror symmetry constructions. In *B-model Gromov–Witten theory*, pp. 1–77, Trends Math., Birkhäuser/Springer, Cham, 2018 Zbl 1423.81144 MR 3965408

[13] T. Coates, A. Corti, H. Iritani, and H.-H. Tseng, Computing genus-zero twisted Gromov–Witten invariants. *Duke Math. J.* **147** (2009), no. 3, 377–438 Zbl 1176.14009 MR 2510741

[14] T. Coates, A. Corti, H. Iritani, and H.-H. Tseng, Hodge-theoretic mirror symmetry for toric stacks. *J. Differential Geom.* **114** (2020), no. 1, 41–115 Zbl 1464.14044 MR 4047552

[15] D. A. Cox and S. Katz, *Mirror Symmetry and Algebraic Geometry*. Math. Surveys Monogr. 68, American Mathematical Society, Providence, RI, 1999 Zbl 0951.14026 MR 1677117

[16] P. Deligne, *Équations différentielles à points singuliers réguliers*. Lecture Notes in Math. 163, Springer, Berlin, 1970 Zbl 0244.14004 MR 0417174

[17] W. Donovan, Perverse schobers on Riemann surfaces: constructions and examples. *Eur. J. Math.* **5** (2019), no. 3, 771–797 Zbl 1423.14120 MR 3993263

[18] W. Donovan and E. Segal, Mixed braid group actions from deformations of surface singularities. *Comm. Math. Phys.* **335** (2015), no. 1, 497–543 Zbl 1327.14185 MR 3314511

[19] I. M. Gel'fand, M. M. Kapranov, and A. V. Zelevinsky, Generalized Euler integrals and A-hypergeometric functions. *Adv. Math.* **84** (1990), no. 2, 255–271 Zbl 0741.33011 MR 1080980

[20] I. M. Gel'fand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, Resultants and Multidimensional Determinants*. Modern Birkhäuser Classics, Birkhäuser, Boston, MA, 2008; Reprint of the 1994 edition Zbl 1138.14001 MR 2394437

[21] I. M. Gel'fand, A. V. Zelevinskiĭ, and M. M. Kapranov, Equations of hypergeometric type and Newton polyhedra. *Dokl. Akad. Nauk SSSR* **300** (1988), no. 3, 529–534 Zbl 0667.33010 MR 948812

[22] I. M. Gel'fand, A. V. Zelevinskiĭ, and M. M. Kapranov, Hypergeometric functions and toric varieties. *Funktsional. Anal. i Prilozhen.* **23** (1989), no. 2, 12–26 Zbl 0721.33006 MR 1011353

[23] D. Halpern-Leistner, The derived category of a GIT quotient. *J. Amer. Math. Soc.* **28** (2015), no. 3, 871–912 Zbl 1354.14029 MR 3327537

[24] D. Halpern-Leistner and S. V. Sam, Combinatorial constructions of derived equivalences. *J. Amer. Math. Soc.* **33** (2020), no. 3, 735–773 Zbl 1454.14045 MR 4127902

[25] D. Hilbert, Mathematische Probleme. *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.* **1900** (1900), 253–297 Zbl 31.0068.03

[26] D. Hilbert, Mathematical problems. *Bull. Amer. Math. Soc.* **8** (1902), no. 10, 437–479 Zbl 33.0976.07 MR 1557926

[27] K. Hori and C. Vafa, Mirror symmetry. 2000, arXiv:hep-th/0002222

[28] R. Hotta, K. Takeuchi, and T. Tanisaki, *D-Modules, Perverse Sheaves, and Representation Theory*. Progr. Math. 236, Birkhäuser, Boston, MA, 2008 Zbl 1136.14009 MR 2357361

[29] H. Iritani, Quantum D-modules of toric varieties and oscillatory integrals. In *Handbook for Mirror Symmetry of Calabi–Yau & Fano Manifolds*, pp. 131–147, Adv. Lect. Math. (ALM) 47, Int. Press, Somerville, MA, 2020 Zbl 1440.14190 MR 4237880

[30] M. Kapranov and V. Schechtman, Perverse schobers. 2015, arXiv:1411.2772

[31] M. Kapranov and V. Schechtman, Perverse sheaves over real hyperplane arrangements. *Ann. of Math. (2)* **183** (2016), no. 2, 619–679 Zbl 1360.14062 MR 3450484

[32] M. Kashiwara, Faisceaux constructibles et systèmes holonômes d'équations aux dérivées partielles linéaires à points singuliers réguliers. In *Séminaire Goulaouic–Schwartz, 1979–1980 (French)*, p. Exp. No. 19, École Polytech., Palaiseau, 1980 Zbl 0444.58014 MR 600704

[33] M. Kashiwara, The Riemann–Hilbert problem for holonomic systems. *Publ. Res. Inst. Math. Sci.* **20** (1984), no. 2, 319–365 Zbl 0566.32023 MR 743382

[34] N. M. Katz, An overview of Deligne's work on Hilbert's twenty-first problem. In *Mathematical Developments Arising from Hilbert Problems (Proc. Sympos. Pure Math., Northern Illinois Univ., De Kalb, Ill., 1974)*, pp. 537–557, Amer. Math. Soc., Providence, RI, 1976 Zbl 0347.14010 MR 0432640

[35] A. Kite, Discriminants and quasi-symmetry. 2017, arXiv:1711.08940

[36] M. Kontsevich, Homological algebra of mirror symmetry. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pp. 120–139, Birkhäuser, Basel, 1995 Zbl 0846.53021 MR 1403918

[37] Z. Mebkhout, Une équivalence de catégories. *Compositio Math.* **51** (1984), no. 1, 51–62 Zbl 0566.32021 MR 734784

[38] J. Plemelj, Riemannsche Funktionenscharen mit gegebener Monodromiegruppe. *Monatsh. Math. Phys.* **19** (1908), no. 1, 211–245 Zbl 39.0461.01 MR 1547764

[39] M. Salvetti, Topology of the complement of real hyperplanes in $\mathbb{C}^N$. *Invent. Math.* **88** (1987), no. 3, 603–618 Zbl 0594.57009 MR 884802

[40] Š. Špenko and M. Van den Bergh, Non-commutative resolutions of quotient singularities for reductive groups. *Invent. Math.* **210** (2017), no. 1, 3–67 Zbl 1375.13007 MR 3698338

[41] Š. Špenko and M. Van den Bergh, A class of perverse schobers in geometric invariant theory. 2019, arXiv:1908.04213

[42] Š. Špenko and M. Van den Bergh, Perverse schobers and GKZ systems. *Adv. Math.* **402** (2022), Paper No. 108307 Zbl 07524877 MR 4406929

[43] The Stacks project authors, The stacks project. 2021, https://stacks.math.columbia.edu

[44] M. van den Bergh, Non-commutative crepant resolutions. In *The Legacy of Niels Henrik Abel*, pp. 749–770, Springer, Berlin, 2004 Zbl 1082.14005 MR 2077594

Špela Špenko
Département de Mathématique, Université Libre de Bruxelles, Campus de la Plaine CP 213,
Bld du Triomphe, 1050 Bruxelles, Belgium; spela.spenko@ulb.be

DOI 10.4171/8ECM/16

AAA-least squares rational approximation and solution of Laplace problems

Stefano Costa and Lloyd N. Trefethen

Abstract. A two-step method for solving planar Laplace problems via rational approximation is introduced. First, complex rational approximations to the boundary data are determined by AAA approximation, either globally or locally near each corner or other singularity. The poles of these approximations outside the problem domain are then collected and used for a global least-squares fit to the solution. Typical problems are solved in a second of laptop time to 8-digit accuracy, all the way up to the corners, and the conjugate harmonic function is also provided. The AAA-least squares combination also offers a new method for avoiding spurious poles in other rational approximation problems, and for greatly speeding them up in cases with many singularities. As a special case, AAA-LS approximation leads to a powerful method for computing the Hilbert transform or Dirichlet-to-Neumann map.

1. Introduction

The aim of this paper is to introduce a new method for the numerical solution of planar Laplace problems, based on a combination of local complex rational approximations by the AAA algorithm followed by a real linear least-squares problem. This method is an outgrowth of three previous works [8, 16, 23], which we now briefly summarize.

The AAA algorithm (adaptive Antoulas–Anderson, pronounced "triple-A") is a fast and flexible method for near-best complex rational approximation [23]. Given a vector Z of real or complex sample points and a corresponding vector F of data values, it finds a rational function r of specified degree or accuracy such that

$$r(Z) \approx F. \tag{1.1}$$

This is done by developing a barycentric representation for r by alternating a nonlinear step of greedy selection of the next barycentric support point with a linear least-squares approximation step to determine the barycentric weights. If F is obtained by

2020 Mathematics Subject Classification. Primary 41A20; Secondary 30E10, 44A15, 65N35.
Keywords. Rational approximation, AAA algorithm, lightning Laplace solver, conformal mapping, Hilbert transform, Dirichlet-to-Neumann map.

sampling a function $f(z)$ with singularities at certain points of Z, such as logarithms and fractional powers, then root-exponential convergence with respect to the degree n is typically achieved (i.e., errors $O(\exp(-C\sqrt{n}\,))$ for some $C > 0$), with poles of the approximants f clustering exponentially near the singularities [26, 32]. The standard implementation of AAA approximation is the code `aaa` in Chebfun [10].

The lightning Laplace solver is a method for solving Laplace problems

$$\Delta u = 0 \text{ on } \Omega, \quad u = h(z) \text{ on } \partial\Omega \tag{1.2}$$

on a simply connected domain Ω in the plane, which we parametrize for convenience by the complex variable z [16]. It also computes an analytic function $f(z)$ such that $u = \operatorname{Re} f$. This method first fixes poles with exponential clustering near each corner of Ω or other point where a singularity is expected. A real linear least-squares problem is then solved to determine a rational function in Ω with the prescribed poles, plus a polynomial term (i.e., poles at infinity), whose real part matches the boundary data as closely as possible. The method converges root-exponentially with respect to the number of poles and generalizes to Neumann boundary data, multiply-connected domains, and the Stokes and Helmholtz equations [7, 14]. The standard implementation is the MATLAB code `laplace` available at [30].

Although the lightning Laplace solver is fast and effective, one would really like to solve Laplace problems by a method more like the AAA algorithm, which allows the set Z to be completely arbitrary and adapts to the singularities of the solution automatically rather than relying on a priori estimates of pole clustering. Two challenges have held back the development of a AAA method for Laplace problems. First, no barycentric representation is known for real parts of rational functions. Second, even if such a formula were available, there would remain the fundamental problem of achieving approximation in a region Ω based on values on the boundary $\partial\Omega$. A AAA-style approximation does not distinguish interior from exterior and includes no mechanism to restrict poles to the latter.

These considerations led to the third contribution that this work builds upon, published on arXiv by the first author in 2020 [8]. The upper row of Figure 1 illustrates the idea as applied to the “NA Digest model problem” [28], an L-shaped region with boundary data $u(z) = (\operatorname{Re} z)^2$. First, complex AAA is used to approximate the real data on the boundary. The resulting analytic function is complex (though real on $\partial\Omega$, up to the approximation accuracy), with poles both inside and outside Ω. Then the poles in Ω are discarded, leaving a set of poles outside Ω that are often clustered effectively for rational approximation. The Laplace problem is solved by computing such an approximation by linear least-squares fitting on $\partial\Omega$.

In the form just described, the AAA-Laplace method can be quite slow because of depending on AAA approximations with a large number of poles. In this article,

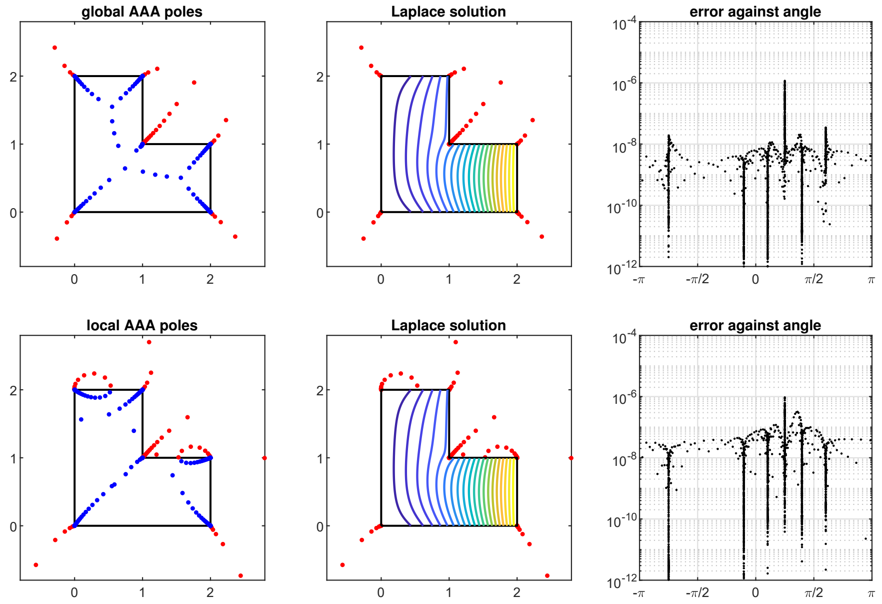


Figure 1. Above, Costa's AAA-Laplace method from [8]. A global AAA approximation gives poles both inside and outside Ω. The poles inside are discarded, and those outside are used for a linear least-squares fit. Errors on the boundary in the rightmost plot are plotted against angle with respect to the point $(1+i)/2$. This computation determines $u(0.99+0.99i) \approx 1.0267919261073$ to 10 digits of accuracy, but it takes 12 s of laptop time because the AAA approximation has 294 poles. Below, the new local variant, in which the poles outside Ω are determined by local AAA approximations near each corner. The computation time falls to 0.67 s because the AAA problems are six times smaller, without much change in accuracy.

we propose a variation that often speeds it up greatly, namely, to use local AAA approximations near each singularity to choose the set of poles. Since the cost of AAA approximation grows with the fourth power of the number of poles, this leads to a speedup potentially by a factor on the order of the cube of the number of corners. For the L-shaped example, the speedup is a factor of about 18.

The AAA-Laplace method as presented in [8] was actually much slower than indicated in Figure 1 for an accidental reason. In that implementation, `aaa` was invoked in its default "cleanup" mode, which led to the removal of many poles close to the singularities and a consequent need to employ AAA approximations involving as many as 1,000 poles. The explanation was proposed in that paper that discarding poles in Ω may tend to halve the number of digits of accuracy, but we recognize now that it is not so, that the loss of accuracy was a consequence of using the cleanup feature. Throughout this paper, we always call `aaa` with "cleanup off."

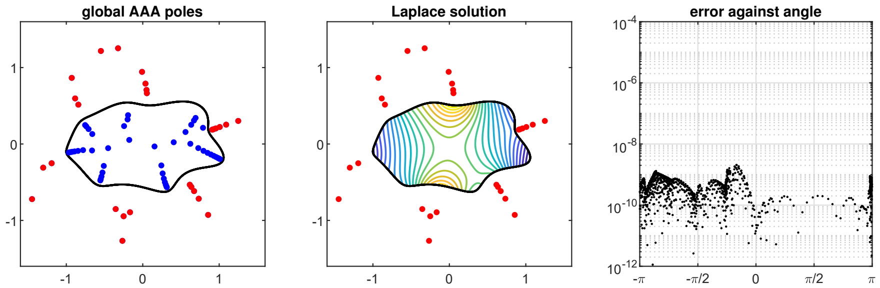


Figure 2. A smooth Laplace problem solved by the global AAA-LS method. A global AAA approximation produces 46 poles inside Ω and 30 outside, and the latter are retained for a real least-squares problem that also includes a polynomial term. 9-digit accuracy is achieved in 0.7 s.

2. Laplace problems

Our main interest is problems with corner singularities, since this is where the power and convenience of rational functions are most decisive. However, the AAA approach can be effective for smooth problems too. Figure 2 presents an example. An irregular domain Ω (bounded by a trigonometric interpolant through 15 complex data points) is given with the Laplace boundary condition $u(z) = -\log|z|$. The vector Z is constructed by sampling $\partial\Omega$ in 1,000 points, and a global AAA fit to the boundary data with tolerance 10^{-8} yields 46 poles in Ω and 30 in $\mathbb{C}\backslash\overline{\Omega}$. The interior poles are discarded, and a least-squares fit to the boundary data is computed via a $1{,}000 \times 102$ matrix: 60 real degrees of freedom for 30 poles and 42 for a polynomial term of degree 20. The computation takes 0.7 s, and the maximum error on Z is 2.1×10^{-9}. A polynomial expansion needs about 10 times as many degrees of freedom to achieve the same accuracy, a ratio that would worsen exponentially for more distorted regions according to the theory of the “crowding phenomenon” in complex analysis [15, Thm. 5].

We now turn to problems with singularities, typically at corners, whose locations are assumed to be known in advance. The local variant of the AAA-LS algorithm proceeds in this manner:

(1) *construct sample point vector Z and fix corresponding data values $H = h(Z)$;*

(2) *for each singularity, run AAA for nearby sample points and data values;*

(3) *discard poles in $\overline{\Omega}$ and retain poles exterior to $\overline{\Omega}$;*

(4) *calculate real least-squares fit to boundary data, including a polynomial term;*

(5) *construct function handles for $u(z)$ and its analytic extension $f(z)$.*

We give some mathematical and MATLAB details of each of these steps. The global variant of the algorithm is the same except that step (2) involves just a single global AAA approximant.

(1) *Construct sample point vector Z and fix corresponding data values $H = h(Z)$.* The problem domain Ω can be quite arbitrary, and it can be multiply connected. Typically, Z will consist of hundreds or thousands of points, which it is simplest to specify in advance with exponential clustering near singularities. In MATLAB, we use constructions like `logspace(-14,0,300)'` for a singularity at one endpoint of [0, 1] and `tanh(linspace(-16,16,600)')` for singularities at both endpoints of [−1, 1]. If AAA-LS software were to be developed analogous to the `laplace` code of [30] for the lightning Laplace method, then it would be worthwhile placing sample points more strategically to avoid having too many more rows in the matrix than necessary.

(2) *For each singularity, run AAA for nearby sample points and data values.* We use the simplest choice: each point of Z is associated with whichever singularity it is closest to (on the same boundary component, if the geometry is multiply connected so there are several boundary components). The Chebfun command `aaa` is invoked with `'cleanup'`, `'off'`, and throughout this paper we specify a AAA tolerance of 10^{-8}.

(3) *Discard poles in $\overline{\Omega}$ and retain poles exterior to $\overline{\Omega}$.* The `aaa` code returns highly accurate pole locations computed via a matrix generalized eigenvalue problem described in [23]. To distinguish those inside and outside Ω, we use the complex variant `inpolygonc = @(z,w) inpolygon(real(z),imag(z),real(w),imag(w))` of the `inpolygon` command.

(4) *Calculate real least-squares fit to boundary data, including a polynomial term.* If `pol` is a row vector of the poles from step (3) and `n` is a small nonnegative integer, then the sequence

```
d = min(abs(Z-pol),[],1);
P = Z.^(0:n); Q = d./(Z-pol);
A = [real(P) real(Q) -imag(P) -imag(Q)];
c = reshape(A\H,[],2)*[1;1i];
```

computes a complex coefficient vector c for the function f in the space spanned by the polynomials of degree n and the given poles such that $u = \operatorname{Re} f$ is the least-squares fit to the data H in the sample points. The vector d contains the distances of the poles to Z and is used to scale the columns of Q to have ∞-norm 1. For n much larger than 10, however, numerical stability requires that the monomials of `Z.^(0:n)` be replaced by orthogonalizations computed by the Vandermonde with Arnoldi procedure of [6]. This can be done by replacing `P = Z.^(0:n)` by `[Hes,P] = VAorthog(Z,n)`, where the code `VAorthog` comes from [7] and is listed in the appendix.

The description and code above apply for bounded, simply-connected domains with Dirichlet boundary conditions. For problems with Neumann boundary conditions on some sides, the corresponding rows of A are modified appropriately. For exterior domains, z is replaced by $(z - z_c)^{-1}$ for some point z_c in the hole. For multiply-connected domains, additional columns of the form $\log|z - z_j|$ must be added, where $\{z_j\}$ are a set of fixed points, one in each hole [2, 29]. In addition, new columns are added corresponding to polynomials in $1/(z - z_j)$ for each j.

(5) *Construct function handles for $u(z)$ and its analytic extension $f(z)$.* For convenience in making plots and other applications, it is desirable to have functions that can be applied to matrices as well as vectors. Following the commands above this can be achieved with

```
f = @(z) reshape([z(:).^(0:n) d./(z(:)-pol)]*c,size(z));
u = @(z) real(f(z)); v = @(z) imag(f(z));
```

When `VAorthog` is used, the first line is replaced by

```
f = @(z) reshape([VAeval(z(:),Hes) d./(z(:)-pol)]*c,size(z));
```

Figure 3 illustrates the method at work on two examples. In the first row, the L shape of Figure 1 has been modified to a square with two circular bites removed. No new issues arise here, as the method does not distinguish between straight and curved sides, so long as they are smooth. The second row shows a doubly-connected problem, and here some new issues do arise. First there is the use of polynomials with respect to both z and $(z - z_c)^{-1}$ as described above (writing z_c instead of z_1 since there is just one hole), as well as the introduction of a $\log|z - z_c|$ term; we take $z_c = -(1 + i)/4$. The domain is discretized by 400 clustered points on each of the eight side segments, and the polynomials in z and $(z - z_c)^{-1}$ are of degree 40. A more fundamental issue also arises in this problem. The boundary data have been taken as 1 on the inner square and 0 on the outer square, a natural situation for a heat flow or electrostatics problem in a doubly connected geometry. Since these boundary conditions are constant on each of the two boundary components, however, the local AAA problems will be trivial and no poles at all will be produced! Clearly that is no route to an accurate solution, so for this computation, poles have been generated by using an artificial boundary condition (the square root of the product of the distances to the eight corners) and then the least-squares problem is solved with the boundary data actually prescribed. The reader is justified if he/she finds this puzzling, and we discuss the matter further in Section 6.

Our final example of this section is an unbounded region with three rectangular holes, as shown in Figure 4. The boundary conditions are $u = 1$ on the rectangle at the left and $u = 0$ on the other two, giving a natural interpretation as the potential around three conductors. Each boundary segment is discretized by 400 clustered points, so

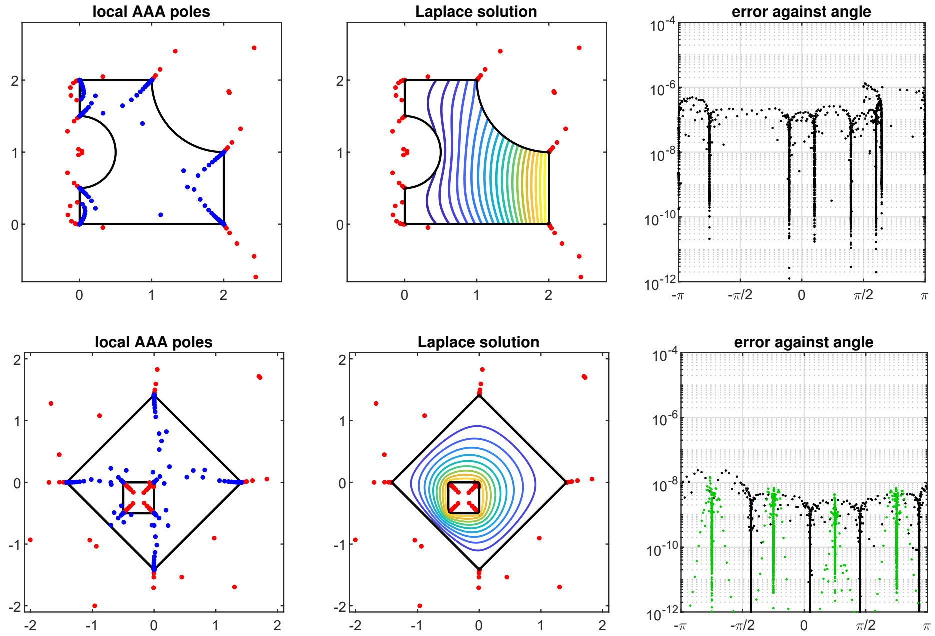


Figure 3. Two examples of Laplace solutions by the local AAA-LS method. Above, a square with two circular bites removed. The computation involves 102 poles outside the domain and a polynomial of degree 20. Below, a multiply-connected domain, solved in 1.7 s with 397 poles outside the domain and a polynomial of degree 40. In the error plot, black dots correspond to the outer boundary and green dots to the inner one. The boundary data used for local-AAA pole location are not those of the Laplace problem, as explained in the text.

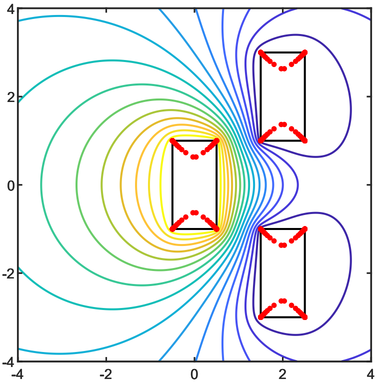


Figure 4. Local AAA-LS solution of a Laplace problem in an unbounded triply-connected domain, requiring reciprocal polynomials with respect to three interior points c_j and also logarithm terms $\log|z - z_j|$. The computation takes 2 s and gives the value $u(1) \approx 0.64357510429036$ to 10-digit accuracy.

the least-squares matrix has 4,800 rows. The AAA fits lead to 52 poles inside a rectangle near each corner, 624 in total, and we also have a reciprocal-polynomial of degree 10 and one real logarithm term in each rectangle, bringing the number of columns of the matrix to $2 \times (624 + 3 \times 11) + 3 = 1{,}317$. A solution is computed in 2 s to 10-digit accuracy as measured by the value at the point $z = 1$ midway between the rectangles, $u(1) \approx 0.64357510429036$.

A fine point to note in this triply-connected example is that the point $z = \infty$ is a point of analyticity, in the interior of the domain, so there should be no logarithmic term there, meaning that the sum of the coefficients of the three log terms centered at the points z_1, z_2, z_3 in the rectangles should be zero. This condition can be enforced by adding one more row to the matrix, or (as was in fact done for the computation in Figure 4) by taking the log columns of the matrix to correspond not to $\log|z - z_j|$ but to $\log|z - z_j| - \log|z - z_{j(\mathrm{mod}\ 3)+1}|$.

3. Conformal mapping

A Laplace solver that also produces the harmonic conjugate of the solution, hence its analytic extension, can be used to compute conformal maps. Details are given in [31], so here we give just one example of construction of the conformal map g of a simply-connected region Ω containing the point $z = 0$ to the unit disk, with $g(0) = 0$ and $g'(0) > 0$. The trick is to write g in the form

$$g(z) = z \exp\bigl(f(z)\bigr), \quad f(z) = \log\bigl(g(z)/z\bigr), \tag{3.1}$$

where f is the unique nonzero analytic function on Ω having real part $-\log|z|$ on $\partial\Omega$ and imaginary part 0 at $z = 0$. Thus f is obtained by solving a Laplace Dirichlet problem, and (3.1) then gives the map g.

Figure 5 illustrates this method for the smooth region of Figure 2, where $-\log|z|$ was already the boundary condition. Thus the conformal map comes from exponentiating the analytic extension of the harmonic function of Figure 2 and multiplying the result by z. As described in [15], this result is then compressed by AAA approximation, and another AAA approximation gives the inverse map. See [31] for extensions to multiply-connected regions.

The speed of these computations is remarkable. After an initial 0.9 s to construct the forward and inverse maps in this example, they can be each then be evaluated in 0.3 μs per point. For example, we take one million random points uniformly distributed in the unit disk, map them conformally to Ω, then map these images back to the unit disk again. The whole back-and-forth process takes 0.6 s, and the maximum error in the million sample points is 1.1×10^{-8}.

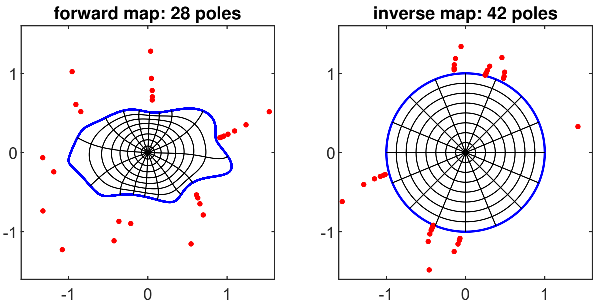


Figure 5. Conformal map of the region of Figure 2 by the global AAA-LS method. The map is computed to 8-digit accuracy in 0.8 s and the rational approximations in both directions are evaluable in less than 1 μs per point. In the left image, the poles differ slightly from those of Figure 2 because a further AAA compression of $z\exp(f(z))$ has taken place.

4. Rational approximation without spurious poles

Though the emphasis in this paper is on Laplace problems, AAA-LS approximation also offers striking advantages for more general rational approximations. It may be much faster than AAA alone for problems with a number of singularities, and since unwanted poles can be discarded, it produces approximations guaranteed to have desired properties of analyticity and stability. Thus AAA-LS may combat what Heather Wilber has called the “spurious poles blues” (discussed in [34], though without this phrase).

We illustrate both the speed and the robustness with an example of approximating a real zigzag function on the interval $[-1, 1]$, as shown in Figure 6. Knowing that poles will need to cluster exponentially at the points $-0.8, -0.6, \ldots, 0.8$, we set up a 3,000-point grid consisting of `-0.9 + 0.2*tanh(linspace(-16,16,300))` and its nine translates at centers $-0.7, -0.5, \ldots, 0.9$. With straight AAA approximation, poles in $[-1, 1]$ virtually always appear. They could be removed for input to a least-squares fit, but the timing would still be very slow for the moderately large degrees needed for effective approximation: 0.3 s, 4.2 s, and 35.3 s on our laptop for degrees 50, 200, and 500. By contrast, with its local AAA fits the AAA-LS method quickly computes a good approximation. In Figure 6, AAA-LS has been run with AAA tolerance 10^{-8}, leading to local fits each of size 51 and 52 and hence quite speedy. This gives 466 poles all together, two of which lie in $[-1, 1]$ and are discarded, as shown in the middle panel of the figure. A least-squares fit with these 464 poles, plus a polynomial of degree 16, then gives the error marked in blue in the bottom figure, with maximum error 3.1×10^{-7}. The whole computation takes half a second, and the resulting approximation can be evaluated in 5 μs per point. By contrast, a polynomial fit with the same 962 degrees of freedom can have error no smaller than 1.6×10^{-3},

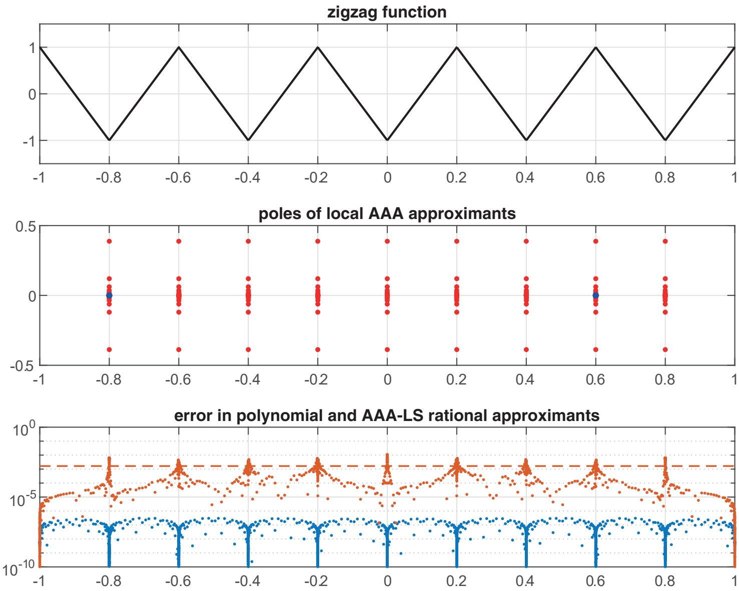


Figure 6. Top, a real zigzag function on $[-1, 1]$ to be approximated over the whole interval by a single rational function. Middle, the 466 poles determined by local AAA fits near each singularity, each of degree 51 or 52. Two poles lie in $[-1, 1]$ and are discarded (blue). Bottom, the resulting errors in the AAA-LS fit show accuracy of 3×10^{-7}. A polynomial with the same 962 degrees of freedom such as a Chebyshev interpolant (dots) could have accuracy at best 10^{-3} (dashed line).

as marked by the orange dashed line. The orange dots show the error for a polynomial Chebyshev interpolant of that degree.

It appears that AAA-LS offers a flexible, fast, and reliable way to compute near-best rational approximations with no unwanted poles. Potential applications lie in many areas of computational science and engineering. An interesting question is “might AAA-LS be further leveraged via a AAA-Lawson iteration as in [24] to lead to truly minimax rational approximations in certain cases?” For this to be possible, it would be necessary first to convert the rational approximation to barycentric form. We have not explored this possibility.

5. Computing the Hilbert transform

If u is a sufficiently smooth real function defined on the real line, its Hilbert transform is the function v defined by the principal value integral

$$v(y) = \frac{1}{\pi} \, \mathrm{PV} \int_{-\infty}^{\infty} \frac{u(x)}{y-x} \, dx. \tag{5.1}$$

The transform can be interpreted as follows: if f is a complex analytic function in the upper half-plane with $\mathrm{Re}\, f(x) = u(x)$ for $x \in \mathbb{R}$, then $v(y) = \mathrm{Im}\, f(y)$. Similar definitions and interpretations apply to the unit circle and other contours. Another name for the Hilbert transform (essentially) is the Dirichlet-to-Neumann map.

It is evident that to compute the Hilbert transform numerically, it suffices to find an analytic function in the upper half-plane whose real part on $\mathbb{R}$ matches that of u to sufficient accuracy. The classical idea of this kind is to use a Fourier transform, perhaps discretized on a finite interval by the fast Fourier transform [20, p. 203]. For example, this is the method used by the `hilbert` command in the MATLAB Signal Processing Toolbox. But it is also possible to use rational approximations instead of trigonometric polynomials, and numerical methods of this kind have been proposed [22, 27, 33].

The AAA-LS method provides another natural approach based on rational approximation, since poles in the upper half-plane can be discarded to ensure the appropriate analyticity. Indeed, all of our AAA-LS Laplace solutions can be regarded as Hilbert transforms, but on more general contours $\partial\Omega$. A prototype code for the real line can be written like this:

```
function [v,f] = ht(u)
X = logspace(-10,10,300)'; X = [X; -X];      % sampling grid
[~,pol] = aaa(u(X),X,'cleanup',0);           % global AAA fit
pol(imag(pol)>=0) = []; pol = pol.';         % discard unwanted poles
d = min(abs(X-pol),[],1);                    % for column normalization
A = d./(X-pol); A = [real(A) -imag(A)];      % fitting matrix
c = reshape(A\u(X),[],2)*[1;1i];             % least-squares solve
f = @(x) reshape((d./(x(:)-pol))*c,size(x)); % analytic extension
v = @(x) imag(f(x));                         % Hilbert transform
```

This is not an item of software—it is a proof of concept. Note that the sampling grid has been taken as 300 points exponentially spaced from 10^{-10} to 10^{10} and their negatives, 600 points all together. This would not be appropriate for all functions, but it is a good starting point for a function which loses analyticity possibly at 0 and at ∞. The code does well at computing Hilbert transforms of the seven example functions in Weideman's list in Table 1 of his paper [33]. In 0.6 s total on a laptop, it produces results for these seven example problems with relative accuracy in the range of 5–13 digits, as detailed in Table 1. We shall not attempt systematic comparisons with other algorithms, but as an indication of the nontriviality of these computations, we mention that applying the MATLAB `hilbert` command for $u(x) = \exp(-x^2)$ on a grid of 1,024 equispaced points in $[-20, 20]$ gives an estimate of $v(2)$ with an error of 1.4×10^{-2}, 11 orders of magnitude greater than the figure in Table 1.

Function u	Hilbert transform $v(2)$	AAA-LS error
$1/(1+x^2)$	0.400000000000000	−1.3 e–12
$1/(1+x^4)$	0.415945165403851	−4.3 e–14
$\sin(x)/(1+x^2)$	0.156805255543717	3.4 e–06
$\sin(x)/(1+x^4)$	0.121897775700258	−1.7 e–07
$\exp(-x^2)$	0.340026217066066	1.0 e–13
$\operatorname{sech}(x)$	0.506584586167368	1.3 e–10
$\exp(-\lvert x\rvert)$	0.328435745958114	−1.4 e–12

Table 1. The example functions $u(x)$ from Table 1 by Weideman [33] together with their Hilbert transforms $v(x)$ evaluated at the arbitrary point $x = 2$. In a total time of 0.6 s, the prototype AAA-LS code `ht` computes these numbers to 5–13 digits of relative accuracy.

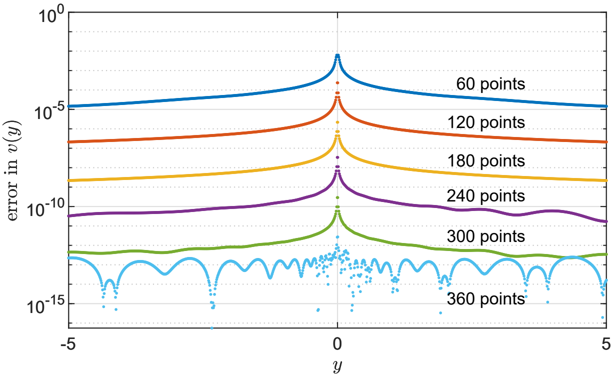


Figure 7. Error at 1,000 points $y \in [-5, 5]$ in the Hilbert transform of $u(x) = \exp(-|x|)$ computed by the global AAA-LS method from $60, 120, \dots, 360$ exponentially spaced samples. This plot was produced in 2 s on a laptop.

Figure 7 illustrates AAA-LS computation of the Hilbert transform graphically for Weideman's final example,

$$u(x) = e^{-|x|}, \quad v(y) = \pi^{-1}\operatorname{sign}(y)\big[e^{|y|}E_1\big(|y|\big) + e^{-|y|}Ei\big(|y|\big)\big], \tag{5.2}$$

where E_1 and Ei are the exponential integrals computed in MATLAB by `expint` and `ei`. For each of the values $L = 1, 2, \dots, 6$, a sample grid of $60L$ points y has been used consisting of $30L$ points exponentially spaced from 10^{-L} to 10^{L} and their negatives. Rapid convergence is observed to an accuracy of better than 10 digits, despite the singularity of u at $x = 0$.

The great flexibility of the AAA-LS method for computing the Hilbert transform is to be noted. It can work with arbitrary data points, which need not be regularly spaced, and it delivers a result as a global representation speedily evaluated via a function handle. No interpolation of data is required (see discussion of this problem in [9]), and singularities in $u(x)$ cause little degradation of accuracy so long as there are sample points clustered nearby, as illustrated in the example of Figure 7.

Many generalizations of this AAA-LS Hilbert transform computation are possible, including other contours both open and closed and more general Riemann–Hilbert problems.

6. Theoretical observations

The core of the AAA-LS method (in its global form) is the following idea, which we shall call the *pole symmetry principle*. Suppose r is a complex rational approximation that closely approximates a real function h on the boundary $\partial\Omega$ of a region Ω. Then there is another complex rational function r_+, *with poles only at the locations of the poles of r outside Ω*, such that $\operatorname{Re} r_+$ also closely approximates h on $\partial\Omega$. The AAA-LS method finds r by AAA approximation on $\partial\Omega$, extracts its poles outside Ω, and then finds r_+ by linear least-squares fitting on $\partial\Omega$.

In particular, for cases with singularities on $\partial\Omega$, rational functions r exist with root-exponential convergence to h as $n \to \infty$ [16]. Such approximations will usually have poles that cluster exponentially on both sides of $\partial\Omega$ near each singularity. The pole symmetry principle proposes that we can discard all the poles inside Ω, retaining only the ones outside Ω, and still get essentially the same root-exponential convergence.

In this section, we assess this idea. Our conclusions can be summarized as follows.

(1) If Ω is a half-plane or a disk, the pole symmetry principle holds exactly (Theorems 6.1 and 6.2).

(2) If Ω is a simply-connected domain with corners, the pole symmetry principle fails in the worse case in that r_+ may have no poles near $\partial\Omega$ even though they are needed to resolve singularities; conversely it may have clusters of poles near $\partial\Omega$ when they are not needed (examples shown in Figure 8). However, both of these situations are nongeneric. For most problems, the principle holds also on regions with corners.

(3) If Ω is a simply-connected domain bounded by an analytic curve, then in a certain theoretical sense it can be reduced to the case of a disk. However, the constants involved may be sufficiently adverse that in practice; it may be more appropriate to think of Ω as a domain with corners. Again the pole symmetry principle will usually hold even if this cannot be guaranteed in the worst case.

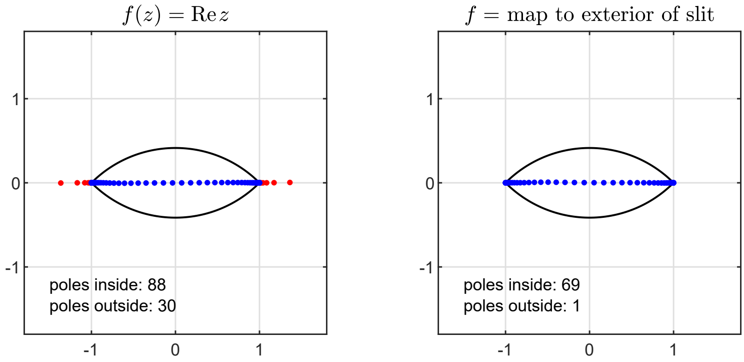


Figure 8. Examples showing that in the worst case, the pole symmetry principle underlying the global AAA-LS method may fail. On the left, AAA approximation gives "too many poles," with poles exponentially clustered outside Ω near ± 1 even though the singularity-free function $u(z) = \mathrm{Re}\, z$ solves the Laplace problem. On the right, it gives "too few poles," providing no poles at all outside Ω near the boundary even though the rational approximation of the solution of the Laplace problem will need them to approximate the branch point singularities at ± 1. Both these situations are nongeneric and unlikely to appear in practice.

(4) If Ω is a multiply-connected domain, then harmonic functions in Ω can in general not be approximated by rational functions: logarithmic terms are needed too. Thus the pole symmetry principle is inapplicable and a local rather than global variant of AAA-LS should be used.

To establish conclusion (1), let $\mathbb{C}_-$ and $\mathbb{C}_+$ denote the open lower and upper complex half-planes, respectively, and let $\|\cdot\|_E$ denote the supremum norm over a set E. The two assertions of the following theorem ensure that complex rational approximation on $\mathbb{R}$ produces "enough poles" to solve the Laplace problem on $\mathbb{C}_+$, and that it does not produce "too many poles" to be efficient.

Theorem 6.1. *Given a bounded real continuous function h on $\mathbb{R}$, let u be the bounded harmonic function in $\mathbb{C}_+$ with $u(x) = h(x)$ for $x \in \mathbb{R}$. Suppose there exists a rational function r, also real on $\mathbb{R}$, such that $\|r - h\|_{\mathbb{R}} \le \varepsilon$ for some $\varepsilon \ge 0$. Then there exists a rational function r_+ whose poles are precisely the poles of r in $\mathbb{C}_-$ such that $\|\mathrm{Re}\, r_+ - h\|_{\mathbb{R}} \le \varepsilon$, and thus by the maximum principle also $\|\mathrm{Re}\, r_+ - u\|_{\mathbb{C}_+} \le \varepsilon$. Conversely, if r_+ is a rational function analytic in $\mathbb{C}_+$ such that $\|\mathrm{Re}\, r_+ - u\|_{\mathbb{C}_+} \le \varepsilon$, then there exists a rational function r whose poles are the poles of r_+ and their reflections in $\mathbb{C}_+$ such that $\|r - h\|_{\mathbb{R}} \le \varepsilon$.*

Proof. Given r as indicated in the first assertion, write $r(z) = (r_+(z) + r_-(z))/2$, where r_+ has its poles in $\mathbb{C}_-$ and r_- has its poles in $\mathbb{C}_+$. By the Schwarz reflection principle, $r(\overline{z}) = \overline{r(z)}$ for all $z \in \mathbb{C}$, and thus the poles of r_- must be the conjugates of the poles of r_+. Symmetry further implies that

$$r_-(z) = \overline{r_+(\overline{z})} \quad \forall z \in \mathbb{C}, \qquad r(x) = \operatorname{Re} r_+(x) \quad \forall x \in \mathbb{R}, \tag{6.1}$$

assuming that the constant $r(\infty)$, if it is nonzero, is split equally between r_- and r_+. Thus $\operatorname{Re} r_+(z)$ is a bounded harmonic function in $\mathbb{C}_+$ with $\|\operatorname{Re} r_+ - h\|_{\mathbb{R}} \le \varepsilon$, hence also $\|\operatorname{Re} r_+ - u\|_{\mathbb{C}_+} \le \varepsilon$ by the maximum principle. Moreover, the poles of r_+ are exactly the poles of r in $\mathbb{C}_-$. Conversely, given r_+ as indicated in the second assertion, the function $r(z) = (r_+(z) + \overline{r_+(\overline{z})})/2$ has the required properties. ■

The other half of conclusion (1) concerns the case of the open unit disk Δ. Let S denote the unit circle and Δ_- the complement of $\overline{\Delta}$ in $\mathbb{C} \cup \{\infty\}$. We get essentially the same theorem as before.

Theorem 6.2. *Given a real continuous function h on S, let u be the harmonic function in Δ with $u(x) = h(x)$ for $x \in S$. Suppose there exists a rational function r, also real on S, such that $\|r - h\|_S \le \varepsilon$ for some $\varepsilon \ge 0$. Then there exists a rational function r_+ whose poles are precisely the poles of r in Δ_- such that $\|\operatorname{Re} r_+ - h\|_S \le \varepsilon$ and thus also $\|\operatorname{Re} r_+ - u\|_\Delta \le \varepsilon$. Conversely, if r_+ is a rational function analytic in Δ such that $\|\operatorname{Re} r_+ - u\|_\Delta \le \varepsilon$, then there exists a rational function r whose poles are the poles of r_+ and their reflections in Δ such that $\|r - h\|_S \le \varepsilon$.*

Proof. One can argue as before or, alternatively, derive this as a corollary of Theorem 6.1 by a Möbius transformation. ■

We now turn to conclusion (2), concerning the case where Ω has corners. As mentioned, in the worst case rational approximation may give "too many poles," meaning poles that are not needed for approximation of the solution of the Laplace problem, and it may give "not enough poles," meaning poles that are inadequate to approximate the solution of the Laplace problem. To explain this, we present a pair of examples in Figure 8, both showing poles of AAA approximations with tolerance 10^{-8} on the boundary of the bounded symmetric "lens" domain Ω bounded by two circular arcs meeting at right angles at $z = \pm 1$.

The first image illustrates "too many poles." When the function $h(z) = \operatorname{Re} z$ is approximated by a rational function on $\partial\Omega$, many poles appear both inside and outside Ω; this will be the rule almost always when a region has corners. And yet this boundary data can be exactly matched by the harmonic function $u(z) = \operatorname{Re} z$, which has just a single pole at ∞. So the clusters of poles obtained by AAA are unnecessary for the Laplace problem in the interior of Ω.

The second image illustrates "too few poles." Here h is taken as the values on $\partial\Omega$ of the analytic function f that maps the exterior of Ω conformally to the exterior of the slit $[-1, 1]$ while leaving the points ±1 and ∞ fixed:

$$f(z) = \frac{1+v^2}{1-v^2}, \quad v = -\left(\frac{z-1}{z+1}\right)^{2/3}. \tag{6.2}$$

With the standard branch of the $2/3$ power, f has a branch cut along $[-1, 1]$, and AAA finds a rational approximation r whose poles lie approximately on this slit. In particular, they all lie within Ω apart from one pole of magnitude 10^{10}, approximating the pair $f(\infty) = \infty$. Thus there are no poles near $\partial\Omega$ for the AAA-LS method to work with in approximating the solution in the interior of Ω, yet this solution has singularities at ± 1 involving fractional powers $(z \pm 1)^{4/3}$, so it would need such poles to get high accuracy.

Thus we see that on domains with corners, failure of the pole symmetry principle is possible. However, the failures we have identified are atypical, at least in these extreme forms. The example on the left in Figure 8 is special in that despite the corners in the domain, the solution to the Laplace problem has no singularities thanks to special boundary data. This is hardly the generic situation (though picking such examples is a common mistake beginners make when testing their Laplace codes!). As for the example on the right, it has the unusual property of involving data h that can be analytically continued to all of $\mathbb{C} \cup \{\infty\} \backslash \overline{\Omega}$. This is another very special situation. Generically, a function h on a domain boundary with corners will only be analytically continuable with branch cuts on both sides, and rational approximations will need to have poles approximating those branch cuts on both sides of the domain. Configurations like that of the second image of Figure 8 are unlikely to appear in applications.

Now we turn to conclusion (3). Suppose Ω is a simply-connected domain bounded by an analytic curve that is not simply a circle or a straight line. For such a problem, Schwarz reflection no longer gives a symmetry equivalence between Ω and $\mathbb{C} \cup \{\infty\} \backslash \overline{\Omega}$. What happens to the pole symmetry principle?

The "pure mathematics answer" is that everything works essentially as before, modified only by the need for a fast exponentially-convergent polynomial term to be added into the rational approximations. The reasoning here can be based on the technique of considering a conformal map $w = \phi(z)$ of $\mathbb{C} \cup \{\infty\} \backslash \overline{\Omega}$ to $\mathbb{C} \cup \{\infty\} \backslash \overline{\Delta}$ with $\phi(\infty) = \infty$ and its inverse map $z = \psi(w)$ [12]. If $\partial\Omega$ is analytic, then ϕ and ψ extend analytically to larger domains, implying that they can be approximated by polynomials in z^{-1} and w^{-1}, respectively, with exponential convergence. It follows that rational approximation of a function h defined on $\partial\Omega$, for example, is equivalent to rational approximation of its transplant $\tilde{h}(w) = h(\psi(w))$ on S, up to exponentially convergent polynomial terms. If h has singularities, then root-exponential convergence of rational approximations in z is ensured by the same property for rational approximation of $\tilde{h}$ in w. By this kind of reasoning, one can argue that AAA-LS in a smooth domain is like AAA-LS in a disk, up to constants associated with polynomial approximations.

The "applied mathematics answer" is not so simple. All across complex analysis, the constants that appear in estimates of interest tend to grow exponentially as functions of geometric parameters such as the aspect ratios of reentrant or salient fingers in boundary curves, and this applies here. So the practical status of the pole symmetry principle for regions with curved boundaries may not be so different from that for regions with corners.

All the discussion above pertains to the global variant of AAA-LS. For local variants, as illustrated in the discussion around the multiply-connected domain of Figure 3, failures of the algorithm are more likely to appear in practice if the AAA step of the algorithm is applied with the data h given. In such cases, we recommend the method used in that figure: replace the actual boundary data h by a function $\hat{h}$ targeted to generate singularities at each corner, such as the product of the square roots of the distances to the corners. Our experience shows that as a practical matter, this strategy is highly effective. The reason for this is that, though not all singularities look alike, a wide range of them can be approximated with root-exponential convergence by exponentially clustered poles, whose configurations need not be tuned to the singularities [16, 32]. So the set of poles utilized by AAA to approximate one function will generally also do well for another.

In the case of a multiply-connected domain, to turn to point (4) of our summary, one should always use a local variant of the AAA-LS method. The reason is that approximating harmonic functions in such a domain will require logarithmic terms since their conjugates are in general multivalued [2]. One can use AAA to approximate a real function h on the boundary $\partial\Omega$ of such a domain by a rational function r, but r will not have the right properties interior to Ω. As illustrated in Figure 9, typically it will approximate different analytic functions near the different boundary components, separated by strings of poles approximating branch cuts (compare Figure 6.9 of [23]). These poles have nothing to do with the harmonic function u in Ω one wants to approximate, so in such a case global rational approximations should not be used.

In discussing local rational approximations above, we alluded to a kind of approximate university of pole distributions for resolving singularities. This suggests that in the end, AAA approximation should not really be necessary; one could equally well use a "lightning" strategy in which poles are positioned a priori rather than determined from the data. Indeed we think this is likely to be the case for problems dominated by singular corners, though the great convenience of starting from AAA approximations remains an advantage. For problems less controlled by corners, global or partially-global variants of AAA-LS will have a power not easily matched by lightning solvers.

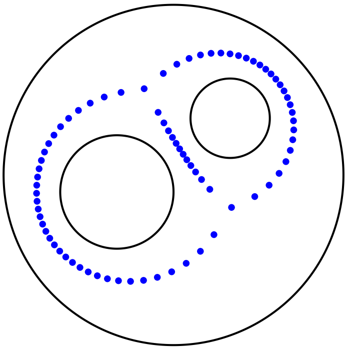


Figure 9. Poles of a global AAA rational approximant r with tolerance 10^{-8} on the boundary of a triply-connected domain with boundary data 0, 1, and 2 on the smaller, larger, and outer circles, respectively. The function r matches the data accurately on all three parts of $\partial\Omega$, but achieves this only by introducing strings of poles that effectively split Ω into subdomains with separate analytic functions. Here, these are the constant functions 0, 1, and 2, though the configuration would be much the same for any analytic boundary data. Effective approximation by a single harmonic function throughout Ω would require an additional logarithmic term in each hole, so for Laplace problems in domains like this, a local rather than global variant of AAA-LS should be used.

7. Discussion

AAA-LS offers a remarkably fast and accurate way to solve Laplace problems in planar domains with corners. Typical examples give 8-digit accuracy in a fraction of a second, and the resulting representation of the solution as the real part of a rational function can be evaluated in microseconds per point. Not just the harmonic function but also its harmonic conjugate are obtained, thereby giving the analytic extension of the solution in the problem domain as well as the solution itself—the Hilbert transform or Dirichlet-to-Neumann map. For domains with holes, this analytic extension is a multivalued analytic function, which consists of a single-valued function plus multivalued log terms, one for each hole [2].

A feature of all these expansion-based methods is that the representations of the solution they compute are numerically nonunique and, a fortiori, nonoptimal. The matrices involved have enormous or infinite condition numbers, and the coefficient vectors they deliver may depend in unpredictable ways on details of boundary discretization and other parameters. If we solve a Laplace problem and obtain 8-digit accuracy with 112 poles, for example, it must not be supposed that these poles are in truly optimal locations or that 112 is the precise minimal number for this accuracy. Despite that, the 8 digits are solid, as can be verified a posteriori by applying the maximum principle on a finer boundary grid, and they are achieved thanks to the regularizing effects of least-squares solvers as realized in the MATLAB backslash command.

Some other methods for computing rational approximations, such as vector fitting [18], IRKA [17], RKFIT [5], IRF [21], AGH [1], and the Haut–Beylkin–Monzón reduction algorithm [19], have optimality as a more central part of their design concept than AAA-LS, though they too will often terminate before optimality is achieved. As a rule, one cannot count on achieving optimality in rational approximation problems, in view of their extreme sensitivities, which are reflected both theoretically and computationally in longstanding complications of spurious poles or "Froissart doublets." For example, it is well known that Padé approximants, which are defined by optimality in approximating a function and its derivatives at a single point, do not in general converge to the function being approximated [4, 13].

Continuing on the matter of optimality in rational approximation, we offer an analogy from the field of matrix iterations for large linear systems of equations $Ax = b$, the core problem of computational science. (Actually it is more than an analogy, since matrix iterations are closely connected with rational approximations.) In theory, one might seek to generate an approximation to the solution vector x at each step of iteration that was truly optimal by some criterion. In a sense this is what certain forms of pure Lanczos or biconjugate gradient iterations do. However, it is well known that such an attempt brings risks of breakdowns and near-breakdowns that interfere with performance [11]. In practice, iterative methods aim for speed rather than optimality, and the idea of trying to solve $Ax = b$ to a certain accuracy in exactly the minimal number of steps is not part of the discussion.

In the past few years about a dozen papers have appeared related to AAA and lightning solution of Laplace problems via rational approximation and its variants; an impressive example we have not mentioned is [3], and an important earlier work is [21]. Most of the methods proposed in these works approximate continuous boundaries by discrete sets, typically with thousands of clustered points, and it is an interesting question to what extent such discretization is necessary. Even if the least-squares problem ultimately solved will involve a matrix with discrete rows, one may wonder whether the discretization can be deferred or hidden away in "continuous-mode" AAA or AAA-LS methods, as is done by the MATLAB code `laplace` [30] and in Chebfun codes such as `minimax`. This is one of many areas in which AAA and lightning methods, which are very young, can be expected to improve with further investigation in the years ahead. We are also exploring speedups to the linear algebra, and the possibility of "log-lightning" AAA-LS approximation as in [25].

Appendix: Sample code

As templates for further explorations, Figures 10 and 11 list the MATLAB codes used to generate the second row of Figure 1.

```
%% Set up
s = tanh(linspace(-12,12,300)');
Z = [1+s; 2+.5i+.5i*s; 1.5+1i+.5*s; 1+1.5i+.5i*s; .5+2i+.5*s; 1i+1i*s];
w = [0 2 2+1i 1+1i 1+2i 2i].';
h = @(z) real(z).^2; H = h(Z);
LW = 'linewidth'; MS = 'markersize'; ms = 6; PO = 'position'; FS = 'fontsize';

%% Local AAA fits
axes(PO,[.02 .6 .35 .35])
inpolygonc = @(z,w) inpolygon(real(z),imag(z),real(w),imag(w));
tol = 1e-8; pol_in = []; pol_out = [];
for k = 1:6
   ii = find(abs(Z-w(k)) == min(abs(Z-w.'),[],2));
   [~,polk] = aaa(H(ii),Z(ii),'tol',tol,'cleanup',0);
   polk_in = polk(inpolygonc(polk,w)); pol_in = [pol_in; polk_in];
   polk_out = polk(~inpolygonc(polk,w)); pol_out = [pol_out; polk_out];
end
plot(w([1:end 1]),'k',LW,.9), axis([-.8 2.8 -.8 2.8]), axis square, hold on
plot(pol_out,'.r',MS,ms), plot(pol_in,'.b',MS,ms), hold off, set(gca,'ytick',0:2)
title('local AAA poles'), set(gca,FS,6)

%% Solution
pol = pol_out.';
d = min(abs(w-pol),[],1);
[Hes,P] = VAorthog(Z,20); Q = d./(Z-pol);
A = [real(P) real(Q) -imag(P) -imag(Q)];
c = reshape(A\H,[],2)*[1; 1i];
F = [P Q]*c; U = real(F);
f = @(z) reshape([VAeval(z(:),Hes) d./(z(:)-pol)]*c,size(z));
u = @(z) real(f(z));

%% Contour and error plots
axes(PO,[.35 .6 .35 .35])
plot(pol,'.r',MS,ms), hold on
x = linspace(0,2,150); [xx,yy] = meshgrid(x,x); zz = xx+1i*yy;
uu = u(zz); uu(~inpolygonc(zz,w)) = NaN;
plot(w([1:end 1]),'k',LW,.9), axis([-.8 2.8 -.8 2.8]), axis square
contour(x,x,uu,20,LW,1), hold off, set(gca,'ytick',0:2)
u99err = u(.99+.99i) - 1.0267919261073
title('Laplace solution'), set(gca,FS,6)
axes(PO,[.73 .6 .25 .35])
semilogy(angle(Z-(.5+.5i)),abs(U-H),'.k',MS,3), grid on
set(gca,FS,6), axis([-pi pi 1e-12 1e-4])
set(gca,'xtick',pi*(-1:.5:1),'xticklabel',{'-\pi','-\pi/2','0','\pi/2','\pi'})
title('error against angle'), set(gca,FS,6)
```

Figure 10. MATLAB code to generate the second row of Figure 1.

```
function [Hes,R] = VAorthog(Z,n,varargin)  % Vand.+Arnoldi orthogonalization
%  Input:   Z = column vector of sample points
%           n = degree of polynomial (>= 0)
%         Pol = cell array of vectors of poles (optional)
% Output: Hes = cell array of Hessenberg matrices (length 1+length(Pol))
%           R = matrix of basis vectors
M = length(Z); Pol = []; if nargin == 3, Pol = varargin{1}; end
% First orthogonalize the polynomial part
Q = ones(M,1); H = zeros(n+1,n);
for k = 1:n
   q = Z.*Q(:,k);
   for j = 1:k, H(j,k) = Q(:,j)'*q/M; q = q - H(j,k)*Q(:,j); end
   H(k+1,k) = norm(q)/sqrt(M); Q(:,k+1) = q/H(k+1,k);
end
Hes{1} = H; R = Q;
% Next orthogonalize the pole parts, if any
while ~isempty(Pol)
   pol = Pol{1}; Pol(1) = [];
   np = length(pol); H = zeros(np,np-1); Q = ones(M,1);
   for k = 1:np
      q = Q(:,k)./(Z-pol(k));
      for j = 1:k, H(j,k) = Q(:,j)'*q/M; q = q - H(j,k)*Q(:,j); end
      H(k+1,k) = norm(q)/sqrt(M); Q(:,k+1) = q/H(k+1,k);
   end
   Hes{length(Hes)+1} = H; R = [R Q(:,2:end)];
end

function [R0,R1] = VAeval(Z,Hes,varargin)  % Vand.+Arnoldi basis construction
%  Input:   Z = column vector of sample points
%         Hes = cell array of Hessenberg matrices
%         Pol = cell array of vectors of poles, if any
% Output:  R0 = matrix of basis vectors for functions
%          R1 = matrix of basis vectors for derivatives
M = length(Z); Pol = []; if nargin == 3, Pol = varargin{1}; end
% First construct the polynomial part of the basis
H = Hes{1}; Hes(1) = []; n = size(H,2);
Q = ones(M,1); D = zeros(M,1);
for k = 1:n
   hkk = H(k+1,k);
   Q(:,k+1) = ( Z.*Q(:,k) - Q(:,1:k)*H(1:k,k)           )/hkk;
   D(:,k+1) = ( Z.*D(:,k) - D(:,1:k)*H(1:k,k) + Q(:,k) )/hkk;
end
R0 = Q; R1 = D;
% Next construct the pole parts of the basis, if any
while ~isempty(Pol)
   pol = Pol{1}; Pol(1) = [];
   H = Hes{1}; Hes(1) = []; np = length(pol); Q = ones(M,1); D = zeros(M,1);
   for k = 1:np
      Zpki = 1./(Z-pol(k)); hkk = H(k+1,k);
      Q(:,k+1) = ( Q(:,k).*Zpki - Q(:,1:k)*H(1:k,k)                   )/hkk;
      D(:,k+1) = ( D(:,k).*Zpki - D(:,1:k)*H(1:k,k) - Q(:,k).*Zpki.^2 )/hkk;
   end
   R0 = [R0 Q(:,2:end)]; R1 = [R1 D(:,2:end)];
end
```

Figure 11. Codes for Vandermonde with Arnoldi orthogonalization and evaluation, from [7].

Acknowledgments. We are grateful for advice and collaboration over the past few years from Peter Baddoo, Pablo Brubeck, Abi Gopal, Yuji Nakatsukasa, Kirill Serkh, André Weideman, and Heather Wilber.

References

[1] B. Alpert, L. Greengard, and T. Hagstrom, Rapid evaluation of nonreflecting boundary kernels for time-domain wave propagation. *SIAM J. Numer. Anal.* **37** (2000), no. 4, 1138–1164 Zbl 0963.65104 MR 1756419

[2] S. Axler, Harmonic functions from a complex analysis viewpoint. *Amer. Math. Monthly* **93** (1986), no. 4, 246–258 Zbl 0604.31001 MR 835293

[3] P. J. Baddoo, Lightning solvers for potential flows. *Fluids* **5** (2020), 227

[4] G. A. Baker Jr. and P. Graves-Morris, *Padé Approximants*. 2nd edn., Encyclopedia Math. Appl. 59, Cambridge University Press, Cambridge, 1996 Zbl 0923.41001 MR 1383091

[5] M. Berljafa and S. Güttel, The RKFIT algorithm for nonlinear rational approximation. *SIAM J. Sci. Comput.* **39** (2017), no. 5, A2049–A2071 Zbl 1373.65037 MR 3702872

[6] P. D. Brubeck, Y. Nakatsukasa, and L. N. Trefethen, Vandermonde with Arnoldi. *SIAM Rev.* **63** (2021), no. 2, 405–415 Zbl 07357074 MR 4253796

[7] P. D. Brubeck and L. N. Trefethen, Lightning Stokes solver. *SIAM J. Sci. Comput.* **44** (2022), no. 3, A1205–A1226 Zbl 07537259 MR 4418039

[8] S. Costa, Solving Laplace problems with the AAA algorithm. 2020, arXiv:2001.09439v1

[9] S. Costa and E. Costamagna, An alternative method for field analysis in inhomogeneous domains. *COMPEL* **40** (2021), no. 2, 223–237

[10] T. A. Driscoll, N. Hale, and L. N. Trefethen, *Chebfun Guide*. Pafnuty Press, Oxford, 2014

[11] R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal, An implementation of the look-ahead Lanczos algorithm for non-Hermitian matrices. *SIAM J. Sci. Comput.* **14** (1993), no. 1, 137–158 Zbl 0770.65022 MR 1201315

[12] D. Gaier, *Lectures on Complex Approximation. Translated from the German by Renate Mclaughlin*. Birkhäuser, Boston, MA, 1987 Zbl 0612.30003 MR 894920

[13] P. Gonnet, S. Güttel, and L. N. Trefethen, Robust Padé approximation via SVD. *SIAM Rev.* **55** (2013), no. 1, 101–117 Zbl 1266.41009 MR 3089442

[14] A. Gopal and L. N. Trefethen, New Laplace and Helmholtz solvers. *Proc. Natl. Acad. Sci. USA* **116** (2019), no. 21, 10223–10225 Zbl 1431.65224 MR 3956366

[15] A. Gopal and L. N. Trefethen, Representation of conformal maps by rational functions. *Numer. Math.* **142** (2019), no. 2, 359–382 Zbl 1414.30011 MR 3941934

[16] A. Gopal and L. N. Trefethen, Solving Laplace problems with corner singularities via rational functions. *SIAM J. Numer. Anal.* **57** (2019), no. 5, 2074–2094 Zbl 1431.65223 MR 4000218

[17] S. Gugercin, A. C. Antoulas, and C. Beattie, $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30** (2008), no. 2, 609–638 Zbl 1159.93318 MR 2421462

[18] B. Gustavsen and A. Semlyen, Rational approximation of frequency domain responses by vector fitting. *IEEE Trans. Power Deliv.* **14** (1999), 1052–1061

[19] T. Haut, G. Beylkin, and L. Monzón, Solving Burgers' equation using optimal rational approximations. *Appl. Comput. Harmon. Anal.* **34** (2013), 83–95 Zbl 1255.65183 MR 2981334

[20] P. Henrici, *Applied and Computational Complex Analysis. Vol. 3. Discrete Fourier Analysis—Cauchy Integrals—Construction of Conformal Maps—Univalent Functions*. Pure Appl. Math. (N.Y.), John Wiley & Sons, New York, 1986 Zbl 0578.30001 MR 822470

[21] A. Hochman, Y. Leviatan, and J. K. White, On the use of rational-function fitting methods for the solution of 2D Laplace boundary-value problems. *J. Comput. Phys.* **238** (2013), 337–358 MR 3028359

[22] Y. Mo, T. Qian, W. Mai, and Q. Chen, The AFD methods to compute Hilbert transform. *Appl. Math. Lett.* **45** (2015), 18–24 Zbl 1325.42002 MR 3316955

[23] Y. Nakatsukasa, O. Sète, and L. N. Trefethen, The AAA algorithm for rational approximation. *SIAM J. Sci. Comput.* **40** (2018), no. 3, A1494–A1522 Zbl 1390.41015 MR 3805855

[24] Y. Nakatsukasa and L. N. Trefethen, An algorithm for real and complex rational minimax approximation. *SIAM J. Sci. Comput.* **42** (2020), no. 5, A3157–A3179 Zbl 1452.65035 MR 4161312

[25] Y. Nakatsukasa and L. N. Trefethen, Reciprocal-log approximation and planar PDE solvers. *SIAM J. Numer. Anal.* **59** (2021), no. 6, 2801–2822 Zbl 07423174 MR 4332971

[26] D. J. Newman, Rational approximation to $|x|$. *Michigan Math. J.* **11** (1964), 11–14 Zbl 0138.04402 MR 171113

[27] V. Y. Protasov, Approximations by simple partial fractions and the Hilbert transform. *Izv. Ross. Akad. Nauk Ser. Mat.* **73** (2009), no. 2, 123–140 Zbl 1178.41010 MR 2532449

[28] L. N. Trefethen, 8-digit Laplace solutions on polygons? Posting on NA Digest. 2018, http://www.netlib.org/na-digest-html

[29] L. N. Trefethen, Series solution of Laplace problems. *ANZIAM J.* **60** (2018), no. 1, 1–26 Zbl 1400.31001 MR 3853349

[30] L. N. Trefethen, Lightning Laplace code `laplace.m`. 2020, https://people.maths.ox.ac.uk/trefethen/laplace.m

[31] L. N. Trefethen, Numerical conformal mapping with rational functions. *Comput. Methods Funct. Theory* **20** (2020), no. 3-4, 369–387 Zbl 1461.30031 MR 4175490

[32] L. N. Trefethen, Y. Nakatsukasa, and J. A. C. Weideman, Exponential node clustering at singularities for rational approximation, quadrature, and PDEs. *Numer. Math.* **147** (2021), no. 1, 227–254 Zbl 1467.65021 MR 4207522

[33] J. A. C. Weideman, Computing the Hilbert transform on the real line. *Math. Comp.* **64** (1995), no. 210, 745–762 Zbl 0830.65127 MR 1277773

[34] H. Wilber, A. Damle, and A. Townsend, Data-driven algorithms for signal processing with rational functions. 2021, arXiv:2105.07324v1

Stefano Costa
IEEE, Piacenza, Italy; stefano.costa@ieee.org

Lloyd N. Trefethen
Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK;
trefethen@maths.ox.ac.uk

EMS prize lectures

DOI 10.4171/8ECM/36

Smooth compactifications in derived non-commutative geometry

Alexander I. Efimov

Abstract. This is a short overview of the author's results related to the notion of a smooth categorical compactification. We cover the construction of a categorical smooth compactification of the derived categories of coherent sheaves, using the categorical resolution of Kuznetsov and Lunts. We also mention examples of homotopically finitely presented DG categories which do not admit a smooth compactification. This is closely related to Kontsevich's conjectures on the generalized versions of categorical Hodge-to-de Rham degeneration, which we disproved. Finally, we mention our new result on the DG categorical analogue of Wall's finiteness obstruction, which in particular gives a criterion for existence of a smooth compactification of a homotopically finite DG category.

1. Introduction

We give a short overview of some of our results concerning smooth compactifications of differential graded categories [8–10].

Suppose that $X \subset \bar{X}$ is a smooth compactification, i.e., X is open in $\bar{X}$ and $\bar{X}$ is smooth and proper over a base field k. Then the restriction functor

$$D^b_{\mathrm{coh}}(\bar{X}) \to D^b_{\mathrm{coh}}(X)$$

is a *localization*. Namely, the induced functor

$$D^b_{\mathrm{coh}}(\bar{X})/D^b_{\mathrm{coh},\bar{X}-X}(\bar{X}) \to D^b_{\mathrm{coh}}(X)$$

is an equivalence of categories.

This motivates a general categorical notion of a smooth compactification. There are notions of smoothness and properness for DG categories, which are defined in terms of the diagonal bimodule. By definition, a categorical smooth compactification of a pre-triangulated DG category $\mathcal{A}$ is given by a smooth and proper pre-triangulated

2020 Mathematics Subject Classification. Primary 14F08; Secondary 18G35, 14E15.
Keywords. Derived categories, differential graded categories, homotopy finiteness, Verdier localization, resolution of singularities.

DG category $\mathcal{C}$, with a functor $\Phi : \mathcal{C} \to \mathcal{A}$, such that Φ is a localization up to direct summands, with an additional assumption that $\ker(\Phi)$ is generated by a single object (see Definition 3.5). Here being a localization means that the induced functor $\overline{\Phi} : \mathcal{C}/\ker(\Phi) \to \mathcal{A}$ is fully faithful, and it is essentially surjective up to direct summands.

Existence of a categorical smooth compactification of a DG category $\mathcal{A}$ automatically implies that $\mathcal{A}$ is smooth. Moreover, $\mathcal{A}$ is actually homotopically finitely presented (hfp); see Definition 3.3.

The following result has been proved in [9].

Theorem 1.1 ([9, Theorem 1.8, part (1)]). *Let X be a separated scheme of finite type over a field* k *of characteristic zero. Then there exists a categorical smooth compactification of the form $D^b_{\mathrm{coh}}(Y) \to D^b_{\mathrm{coh}}(X)$, where Y is smooth and proper.*

In [9], Theorem 1.1 was used to prove the homotopy finiteness for derived categories of coherent sheaves over a field of characteristic zero, confirming a conjecture of Kontsevich.

The construction of a smooth compactification in Theorem 1.1 uses the categorical resolution of singularities of Kuznetsov and Lunts [15], as well as Orlov's results on semi-orthogonal gluings of geometric DG categories [21].

The statement of Theorem 1.1 is conceptually very closely related with the following conjecture of Bondal and Orlov.

Conjecture 1.2 ([2]). *Let Y be a variety with rational singularities, and $f : X \to Y$ a resolution of singularities. Then the functor $\mathbf{R}f_* : D^b_{\mathrm{coh}}(X) \to D^b_{\mathrm{coh}}(Y)$ is a localization.*

The methods of the proof of Theorem 1.1 allow to prove Conjecture 1.2 in a certain class of cases.

Theorem 1.3 ([9, Theorem 1.10]). *Suppose that Y has rational singularities, $Z \subset Y$ is a closed smooth subscheme, and $X = Bl_Z Y$ is smooth, so that $f : X \to Y$ is a resolution of singularities. Denote by $T = f^{-1}(Z)$ the exceptional divisor, by $p : T \to Z$ the induced morphism, and by $j : T \to X$ the embedding. Suppose that $\mathbf{R}f_* I_T^n = I_Z^n$ for $n \geq 1$. Then the functor $\mathbf{R}f_* : D^b_{\mathrm{coh}}(X) \to D^b_{\mathrm{coh}}(Y)$ is a localization, and its kernel is generated by $j_*((p^* D^b_{\mathrm{coh}}(Z))^{\perp})$.*

In particular, Theorem 1.3 applies in the case when Y is a cone over a projectively normal embedding of a smooth Fano variety, and Z is the origin.

The following question for general homotopically finite DG categories was formulated by Toën.

Question 1.4. Is it true that any homotopically finite DG category admits a categorical smooth compactification?

It turns out surprisingly that the answer is “no”, and a counterexample has been obtained in [8]. Question 1.4 is closely related with two (unpublished) conjectures of Kontsevich on the generalized versions of Hodge-to-de Rham degeneration, which we disproved in [8] (these are Conjectures 5.3 and 5.4).

One can further ask “what are the necessary and sufficient conditions for an hfp DG category to have a categorical smooth compactification?”. We have the following (new) result.

Theorem 1.5 ([10]). *Let $\mathcal{A}$ be an hfp pre-triangulated DG category. The following are equivalent.*

(1) *$\mathcal{A}$ admits a smooth categorical compactification.*

(2) *There exists a DG functor $\mathcal{C} \to \mathcal{A}$, where $\mathcal{C}$ is smooth and proper, such that*

$$[I_{\mathcal{A}}] \in \operatorname{Im}\big(K_0(\mathcal{C} \otimes \mathcal{C}^{\mathrm{op}}) \to K_0(\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}})\big).$$

Here $I_{\mathcal{A}}$ is the diagonal $\mathcal{A}$-$\mathcal{A}$-bimodule.

This theorem is closely related with a certain DG categorical analogue of Wall’s finiteness obstruction theorem; see Section 6.

The paper is organized as follows.

In Section 2, we briefly recall some basic notions and statements about triangulated categories and DG categories.

In Section 3, we discuss the general notion of a categorical smooth compactification.

In Section 4, we formulate our result on smooth compactifications of derived categories of coherent sheaves, and briefly explain the idea of the proof.

Section 5 discusses the question of existence of smooth compactifications, and the closely related Conjectures 5.3 and 5.4.

Finally, in Section 6 we briefly mention our new results on the DG categorical analogue of Wall’s finiteness obstruction theorem about finitely dominated spaces. This in particular gives a criterion for when a homotopically finite DG category has a smooth compactification.

2. Some preliminaries on triangulated categories and DG categories

For a very nice introduction to DG categories and their derived categories, we refer to [12]. For triangulated categories, we refer to Neeman’s book [19]. The notion of a DG enhancement of a triangulated category has been introduced in [3]. The notion of a DG quotient of DG categories has been introduced in [13] and an explicit construction has been given in [6]. For model structures on the categories of DG algebras and DG categories we refer to [22, 23].

Fix some base field k. For a quasi-projective scheme X over k, we have the category of finite rank vector bundles on X, or equivalently the category of locally free sheaves of finite rank. After adding to it the cokernels, we get the abelian category $\mathrm{Coh}(X)$ of coherent sheaves. More generally, the abelian category $\mathrm{Coh}(X)$ can be defined for any noetherian (or even locally coherent) scheme X. In this note, we deal only with separated schemes of finite type over k.

The objects of the derived category $D^b(\mathrm{Coh}(X)) = D^b_{\mathrm{coh}}(X)$ are bounded complexes of coherent sheaves. The morphisms are more complicated: they are obtained from the naive category of complexes by inverting the quasi-isomorphisms. A quasi-isomorphism is a morphism of complexes that induces an isomorphism in cohomology.

The derived category $D^b_{\mathrm{coh}}(X)$ is always triangulated. It has a full triangulated subcategory of perfect complexes $D_{\mathrm{perf}}(X) \subset D^b_{\mathrm{coh}}(X)$, which is formed by bounded complexes of locally free sheaves (that is, of vector bundles). More precisely, if X is not necessarily quasi-projective, an object $\mathcal{F} \in D^b_{\mathrm{coh}}(X)$ is a perfect complex if it is locally quasi-isomorphic to a bounded complex of locally free sheaves.

A DG category $\mathcal{A}$ is given by the following data:

- a class of objects $\mathrm{Ob}(\mathcal{A})$;
- for any pair of objects $x, y \in \mathrm{Ob}(\mathcal{A})$, a complex of vector spaces $\mathcal{A}(x, y) = \mathrm{Hom}_{\mathcal{A}}(x, y)$;
- for any objects $x, y, z \in \mathrm{Ob}(\mathcal{A})$, a composition map $\mathcal{A}(y, z) \otimes \mathcal{A}(x, y) \to \mathcal{A}(x, z)$.

The composition maps are required to be morphisms of complexes: they are homogeneous of degree zero and satisfy the (super-)Leibniz rule. They are also required to be associative. For each object $x \in \mathrm{Ob}(\mathcal{A})$, it is required that there is a unit morphism 1_x of degree zero (and automatically $d(1_x) = 0$).

The homotopy category of a DG category $\mathcal{A}$ is a k-linear category $H^0(\mathcal{A})$ which has the same objects as $\mathcal{A}$, and the morphisms are given by

$$H^0(\mathcal{A})(x, y) = H^0\big(\mathcal{A}(x, y)\big).$$

It is also convenient to define similarly the k-linear category $Z^0(\mathcal{A})$ with the same objects as $\mathcal{A}$, and with the morphisms given by $Z^0(\mathcal{A})(x, y) = Z^0(\mathcal{A}(x, y))$.

For a small DG category $\mathcal{A}$, just as for DG algebras, there is a notion of a right DG $\mathcal{A}$-module: it is a DG functor $\mathcal{A}^{\mathrm{op}} \to \text{Mod-k}$, where Mod-k is the DG category of complexes of vector spaces. DG $\mathcal{A}$-modules form a DG category Mod-$\mathcal{A}$. The derived category $D(\mathcal{A})$ of right $\mathcal{A}$-modules is obtained from $H^0(\text{Mod-}\mathcal{A})$ by inverting quasi-isomorphisms. Equivalently, $D(\mathcal{A})$ is obtained from $Z^0(\text{Mod-}\mathcal{A})$ by inverting quasi-isomorphisms. Again, as for DG algebras, $D_{\mathrm{perf}}(\mathcal{A}) \subset D(\mathcal{A})$ is the full subcategory of compact objects.

The Yoneda embedding $\mathcal{A} \hookrightarrow \text{Mod-}\mathcal{A}$ induces a fully faithful functor $H^0(\mathcal{A}) \hookrightarrow D(\mathcal{A})$. If its image is a triangulated subcategory of $D(\mathcal{A})$, then we call the DG category $\mathcal{A}$ *pre-triangulated*. In this case, we have $D_{\text{perf}}(\mathcal{A}) \simeq H^0(\mathcal{A})^{\text{Kar}}$ – Karoubi completion.

The basic example is the following: for a separated scheme X of finite type over k we take the DG category $\mathfrak{D}^b_{\text{coh}}(X)$ of bounded below complexes of injective quasi-coherent sheaves with bounded coherent cohomology. Then $\mathfrak{D}^b_{\text{coh}}(X)$ is pre-triangulated and $H^0(\mathfrak{D}^b_{\text{coh}}(X))$ is equivalent to $D^b_{\text{coh}}(X)$. We denote by $\text{Perf}(X) \subset \mathfrak{D}^b_{\text{coh}}(X)$ the full DG subcategory of perfect complexes.

If $\mathcal{T}$ is a small triangulated category and $\mathcal{S} \subset \mathcal{T}$ is a full triangulated subcategory, then there is a notion of a quotient category $\mathcal{T}/\mathcal{S}$, due to Verdier [26,27]. The category $\mathcal{T}/\mathcal{S}$ is again triangulated, and we have an exact quotient functor $\mathcal{T} \to \mathcal{T}/\mathcal{S}$. The category $\mathcal{T}/\mathcal{S}$ is obtained from $\mathcal{T}$ by inverting the morphisms $f : x \to y$ such that $\text{Cone}(f) \in \mathcal{S}$.

The basic example is coming from geometry: let X be as above, $Z \subseteq X$ a closed subscheme, and $U = X - Z$. Denote by $D^b_{\text{coh},Z}(X) \subseteq D^b_{\text{coh}}(X)$ the full subcategory of complexes whose cohomology is supported on Z. Then we have an equivalence

$$D^b_{\text{coh}}(X)/D^b_{\text{coh},Z}(X) \simeq D^b_{\text{coh}}(U);$$

see [20, Lemma 2.2].

There is a notion of a DG quotient $\mathcal{A}/\mathcal{B}$ of a small DG category $\mathcal{A}$ by a full DG subcategory $\mathcal{B} \subseteq \mathcal{A}$, which was first defined by Keller [13], and then an explicit construction has been given by Drinfeld [6]. The main property of the DG quotient is its compatibility with the Verdier quotient of triangulated categories. Namely, if $\mathcal{A}$ is a pre-triangulated small DG category, and $\mathcal{B} \subset \mathcal{A}$ is a full pre-triangulated DG subcategory, then we have an equivalence $H^0(\mathcal{A}/\mathcal{B}) \simeq H^0(\mathcal{A})/H^0(\mathcal{B})$.

In particular, within the above notation we have a quasi-equivalence

$$\mathfrak{D}^b_{\text{coh}}(X)/\mathfrak{D}^b_{\text{coh},Z}(X) \simeq \mathfrak{D}^b_{\text{coh}}(U).$$

3. Categorical smooth compactifications

The following notions of smoothness and properness for DG categories are due to Kontsevich.

A DG category $\mathcal{A}$ is called *proper* (over k) if for any $x, y \in \mathcal{A}$ the complex $\mathcal{A}(x, y)$ has finite-dimensional total cohomology, and the triangulated category $D_{\text{perf}}(\mathcal{A})$ is generated by a single object. Here and below we say that a triangulated category T is generated by an object x if T is the smallest *idempotent complete* triangulated subcategory of T containing x. Equivalently, any (isomorphism class of an) object of T can be obtained from x using cones and direct summands.

A DG category $\mathcal{A}$ is called *smooth* (over k) if the diagonal $\mathcal{A}$-$\mathcal{A}$-bimodule $I_{\mathcal{A}}$ is perfect over $\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}}$. Here $I_{\mathcal{A}}(x, y) = \mathcal{A}(x, y)$.

These properties are compatible with the corresponding properties of schemes. Namely, the following holds.

Proposition 3.1 ([21, Proposition 3.30] and [17, Proposition 3.13]). *If X is a separated scheme of finite type over* k*, then the DG category* $\mathrm{Perf}(X)$ *is smooth (resp. proper) if and only if X is smooth (resp. proper).*

Much more surprising is the following theorem of Lunts.

Theorem 3.2 ([17, Theorem 6.3]). *For any separated scheme X of finite type over a perfect field* k*, the DG category $\mathfrak{D}^b_{\mathrm{coh}}(X)$ is smooth.*

There is a notion of an hfp DG category. Before giving its formal definition, we mention that it is an analogue of the notion of a *finitely dominated* topological space. Recall that a (possibly infinite) CW complex X is called finitely dominated if there exists a finite CW complex Y and continuous maps $f : X \to Y$, $g : Y \to X$ such that $gf \sim \mathrm{id}_X$. Equivalently, the identity map id_X is homotopic to a map $r : X \to X$ such that the closure $\overline{r(X)}$ is compact.

Formal definition of hfp DG algebras and DG categories is the following.

Definition 3.3 ([25]). (1) A finite cell DG algebra B is a DG algebra which is isomorphic as a graded algebra to the free finitely generated associative algebra:

$$B^{gr} \cong \mathrm{k}\langle x_1, \dots, x_n\rangle,$$

and moreover we have

$$dx_i \in \mathrm{k}\langle x_1, \dots, x_{i-1}\rangle, \quad 1 \le i \le n. \tag{3.1}$$

(2) A DG algebra A is hfp if in the homotopy category $\mathrm{Ho}(dgalg_{\mathrm{k}})$ the object A is a retract of some finite cell DG algebra B.

(3) A DG category $\mathcal{A}$ is hfp if it is Morita equivalent to an hfp DG algebra.

Recall that in any category $\mathcal{C}$ an object X is a retract of Y iff there exists morphisms $f : X \to Y$, $g : Y \to X$ such that $gf = \mathrm{id}_X$.

Proposition 3.4 ([25]). *Let $\mathcal{A}$ be a small DG category over* k.

(1) *If $\mathcal{A}$ is hfp, then $\mathcal{A}$ is smooth.*

(2) *If $\mathcal{A}$ is smooth and proper, then $\mathcal{A}$ is hfp.*

Informally, an hfp DG category is a smooth DG category “given by a finite amount of data”. For example, the k-algebra of rational functions $\mathrm{k}(x)$ is smooth but not hfp.

An equivalent definition of an hfp DG category is the following. First, there is a notion of a finite cell DG category: as a k-linear graded category, it is a path category of a finite graded quiver with arrows $x_1, \dots, x_n$ such that the differential satisfied the condition analogous to (3.1). Now, a DG category $\mathcal{A}$ is hfp if it is a retract of a finite cell DG category in the Morita homotopy category of DG categories $\mathrm{Ho}_M(\mathrm{dgcat}_{\mathrm{k}})$ (which is obtained by inverting Morita equivalences).

Recall that a usual smooth compactification of a smooth algebraic variety X is given by a smooth and proper variety $\bar{X}$ and an open embedding $X \hookrightarrow \bar{X}$. Denote by $Z = \bar{X} - X$ the infinity locus. As already mentioned in the previous section, we have an equivalence $D^b_{\mathrm{coh}}(X) \simeq D^b_{\mathrm{coh}}(\bar{X})/D^b_{\mathrm{coh},Z}(\bar{X})$. Hence, we also have a quasi-equivalence of DG categories

$$\mathfrak{D}^b_{\mathrm{coh}}(X) \simeq \mathfrak{D}^b_{\mathrm{coh}}(\bar{X})/\mathfrak{D}^b_{\mathrm{coh},Z}(\bar{X}).$$

This motivates a general definition of a categorical smooth compactification, which we already mentioned in the introduction.

Definition 3.5. A smooth categorical compactification of a DG category $\mathcal{A}$ is a DG functor $F : \mathcal{C} \to \mathcal{A}$, where the DG category $\mathcal{C}$ is smooth and proper, the extension of scalars functor $F : \mathrm{Perf}(\mathcal{C}) \to \mathrm{Perf}(\mathcal{A})$ is a localization (up to direct summands), and its kernel is generated by a single object.

We have the following implication, which is quite easy to prove.

Proposition 3.6 ([9, Corollary 2.9]). *If a DG category $\mathcal{A}$ has a smooth categorical compactification, then it is hfp.*

4. Smooth compactifications of derived categories of coherent sheaves

We have the following general result.

Theorem 4.1 ([9, Theorem 1.8]). *For any separated scheme X of finite type over a field* k *of characteristic zero, there exists a smooth projective variety Y and a quasi-equivalence $\mathfrak{D}^b_{\mathrm{coh}}(Y)/\mathcal{S} \simeq \mathfrak{D}^b_{\mathrm{coh}}(X)$, where $\mathcal{S} \subset \mathfrak{D}^b_{\mathrm{coh}}(Y)$ is a pre-triangulated subcategory generated by a single object. In particular, the DG category $\mathfrak{D}^b_{\mathrm{coh}}(Y)$ is hfp.*

This confirms a conjecture of Kontsevich on the homotopy finiteness of the DG category $\mathfrak{D}^b_{\mathrm{coh}}(X)$.

Remark. A similar result is expected to hold over any perfect field. In our proof, we cannot get rid of the characteristic zero assumption: we use the categorical resolution of singularities of Kuznetsov and Lunts, which in turn uses the classical Hironaka's theorem.

We now explain the idea of the proof of Theorem 4.1. It is based on the following constructions.

The first one is the categorical resolution of singularities due to Kuznetsov and Lunts [15]. Let us restrict to proper schemes. For any proper scheme X over k, they construct a smooth and proper DG category $\mathscr{C}$ together with a fully faithful functor $\mathrm{Perf}(X) \hookrightarrow \mathscr{C}$. Moreover, this DG category $\mathscr{C}$ has a semi-orthogonal decomposition into derived categories of some smooth and proper varieties:

$$\mathscr{C} = \langle \mathfrak{D}^b_{\mathrm{coh}}(Y_1), \dots, \mathfrak{D}^b_{\mathrm{coh}}(Y_m) \rangle.$$

More precisely, one chooses a resolution $Z \to X_{\mathrm{red}}$ by a sequence of blow-ups with smooth centers. Then the varieties $Y_1, \dots, Y_m$ are exactly the centers of the blow-ups and the resolution Z (each of these varieties can appear in the list several times).

Another general construction due to Orlov [21] allows to embed such a semi-orthogonal gluing of $\mathfrak{D}^b_{\mathrm{coh}}(Y_i)$ into a single derived category $\mathfrak{D}^b_{\mathrm{coh}}(Y)$ (here Y_i and Y are smooth and proper). Taking such embedding $\mathscr{C} \hookrightarrow \mathfrak{D}^b_{\mathrm{coh}}(Y)$ (where $\mathscr{C}$ is as above), we obtain the fully faithful composition functor $\mathrm{Perf}(X) \hookrightarrow \mathscr{C} \hookrightarrow \mathfrak{D}^b_{\mathrm{coh}}(Y)$. Passing to large categories (i.e., categories of ind-objects), we can take a right adjoint to this embedding, which restricts to a functor $\Phi : \mathfrak{D}^b_{\mathrm{coh}}(Y) \to \mathfrak{D}^b_{\mathrm{coh}}(X)$. It turns out (but it is not easy to prove) that this functor is actually a desired localization functor promised by Theorem 4.1.

Remark. Strictly speaking, in [9] it is proved that the functor $\Phi : D^b_{\mathrm{coh}}(Y) \to D^b_{\mathrm{coh}}(X)$ is a localization under some assumptions on the choices of integer parameters in the construction of the category $\mathscr{C}$ in [15]. We do not discuss these details in the present note.

The construction of the categorical resolution from [15] uses two general methods to “partially resolve” the category $\mathrm{Perf}(X)$. The first one allows to deal with nilpotents in the structure sheaf $\mathscr{O}_X$. Namely, assuming that the reduced part $X_{\mathrm{red}} \subset X$ is smooth, one can find a categorical resolution by a certain ringed space $(X, \mathscr{A}_X)$, where $\mathscr{A}_X$ is a sheaf of associative algebras (and non-commutative unless $X = X_{\mathrm{red}}$). This ringed space is equipped with a morphism $(X, \mathscr{A}_X) \xrightarrow{\rho_X} X$, and the pullback functor

$$\rho_X^* : D_{\mathrm{perf}}(X) \to D_{\mathrm{perf}}(X, \mathscr{A}_X)$$

is fully faithful. It is not hard to show that the pushforward functor

$$\rho_{X*} : D^b_{\mathrm{coh}}(X, \mathscr{A}_X) \to D^b_{\mathrm{coh}}(X)$$

is a localization.

Remark. The ringed space $(X, \mathcal{A}_X)$ is given by a certain generalization of algebras considered by Auslander [1]. Namely, for a finite-dimensional algebra A of finite representation type, Auslander constructs an algebra $B = \operatorname{End}_A(\bigoplus_i M_i)$, where M_i are representatives of all isomorphism classes of indecomposable finite-dimensional A-modules. In fact, for $X = \operatorname{Spec} \mathrm{k}[x]/x^n$ we have

$$(X, \mathcal{A}_X) = \left(\mathrm{pt}, \operatorname{End}_{\mathrm{k}[x]/x^n}\left(\bigoplus_{i=1}^{n} \mathrm{k}[x]/x^i\right)\right).$$

Another more interesting construction involved in the categorical resolution is the "categorical blow-up". Without going into details, this is a certain categorical modification of the usual blow-up. Given any noetherian scheme X and a closed subscheme S, consider the blow-up $f : Y \to X$, i.e., $Y = \operatorname{Proj}_X(\bigoplus_{n\geq 0} I_S^n)$. Then under some assumptions on S (always achievable by replacing S with its sufficiently large infinitesimal neighborhood), one can define a certain semi-orthogonal gluing of $\mathfrak{D}^b_{\mathrm{coh}}(Y)$ and $\mathfrak{D}^b_{\mathrm{coh}}(S)$, denoted by $\mathcal{D}_{\mathrm{coh}}(Y, S)$, with a functor

$$\pi_* : \mathcal{D}_{\mathrm{coh}}(Y, S) \to \mathfrak{D}^b_{\mathrm{coh}}(X).$$

It is proved in [9] that under some additional assumptions on S (again they always hold after infinitesimally enlarging S) this functor π_* is a localization. This is the most difficult part of the proof of Theorem 4.1. Note that if we use $\mathfrak{D}^b_{\mathrm{coh}}(Y)$ instead of $\mathfrak{D}_{\mathrm{coh}}(Y, S)$, then

(1) the pushforward functor $\mathbf{R}f_* : \mathfrak{D}^b_{\mathrm{coh}}(Y) \to \mathfrak{D}^b_{\mathrm{coh}}(X)$ is usually not a localization, and a necessary condition is that $\mathbf{R}f_*(\mathcal{O}_Y) = \mathcal{O}_X$;

(2) if we assume that this condition is satisfied, we are not able in general to prove that $\mathbf{R}f_*$ is a localization (this is a generalization of Conjecture 1.2). So even in this case we use $\mathfrak{D}_{\mathrm{coh}}(Y, S)$ instead of $\mathfrak{D}^b_{\mathrm{coh}}(Y)$.

Using these localization statements as building blocks, the proof of Theorem 4.1 is obtained by induction of the number of blow-ups of smooth centers in the resolution process of X_{red}.

5. Existence of smooth compactifications

In this section, we assume that the characteristic of the base field k is zero.

Recall the question of Toën, mentioned in the introduction.

Question 5.1. Let $\mathcal{A}$ be a homotopically finite DG category. Does it admit a smooth compactification?

Quite surprisingly, the paper [8] gives a negative answer. We briefly explain the idea of a counterexample, and its close relation with generalized versions of the noncommutative Hodge-to-de Rham degeneration.

Recall that the classical Hodge theory implies that for any smooth and proper algebraic variety X over a field k of characteristic zero the spectral sequence

$$E_1^{p,q} = H^q_{Zar}(X, \Omega^p_X) \Rightarrow H^{p+q}_{DR}(X)$$

degenerates. Here the limit of the spectral sequence is the algebraic de Rham cohomology.

The following categorical generalization was conjectured by Kontsevich and Soibelman [14], and proved by Kaledin [11].

Theorem 5.2 ([11, Theorem 5.4]). *Let A be a smooth and proper DG algebra over a field of characteristic zero. Then the Hochschild-to-cyclic spectral sequence degenerates, so that we have an isomorphism $HP_\bullet(A) \cong HH_\bullet(A)((u))$.*

The following conjectures were formulated by Kontsevich (unpublished).

Conjecture 5.3 (Kontsevich). *Let A be a smooth DG algebra over a field of characteristic zero. Then the composition*

$$K_0(A \otimes A^{\mathrm{op}}) \xrightarrow{\mathrm{ch}} \big(HH_\bullet(A) \otimes HH_\bullet(A^{\mathrm{op}})\big)_0 \xrightarrow{\mathrm{id}\otimes\delta^-} \big(HH_\bullet(A) \otimes HC^-_\bullet(A^{\mathrm{op}})\big)_1$$

vanishes on the class $[A]$ of the diagonal bimodule.

Here $\delta^- : HH_\bullet(A^{\mathrm{op}}) \to HC^-_\bullet(A^{\mathrm{op}})[-1]$ denotes the boundary map in the long exact sequence

$$\cdots \to HC^-_{n+1}(A^{\mathrm{op}}) \to HC^-_{n-1}(A^{\mathrm{op}}) \to HH_{n-1}(A^{\mathrm{op}}) \xrightarrow{\delta^-} HC^-_n(A^{\mathrm{op}}) \to \cdots;$$

see for example [7, Section 3].

Conjecture 5.4 (Kontsevich). *Let B be a proper DG algebra over a field k of characteristic zero. Then the composition map*

$$\big(HH_\bullet(B) \otimes HC_\bullet(B^{\mathrm{op}})\big)[1] \xrightarrow{\mathrm{id}\otimes\delta^+} HH_\bullet(B) \otimes HH_\bullet(B^{\mathrm{op}}) \to \mathrm{k} \qquad (5.1)$$

is zero.

Here $\delta^+ : HC_\bullet(B^{\mathrm{op}})[1] \to HH_\bullet(B^{\mathrm{op}})$ denotes the boundary map in a similar long exact sequence

$$\cdots \to HH_{n+1}(B^{\mathrm{op}}) \to HC_{n+1}(B^{\mathrm{op}}) \to HC_{n-1}(B^{\mathrm{op}}) \xrightarrow{\delta^+} HH_n(B^{\mathrm{op}}) \to \cdots;$$

see [16, Section 2.2]. The second map in (5.1) is given by the composition

$$HH_\bullet(B) \otimes HH_\bullet(B^{\mathrm{op}}) \cong HH_\bullet(B \otimes B^{\mathrm{op}}) \to HH_\bullet\big(\mathrm{End}_{\mathrm{k}}(B)\big) \cong HH_\bullet(\mathrm{k}) = \mathrm{k},$$

where we used the Künneth isomorphism for HH, the diagonal bimodule structure on B, and the (derived) Morita equivalence between $\mathrm{End}_{\mathrm{k}}(B)$ and k.

Both Conjectures 5.3 and 5.4 had a strong motivation. Namely, in the case of smooth DG algebras, the following holds.

Proposition 5.5 ([8, Proposition 4.1]). *Let B be a smooth DG algebra and $F : \mathrm{Perf}(A) \to \mathrm{Perf}(B)$ a localization functor, where A is a smooth and proper DG algebra. Then Conjecture 5.3 holds for B.*

This is easy to prove (of course, assuming Kaledin's theorem (Theorem 5.2)). Namely, it almost immediately follows from the commutative diagram

$$\begin{tikzcd}
HH_\bullet(A) \otimes HH_\bullet(A^{\mathrm{op}}) \arrow[r, "\mathrm{id}\otimes\delta^-"] \arrow[d] & HH_\bullet(A) \otimes HC_\bullet^-(A^{\mathrm{op}})[-1] \arrow[d] \\
HH_\bullet(B) \otimes HH_\bullet(B^{\mathrm{op}}) \arrow[r, "\mathrm{id}\otimes\delta^-"] & HH_\bullet(B) \otimes HC_\bullet^-(B^{\mathrm{op}})[1],
\end{tikzcd}$$

and from the degeneration of the Hochschild-to-cyclic spectral sequence for A. We have the following corollary.

Corollary 5.6 ([8, Corollary 4.2]). *Let X be a separated scheme of finite type over* k, *and $\mathcal{G} \in D^b_{\mathrm{coh}}(X)$ – a generator. Then Conjecture 5.3 holds for the smooth DG algebra $A = \mathbf{R}\,\mathrm{End}(\mathcal{G})$.*

Indeed, this follows from Proposition 5.5 and from Theorem 1.1 (in fact, a weakened version of Theorem 1.1 is sufficient; see [8, Remark 4.3]).

Similar (dual) statements hold for proper DG algebras.

Proposition 5.7 ([8, Proposition 5.1]). *Let B be a proper DG algebra and $\mathrm{Perf}(B) \hookrightarrow \mathrm{Perf}(A)$ a fully-faithful functor, where A is a smooth and proper DG algebra. Then Conjecture 5.4 holds for B.*

Corollary 5.8 ([8, Corollary 5.2]). *Let X be a separated scheme of finite type over* k, *and $Z \subset X$ a closed proper subscheme. For any object $\mathcal{F} \in \mathrm{Perf}_Z(X)$, Conjecture 5.4 holds for the proper DG algebra $B = \mathbf{R}\,\mathrm{End}(\mathcal{F})$.*

However, we disproved both Conjectures 5.3 and 5.4. The counterexamples are provided by [8, Theorems 4.5 and 5.4]. The counterexample to Conjecture 5.3 is in fact hfp, hence by Proposition 3.6 it gives a negative answer to Question 1.4.

We briefly describe the counterexample to Conjecture 5.4. Recall that given DG algebras A and B, together with an A-B-bimodule M, we can form a gluing $C = \begin{pmatrix} B & 0 \\ M & A \end{pmatrix}$. This is a DG algebra which equals $A \oplus B \oplus M$ as a complex of vector spaces, and the multiplication is given by

$$(a, b, m) \cdot (a', b', m') = (aa', bb', am' + mb').$$

Let us take $A = \mathrm{k}[x]/x^6$ and $B = \mathrm{k}[y]/y^3$, where $\deg(x) = 0$, $\deg(y) = 1$, and $dx = 0$, $dy = 0$. Then one can show that there exists a DG A-B-bimodule M such that $H^\bullet(M) = \mathrm{k}[0]$ and the DG algebra $C = \left(\begin{smallmatrix} B & 0 \\ M & A \end{smallmatrix}\right)$ is a counterexample to Conjecture 5.4. Namely, the DG algebra C is proper (but not smooth), and its cohomology $H^\bullet(C)$ is 10-dimensional. Further, we have the elements $x \in H^0(C)$, $y \in H^1(C)$. Using natural maps $H^n(C) \to HH_{-n}(C) \to HC_{-n}(C)$ and similarly for C^{op}, we can consider x and y as classes in Hochschild and cyclic homology, respectively: $x \in HH_0(C)$, $y \in HC_{-1}(C^{\mathrm{op}})$. Now, a bimodule M is constructed in such a way that $\langle x, \delta^+(y)\rangle \neq 0$, disproving conjecture 5.4. For details see [8, Theorem 5.4].

6. Wall's finiteness obstruction for DG categories

Here we mention some new results, to appear in [10]. In particular, we formulate a criterion for a homotopically finite DG category to have a smooth compactification.

As we already mentioned, the notion of an hfp DG category is analogous to the notion of a finitely dominated CW complex.

In 1959, Milnor [18] asked if every finitely dominated CW complex X is homotopy equivalent to a finite CW complex. This was already known in the case when each connected component of X is simply connected, but it was considered to be a difficult problem in general.

For simplicity, let us assume that X is connected. In 1965, C. T. C. Wall defined an invariant $w(X) \in \widetilde{K_0}(\mathbb{Z}[\pi_1(X)])$ (an element of the reduced Grothendieck group of $\mathbb{Z}[\pi_1(X)]$) for any finitely dominated space X. Recall that for an associative unital ring A the group $K_0(A)$ is generated by isomorphism classes of finitely generated projective (right) A-modules $[P]$, subject to relations $[P \oplus Q] = [P] + [Q]$. If a ring A is equipped with a unital homomorphism $A \to \mathbb{Z}$ (i.e., A is augmented), its reduced Grothendieck group $\widetilde{K_0}(A)$ is defined to be the kernel $\ker(K_0(A) \to K_0(\mathbb{Z}) = \mathbb{Z})$. In fact, we have a decomposition $K_0(A) \cong \mathbb{Z} \oplus \widetilde{K_0}(A)$. Note that for any group G the group ring $\mathbb{Z}[G]$ is naturally augmented. Wall proved the following result.

Theorem 6.1 ([28, Theorem F]). *A connected finitely dominated space X has a homotopy type of a finite CW complex if and only if $w(X) = 0$.*

Probably the simplest description (and different from the original one) of the class $w(X)$ is the following. Recall that for a DG ring B the group $K_0(B)$ is defined to be the Grothendieck group $K_0(D_{\mathrm{perf}}(B))$. Here for a small triangulated category T the group $K_0(T)$ is generated by the isomorphism classes of objects $[X]$, $X \in T$, subject to relations $[Y] = [X] + [Z]$ for an exact triangle

$$X \to Y \to Z \to X[1]$$

in T. If the DG ring B is concentrated in degree zero, i.e., $B = H^0(B)$, then the two definitions of $K_0(B)$ agree. Moreover, if B is (cohomologically) non-positively graded, then $K_0(B) \cong K_0(H^0(B))$; see [4, Theorem 5.3.1 and Proposition 6.2.1].

Now choose a base point $x_0 \in X$. Consider the DG ring $C_\bullet(\Omega_{x_0}X)$ of singular chains on the based loop space. By the result of Brav–Dyckerhoff [5, Proposition 5.1], the DG ring $C_\bullet(\Omega_{x_0}X)$ is smooth over $\mathbb{Z}$ (and moreover it is hfp). It follows that the augmentation module $\mathbb{Z}$ is perfect: $\mathbb{Z} \in \operatorname{Perf}(C_\bullet(\Omega_{x_0}X))$. Any perfect module defines a class in K_0, hence we have a well-defined class

$$\widetilde{w}(X) := [\mathbb{Z}] \in K_0\big(C_\bullet(\Omega_{x_0}X)\big) \cong K_0\big(\mathbb{Z}\big[\pi_1(X, x_0)\big]\big),$$

since $H_0(\Omega_{x_0}X) \cong \mathbb{Z}[\pi_1(X, x_0)]$. The class $w(X) \in \widetilde{K_0}(\mathbb{Z}[\pi_1(X, x_0)])$ is simply the projection of $\widetilde{w}(X)$.

Remark. The class $\widetilde{w}(X) \in K_0(\mathbb{Z}[\pi_1(X, x_0)])$ contains essentially the same information as $w(X) \in \widetilde{K_0}(\mathbb{Z}[\pi_1(X, x_0)])$. Namely, under the identification

$$K_0\big(\mathbb{Z}\big[\pi_1(X, x_0)\big]\big) \cong \mathbb{Z} \oplus \widetilde{K_0}\big(\mathbb{Z}\big[\pi_1(X, x_0)\big]\big)$$

the class $\widetilde{w}(X)$ is given by $(\chi(X), w(X))$, where $\chi(X)$ is the Euler characteristic.

Equivalent formulation of Wall's theorem is thus the following: a finitely dominated connected space X has a homotopy type of a finite CW complex if and only if the class $[\mathbb{Z}] \in K_0(C_\bullet(\Omega_{x_0}X))$ is an integer multiple of the class $[C_\bullet(\Omega_{x_0}X)]$.

Now fix some base field k of arbitrary characteristic. For a small DG category $\mathcal{A}$, we put $K_0(\mathcal{A}) := K_0(D_{\mathrm{perf}}(\mathcal{A}))$. Recall that we denote by $I_{\mathcal{A}}$ the diagonal $\mathcal{A}$-$\mathcal{A}$-bimodule.

Theorem 6.2 ([10]). *For a small DG category $\mathcal{A}$, the following are equivalent:*

(i) *$\mathcal{A}$ is Morita equivalent to a finite cell DG category;*

(ii) *$\mathcal{A}$ is hfp, and moreover $[I_{\mathcal{A}}] \in \operatorname{Im}(K_0(\mathcal{A}) \otimes K_0(\mathcal{A}^{\mathrm{op}}) \to K_0(\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}}))$;*

(iii) *$\mathcal{A}$ is Morita equivalent to a DG quotient $\mathcal{E}/\mathcal{S}$, where $\mathcal{E}$ is a pre-triangulated proper DG category with a full exceptional collection, and $\mathcal{S}$ is a subcategory generated by a single object.*

Remark. To explain the analogy between Theorem 6.2 and Wall's theorem, let us consider the following three categories with a class of morphisms called weak equivalences (the most important part of the model structure):

(1) the category Top of topological spaces, with a class of weak homotopy equivalences;

(2) the category $Z^0(\text{Mod-}\mathcal{A})$ (see Section 2 for this notation) of right DG modules over a fixed DG category $\mathcal{A}$, with a class of quasi-isomorphisms;

(3) the category $\mathrm{dgcat}_\mathrm{k}$ of small DG categories over a field k, with a class of Morita equivalences.

For each of these categories, one has the class of "finite cell" objects, namely: finite CW complexes in Top, semi-free finitely generated $\mathcal{A}$-modules in $Z^0(\text{Mod-}\mathcal{A})$, and finite cell DG categories in $\mathrm{dgcat}_\mathrm{k}$. Then we have the classes of hfp objects: these are homotopy retracts of finite cell objects. Thus, the hfp objects are as follows: finitely dominated spaces in Top, perfect $\mathcal{A}$-modules in $Z^0(\text{Mod-}\mathcal{A})$, hfp DG categories in $\mathrm{dgcat}_\mathrm{k}$.

Now, Wall's theorem (more precisely, an analogue of Theorem 6.1 for not necessarily connected spaces) gives a K-theoretic criterion for a finitely dominated space to have a homotopy type of a finite CW complex.

Further, Thomason's classification of dense subcategories of triangulated categories [24, Theorem 2.1] gives a K-theoretic criterion for a perfect $\mathcal{A}$-module M to be quasi-isomorphic to a semi-free finitely generated $\mathcal{A}$-module. This happens if and only if the class $[M] \in K_0(\mathcal{A})$ is contained in the subgroup generated by the classes of representable $\mathcal{A}$-modules.

From this point of view, our theorem (Theorem 6.2) is an analogue of the results of Wall and Thomason for DG categories, plus also an alternative characterization of finite cell DG categories (equivalence (i)⇔(iii)). The following table summarizes the above discussion.

Topological spaces	DG modules over a small DG category $\mathcal{A}$	Small DG categories over k
Weak homotopy equivalences	Quasi-isomorphisms	Morita equivalences
Finite CW complexes	Semi-free finitely generated $\mathcal{A}$-modules	Finite cell DG categories
Finitely dominated spaces	Perfect $\mathcal{A}$-modules	hfp DG categories
Wall's finiteness obstruction theorem	Thomason's classification of dense subcategories	Theorem 6.2

There are different ways to formulate a "relative" version of Theorem 6.2. We choose the following "minimalistic" formulation.

Theorem 6.3 ([10]). *Let $\mathcal{A}$ and $\mathcal{B}$ be hfp, pre-triangulated, Karoubi complete DG categories, and $\mathcal{B} \neq 0$. The following are equivalent.*

(i) *The class* $[I_{\mathcal{A}}] \in K_0(\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}})$ *is contained in the subgroup generated by the images* $\mathrm{Im}(K_0(\mathcal{B} \otimes \mathcal{B}^{\mathrm{op}}) \to K_0(\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}}))$ *under various pairs of quasi-functors* $(\mathcal{B} \to \mathcal{A}, \mathcal{B}^{\mathrm{op}} \to \mathcal{A}^{\mathrm{op}})$.

(ii) *We have a Morita equivalence* $\mathcal{A} \simeq \mathcal{C}/\mathcal{S}$, *where* $\mathcal{C} = \langle \mathcal{B}, \dots, \mathcal{B} \rangle$ *is a (smooth) semi-orthogonal gluing of a finite number of copies of* $\mathcal{B}$, *and* $\mathcal{S} \subset \mathcal{C}$ *is generated by a single object.*

Using this relative version of Wall's finiteness obstruction for DG categories, we prove the following criterion for existence of a categorical smooth compactification.

Theorem 6.4 ([10]). *Let* $\mathcal{A}$ *be an hfp pre-triangulated DG category. The following are equivalent.*

(1) $\mathcal{A}$ *admits a smooth categorical compactification.*

(2) *There exists a DG functor* $\mathcal{C} \to \mathcal{A}$, *where* $\mathcal{C}$ *is smooth and proper, such that*

$$[I_{\mathcal{A}}] \in \mathrm{Im}\left(K_0(\mathcal{C} \otimes \mathcal{C}^{\mathrm{op}}) \to K_0(\mathcal{A} \otimes \mathcal{A}^{\mathrm{op}})\right).$$

For example, this allows to show existence of a smooth compactification of the derived category of coherent D-modules on a separated scheme of finite type X over a field of characteristic zero (although it is not clear how to construct such compactification explicitly).

Acknowledgments. I am grateful to Dmitry Kaledin, Maxim Kontsevich, Alexander Kuznetsov, Valery Lunts, Dmitri Orlov, Bertrand Toën for useful discussions. I am also grateful to an anonymous referee for useful remarks and suggestions.

Funding. The author is partially supported by the HSE University Basic Research Program.

References

[1] M. Auslander, Representation theory of Artin algebras. I, II. *Comm. Algebra* **1** (1974), 177–268; ibid. 1 (1974), 269–310 Zbl 0285.16029 MR 349747

[2] A. Bondal and D. Orlov, Derived categories of coherent sheaves. In *Proceedings of the International Congress of Mathematicians, Vol. II (Beijing, 2002)*, pp. 47–56, Higher Ed. Press, Beijing, 2002 Zbl 0996.18007 MR 1957019

[3] A. I. Bondal and M. M. Kapranov, Framed triangulated categories. *Mat. Sb.* **181** (1990), no. 5, 669–683 MR 1055981

[4] M. V. Bondarko, Weight structures vs. t-structures; weight filtrations, spectral sequences, and complexes (for motives and in general). *J. K-Theory* **6** (2010), no. 3, 387–504 Zbl 1303.18019 MR 2746283

[5] C. Brav and T. Dyckerhoff, Relative Calabi–Yau structures. *Compos. Math.* **155** (2019), no. 2, 372–412 Zbl 1436.18009 MR 3911626

[6] V. Drinfeld, DG quotients of DG categories. *J. Algebra* **272** (2004), no. 2, 643–691 Zbl 1064.18009 MR 2028075

[7] A. I. Efimov, Generalized non-commutative degeneration conjecture. *Proc. Steklov Inst. Math.* **290** (2015), no. 1, 1–10 Zbl 1338.14006 MR 3488776

[8] A. I. Efimov, Categorical smooth compactifications and generalized Hodge-to-de Rham degeneration. *Invent. Math.* **222** (2020), no. 2, 667–694 Zbl 07269005 MR 4160877

[9] A. I. Efimov, Homotopy finiteness of some DG categories from algebraic geometry. *J. Eur. Math. Soc. (JEMS)* **22** (2020), no. 9, 2879–2942 Zbl 1478.14010 MR 4127943

[10] A. I. Efimov, Wall's finiteness obstruction for DG categories. In preparation

[11] D. Kaledin, Spectral sequences for cyclic homology. In *Algebra, Geometry, and Physics in the 21st Century*, pp. 99–129, Progr. Math. 324, Birkhäuser/Springer, Cham, 2017 Zbl 1390.14011 MR 3702384

[12] B. Keller, Deriving DG categories. *Ann. Sci. École Norm. Sup. (4)* **27** (1994), no. 1, 63–102 Zbl 0799.18007 MR 1258406

[13] B. Keller, On the cyclic homology of exact categories. *J. Pure Appl. Algebra* **136** (1999), no. 1, 1–56 Zbl 0923.19004 MR 1667558

[14] M. Kontsevich and Y. Soibelman, Notes on A_∞-algebras, A_∞-categories and non-commutative geometry. In *Homological Mirror Symmetry*, pp. 153–219, Lecture Notes in Phys. 757, Springer, Berlin, 2009 Zbl 1202.81120 MR 2596638

[15] A. Kuznetsov and V. A. Lunts, Categorical resolutions of irrational singularities. *Int. Math. Res. Not. IMRN* **2015** (2015), no. 13, 4536–4625 Zbl 1338.14020 MR 3439086

[16] J.-L. Loday, *Cyclic Homology*. Grundlehren Math. Wiss. 301, Springer, Berlin, 1992 Zbl 0780.18009 MR 1217970

[17] V. A. Lunts, Categorical resolution of singularities. *J. Algebra* **323** (2010), no. 10, 2977–3003 Zbl 1202.18006 MR 2609187

[18] J. Milnor, On spaces having the homotopy type of a CW-complex. *Trans. Amer. Math. Soc.* **90** (1959), 272–280 Zbl 0084.39002 MR 100267

[19] A. Neeman, *Triangulated Categories*. Ann. of Math. Stud. 148, Princeton University Press, Princeton, NJ, 2001 Zbl 0974.18008 MR 1812507

[20] D. Orlov, Formal completions and idempotent completions of triangulated categories of singularities. *Adv. Math.* **226** (2011), no. 1, 206–217 Zbl 1216.18012 MR 2735755

[21] D. Orlov, Smooth and proper noncommutative schemes and gluing of DG categories. *Adv. Math.* **302** (2016), 59–105 Zbl 1368.14031 MR 3545926

[22] G. N. Tabuada, Une structure de catégorie de modèles de Quillen sur la catégorie des dg-catégories. *C. R. Math. Acad. Sci. Paris* **340** (2005), no. 1, 15–19 Zbl 1060.18010 MR 2112034

[23] G. N. Tabuada, *Théorie homotopique des DG-catégories*. Thèse de doctorat, Université Paris-Diderot - Paris 7, 2007

[24] R. W. Thomason, The classification of triangulated subcategories. *Compos. Math.* **105** (1997), no. 1, 1–27 Zbl 0873.18003 MR 1436741

[25] B. Toën and M. Vaquié, Moduli of objects in dg-categories. *Ann. Sci. École Norm. Sup. (4)* **40** (2007), no. 3, 387–444 Zbl 1140.18005 MR 2493386

[26] J.-L. Verdier, Catégories dérivées: quelques résultats (état 0). In *Cohomologie étale*, pp. 262–311, Lecture Notes in Math. 569, Springer, Berlin, 1977 Zbl 0407.18008 MR 3727440

[27] J.-L. Verdier, Des catégories dérivées des catégories abéliennes. *Astérisque* **239** (1996), xii+253 pp. Zbl 0882.18010 MR 1453167

[28] C. T. C. Wall, Finiteness conditions for CW-complexes. *Ann. of Math. (2)* **81** (1965), 56–69 Zbl 0152.21902 MR 171284

Alexander I. Efimov
Steklov Mathematical Institute of RAS, Gubkin str. 8, GSP-1, Moscow 119991; and National Research University Higher School of Economics, Myasnitskaya Ulitsa 20, Moscow 101000, Russia; efimov@mccme.ru

DOI 10.4171/8ECM/38

Global properties of some weight 3 variations of Hodge structure

Simion Filip

Abstract. We survey results on the global geometry of variations of Hodge structure with Hodge numbers $(1, 1, 1, 1)$. Included are uniformization results of domains in flag manifolds, a strong Torelli theorem, as well as the formula for the sum of Lyapunov exponents conjectured by Eskin, Kontsevich, Möller, and Zorich. Additionally, we establish the Anosov property of the monodromy representation, using gradient estimates of certain functions derived from the Hodge structure.

1. Introduction

Consider the following family of algebraic 3-manifolds in $\mathbb{P}^4$:

$$\left\{t \cdot (x_1^5 + x_1^5 + x_2^5 + x_3^5 + x_4^5 + x_5^5) - x_1 x_2 x_3 x_4 x_5 = 0\right\} =: Q_t.$$

This is the famous mirror quintic family, used by string theorists in [4] to make predictions about the number of rational curves on a generic quintic 3-fold. Every member of the family admits a nowhere vanishing 3-form Ω_t, which, integrated over an explicit cycle (near $t = 0$), yields the hypergeometric series

$$\psi_0(t) = \sum_{n \geq 0} \frac{(5n)!}{(n!)^5} t^n. \tag{1.1}$$

This function satisfies (after rescaling t to $5^5 t$) the hypergeometric differential equation

$$\left[D^4 - t\left(D + \frac{1}{5}\right)\left(D + \frac{2}{5}\right)\left(D + \frac{3}{5}\right)\left(D + \frac{4}{5}\right)\right]\psi_0 = 0, \quad \text{where } D := t\frac{d}{dt}.$$

All the results below are of interest already in this particular example, although they apply to a much larger class of variations of Hodge structure.

2020 Mathematics Subject Classification. Primary 14D07; Secondary 34D08, 37D20, 14L24.
Keywords. Variations of Hodge structure, Calabi–Yau, Anosov representation, GIT, Lyapunov exponents.

Monodromy. Let $X := X(\infty, \infty, 5)$ denote the orbifold Riemann surface $\mathbb{P}^1 \setminus \{0, 1\}$ with an orbifold point of order 5 at ∞; the notation is meant to suggest that it comes from the hyperbolic triangle group with angles $(\pi/\infty, \pi/\infty, \pi/5)$. The cohomology $H^3(Q_t; \mathbb{Z})$ has rank 204, but of interest to us is a piece invariant by a natural finite abelian group of roots of unity. This invariant subspace has rank 4, and in fact gives a local system over X (the monodromy around ∞ has order 5).

The explicit matrices are, in an appropriate choice of basis, around 1 and ∞:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 5 & 5 & 1 & 0 \\ 0 & -5 & -1 & 1 \end{bmatrix}.$$

It can be checked that the last matrix has order 5, and that both preserve the standard symplectic form on $\mathbb{R}^4$. Here is the first result:

Theorem 1.1 (Log-Anosov monodromy). *The monodromy representation*

$$\rho\colon \pi_1(X) \to \mathbf{Sp}_4(\mathbb{R})$$

is log-Anosov. There exists a continuous, dynamics-preserving, ρ-equivariant map

$$\xi\colon \partial\widetilde{X} \to \mathbb{P}(\mathbb{R}^4).$$

Let us explain the terms. The universal cover of X, denoted by $\widetilde{X}$, is isometric to the hyperbolic plane and as such has a boundary, isomorphic to $\mathbb{P}^1(\mathbb{R})$, and denoted by $\partial\widetilde{X}$. The notion of Anosov representation, introduced by Labourie [15], requires a quantitative divergence of the singular values of the monodromy matrices; see Definition 3.1 for the details. In the present context, we need to take into account the presence of unipotent elements in the group, hence the term "log-Anosov". Under the name relatively Anosov, or relatively dominated, such representations were studied by Kapovich–Leeb [12] and Zhu [21], respectively.

Integral vectors. The "limit set" curve, i.e., the image of ξ, provided by Theorem 1.1 is a fractal curve (in fact it is possible and not hard to prove that because of the rank 1 unipotent element, the limit curve cannot be rectifiable). See Figure 1 for some illustrations. Nonetheless, we can classify the rational points on the curve, and similarly an analogous limit set in the Lagrangian Grassmannian.

Theorem 1.2 (Rational directions on limit curve). *Let $\Gamma \subset \mathbf{Sp}_4(\mathbb{Z})$ denote the image of the monodromy representation.*

(1) *A line $[v] \in \xi(\partial\widetilde{X}) \subset \mathbb{P}(\mathbb{Q}^4)$ has rational coordinates if and only if there exists a unipotent transformation $\gamma \in \Gamma$ such that v is both in the kernel and*

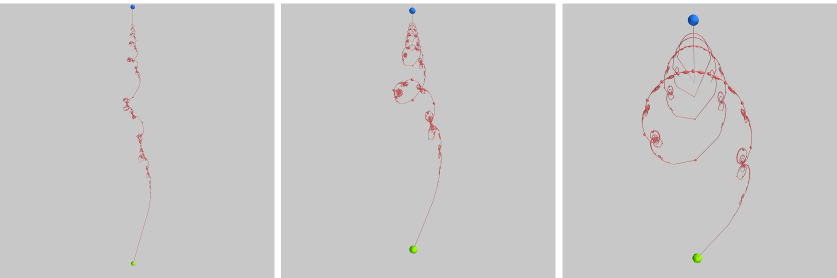


Figure 1. Sample images of the map ξ from Theorem 1.1 for several families of Calabi–Yau manifolds. The middle curve corresponds to the mirror quintic.

image of $\gamma - \mathbf{1}$. In particular, the rational vectors on $\xi(\partial\widetilde{X})$ fall into finitely many orbits under the action of Γ, *corresponding to the cusps of X.*

(2) *Suppose that a Lagrangian $L \subset \mathbb{Q}^4$ contains $\xi(p)$, for some point $p \in \partial\widetilde{X}$. Then there exists a unipotent $\gamma \in \Gamma$ and $[v] = \xi(p)$ as in the first part, in the kernel and image of γ, such that $[v] \subset L$.*

Note that the property of being both in the kernel, and in the image, of $\gamma - \mathbf{1}$ is equivalent (in $\mathbf{Sp}_4$) to v belonging to the deepest part of the monodromy weight filtration.

Torelli theorems. Our methods also allow us to gain some insight into the global structure of some moduli spaces of Calabi–Yau 3-folds. Before discussing it, let us introduce some notation. Let $\mathbb{V} \to X$ denote the rank 4 local system of the cohomology of interest of Q_t. It admits a variation of Hodge structure, i.e., a decomposition of the complexification:

$$\mathbb{V}_{\mathbb{C}} = \mathcal{V}^{3,0} \oplus \mathcal{V}^{2,1} \oplus \mathcal{V}^{1,2} \oplus \mathcal{V}^{0,3}$$

which depends on the point on X (with additional properties). The symplectic pairing on $\mathbb{V}$ induces an indefinite Hermitian pairing on $\mathbb{V}_{\mathbb{C}}$, for which $F^2 = \mathcal{V}^{3,0} \oplus \mathcal{V}^{2,1}$ is of signature $(1,1)$. Note that it is a Lagrangian subspace for the symplectic form, and we denote by $\mathrm{LGr}^{1,1}(V_{\mathbb{C}})$ the space of all such Lagrangians in a fixed vector space $V_{\mathbb{C}}$.

Theorem 1.3 (Strong Torelli using Lagrangians). (1) *The maps induced by the Hodge filtration:*

$$\widetilde{X} \xrightarrow{F^2} \mathrm{LGr}^{1,1}(V_{\mathbb{C}})$$

and taking the quotient by π_1:

$$X \xrightarrow{F^2} \Gamma \backslash \mathrm{LGr}^{1,1}(V_{\mathbb{C}})$$

are injective.

(2) *Furthermore, for $x, y \in \widetilde{X}$, if L is a real Lagrangian such that $(F^2(x) \cap L_{\mathbb{C}}) \neq 0 \neq (F^2(y) \cap L_{\mathbb{C}})$, then $x = y$.*

A further dichotomy, related to rational Lagrangian subspaces, is contained in Corollary 1.4. For the mirror quintic family, a generic Torelli theorem using the full Griffiths period domain was proved by Usui [19].

Remark (On Griffiths' intermediate Jacobians). When the real weight 3 Hodge structure has an underlying integral structure, one can associate to it the Griffiths intermediate Jacobian $V_{\mathbb{Z}} \backslash V_{\mathbb{C}} / F^2$. In general, this is not an abelian variety, and the period domain of such objects is $\mathrm{LGr}^{1,1}(V_{\mathbb{C}})$, a pseudo-Hermitian homogeneous space, in contrast to Siegel spaces parametrizing marked abelian varieties.

It is not possible to take a Hausdorff quotient of $\mathrm{LGr}^{1,1}(V_{\mathbb{C}})$ by $\mathbf{Sp}_4(\mathbb{Z})$, or any lattice in $\mathbf{Sp}_4(\mathbb{R})$. However, Theorem 4.3 implies that for the monodromy Γ of a VHS satisfying assumption A, *it is* possible to take the quotient. Theorem 1.3 then says that the period map to this quotient is injective. So it can be viewed a Torelli theorem in the classical sense.

Theorem 1.2 above, combined with the strong Torelli theorem provides an interesting property of rational Lagrangian subspaces:

Corollary 1.4 (Dichotomy for rational Lagrangian subspaces). *With assumptions as in Theorem 1.2, for every rational Lagrangian subspaces $L \subset \mathbb{Q}^4$, precisely one of the following holds:*

- *either there exists a unipotent transformation $\gamma \in \Gamma$ and vector $v \in L$ such that v is both in the kernel and in the image of $\gamma - \mathbf{1}$,*
- *or there exists a unique $x \in \widetilde{X}$ such that $L_{\mathbb{C}} \cap F^2(x) \neq \{0\}$.*

Lyapunov exponents. The original reason that prompted the above results was a conjecture of Eskin, Kontsevich, Möller, and Zorich from [6] relating Lyapunov exponents, which are invariants coming from dynamical systems, with the degrees of certain line bundles. Using the above results, we can establish their conjecture:

Theorem 1.5 (Formula for the sum of Lyapunov exponents). *Let $\lambda_1 \geq \lambda_2 \geq 0$ be the nonnegative Lyapunov exponents of the geodesic flow on the unit tangent, for the cocycle induced by $\mathbb{V}$. Then*

$$\lambda_1 + \lambda_2 = \frac{\deg \mathcal{V}^{0,3}_{\mathrm{ext}} + \deg \mathcal{V}^{1,2}_{\mathrm{ext}}}{\chi(X)} = \frac{6}{5},$$

where $\chi(X)$ is the (orbifold) Euler characteristic of X, deg *denotes the (orbifold) degree of a complex line bundle, and the subscript* ext *denotes the Deligne extension of a bundle across punctures.*

As we shall explain below, in fact the stronger conjecture, Conjecture 6.4 of [6], of a number-theoretic flavor, also holds. Note that the degrees of the bundles were computed in loc. cit. and come from the parameters of the corresponding hypergeometric differential equation.

Explicit nonvanishing. The conjecture alluded to above is formulated in terms of the subspace invariant by the monodromy near the singularity at 0. It has an explicit statement in terms of power series, which we now explain.

Recall from equation (1.1) that we defined one solution of the hypergeometric equation, and consider the second one

$$\psi_1(t) := \sum_{n\geq 0} \frac{(5n)!}{(n!)^5}\left(\sum_{k=n+1}^{5n} \frac{1}{k}\right) \cdot t^n.$$

There are two more (with further logarithmic terms), but we are interested in the following Wronskian determinant (which is a 2 × 2 minor of the full matrix of solutions)

$$Wr(t) := \psi_0(t)\psi_1'(t) - \psi_0'(t)\psi_1(t).$$

To describe the uniformization of the orbifold X, we make use again of classical hypergeometric functions:

$$F_0(t) = \sum_{n\geq 0} \frac{(a)_n(b)_n}{(n!)^2} t^n,$$

$$F_1(t) = F_0(t)\log t + \sum_{n\geq 0} \frac{(a)_n(b)_n}{(n!)^2}\left(\sum_{k=1}^{n} \frac{1}{a+k-1} + \frac{1}{b+k-1} - \frac{2}{k}\right)t^n,$$

where $a = \frac{2}{5}$ and $b = \frac{3}{5}$.

Define now the map

$$q := \exp\left(\frac{F_1(t)}{F_0(t)}\right) = t \cdot \exp\left(\frac{P_1(t)}{P_0(t)}\right),$$

where $P_1(t)$, $P_0(t)$ are the power series appearing under the summation sign in the definition of F_0, F_1. Finally, define the inverse power series $\lambda_5(q) = \sum_{n\geq 0} \Lambda_n q^n$ such that

$$\lambda_5\left(\exp\left(\frac{F_1(t)}{F_0(t)}\right)\right) = t.$$

This yields the uniformization cover

$$\left\{0 < |q| < e^{D(1)}\right\} \xrightarrow{\lambda_5} X,$$

where $D(1)$ denotes a sum of values of the logarithmic derivative of the gamma-function.

The explicit nonvanishing, conjectured in [6, Conj. 6.4], then reads

$$Wr\left(\frac{1}{5^5}\lambda_5(q)\right) \text{ never vanishes.}$$

Domains of discontinuity. A consequence of the Anosov property of the monodromy representation is that the image group $\Gamma \subset \mathbf{Sp}_4(\mathbb{R})$ has a large class of domains of discontinuity in real and complex flag manifolds associated to $\mathbf{Sp}_4$. In Theorem 4.3, we list some of them.

Thin groups. No lattice in $\mathbf{Sp}_4(\mathbb{R})$ can have a domain of discontinuity as in Theorem 4.3. It follows that the monodromy group Γ, when it has an integral structure, is necessarily a "thin group" in the sense of Sarnak [17]. Let us note that the proofs in the present paper offer an alternate route to some of the results from [2, 9], where thinness is established using ping-pong and an explicit construction of cones. These explicit cones have, nonetheless, other applications to a more detailed understanding of the monodromy groups.

Below we outline the main notions that go into the proof of the above results, in greater generality. A detailed account is in [8].

2. Variations of Hodge structure and hypergeometric equations

2.1. Variations of Hodge structure

Let X be a complex manifold and $\mathbb{V} \to X$ a local system of real vector spaces. Equivalently, this is a bundle with flat connection ∇, also called the Gauss–Manin connection.

Definition 2.1 (Variation of Hodge structure). A *variation of Hodge structure* (or VHS) on $\mathbb{V}$ of weight n is a decomposition of the complexification

$$\mathbb{V}_{\mathbb{C}} = \bigoplus_{p+q=n} \mathcal{V}^{p,q}(x), \quad x \in X$$

such that the following hold.

- The *Hodge filtration* $\mathcal{F}^p := \oplus_{s \geq p} \mathcal{V}^{s,n-s}$ varies holomorphically.
- The *Griffiths transversality condition*

  $$\nabla(\mathcal{F}^p) \subset \mathcal{F}^{p-1} \otimes \Omega^1_X$$

  is satisfied.
- Under complex conjugation, we have that $\mathcal{V}^{p,q} = \overline{\mathcal{V}^{q,p}}$.

Additionally, the variation is *polarized* if there exists a $(-1)^n$-symmetric nondegenerate bilinear form on $\mathbb{V}$, parallel for the Gauss–Manin connection, and such that the induced Hermitian pairing on $V_{\mathbb{C}}$ has signature $(-1)^q$ on $\mathcal{V}^{p,q}$.

The classical case is that of a weight 2 variation, when there are two bundles $\mathcal{V}^{1,0} \oplus \mathcal{V}^{0,1}$. Such variations come from holomorphic families of abelian varieties or Riemann surfaces.

Second fundamental form. The Griffiths transversality condition allows us to define a second fundamental form, by taking $\nabla(\mathcal{F}^p)/\mathcal{F}^p$ to obtain a well-defined *linear* map of bundles

$$\sigma_\bullet = \oplus \sigma_{p,q} \quad \text{with } \sigma_{p,q} \colon \mathcal{V}^{p,q} \to \mathcal{V}^{p-1,q+1} \otimes \Omega^1_X.$$

Tensor constructions. We will be interested in variations with $\dim \mathcal{V}^{p,q} = 1$ and weight 3, or said differently with Hodge numbers $(1,1,1,1)$. These admit an invariant symplectic form and the monodromy of the local system is valued in $\mathbf{Sp}_4(\mathbb{R})$. We can also perform natural tensor constructions, of which the most useful is the (reduced) second exterior power $\mathbb{W} := \Lambda^2_\circ \mathbb{V}$, where reduced means that we remove a 1-dimensional invariant subspace generated by the symplectic form. Therefore, $\mathbb{W}$ has rank 5 and Hodge numbers $(1,1,1,1,1)$ (we Tate-twist the construction so that it has weight 4, not 6). In fact, up to passing to a finite cover, it is possible to recover $\mathbb{V}$ from $\mathbb{W}$.

We assume from now on that X is a finite volume complete hyperbolic Riemann surface, and denote by the subscript ext the Deligne extension of bundles across the punctures (the reader unfamiliar with these terms can just assume that X is compact).

Definition 2.2 (Assumption A). We will say that $\mathbb{V}$ satisfies *assumption A* if the component of the second fundamental form

$$\sigma_{2,1} \colon \mathcal{V}^{2,1}_{\text{ext}} \to \mathcal{V}^{1,2}_{\text{ext}} \otimes \Omega^1_X$$

is an isomorphism.

Equivalently on $\mathbb{W}$, the requirement is that $\sigma_{4,0}$ is an isomorphism.

All the theorems described below apply as soon as the VHS satisfies assumption A.

2.2. Hypergeometric equations

For general information and the results below on hypergeometric equations, we refer to the texts of Beukers–Heckman [1] or Yoshida [20].

Let $\alpha_1, \beta_1, \ldots, \alpha_n, \beta_n$ be a list of $2n$ numbers (while we will always take them real later, at this stage they can be complex). Define the differential operator

$$D_{\alpha,\beta} := \prod_{i=1}^{n} (D - \beta_i) - z \prod_{i=1}^{n} (D + \alpha_i), \quad \text{where } D = z\partial_z.$$

$\alpha_\bullet$	$\beta_\bullet$	Conditions
$\left(\mu, \frac{1}{2}, \frac{1}{2}, 1-\mu\right)$	$(0,0,0,0)$	$\mu \in \left(0, \frac{1}{2}\right]$
	$(0,0,\nu,1-\nu)$	$0 < \nu < \mu \leq \frac{1}{2}$

Table 1. Parameters μ, ν are real.

$\beta_\bullet$	Conditions
$\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$	
$(0,0,0,0)$	
$(0,0,\mu,1-\mu)$	$0 < \mu < \frac{2k-1}{2N}$
$\left(\mu, \frac{1}{2}, \frac{1}{2}, 1-\mu\right)$	$\frac{N-1}{2N} < \mu < \frac{1}{2}$
$\left(\frac{M-(2k_M+1)}{2M}, \frac{M-1}{2M}, \frac{M+1}{2M}, \frac{M+(2k_M+1)}{2M}\right)$	$\frac{2k_M+1}{M} < \frac{1}{N}$

Table 2. Parameters M, k_M are integers; μ is real. Throughout, $\alpha_\bullet = (\frac{N-(2k+1)}{2N}, \frac{N-1}{2N}, \frac{N+1}{2N}, \frac{N+(2k+1)}{2N})$ with arbitrary integers $k \geq 1$, $N > 2k+1$.

With these conventions, the exponents (or Riemann scheme) of the operator are:

- at 0: $\beta_1, \ldots, \beta_n$,
- at ∞: $\alpha_1, \ldots, \alpha_n$,
- at 1: $0, 1, \ldots, n-2, \gamma := (n-1) - \sum_{i=1}^{n}(\alpha_i + \beta_i)$.

The parameters are slightly different from those of [1], specifically $\beta_i^{BK} = 1 - \beta_i$, and there are plenty of other variations in the literature.

The local system of solutions of the hypergeometric equations is *rigid*, meaning that it has no nontrivial deformations when we require the conjugacy classes at the cusps to be fixed. By a theorem of Simpson [18, Cor. 8.1], it follows that the local system underlies a variation of Hodge structure. Fedorov [7] provided a recipe for computing the Hodge numbers of the corresponding VHS.

This allows us to tabulate the values of the hypergeometric parameters which satisfy assumption A, listed in Tables 1 and 2.

Assumption B/maximal representations. It is natural to consider also a variant of assumption A, asking instead that the second fundamental form $\sigma_{3,0}$ is an isomorphism. This leads to the class of *maximal monodromy representations*, in the sense introduced in [3]. Following an analogous algorithm as in the case of assumption A, it is possible to tabulate the hypergeometric rank 5 parameters such that the local system $\mathbb{W}$, with monodromy in $\mathbf{SO}_{2,3}(\mathbb{R})$, is maximal. The results are listed in Table 3.

Schwarz reflection. Hypergeometric equations with real parameters satisfy a complex conjugation symmetry and their monodromy groups can be embedded with

$\alpha_\bullet$	$\beta_\bullet$
$\left(\mu, \frac{1}{2}, \frac{1}{2}, 1-\mu, \frac{1}{2}\right)$	$\left(0, 0, 0, \frac{M}{2M+1}, \frac{M+1}{2M+1}\right)$
or	or
$\left(\frac{N-k_N}{2N}, \frac{N-1}{2N}, \frac{1}{2}, \frac{N+1}{2N}, \frac{N+k_N}{2N}\right)$	$\left(0, \frac{k_M}{M}, \frac{k_M+1}{M}, \frac{M-(k_M+1)}{M}, \frac{M-k_M}{M}\right)$

Table 3. Any set choice from the first column is compatible with any choice from the second, subject to the condition $\alpha_{\min} > \beta_{\text{med}}$, where $\alpha_{\min} := \mu$ or $\frac{N-k_N}{2N}$, and $\beta_{\text{med}} := \frac{M}{2M+1}$ or $\frac{k_M+1}{M}$ depending on the choices. The parameter μ is real, while M, N, k_M, k_N are positive integers with $1 < k_N < N$ and $2(k_M+1) < M$.

index 2 into a group generated by three order 2 involutions. Geometrically, this corresponds to the following construction. Take a basis of solutions in the upper half plane. Analytically continue it into the lower half-plane by choosing one of the three segments formed by removing 0, 1 from $\mathbb{R}$. Then, apply the Schwarz reflection, i.e., map $f(z)$ to $\overline{f(\bar{z})}$, to obtain another basis of solutions in the upper half-plane.

The operation of the Schwarz reflection is just complex-conjugating the coefficients of a Taylor expansion of $f(z)$. Because the hypergeometric equation has real coefficients, the resulting functions are still solutions of the hypergeometric equation. Note that on each of the three segments of the real axis, there is a basis of solutions with real coefficients, but on each segment the basis is different. For instance, $\log z$ is real-valued on $(0, \infty)$ but has also an imaginary component on $(-\infty, 0)$.

It is possible to choose a basis and explicitly give the matrices of the three reflections generating the above construction. Let $\mathcal{R}_A$, $\mathcal{R}_B$, $\mathcal{R}_C$ be the transformations corresponding to crossing along $(1, \infty)$, $(0, 1)$, and $(\infty, 0)$ on the real axis. Then the matrices giving the transformations are listed in equation (2.1). To obtain the action of $\mathcal{R}_X$ on the space of solutions, one must apply complex conjugation to the coordinates after applying the matrix R_X:

$$R_A := \begin{bmatrix} 0 & \cdots & 0 & 1 & -A_1 \\ 0 & \cdots & 1 & 0 & -A_2 \\ & \iddots & & & \\ 1 & \cdots & 0 & 0 & -A_{n-2} \\ 0 & \cdots & 0 & 0 & -A_n \end{bmatrix} \quad R_B := \begin{bmatrix} 0 & \cdots & 0 & 1 & -\overline{B_1} \\ 0 & \cdots & 1 & 0 & -\overline{B_2} \\ & \iddots & & & \\ 1 & \cdots & 0 & 0 & -\overline{B_{n-2}} \\ 0 & \cdots & 0 & 0 & -\overline{B_n} \end{bmatrix}$$

$$\text{as well as } R_C := \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ & & \iddots & & \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}. \tag{2.1}$$

Recall that the hypergeometric equation has parameters α_i, β_j and we set $a_j := \exp(2\pi\sqrt{-1}\alpha_j)$ and $b_j := \exp(2\pi\sqrt{-1}\beta_j)$ as well as

$$p_A(t) := \prod_{j=1}^{n}(t - a_j) = t^n + A_1 t^{n-1} + \cdots + A_n,$$
$$p_B(t) := \prod_{j=1}^{n}(t - b_j) = t^n + B_1 t^{n-1} + \cdots + B_n$$

to obtain the entries of the matrices. This should be compared to the Levelt presentation of the monodromy matrices of the hypergeometric equation from [1, Thm. 3.5].

Tiling by polyhedra. For the domains of discontinuity constructed in Theorem 4.3, it is possible to give a detailed description of a tiling by polyhedra, in direct analogy with the tiling of the hyperbolic plane by hyperbolic triangles. Crucially, the variation of Hodge structure underlying the hypergeometric local system provides the data needed to construct the edges of the polyhedron.

3. Log-Anosov representations

3.1. Lie theory and Anosov representations

The notion of Anosov representation was introduced by Labourie in [15]. We refer to the works of Guichard–Wienhard [11, 13] for extensive references and a general introduction to the subject. Some general background in Lie theory can be gathered from [14].

It is worth keeping in mind that there is both an extrinsic and an intrinsic approach to Anosov representations, and that both are useful. Fix a semisimple real algebraic group G. One can look at a representation $\rho\colon \pi_1(X) \to G$, or consider a (possibly reducible) algebraic representation $\phi\colon G \to \mathbf{GL}(V)$. It is possible to go back and forth between the properties of ρ and the properties of $\phi \circ \rho$, and it is useful to do so.

Lie theory preliminaries. Continuing with the semisimple Lie group G as above, let $K \subset G$ be maximal compact and let $\mathfrak{a} = \operatorname{Lie} A$ be the split part of a Cartan subalgebra of $\mathfrak{g}$. Let $\Phi \subset \mathfrak{a}^\vee$ be the roots and $\Delta \subset \Phi$ the subset of simple roots, for a choice of ordering, which also yields the Weyl chamber $\mathfrak{a}^+ \subset \mathfrak{a}$. Finally, let $\theta \subset \Delta$ be a (nonempty) set of roots.

KAK, or polar decomposition. One way to approach the coarse geometry of the group G is via its KAK decomposition. Namely, any $g \in G$ can be written as

$$g = k_-(g) \cdot e^{\mu(g)} k_+(g) \quad \text{with } k_\pm(g) \in K \text{ and } \mu(g) \in \mathfrak{a}^+.$$

The K-components are not necessarily unique, but $\mu(g)$ is.

Let also $\| - \| \colon \pi_1(X) \to \mathbb{R}_{\geq 1}$ be the matrix norm induced by some Fuchsian representation $\pi_1(X) \to \mathbf{SL}_2(\mathbb{R})$.

Definition 3.1 (Log-Anosov representation). A representation $\rho \colon \pi_1(X) \to G$ is called *log-Anosov* if there exist $C, \varepsilon > 0$ such that

$$\alpha\big(\mu\big(\rho(\gamma)\big)\big) \geq \varepsilon \cdot \log \|\gamma\| - C \quad \forall \alpha \in \theta,\ \forall \gamma \in \pi_1(X),$$

where μ denotes the element in $\mathfrak{a}^+$ from the KAK decomposition.

In the more general setting of relatively hyperbolic groups, this notion was studied under the name relatively Anosov, or relatively dominated, by Kapovich–Leeb [12] and Zhu [21].

Boundary map. A consequence of the (log)-Anosov property is the existence of boundary maps; see for instance [10, Thm. 1.1] and [21, Thm. 1.2]. Specifically, recall that the universal cover $\widetilde{X}$ is isometric to the hyperbolic plane, and it has a visual boundary $\partial \widetilde{X}$. Then there exists a continuous, ρ-equivariant map

$$\xi \colon \partial \widetilde{X} \to \mathscr{F}_\theta,$$

where $\mathscr{F}_\theta$ is the manifold of flags associated to θ, so $\mathscr{F}_\theta \cong G/P_\theta$, where P_θ is a parabolic subgroup corresponding to θ.

3.2. Proper discontinuity, stability, and GIT

Kapovich, Leeb, and Porti in [13] have emphasized the analogy between the action of discrete subgroups of Lie groups, and those of algebraic groups in linear representations, viewed from the lens of Mumford's Geometric Invariant Theory [16]. We make some further definitions and take it as a viewpoint to perform some of the constructions appearing later.

Suppose for this section that $\Gamma \subset G$ is a closed subgroup of G. Associated to it are various notions of limit sets, similarly to how there are various notions of boundaries for G, or G/K. The one of interest to us, which we will denote by $\mathscr{L}_+ \subset K \times \mathbb{P}\mathfrak{a}^+$, consists of all possible accumulation points of the coordinates $k_+(\gamma) \in K$ and $[\mu(\gamma)] \in \mathbb{P}\mathfrak{a}^+$ (the projectivized cone) as $\mu(\gamma) \to +\infty$. Suppose now that V is a G-representation.

Definition 3.2 (Stable and semistable points). Let $[v] \in \mathbb{P}(V)$ be a point and $v \in V$ a lift of it to the vector space. Then $[v]$ is *stable* if

$$\|e^{t\mu} k_+ v\| \to +\infty \quad \forall \big(k, [\mu]\big) \in \mathscr{L}_+$$

and $[v]$ is *semistable* if

$$\liminf \|e^{t\mu} k_+ v\| > 0 \quad \forall \big(k, [\mu]\big) \in \mathscr{L}_+.$$

This definition is the direct analogue of the Hilbert–Mumford numerical criterion. It is then possible to show that the following theorem holds.

Theorem 3.3 (Proper discontinuity). (1) *The set of stable points is open in* $\mathbb{P}(V)$.

(2) *The action of* Γ *on the set of stable points is properly discontinuous.*

(3) *A stable point and a semistable point cannot be dynamically related.*

Two points x, y are said to be dynamically related if there exist sequences x_i and $\gamma_i \in \Gamma$ such that $x_i \to x$ and $\gamma_i x_i \to y$. Note that this is an equivalence relation (take $y_i = \gamma_i x_i$ and γ_i^{-1} as the sequence).

If the group Γ is the image of a log-Anosov representation, then the boundary map ξ gives useful control on the limit set $\mathscr{L}_+$, and hence on the set of stable points in various linear representations of G.

4. Variations of Hodge structure and log-Anosov representations

In this section, we combine some ideas from Hodge theory with those coming from Anosov representations.

4.1. Growth of vectors

Let $\mathbb{V} \to X$ be a VHS satisfying assumption A from Definition 2.2. Let $\mathbb{W} := \Lambda^2_\circ \mathbb{V}$ be the corresponding reduced second exterior power. We will work on a universal cover $\widetilde{X}$, where both local systems become trivial and can be identified with fixed vector spaces V, W, and the VHS gives a Hodge decomposition of these fixed vector spaces. Recall also that V is symplectic whereas W carries an indefinite pairing of signature $(2,3)$.

Pick a nonzero vector $w \in W_{\mathbb{R}}$ and write its Hodge decomposition:

$$w = w^{4,0} \oplus w^{3,1} \oplus w^{2,2} \oplus w^{1,3} \oplus w^{0,4}$$

Theorem 4.1 (Growth of vectors). *Let* $f_w(x) := \|w^{0,4}\|^2$, *where* $\| - \|$ *is computed with respect to the Hodge norm at* $x \in \widetilde{X}$.

(1) *Suppose that* w *is isotropic. Then* f_w *has at most one critical point, which can only be a local minimum.*

(2) *Suppose that* w *is positive-definite, for the indefinite metric. Then* f_w *has precisely one critical point, which is a local minimum.*

The proof of this result is based on a gradient estimate for f_w, combined with an argument using Palais–Smale sequences to exclude multiple local minima. In fact, the gradient estimate can be strengthened and used to show that when w is positive-definite, it has exponential growth:

Theorem 4.2 (Exponential growth). *Suppose that w is positive-definite and f_w has a minimum at x_0. Then there exist $C, \varepsilon > 0$ such that*

$$f_w(x) \geq \frac{1}{C} e^{\varepsilon \cdot \operatorname{dist}(x, x_0)} - C.$$

With this information in hand, it is not hard to obtain the log-Anosov condition on the monodromy representation from Definition 3.1.

4.2. Uniformization results

With the log-Anosov property in hand, the formalism of stable vectors in representations from Section 3.2 provides a wealth of domains of discontinuity for the monodromy.

Theorem 4.3 (Domains of discontinuity). *Let $\Gamma \subset \mathbf{Sp}_4(V_{\mathbb{R}})$ be the image of the monodromy group.*

(1) *In the real Lagrangian Grassmannian $\operatorname{LGr}(V_{\mathbb{R}})$, there exists an open nonempty set Ω_L on which Γ acts properly discontinuously.*

(2) *The pseudo-sphere $\mathbb{S}^{1,3}$ of unit vectors in $W_{\mathbb{R}}$ is a domain of discontinuity for Γ.*

(3) *The complex Grassmannian of Lagrangians of signature $(1,1)$ for the indefinite Hermitian metric on $V_{\mathbb{C}}$, denoted by $\operatorname{LGr}^{1,1}(V_{\mathbb{C}})$, is also a domain of discontinuity for Γ.*

(4) *In the complex projective space $\mathbb{P}(V_{\mathbb{C}})$, there exists an open, nonempty set Ω_P on which Γ acts properly discontinuously.*

More interestingly, it is possible to obtain a uniformization result for the domain of discontinuity in the Lagrangian Grassmannian. For an element $F^2 = \mathcal{V}^{3,0} \oplus \mathcal{V}^{2,1}$ of the Hodge filtration, set

$$\beta(F^2) := \big\{\text{real Lagrangians } L \text{ s.t. } L_{\mathbb{C}} \cap F^2 \neq \{0\}\big\}.$$

In other words, we consider the real Lagrangians which are not transverse to F^2, after complexification. It can be directly checked that for a fixed F^2, these form a circle inside the real 3-dimensional manifold $\operatorname{LGr}(V_{\mathbb{R}})$. Let $\widetilde{\mathrm{Bad}} \to \widetilde{X}$ denote this circle bundle over the universal cover, and $\mathrm{Bad} \to X$ its quotient by $\pi_1(X)$. The reason for the name “Bad” will be explained in the section on Lyapunov exponents below. For now, observe that there is a tautological developing map

$$\widetilde{\mathrm{Bad}} \xrightarrow{\mathrm{Dev}} \operatorname{LGr}(V_{\mathbb{R}})$$

since each fiber of the bundle is a circle in that Lagrangian Grassmannian.

Theorem 4.4 (Uniformization). *The developing map* Dev *is a bijection between* $\widetilde{\mathrm{Bad}}$ *and the domain of discontinuity* $\Omega_L \subset \mathrm{LGr}(V_{\mathbb{R}})$ *from Theorem 4.3.*

4.3. Formula for Lyapunov exponents

Theorem 4.4 implies, in a stronger form, Conjecture 6.4 from [6] that the MUM Lagrangian is never "bad". Recall that in that context, to every vector $w \in W_{\mathbb{R}}$ one can associate a "bad locus" corresponding to points in the universal cover where $w^{0,4} = 0$. The emptiness of the bad locus, for at least one vector w, implies the expected formula for the sum of Lyapunov exponents. This is precisely the formula stated in the introduction in Theorem 1.5.

Maximal representations. We end with an observation regarding what we called "assumption B", or equivalently the condition that the VHS is maximal described in Section 2.1. A uniformization result analogous to Theorem 4.4 holds in this case, and was established by Collier, Tholozan, and Toulisse [5, Thm. 1]. It implies the following formula for the *top* Lyapunov exponent of $\mathbb{V}$:

$$\lambda_1(\mathbb{V}) = \frac{\deg \mathcal{V}^{0,3}_{\mathrm{ext}}}{\chi(X)}.$$

Note that the only representations which satisfy both assumption A and assumption B are those which are a symmetric power of the standard Fuchsian representation. In that case, all the above theorems, including the domains of discontinuity and the formula for Lyapunov exponents, are immediate verifications in linear algebra.

Acknowledgments. Most of this work was completed when I visited the Institute for Advanced Study during the academic year 2018-2019. I am sincerely grateful for the excellent working conditions. This research was partially conducted during the period the author served as a Clay Research Fellow.

Funding. This material is based upon work supported by the US National Science Foundation under Grants No. DMS-2005470 (SF), No. DMS-1638352 (IAS), as well as DMS-1107452, 1107263, 1107367 "RNMS: Geometric Structures and Representation Varieties" (the GEAR Network).

References

[1] F. Beukers and G. Heckman, Monodromy for the hypergeometric function ${}_nF_{n-1}$. *Invent. Math.* **95** (1989), no. 2, 325–354 Zbl 0663.30044 MR 974906

[2] C. Brav and H. Thomas, Thin monodromy in Sp(4). *Compos. Math.* **150** (2014), no. 3, 333–343 Zbl 1311.14010 MR 3187621

[3] M. Burger, A. Iozzi, F. Labourie, and A. Wienhard, Maximal representations of surface groups: symplectic Anosov structures. *Pure Appl. Math. Q.* **1** (2005), no. 3, Special Issue: In memory of Armand Borel. Part 2, 543–590 Zbl 1157.53025 MR 2201327

[4] P. Candelas, X. C. de la Ossa, P. S. Green, and L. Parkes, A pair of Calabi–Yau manifolds as an exactly soluble superconformal theory. *Nuclear Phys. B* **359** (1991), no. 1, 21–74 Zbl 1098.32506 MR 1115626

[5] B. Collier, N. Tholozan, and J. Toulisse, The geometry of maximal representations of surface groups into $\mathrm{SO}_0(2,n)$. *Duke Math. J.* **168** (2019), no. 15, 2873–2949 Zbl 07145323 MR 4017517

[6] A. Eskin, M. Kontsevich, M. Möller, and A. Zorich, Lower bounds for Lyapunov exponents of flat bundles on curves. *Geom. Topol.* **22** (2018), no. 4, 2299–2338 Zbl 1386.37036 MR 3784522

[7] R. Fedorov, Variations of Hodge structures for hypergeometric differential operators and parabolic Higgs bundles. *Int. Math. Res. Not. IMRN* **2018** (2018), no. 18, 5583–5608 Zbl 1408.32017 MR 3862114

[8] S. Filip, Uniformization of some weight 3 variations of Hodge structure, Anosov representations, and Lyapunov exponents. 2021, arXiv:2110.07533

[9] S. Filip and C. Fougeron, A cyclotomic family of thin hypergeometric monodromy groups in $\mathrm{Sp}_4(\mathbb{R})$. 2021, arXiv:2106.09181

[10] F. Guéritaud, O. Guichard, F. Kassel, and A. Wienhard, Anosov representations and proper actions. *Geom. Topol.* **21** (2017), no. 1, 485–584 Zbl 1373.37095 MR 3608719

[11] O. Guichard and A. Wienhard, Anosov representations: domains of discontinuity and applications. *Invent. Math.* **190** (2012), no. 2, 357–438 Zbl 1270.20049 MR 2981818

[12] M. Kapovich and B. Leeb, Relativizing characterizations of Anosov subgroups, I. 2018, arXiv:1807.00160

[13] M. Kapovich, B. Leeb, and J. Porti, Dynamics on flag manifolds: domains of proper discontinuity and cocompactness. *Geom. Topol.* **22** (2018), no. 1, 157–234 Zbl 1381.53090 MR 3720343

[14] A. W. Knapp, Structure theory of semisimple Lie groups. In *Representation Theory and Automorphic Forms (Edinburgh, 1996)*, pp. 1–27, Proc. Sympos. Pure Math. 61, Amer. Math. Soc., Providence, RI, 1997 Zbl 0902.17003 MR 1476489

[15] F. Labourie, Anosov flows, surface groups and curves in projective space. *Invent. Math.* **165** (2006), no. 1, 51–114 Zbl 1103.32007 MR 2221137

[16] D. Mumford, J. Fogarty, and F. Kirwan, *Geometric Invariant Theory*. 3rd edn., Ergeb. Math. Grenzgeb. (2) 34, Springer, Berlin, 1994 Zbl 0797.14004 MR 1304906

[17] P. Sarnak, Notes on thin matrix groups. In *Thin Groups and Superstrong Approximation*, pp. 343–362, Math. Sci. Res. Inst. Publ. 61, Cambridge Univ. Press, Cambridge, 2014 Zbl 1365.11039 MR 3220897

[18] C. T. Simpson, Harmonic bundles on noncompact curves. *J. Amer. Math. Soc.* **3** (1990), no. 3, 713–770 Zbl 0713.58012 MR 1040197

[19] S. Usui, Generic Torelli theorem for quintic-mirror family. *Proc. Japan Acad. Ser. A Math. Sci.* **84** (2008), no. 8, 143–146 Zbl 1164.14003 MR 2457802

[20] M. Yoshida, *Fuchsian Differential Equations. With Special Emphasis on the Gauss-Schwarz Theory*. Aspects of Mathematics, E11, Friedr. Vieweg & Sohn, Braunschweig, 1987 Zbl 0618.35001 MR 986252

[21] F. Zhu, Relatively dominated representations. *Ann. Inst. Fourier (Grenoble)* **71** (2021), no. 5, 2169–2235 Zbl 07492562 MR 4398259

Simion Filip
Department of Mathematics, University of Chicago, 5734 S University Ave, Chicago,
IL 60637, USA; sfilip@math.uchicago.edu

DOI 10.4171/8ECM/07

On primes, almost primes, and the Möbius function in short intervals

Kaisa Matomäki

Abstract. In this article, aimed at a general mathematical audience, we have three goals. First, we give a brief account of the classical theory connecting primes, the Riemann zeta function, and the Möbius function. Second, we discuss the state-of-art results concerning primes, almost primes, and the Möbius function in short intervals. Third, we outline the most fundamental concepts underlying the proofs of such results.

1. Introduction

Some of the most prominent topics in analytic number theory include the prime numbers, the Riemann zeta function, and the Möbius function. In this article, aimed at a general mathematical audience, we first introduce some classical results on the primes and their relation to the Riemann zeta function in Section 2. Then we go on to discuss primes and almost primes in short intervals in Section 3, starting with classical results and moving on to very recent works. In Section 4 we make a similar journey with the Möbius function. Finally, in Section 5 we discuss the proof strategies, mostly in rather general terms.

2. Primes and the Riemann zeta function

2.1. Primes

We write $\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, \dots\}$ for the set of primes, i.e., natural numbers > 1 that are only divisible by 1 and themselves. The letter p with or without subscripts will always denote a prime.

One of the first theorems concerning primes is that of Euclid (ca. 300 BC), stating that there are infinitely many prime numbers. This can be quickly proved in various ways. The most classical way is to make a counter assumption that only $p_1, \dots, p_k$

2020 Mathematics Subject Classification. Primary 11N37; Secondary 11M06, 11N05, 11N25.
Keywords. Prime numbers, Riemann zeta function, Möbius function, short intervals.

are primes. Then $p_1 \cdots p_k + 1$ is either a new prime or divisible by a new prime which is a contradiction.

By the fundamental theorem of arithmetic every natural number can be uniquely written as a product of primes, e.g., $2021 = 43 \cdot 47$. In other words, primes are like the building blocks of the natural numbers.

By Euclid's theorem there are infinitely many primes, but we have much more precise information. Hadamard and de la Valleé Poussin showed independently in the end of the 19th century (see e.g. [12, Notes to Chapter 12]) that[1]

$$\#\{p \in \mathbb{P} : p \leq x\} = \big(1 + o(1)\big) \int_2^x \frac{dx}{\log x} = \big(1 + o(1)\big) \frac{x}{\log x}.$$

This is called the prime number theorem (PNT) and it sort of asserts that the "probability" that an integer n is prime is about $1/\log n$.

In light of this, it is convenient to normalize prime p by $\log p$. More precisely, we write $\Lambda(n)$ for the von Mangoldt function

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k \text{ with } k \geq 1; \\ 0 & \text{otherwise.} \end{cases}$$

Now the PNT is equivalent to the fact that

$$\sum_{n \leq x} \Lambda(n) = \big(1 + o(1)\big) x.$$

As for the $o(1)$ error term in the PNT, the best result (see e.g. [12, Theorem 12.2]) currently is that

$$\sum_{n \leq x} \Lambda(n) = x + O\bigg(x \exp\bigg(-\frac{c(\log x)^{3/5}}{(\log\log x)^{1/5}}\bigg)\bigg) \tag{2.1}$$

for some absolute constant $c > 0$.

2.2. The Riemann zeta function

Next we introduce some basic properties of the Riemann zeta function. For a reference to the results in this and the following subsection, and much more, see e.g. [12] or [25].

[1] We use, for $f : \mathbb{R} \to \mathbb{C}$ and $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$, the notation $f(x) = O(g(x))$ when there exists a constant $C > 0$ such that $|f(x)| \leq Cg(x)$ for all x and the notation $f(x) = o(g(x))$ when $\lim_{x \to \infty} f(x)/g(x) = 0$. For instance $O(x^{1/2})$ denotes a quantity which is, for some constant $C > 0$, at most $Cx^{1/2}$ for all x and $o(1)$ denotes a quantity tending to 0 when $x \to \infty$.

Write, for $\Re s > 1$,

$$\zeta(s) = \sum_{n \in \mathbb{N}} \frac{1}{n^s} = \prod_{p \in \mathbb{P}} \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots\right) = \prod_{p \in \mathbb{P}} \left(1 - \frac{1}{p^s}\right)^{-1}. \tag{2.2}$$

The function $\zeta(s)$ can be analytically continued to the whole complex plane except for a simple pole at $s = 1$ with residue 1. The function $\zeta(s)$ is called the Riemann zeta function and it satisfies the functional equation

$$\zeta(s) = 2^s \pi^{s-1} \sin\left(\frac{\pi s}{2}\right) \Gamma(1-s)\zeta(1-s), \tag{2.3}$$

where $\Gamma(s)$ is the gamma function.

The functional equation can be used to obtain some basic information about the zeros of the Riemann zeta function. Notice first that on the right-hand side of the functional equation (2.3) the function $\sin(\pi s/2)$ has a zero at each even integer. For $s = 0$ the zero is cancelled by the pole of $\zeta(1-s)$ whereas for positive even integers the poles of $\Gamma(1-s)$ cancel with the zeros. But for negative even integers there are no poles and hence also $\zeta(s)$ has a zero at each negative even integer $-2, -4, -6, \ldots$. These zeros are called the trivial zeros of $\zeta(s)$.

The remaining zeros of $\zeta(s)$ are called non-trivial. From the Euler product (2.2) one sees that there are no zeros with $\Re s > 1$ and thus, by the functional equation (2.3) there are no non-trivial zeros with $\Re s < 0$. Hence all the non-trivial zeros of the zeta function must lie in the critical strip $0 \leq \Re s \leq 1$.

Writing $N(T)$ for the number of non-trivial zeros with $|\Im s| \leq T$, the Riemann-von Mangoldt formula states that

$$N(T) = \frac{T}{2\pi} \log \frac{T}{2\pi} - \frac{T}{2\pi} + O(\log T). \tag{2.4}$$

The famous Riemann hypothesis (RH) asserts that all these non-trivial zeros actually lie on the critical line $\Re s = 1/2$. This has been numerically verified in [21] for all zeros with $|\Im s| \leq 3 \cdot 10^{12}$. Furthermore we know that the exists a constant $c > 0$ such that, for any zero $s = \beta + it$ of $\zeta(s)$ with $|t| \geq 10$, one has

$$\beta \leq 1 - \frac{c}{\left(\log |t|\right)^{2/3} \left(\log \log |t|\right)^{1/3}}; \tag{2.5}$$

the complement of this region is called the Vinogradov–Korobov zero-free region.

2.3. The relation between primes and the Riemann zeta function

It turns out that the non-trivial zeros of the zeta function are closely related to the prime numbers. The relation between von Mangoldt function and the zeros of the zeta function stem from a Dirichlet series identity; for $\Re s > 1$, one has

$$\begin{aligned} -\frac{\zeta'(s)}{\zeta(s)} &= -\frac{d}{ds}\log\zeta(s) = \frac{d}{ds}\log\prod_{p\in\mathbb{P}}\left(1-\frac{1}{p^s}\right) \\ &= \sum_{p\in\mathbb{P}}\frac{d}{ds}\log\left(1-\frac{1}{p^s}\right) = \sum_{p\in\mathbb{P}}\frac{p^{-s}\log p}{1-\frac{1}{p^s}} = \sum_{n=1}^{\infty}\frac{\Lambda(n)}{n^s}. \end{aligned} \tag{2.6}$$

This identity is one of the reasons why it is more convenient to study $\Lambda(n)$ than the characteristic function of the primes.

To utilize (2.6) to study primes in $[1, x]$, one uses the contour integration formula

$$\frac{1}{2\pi i}\int_{2-i\infty}^{2+i\infty}\frac{y^s}{s}\,ds = \begin{cases} 0 & \text{if } y < 1; \\ 1 & \text{if } y > 1. \end{cases} \tag{2.7}$$

Combining these two observations we obtain (when $x \notin \mathbb{N}$)

$$\sum_{n\le x}\Lambda(n) = \sum_{n}\Lambda(n)\frac{1}{2\pi i}\int_{2-i\infty}^{2+i\infty}\frac{(x/n)^s}{s}\,ds = -\frac{1}{2\pi i}\int_{2-i\infty}^{2+i\infty}\frac{\zeta'}{\zeta}(s)\frac{x^s}{s}\,ds.$$

Moving the integration to the left side of the line $\Re s = 1$, one picks up a pole at $s = 1$ with residue $-x$, so this gives the main term in the PNT. The zeros of the zeta function are also poles of the integrand and one can derive, for any $x \ge T \ge 2$, the explicit formula

$$\sum_{n\le x}\Lambda(n) = x - \sum_{\substack{\rho \\ \zeta(\rho)=0 \\ |\Im(\rho)|\le T}}\frac{x^\rho - 1}{\rho} + O\left(\frac{x}{T}\log^2 x\right).$$

One can now use this and (2.4) to relate the error term in the PNT to the zero-free region for the zeta function. In particular, one can show that

$$\text{PNT} \Leftrightarrow \sum_{n\le x}\Lambda(n) = \big(1+o(1)\big)x \Leftrightarrow \zeta(s) \ne 0 \quad \text{when } \Re s = 1.$$

Furthermore, one obtains (2.1) using the zero-free region (2.5). Finally, it is possible to show this way that

$$\text{RH} \Leftrightarrow \sum_{n\le x}\Lambda(n) = x + O(x^{1/2+\varepsilon}) \quad \text{for all } \varepsilon > 0, \tag{2.8}$$

where the implied constant is allowed to depend on ε.

2.4. Dirichlet L-functions

The Riemann zeta function is the simplest member of a large family of L-functions (for a lot of information about general L-functions, see [14, Chapter 5]). Let us introduce here also Dirichlet L-functions, which are L-functions of degree 1 like $\zeta(s)$.

Let $\chi\colon \mathbb{Z} \to \mathbb{C}$ be a Dirichlet character of modulus q; i.e., a function that

(i) is periodic with period q (i.e., $\chi(a+q) = \chi(a)$ for all $a \in \mathbb{Z}$);

(ii) is completely multiplicative (i.e., $\chi(mn) = \chi(m)\chi(n)$ for all $m, n \in \mathbb{Z}$);

(iii) is such that $\chi(r) = 0$ whenever $(r, q) \neq 1$.

For every $q \in \mathbb{N}$, a trivial instance of Dirichlet character is the principal character $\chi_0(n) = 1_{(n,q)=1}$. For modulus 4 the only non-principal character is χ_4 defined at primes by

$$\chi_4(p) = \begin{cases} 1 & \text{if } p \equiv 1 \pmod 4; \\ -1 & \text{if } p \equiv 3 \pmod 4; \\ 0 & \text{if } p = 2. \end{cases}$$

For a Dirichlet character χ, the corresponding Dirichlet L-function $L(s, \chi)$ is defined for $\Re s > 1$ by

$$L(s,\chi) = \sum_{n\in\mathbb{N}} \frac{\chi(n)}{n^s} = \prod_{p\in\mathbb{P}} \left(1 + \frac{\chi(p)}{p^s} + \frac{\chi(p^2)}{p^{2s}} + \cdots\right) = \prod_{p\in\mathbb{P}} \left(1 - \frac{\chi(p)}{p^s}\right)^{-1}.$$

The Dirichlet L-functions play an important role when studying prime numbers in arithmetic progressions thanks to the orthogonality relation

$$\frac{1}{\varphi(q)} \sum_{\chi \pmod q} \chi(m) = \begin{cases} 1 & \text{if } m \equiv 1 \pmod q; \\ 0 & \text{otherwise.} \end{cases}$$

The Dirichlet L-functions have a very similar theory as the Riemann zeta function, with a functional equation, the generalized RH, etc. The zero-free region for $L(s, \chi)$ is not as good as for $\zeta(s)$. In particular, one has not been able to rule out the possibility of a real exceptional character which has a real zero very close to $s = 1$.

3. Primes and almost primes in short intervals

3.1. Primes in short intervals

The PNT tells us about the behavior of primes in $[1, x]$ but even the best known quantitative result (2.1) is so weak that it does not imply that there exists $\varepsilon > 0$ such that, for sufficiently large x, the interval $(x, x + x^{1-\varepsilon}]$ contains primes. However, Hoheisel [10] showed such a statement already in 1930 and Huxley's [11] PNT from 1972 gives, for any $\varepsilon > 0$,

$$\sum_{x<n\le x+H} \Lambda(n) = \big(1 + o(1)\big)H \quad \text{for } H \ge x^{7/12+\varepsilon}. \tag{3.1}$$

This is based on Huxley's [11] zero-density estimate

$$N(\sigma, T) = O\big(T^{(\frac{12}{5}+\varepsilon)(1-\sigma)}(\log T)^{O(1)}\big) \quad \text{for all } T \ge 2 \text{ and } \sigma \in [1/2, 1], \tag{3.2}$$

where $N(\sigma, T)$ is the number of zeros of the Riemann zeta function in the rectangle $\Re(s) \ge \sigma$, $|\Im(s)| \le T$. Huxley's result has resisted improvements, except that Heath-Brown [9] has shown (3.2) for $H \ge x^{7/12-o(1)}$.

The so-called density hypothesis asserts that

$$N(\sigma, T) = O(T^{2-2\sigma+\varepsilon}) \quad \text{for all } T \ge 2 \text{ and } \sigma \in [1/2, 1], \tag{3.3}$$

and this would imply that (3.2) holds for $H \ge x^{1/2+\varepsilon}$ for any $\varepsilon > 0$ (see e.g. [12, Theorem 12.8]). Note that the density hypothesis is a consequence of the Lindelöf hypothesis (see e.g. [12, Section 1.9]) asserting that $|\zeta(1/2 + it)| \ll |t|^\varepsilon$ for every $\varepsilon > 0$.

If one does not require an asymptotic formula for the number of primes in a short interval but contends with a lower bound of correct order of magnitude, then shorter intervals can be reached. In particular, following the initial breakthrough of Iwaniec and Jutila [13] and a succession of further improvements, Baker–Harman–Pintz [1] showed that, for large enough x and some $\varepsilon > 0$,

$$\sum_{x<n\le x+H} \Lambda(n) \ge \varepsilon H \quad \text{for } H \ge x^{0.525}. \tag{3.4}$$

For shorter intervals one does not even know existence of primes. However, assuming the RH one knows that, for large enough x, the interval $(x, x + x^{1/2}\log x]$ always contains primes (see e.g. [12, Theorem 12.10]). This barely falls short of one of the four famous problems of Landau, asserting that there is always a prime between two consecutive squares, which would follow if one could show that $(x, x + x^{1/2}]$ always contains primes.

Cramer made a probabilistic model based on "probability of n being prime is $1/\log n$". Based on this, one expects that, for a large enough C, the interval $(x, x + C\log^2 x]$ contains primes for all large x; for more precise conjectures, see [5,6]. Here we see a large gap between what is known and what is expected.

3.2. Primes in almost all short intervals

As even under the RH it is not known that $(x, x + x^{1/2}]$ always contains primes, it is natural to ask what if one only requires that almost all intervals contain primes.

A variant of Huxley's PNT says that, for almost all $x \in (X, 2X]$,

$$\sum_{x<n\leq x+H} \Lambda(n) = \bigl(1 + o(1)\bigr)H \quad \text{for } H \geq x^{1/6+\varepsilon},$$

see e.g. [7, Theorem 9.1]. This can be proved using a technique due to Selberg [23] and Huxley's zero-density estimate (3.2). Furthermore also this result has resisted improvements.

Again if one only wants a lower bound for the number of primes, one can do better. By a sieve method Jia [15] has shown that, for some $\varepsilon > 0$,

$$\sum_{x<n\leq x+H} \Lambda(n) \geq \varepsilon H \quad \text{for } H \geq x^{1/20}. \tag{3.5}$$

Assuming the density hypothesis (3.3) (or the Lindelöf hypothesis) one can show that, for every $\varepsilon > 0$, almost all intervals $(x, x + x^{\varepsilon}]$ contain asymptotically the expected number of primes (see [12, Theorem 12.9]).

Based on probabilistic models, one expects that, for any $h \to \infty$ with $X \to \infty$, the interval $(x, x + h \log x]$ contains primes for almost all $x \in (X, 2X]$, so again we are far from the expected truth. Heath-Brown [8] has established this conjecture assuming both the RH and the pair correlation conjecture for zeros of $\zeta(s)$ which concerns the distribution of the gaps between the imaginary parts of the zeros.

The author and Jori Merikoski [17] have worked on studying the distribution of primes under the very unlikely assumption that there exist so-called exceptional characters for which the corresponding L-function has a zero extremely close to $s = 1$. If such an exceptional character existed, it would have some very interesting consequences. Concerning primes in short intervals, as a corollary in our work we obtain the following theorem.

Theorem 3.1 (Matomäki–Merikoski [17]). *Let $C \geq 2$. Let χ be a primitive quadratic character modulo $q \geq 2$ and assume that $L(s, \chi)$ has a real zero β_0 such that*

$$\beta_0 = 1 - \frac{1}{\eta \log q}.$$

for some $\eta \geq 10$.

Let $X \in [q^{10}, q^{\eta^{99/100}}]$ and let $2 \leq H \leq X^{1/3}$. Then

$$\int_X^{2X} \Bigl(\sum_{y<n\leq y+H} \Lambda(n) - H\Bigr)^2 dy = O_C\biggl(H^2 X\Bigl(\frac{\log X}{H} + \exp(-C\sqrt{\log \eta})\Bigr)\biggr).$$

This implies that as soon as

$$\eta \to \infty, \quad \frac{H}{\log X} \to \infty, \quad \text{and} \quad q^{10} \le X \le q^{\eta^{99/100}},$$

we get the asymptotic formula

$$\sum_{y<p\le y+H} 1 = \big(1+o(1)\big)\frac{H}{\log y}$$

for almost all $y \in [X, 2X]$.

Note that it is widely believed that such exceptional characters do not exist. But at least our result allows one to assume they do not exist when attacking primes in almost all short intervals.

3.3. Almost primes in short intervals

As discussed above, one expects that, for any $h \to \infty$ with $X \to \infty$, the interval $(x, x + h \log x]$ contains primes for almost all $x \in (X, 2X]$. This being far out of reach, one can ask similar questions about almost-primes, i.e., P_k numbers that have at most k prime factors or E_k numbers that have exactly k prime factors.

Here P_k numbers are easier to deal with since classical sieve methods can be applied. For instance Wu [26] has shown that, for all sufficiently large x, the interval $(x - x^{101/232}, x]$ contains P_2 numbers. This is significantly better than the corresponding result for the primes, where one could not cross the $1/2$ barrier even assuming the RH.

Due to the so-called parity barrier (see e.g. [2, Section 16.4]), classical sieves are unable to distinguish between numbers having an even and an odd number of prime factors. In particular, a sieve can be used to find P_2 numbers but, without additional input, it is impossible to tell whether it found primes or E_2-numbers.

However, E_k numbers are still easier to deal with than the primes, thanks to sums over them having a multilinear structure. Teräväinen has shown that for $k \ge 2$, there exists a constant C_k such that, for almost all $x \in (X, 2X]$, the interval $(x, x + (\log_{k-1} X)^{C_k} \log X]$ contains an E_k-number, where $\log_m X$ is m times iterated logarithm. Furthermore, in Teräväinen's result one can take $C_2 = 2.51$ and $C_3 = 6 + \varepsilon$.

Let us turn into discussing P_k numbers in almost all intervals. Following Friedlander [3,4], Friedlander and Iwaniec [2, Section 6.10] showed that as soon as $h \to \infty$ with $X \to \infty$, the interval $(x - h \log X, x]$ contains P_{19}-numbers for almost all $x \in (X/2, X]$. Furthermore, they say that, using more advanced techniques, one could obtain P_3 numbers. The author improved this in a recent preprint [16].

Theorem 3.2 (Matomäki [16]). *Let $h \to \infty$ with $X \to \infty$. Then the interval $(x - h \log X, x]$ contains P_2 numbers for almost all $x \le X$.*

4. The Möbius function

4.1. Introducing the Möbius function

Let $\mu(n)$ denote the Möbius function

$$\mu(n) = \begin{cases} (-1)^k & \text{if } n = p_1 \cdots p_k \text{ with } p_i \text{ distinct;} \\ 0 & \text{otherwise.} \end{cases}$$

Now, for $\Re s > 1$,

$$\sum_{n=1}^{\infty} \frac{\mu(n)}{n^s} = \prod_{p \in \mathbb{P}} \left(1 - \frac{1}{p^s}\right) = \frac{1}{\zeta(s)},$$

so $\mu(n)$ is closely related to $\Lambda(n)$ whose generating Dirichlet series was $-\zeta'/\zeta(s)$.

In particular, using similar contour integration arguments as in Section 2.3 one can show that

$$\text{PNT} \Leftrightarrow \zeta(s) \text{ has no zeros with } \Re s = 1 \Leftrightarrow \sum_{n \le x} \mu(n) = o(x)$$

$$\text{RH} \Leftrightarrow \sum_{n \le x} \mu(n) = O(x^{1/2+\varepsilon}) \quad \text{for all } \varepsilon > 0,$$

where the implied constant may depend on ε.

4.2. Möbius in short intervals

Until 2014 the story about the Möbius function in short intervals was exactly the same as for $\Lambda(n)$. In particular, Motohashi [20] and Ramachandra [22] independently adapted Huxley's proof of [11] to show that

$$\sum_{x < n \le x+H} \mu(n) = o(H) \quad \text{for } H \ge x^{7/12+\varepsilon}. \tag{4.1}$$

Analogously it was known by [22] that, for almost all $x \in (X, 2X]$,

$$\sum_{x < n \le x+H} \mu(n) = o(H) \quad \text{for } H \ge x^{1/6+\varepsilon}.$$

This almost-all interval result was significantly improved in the author's work with Radziwiłł [18] showing the following theorem.

Theorem 4.1 (Matomäki–Radziwiłł [18]). *Let $H \to \infty$ with $x \to \infty$. Then, for almost all $x \in (X, 2X]$, one has*

$$\sum_{x < n \le x+H} \mu(n) = o(H).$$

Our result is more general and has led to numerous advancements, e.g., concerning Chowla's conjecture (see e.g. [24]). In the proof we crucially used the fact that a typical n has prime factors from certain convenient intervals—something that is certainly not true for $n \in \mathbb{P}$.

A natural question is whether one can improve also on (4.1) along similar lines. Recently, the author and J. Teräväinen [19] obtained such a result.

Theorem 4.2 (Matomäki–Teräväinen [19]). *One has*

$$\sum_{x<n\le x+H} \mu(n) = o(H) \quad \textit{for } H \ge x^{0.55+\varepsilon}.$$

Note that $7/12 = 0.5833\ldots$, and that even under RH one cannot get beyond $1/2$, so we get significantly closer to this natural barrier.

5. Proof strategy

5.1. The general strategy

We have already discussed how contour integration can be used to relate questions about primes and the Möbius function to questions about the Riemann zeta function. However, there is another more flexible way to go which we will describe in this section.

In this strategy for proving results on primes or the Möbius function there are two steps: a combinatorial step and an analytic step. In the combinatorial step a combinatorial identity or a sieve method is used to reduce the problem to that of estimating so-called type I and type II sums. In the analytic step these type I and type II sums are estimated.

This overall strategy works for various problems concerning primes, including problems for which no other strategy is known. On the other hand, it can also be used e.g. to reprove Huxley's PNT (3.1) without appealing to zero density results; see e.g. [7, Section 7.3].

5.2. The combinatorial step

When one is looking for an asymptotic formula for the number of primes in some interesting set, the combinatorial step is often done using Vaughan's identity or Heath-Brown's identity (see [14, Sections 13.3–13.4]). A special case of Vaughan's identity (see e.g. [14, Proposition 13.4]) implies that, for any (α_n),

$$\sum_{X<n\le 2X} \alpha_n \Lambda(n) = \sum_{\substack{X<bc\le 2X \\ b\le X^{1/3}}} \alpha_{bc}\mu(b)\log c$$

$$- \sum_{\substack{X<abc\le 2X \\ b,c\le X^{1/3}}} \alpha_{abc}\mu(b)\Lambda(c) + \sum_{\substack{X<abc\le 2X \\ b,c> X^{1/3}}} \alpha_{abc}\mu(b)\Lambda(c).$$

From this one can see that instead of $\sum_{X<n\le 2X} \alpha_n \Lambda(n)$ it suffices to study type I sums

$$\sum_{\substack{X<mn\le 2X \\ m\le X^{1/3}}} \alpha_{mn} a_m \quad \text{and} \quad \sum_{\substack{X<mn\le 2X \\ m\le X^{1/3}}} \alpha_{mn} a_m \log n \tag{5.1}$$

with arbitrary bounded coefficients a_m and type II sums

$$\sum_{\substack{X<mn\le 2X \\ X^{1/3}\le m\le X^{2/3}}} \alpha_{mn} a_m b_n$$

with arbitrary bounded coefficients a_m and b_n.

Heath-Brown's identity is a more flexible variant of Vaughan's identity in terms of the different sums it produces, and it is of benefit to be able to deal e.g. with type I_2 sums

$$\sum_{\substack{X<\ell mn\le 2X \\ m\sim M \\ n\sim N}} \alpha_{\ell mn} a_\ell.$$

5.3. The analytic step

In the analytic step one estimates the resulting type I and type II sums. In type I sums (5.1) there is a smooth variable n and one often wants to bring the sum over n inside. For instance if $\alpha_n = 1_{n\in(X,X+X^{3/4}]}$, then

$$\sum_{\substack{X<mn\le 2X \\ m\le X^{1/3}}} \alpha_{mn} a_m = \sum_{m\le X^{1/3}} a_m \sum_{X/m<n\le (X+X^{3/4})/m} 1 = X^{3/4} \sum_{m\le X^{1/3}} \frac{a_m}{m} + O(X^{1/3}),$$

so that we get an asymptotic formula for such a type I sums.

In type II sums we have genuine bilinear structure and quite often one applies Cauchy–Schwarz at some point, either to separate the variables or to dispose of some of the coefficients.

For instance when working on problems concerning short intervals, one can use Dirichlet polynomials through contour integration (2.7). One gets that

$$\frac{1}{H} \sum_{x<mn\le x+H} a_m b_n \approx \frac{1}{X} \sum_{X<mn\le 2X} a_m b_n$$

essentially if

$$\int_{(\log X)^{100}}^{X/H} \left| \sum_{mn\sim X} \frac{a_m b_n}{(mn)^{1/2+it}} \right| dt = O\Big(\frac{X^{1/2}}{(\log X)^{100}}\Big).$$

Such mean values can be estimated through mean and large value results for Dirichlet polynomials; see e.g. [7, Chapter 7].

5.4. Sieve methods

If one does not require an asymptotic formula, one can use a sieve method. The most popular prime-detecting sieve is Harman's sieve (for a comprehensive account, see [7]) that has been used e.g. in proofs of (3.4) and (3.5).

For $\mathcal{A} \subset \mathbb{N}$ and $z \geq 2$, write $P(z) = \prod_{p<z} p$ and

$$S(\mathcal{A}, z) = \sum_{\substack{n \in \mathcal{A} \\ (n, P(z))=1}} 1.$$

If now $\mathcal{A} \subseteq (X, 2X] \cap \mathbb{N}$, then

$$\mathcal{A} \cap \mathbb{P} = S(\mathcal{A}, 2X^{1/2}).$$

Writing also $\mathcal{A}_d = \{n \in \mathcal{A} : d \mid n\}$, one has the Buchstab identity

$$S(\mathcal{A}, z) = S(\mathcal{A}, w) - \sum_{w \leq p < z} S(\mathcal{A}_p, p).$$

Harman's sieve method consists of consecutive applications of Buchstab's identity to reach type I and type II sums. Some sums with a positive sign can be dropped if one looks for a lower bound.

When one is looking for P_k numbers, one can use more classical sieve methods that require only type I information. For instance in a lower bound sieve one replaces the identity

$$S(\mathcal{A}, z) = \sum_{\substack{n \in \mathcal{A} \\ (n, P(z))=1}} 1 = \sum_{n \in \mathcal{A}} \sum_{d \mid (n, P(z))} \mu(d)$$

by an inequality

$$S(\mathcal{A}, z) = \sum_{\substack{n \in \mathcal{A} \\ (n, P(z))=1}} 1 \geq \sum_{n \in \mathcal{A}} \sum_{d \mid (n, P(z))} \mu^-(d)$$

for an appropriate chosen sequence $\mu^-(d)$ which is supported only on $d \leq D$. Now one encounters a type I sum

$$\sum_{\substack{dn \in \mathcal{A} \\ d \mid P(z) \\ d \leq D}} \mu^-(d).$$

Unfortunately, such a sieve can produce a non-trivial lower bound only when $D > z^2$.

5.5. Implementation of the strategy

In this subsection we briefly discuss the combinatorial and analytic steps in the proofs of Theorems 3.1, 3.2, 4.1, and 4.2.

In the proof of Theorem 3.2 on P_2 numbers in almost all short intervals, the combinatorial tool used is Richert's weighted sieve with β-sieve (for a comprehensive account of these sieves, see [2, Chapters 25 and 11]). These sieves reduce the problem to understanding type I sums. As mentioned in Section 5.4, sieves using only type I information as input are incapable of catching primes, but here our goal is P_2 numbers. Then in the analytic step we reduce estimating the type I sums in almost all intervals into estimating averages of Kloosterman sums which can be done by known results.

Let us now turn to the proof of Theorem 4.1 on the Möbius function for almost all intervals. The combinatorial step uses Ramaré's identity in the form saying that, for $(P, Q] \subseteq (1, H]$, one has

$$\sum_{x<n\leq x+H} \mu(n) = -\sum_{\substack{x<pm\leq x+H \\ P<p\leq Q}} \frac{\mu(pm)}{\#\{P<q\leq Q : q \mid m\} + 1_{p\nmid m}} + O\left(H\frac{\log P}{\log Q}\right),$$

where the error term comes from those n that do not have a prime factor in the interval $(P, Q]$. This combinatorial step leads to type II sums with one of the variables (i.e., p) being very small. In the analytic step we reduce estimating such sums to mean square estimates for Dirichlet polynomials. In order to reach very short intervals, we need to use an iterative argument, with several applications of Ramaré's identity.

In the proof of Theorem 4.2 on the Möbius function in all short intervals, in the combinatorial step we use both Ramaré's identity and Heath-Brown's identity. Ramaré's identity allows us to extract a very small prime factor from the sum over $\mu(n)$ before using the Heath-Brown identity to split into type I, type II, and type I/II sums. In the analytic step we again use estimates on Dirichlet polynomials. This method actually works in greater generality. For instance we obtain also the following theorem.

Theorem 5.1 (Matomäki–Teräväinen [19]).

$$\sum_{\substack{x<p_1p_2\leq x+H \\ p_j\in\mathbb{P}}} 1 = H\frac{\log\log x}{\log x} + O\left(H\frac{\log\log\log x}{\log x}\right), \quad H \geq x^{0.55+\varepsilon}.$$

The proof of Theorem 3.1 works somewhat differently though there are similar steps. Thanks to the assumption on the existence of exceptional characters, the relevant type II sums become quite easy to bound and then one just needs to obtain enough type I information to find primes. In the analytic step for the type I sums one again reduces the problem to that of bounding Kloosterman sums.

Acknowledgments. The author would like to thank Juho Salmensuu and the referee for helpful comments on the manuscript.

Funding. Research supported by the Academy of Finland grant no. 285894.

References

[1] R. C. Baker, G. Harman, and J. Pintz, The difference between consecutive primes. II. *Proc. London Math. Soc. (3)* **83** (2001), no. 3, 532–562 Zbl 1016.11037 MR 1851081

[2] J. Friedlander and H. Iwaniec, *Opera de cribro*. Amer. Math. Soc. Colloq. Publ. 57, American Mathematical Society, Providence, RI, 2010 Zbl 1226.11099 MR 2647984

[3] J. B. Friedlander, Sifting short intervals. *Math. Proc. Cambridge Philos. Soc.* **91** (1982), no. 1, 9–15 Zbl 0477.10036 MR 633251

[4] J. B. Friedlander, Sifting short intervals. II. *Math. Proc. Cambridge Philos. Soc.* **92** (1982), no. 3, 381–384 Zbl 0503.10032 MR 677462

[5] A. Granville, Harald Cramér and the distribution of prime numbers. *Scand. Actuar. J.* **1995** (1995), no. 1, 12–28 Zbl 0833.01018 MR 1349149

[6] A. Granville and A. Lumley, Primes in short intervals: Heuristics and calculations. 2020, arXiv:2009.05000

[7] G. Harman, *Prime-Detecting Sieves*. London Math. Soc. Monogr. Ser. 33, Princeton University Press, Princeton, NJ, 2007 Zbl 1220.11118 MR 2331072

[8] D. R. Heath-Brown, Gaps between primes, and the pair correlation of zeros of the zeta function. *Acta Arith.* **41** (1982), no. 1, 85–99 Zbl 0414.10044 MR 667711

[9] D. R. Heath-Brown, The number of primes in a short interval. *J. Reine Angew. Math.* **389** (1988), 22–63 Zbl 0646.10032 MR 953665

[10] G. Hoheisel, Primzahlprobleme in der Analysis. *Sitzungsber. Preuß. Akad. Wiss. Phys.-Math. Kl.* **1930** (1930), 580–588 Zbl 56.0172.02

[11] M. N. Huxley, On the difference between consecutive primes. *Invent. Math.* **15** (1972), 164–170 Zbl 0241.10026 MR 292774

[12] A. Ivić, *The Riemann Zeta-Function. Theory and Applications*. Dover Publications, Mineola, NY, 2003 Zbl 1034.11046 MR 1994094

[13] H. Iwaniec and M. Jutila, Primes in short intervals. *Ark. Mat.* **17** (1979), no. 1, 167–176 Zbl 0408.10029 MR 543511

[14] H. Iwaniec and E. Kowalski, *Analytic Number Theory*. Amer. Math. Soc. Colloq. Publ. 53, American Mathematical Society, Providence, RI, 2004 Zbl 1059.11001 MR 2061214

[15] C. Jia, Almost all short intervals containing prime numbers. *Acta Arith.* **76** (1996), no. 1, 21–84 Zbl 0841.11043 MR 1390568

[16] K. Matomäki, Almost primes in almost all very short intervals. *J. Lond. Math. Soc. (2)* **106** (2022), no. 2, 1061–1097 MR 4477211

[17] K. Matomäki and J. Merikoski, Siegel zeros, twin primes, Goldbach's conjecture, and primes in short intervals. 2022, arXiv:2112.11412

[18] K. Matomäki and M. Radziwiłł, Multiplicative functions in short intervals. *Ann. of Math. (2)* **183** (2016), no. 3, 1015–1056 Zbl 1339.11084 MR 3488742

[19] K. Matomäki and J. Teräväinen, On the Möbius function in all short intervals. *J. Eur. Math. Soc. (JEMS)* **25** (2023), no. 4, 1207–1225 Zbl 07683509 MR 4577962

[20] Y. Motohashi, On the sum of the Möbius function in a short segment. *Proc. Japan Acad.* **52** (1976), no. 9, 477–479 Zbl 0372.10033 MR 424726

[21] D. Platt and T. Trudgian, The Riemann hypothesis is true up to $3 \cdot 10^{12}$. *Bull. Lond. Math. Soc.* **53** (2021), no. 3, 792–797 Zbl 07381909 MR 4275089

[22] K. Ramachandra, Some problems of analytic number theory. *Acta Arith.* **31** (1976), no. 4, 313–324 Zbl 0291.10034 MR 424723

[23] A. Selberg, On the normal density of primes in small intervals, and the difference between consecutive primes. *Arch. Math. Naturvid.* **47** (1943), no. 6, 87–105 Zbl 0063.06869 MR 12624

[24] T. Tao, The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi* **4** (2016), e8, 36 Zbl 1383.11116 MR 3569059

[25] E. C. Titchmarsh, *The Theory of the Riemann Zeta-Function*. 2nd edn., The Clarendon Press, New York, 1986 Zbl 0601.10026 MR 882550

[26] J. Wu, Almost primes in short intervals. *Sci. China Math.* **53** (2010), no. 9, 2511–2524 Zbl 1221.11196 MR 2718844

Kaisa Matomäki
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland;
ksmato@utu.fi

DOI 10.4171/8ECM/04

Bogoliubov excitation spectrum of Bose gases

Phan Thành Nam

Abstract. We review some rigorous results on the derivation of Bogoliubov excitation spectrum of interacting Bose gases from many-body Schrödinger equations.

1. Introduction

The Bose–Einstein condensation (BEC) has been an important topic in quantum physics for a long time since the first predictions in 1924 [11, 24], and especially after the experimental observations in 1995 [2, 19]. Roughly speaking, BEC is the phenomenon when many bosons occupy a common quantum state at very low temperatures, thus allowing to observe in our macroscopic scales many interesting quantum phenomena such as superfluidity and quantized vortices.

While the pioneer works of Bose and Einstein [11, 24] concern only the *non-interacting* gas, in reality the particles do interact and the rigorous understanding of *interacting* systems remains a very challenging problem in mathematical physics. The theory of interacting Bose gases essentially started in 1947 when Bogoliubov [10] proposed an approximation theory and used it to predict the excitation spectrum of Bose gases. In particular, Bogoliubov's theory gives a satisfactory explanation of Landau's criterion for superfluidity [32]. Since then, there have been several attempts to justify Bogoliubov's theory from first principles, namely from many-body Schrödinger equations, and some rigorous results will be reviewed below.

Heuristically, Bogoliubov's theory based on the key assumption that *the interaction is sufficiently weak*. In this case, the total interaction felt by each particle can be effectively replaced by a one-body mean-field potential, in the spirit of the law of large number in probability theory. This so-called *mean-field approximation* leads to Hartree's theory (or the Gross–Pitaevskii theory) which has been used widely to study the condensate. Moreover, the weak interaction ansatz also allows to treat excited particles by the second-order perturbation method. Consequently, Bogoliubov's theory

2020 Mathematics Subject Classification. 81V73.
Keywords. Bose gases, Bose–Einstein condensation, Bogoliubov's theory.

gives an effective description for the fluctuations around the condensate, as some sort of the central limit theorem.

In this review, we will focus on two specific scaling regimes where the interactions are weak but still play a leading order role.

- The *mean-field regime*: the interaction range is long, but the interaction strength is weak. Thus there are many but weak collisions, which is an ideal situation to apply the mean-field approximation.
- The *Gross–Pitaevskii regime*: the interaction range is short, but the interaction strength is strong. Thus there are few but strong collisions, making the mean-field behavior less obvious.

Although the mean-field and Gross–Pitaevskii regimes correspond to different physical systems, it turns out that Bogoliubov's arguments apply successfully to both cases. In fact, thanks to a series of works by many authors in the last 10 years, the validity of Bogoliubov excitation spectrum has been proved in both regimes. In 2011, Seiringer [55] for the first time justified the Bogoliubov excitation spectrum in the mean-field regime for the homogeneous Bose gas in the torus $\mathbb{T}^3$. Later his result was extended to general trapped systems in $\mathbb{R}^3$ in [30,36]; see also [8,13,21,46,49,52,54] for various extensions. On the other hand, in the Gross–Pitaevskii regime, which is most relevant to the physical setup in [2, 19], the analysis is significantly more challenging since Bogoliubov's theory admits a subtle correction. The correction to Bogoliubov's theory in the Gross–Pitaevskii regime was established by Boccato, Brennecke, Cenatiempo, and Schlein [7] for the homogeneous gas. Very recently, this result was finally extended to general trapped systems in $\mathbb{R}^3$ in [17,50].

In the following, I will explain in detail Bogoliubov's theory and review the results obtained in [36, 50]. I will also discuss some possible extensions and open problems in the end.

2. Bogoliubov's theory

To make the idea transparent, let us start with a trapped system in the mean-field regime. We consider a system of N bosons in $\mathbb{R}^3$ described by the Hamiltonian

$$H_N = \sum_{i=1}^{N} \big(-\Delta_{x_i} + V_{\mathrm{ext}}(x_i) \big) + \frac{1}{N-1} \sum_{1 \le i < j \le N} W(x_i - x_j) \tag{2.1}$$

which acts on the symmetric space $\mathfrak{H}^N = \bigotimes_{\mathrm{sym}}^N L^2(\mathbb{R}^3)$. Here $x_i \in \mathbb{R}^3$ stands for the coordinate of the ith particle (we ignore the spin for simplicity) and $\mathfrak{H}^N$ consists of functions in $L^2((\mathbb{R}^3)^N)$ satisfying

$$\Psi(x_1, \dots, x_N) = \Psi(x_{\tau(1)}, \dots, x_{\tau(N)}), \quad \forall \tau \in S_N.$$

We assume that the external potential $V_{\text{ext}} : \mathbb{R}^3 \to \mathbb{R}$ satisfies

$$(V_{\text{ext}})_- \in L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3), \quad (V_{\text{ext}})_+ \in L^1_{\text{loc}}(\mathbb{R}^3), \quad \lim_{|x|\to\infty} V_{\text{ext}}(x) = \infty \quad (2.2)$$

and that the interaction potential $W : \mathbb{R}^3 \to \mathbb{R}$ satisfies

$$W^2 \in L^{3/2}(\mathbb{R}^3) + L^\infty(\mathbb{R}^3). \tag{2.3}$$

Under these conditions, H_N is well defined on the core domain $\bigotimes_{\text{sym}}^N C_c^\infty(\mathbb{R}^3)$ and it is bounded from below. Consequently, H_N can be extended to be a self-adjoint operator on $\mathfrak{H}^N$ by Friedrichs' method. The trapping condition $\lim_{|x|\to\infty} V_{\text{ext}}(x) = \infty$ ensures that H_N has a compact resolvent, and hence it has eigenvalues

$$\lambda_1(H_N) \le \lambda_2(H_N) \le \cdots, \quad \lim_{j\to\infty} \lambda_j(H_N) = \infty.$$

We are interested in the asymptotic behavior of the eigenvalues of H_N when $N \to \infty$.

In the non-interacting gas, namely $W = 0$, the spectrum of H_N can be computed explicitly from the spectrum of the one-body operator $-\Delta + V_{\text{ext}}$ as follows:

$$\sigma(H_N) = \left\{ \sum_{i\ge 1} n_i e_i \mid e_i \in \sigma(-\Delta + V_{\text{ext}}),\ n_i \in \{0,1,2,\ldots\},\ \sum_{i\ge1} n_i = N \right\}.$$

On the other hand, for the interacting gas, namely $W \neq 0$, it is in general impossible to compute the spectrum of H_N when N becomes large, even numerically. Therefore, it is important to derive effective theories, which are less precise (describing only some collective properties of the system) but easier to deal with.

One of the most popular approximation methods used in computational quantum physics and chemistry is the *mean-field approximation*, which was first introduced by Curie and Weiss to describe phase transitions in statistical mechanics. Heuristically, the mean-field theory is based on the assumption that the particles are *independent*, leading to a replacement of the *linear* problem of N particles by a *non-linear* problem of one particle. Mathematically, N independent and identical particles can be described by the Hartree state

$$\Psi(x_1,\ldots,x_N) = u^{\otimes N}(x_1,\ldots,x_N) = u(x_1)\cdots u(x_N),$$

where u is a normalized function in $L^2(\mathbb{R}^3)$. The energy per particle of the factorized wave function $u^{\otimes N}$ is given by the *Hartree functional*

$$\begin{aligned}\mathcal{E}_{\text{H}}(u) = &\int_{\mathbb{R}^3} \left(\left|\nabla u(x)\right|^2 + V_{\text{ext}}(x)\left|u(x)\right|^2\right) \mathrm{d}x \\ &+ \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \left|u(x)\right|^2 \left|u(y)\right|^2 W(x-y)\, \mathrm{d}x\, \mathrm{d}y.\end{aligned}$$

In Hartree's theory, the lowest energy per particle is

$$e_{\mathrm{H}} = \inf_{\|u\|_{L^2(\mathbb{R}^3)}=1} \mathcal{E}_{\mathrm{H}}(u).$$

It is not difficult to show that e_{H} has a minimizer u_0 which is non-negative and solves the self-consistent equation

$$Du_0 = 0, \quad D = -\Delta + V_{\mathrm{ext}} + |u_0|^2 * W - \varepsilon_0, \tag{2.4}$$

where ε_0 is a real constant (it is the Lagrange multiplier associated with the mass constraint $\|u\|_{L^2(\mathbb{R}^3)} = 1$). Thus the mean-field approximation suggests that the ground state energy of H_N in (3.7) satisfies

$$E_N = Ne_{\mathrm{H}} + o(N)_{N\to\infty} \tag{2.5}$$

and that u_0 describes the Bose–Einstein condensate. We refer to [34] and the reviews [33, 53] for rigorous results on the validity of Hartree's theory.

In this review, we are interested in the next order correction to Hartree's theory, which is given by Bogoliubov's theory. We will give below two different heuristic derivations of Bogoliubov's theory: the first is obtained by applying the second-order perturbation method to the Hartree functional, and the second is obtained by manipulating the many-body Hamiltonian in the second quantization language. While the first is shorter and easier to access for a general audience, the second is closer to Bogoliubov's original argument [10] and easier to justify mathematically.

2.1. Bogoliubov's theory from the second-order perturbation

To describe the excited particles, namely the particles outside of the condensate, we can apply the second-order perturbation method to the Hartree functional. More precisely, if u_0 is a Hartree minimizer, then for $v \perp u_0$ we have the Taylor expansion

$$\mathcal{E}_{\mathrm{H}}\left(\frac{u_0+v}{\sqrt{1+\|v\|_{L^2}^2}}\right) = e_{\mathrm{H}} + \frac{1}{2}\left\langle \begin{pmatrix} v \\ v \end{pmatrix}, \mathcal{E}_{\mathrm{H}}''(u_0) \begin{pmatrix} v \\ v \end{pmatrix} \right\rangle + o\big(\|v\|_{H^1(\mathbb{R}^3)}^2\big) \tag{2.6}$$

with the Hessian matrix

$$\mathcal{E}_{\mathrm{H}}''(u_0) = \begin{pmatrix} D+K & K \\ K & D+K \end{pmatrix},$$

where K is the operator on $L^2(\mathbb{R}^3)$ with kernel

$$K(x,y) = u_0(x)u_0(y)w(x-y).$$

Roughly speaking, Bogoliubov's theory suggests that we may lift the Taylor expansion (2.6) to the many-body level, leading to the following refinement of (2.5):

$$\sigma(H_N) = Ne_{\mathrm{H}} + \sigma(\mathbb{H}_{\mathrm{Bog}}) + o(1)_{N\to\infty}, \tag{2.7}$$

where the Bogoliubov Hamiltonian $\mathbb{H}_{\mathrm{Bog}}$ is the *second quantization* of $\frac{1}{2}\mathcal{E}''_{\mathrm{H}}(\varphi)$ that we will introduce later.

Note that we always have $\mathcal{E}''_{\mathrm{H}}(\varphi) \geq 0$ since u_0 is a Hartree minimizer (in particular $D \geq 0$ and u_0 is a ground state of D). Moreover, it is known that if the Hessian matrix is non-degenerate, namely

$$\mathcal{E}''_{\mathrm{H}}(\varphi) \geq \eta > 0 \quad \text{on } \mathfrak{H}_+ \oplus \mathfrak{H}_+ \tag{2.8}$$

with $\mathfrak{H}_+ = \{u_0\}^\perp \subset L^2(\mathbb{R}^3)$ and a constant $\eta > 0$, then it can be diagonalized by a symplectic matrix of the form

$$\mathcal{V} = \begin{pmatrix} \sqrt{1+s^2} & s \\ s & \sqrt{1+s^2} \end{pmatrix}, \quad \mathcal{V}^* \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \mathcal{V} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \tag{2.9}$$

namely

$$\mathcal{V}^* \mathcal{E}''_{\mathrm{H}}(\varphi) \mathcal{V} = \begin{pmatrix} E_\infty & 0 \\ 0 & E_\infty \end{pmatrix}, \tag{2.10}$$

where E_∞ is unitarily equivalent to $(D^{1/2}(D+2K)D^{1/2})^{1/2}$. Consequently, up to a constant, the Bogoliubov Hamiltonian $\mathbb{H}_{\mathrm{Bog}}$ is unitarily equivalent to $\mathrm{d}\Gamma(E_\infty)$, the quantization of E_∞ (see (2.14) below). We refer to [20, 48] for general discussions on the diagonalization procedure, in particular for the emergence of the symplectic structure in (2.9). In summary, (2.8) implies that the excitation spectrum of H_N can be described by the spectrum of E_∞ as follows:

$$\begin{aligned} \sigma(H_N) - \lambda_1(H_N) &\approx \sigma\big(\mathrm{d}\Gamma(E_\infty)\big) \\ &= \Big\{ \sum_{i\geq 1} n_i e_i \mid e_i \in \sigma(E_\infty),\ n_i \in \{0,1,\dots\} \Big\}. \end{aligned} \tag{2.11}$$

2.2. Bogoliubov's theory from the microscopic equation

Now we explain Bogoliubov's theory from the microscopic description of the many-body system, which is closer to the original argument in [10].

Let us recall the Fock space formalism. Let $\mathfrak{K}$ be $L^2(\mathbb{R}^3)$ or a subspace of $L^2(\mathbb{R}^3)$. We define the bosonic Fock space

$$\mathcal{F}(\mathfrak{K}) = \bigoplus_{n=0}^{\infty} \mathfrak{K}^n, \quad \mathfrak{K}^n = \bigotimes_{\mathrm{sym}}^{n} \mathfrak{K}. \tag{2.12}$$

For $g \in \mathfrak{K}$, we define the creation and annihilation operators $a^*(g)$, $a(g)$ on $\mathcal{F}(\mathfrak{K})$ by

$$\big(a^*(g)\Psi\big)(x_1,\dots,x_{n+1}) = \frac{1}{\sqrt{n+1}}\sum_{j=1}^{n+1} g(x_j)\Psi(x_1,\dots,x_{j-1},x_{j+1},\dots,x_{n+1}),$$
$$\big(a(g)\Psi\big)(x_1,\dots,x_{n-1}) = \sqrt{n}\int_{\mathbb{R}^3}\overline{g(x_n)}\Psi(x_1,\dots,x_n)\,\mathrm{d}x_n, \quad \forall\Psi\in\mathfrak{K}^n,\ \forall n.$$

It is also convenient to define the operator-valued distributions

$$a_x^* = \sum_{n=1}^{\infty}\overline{f_n(x)}a^*(f_n), \quad a_x = \sum_{n=1}^{\infty} f_n(x)a(f_n), \quad x\in\mathbb{R}^3,$$

where $\{f_n\}_{n=1}^{\infty}$ is an orthonormal basis of $\mathfrak{K}$ (the definitions of a_x, a_x^* are independent of the choice of the basis). Equivalently, we have

$$a^*(g) = \int_{\mathbb{R}^3} g(x)a_x^*\,\mathrm{d}x, \quad a(g) = \int_{\mathbb{R}^3}\overline{g(x)}a_x\,\mathrm{d}x, \quad \forall g\in\mathfrak{K}.$$

These operators satisfy the canonical commutation relations (CCR)

$$\begin{gathered}\big[a(g_1),a(g_2)\big]=\big[a^*(g_1),a^*(g_2)\big]=0, \quad \big[a(g_1),a^*(g_2)\big]=\langle g_1,g_2\rangle, \quad \forall g_1,g_2\in\mathfrak{K},\\ [a_x^*,a_y^*]=[a_x,a_y]=0, \quad [a_x,a_y^*]=\delta(x-y), \quad \forall x,y\in\mathbb{R}^3. \end{gathered}\tag{2.13}$$

It turns out that many important operators on Fock space can be expressed in the second quantization form using the creation and annihilation operators. For example, for any one-body self-adjoint operator A we can write its second quantization as

$$\mathrm{d}\Gamma(A) := \bigoplus_{n=0}^{\infty}\Big(\sum_{i=1}^{n} A_{x_i}\Big) = \iint_{\mathbb{R}^3} A(x,y)a_x^*a_y\,\mathrm{d}x\,\mathrm{d}y, \tag{2.14}$$

where $A(x,y)$ is the kernel of A. Similarly, the Hamiltonian in (2.1) can be extended to be an operator on $\mathcal{F}(L^2(\mathbb{R}^3))$ as

$$\begin{aligned}H_N = &\int_{\mathbb{R}^3} a_x^*\big(-\Delta_x + V_{\text{ext}}(x)\big)a_x\,\mathrm{d}x\\ &+\frac{1}{2N}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} W(x-y)a_x^*a_y^*a_xa_y\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}\tag{2.15}$$

Roughly speaking, Bogoliubov's theory [10] contains three key steps.

Step 1 (c-number substitution). From the assumption on the complete condensation on the Hartree minimizer u_0, namely

$$\big\langle\Psi_N, a^*(u_0)a(u_0)\Psi_N\big\rangle = N + o(N), \tag{2.16}$$

and the commutation relation

$$\big[a(u_0), a^*(u_0)\big] = 1 \ll \big\langle \Psi_N, a^*(u_0)a(u_0)\Psi_N\big\rangle = N_0 \approx N,$$

we see that $a(u_0)$ and $a^*(u_0)$ "mostly commute." Pushing this idea further, we may heuristically think of $a(u_0)$ and $a^*(u_0)$ as the scalar number $N_0^{1/2}$. Put differently, we may factor out the contribution of the condensate as a scalar field as

$$a_x \approx N_0^{1/2} u_0(x) + c_x, \tag{2.17}$$

where a_x, c_x are annihilation operators on $\mathcal{F}(\mathfrak{H})$, $\mathcal{F}(\mathfrak{H}_+)$, respectively, where $\mathfrak{H} = L^2(\mathbb{R}^3)$ and $\mathfrak{H}_+ = \{u_0\}^\perp \subset \mathfrak{H}$. This allows us to focus on the Fock space $\mathcal{F}(\mathfrak{H}_+)$ which corresponds to excited particles.

Step 2 (Quadratic reduction). Inserting (2.17) in (2.15) and expanding to second order, we obtain

$$H_N \approx N e_{\mathrm{H}} + \mathbb{H}_{\mathrm{Bog}} + o(1)_{N\to\infty}, \tag{2.18}$$

where

$$\begin{aligned}\mathbb{H}_{\mathrm{Bog}} &= \mathrm{d}\Gamma(D)\\ &\quad + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} W(x-y)u_0(x)u_0(y)(2c_x^*c_y + c_x^*c_y^* + c_xc_y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned} \tag{2.19}$$

Here we have ignored all terms containing more than 2 operators c_x or c_x^* thanks to the BEC (heuristically $c_x \ll N^{1/2} \approx N_0^{1/2}$). Moreover, the terms containing only one operator c_x or c_x^* are canceled due to the Hartree's equation (2.4).

Note that the Bogoliubov Hamiltonian in (2.19) can be rewritten as

$$\mathbb{H}_{\mathrm{Bog}} = \int_{\mathbb{R}^3} c_x^*(D+K)_x c_x\,\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} K(x,y)(c_x^*c_y^* + c_xc_y)\,\mathrm{d}x\,\mathrm{d}y$$

which is exactly the second quantized version of the Hessian energy

$$\begin{aligned}\frac{1}{2}\left\langle \begin{pmatrix} v\\ \overline{v}\end{pmatrix}, \mathcal{E}_{\mathrm{H}}''(\varphi)\begin{pmatrix} v\\ \overline{v}\end{pmatrix}\right\rangle &= \int \overline{v(x)}(D+K)v(x)\,\mathrm{d}x\\ &\quad + \frac{1}{2}\iint K(x,y)\big(v(x)v(y) + \overline{v(x)v(y)}\big)\,\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

via the simple rules $\overline{v(x)} \mapsto a_x^*$, $v(x) \mapsto a_x$.

Step 3 (Diagonalization). The Bogoliubov Hamiltonian $\mathbb{H}_{\mathrm{Bog}}$ in (2.18) can be diagonalized by a unitary operator on $\mathcal{F}(\mathfrak{H}_+)$ of the form

$$T = \exp\left(\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \big(k(x,y)c_x^*c_y^* - \mathrm{h.c.}\big)\,\mathrm{d}x\,\mathrm{d}y\right)$$

with an appropriate kernel $k(x, y)$. The actions of T are characterized by

$$
\begin{aligned}
T^* c(v) T &= c(\sqrt{1+s^2} v) + c^*(sv), \\
T^* c^*(v) T &= c^*(\sqrt{1+s^2} v) + c(sv), \quad \forall v \in \mathfrak{H}_+,
\end{aligned}
$$

where

$$
s = \operatorname{sh}(k) = \frac{e^k - e^{-k}}{2}
$$

with k being the operator with kernel $k(x, y)$. If we choose the operator s as in (2.10), then a simple computation using the CCR (2.13) leads to the identity

$$
T^* \mathbb{H}_{\mathrm{Bog}} T = \frac{1}{2} Tr_{\mathfrak{H}_+}(E_\infty - D - K) + \mathrm{d}\Gamma(E_\infty). \tag{2.20}
$$

Thus from (2.18) we deduce that, up to a unitary transformation,

$$
H_N \approx N e_{\mathrm{H}} + \frac{1}{2} Tr_{\mathfrak{H}_+}(E_\infty - D - K) + \mathrm{d}\Gamma(E_\infty) + o(1)_{N \to \infty}, \tag{2.21}
$$

which is consistent with the prediction in (2.11) for the excitation spectrum.

3. Validity of Bogoliubov's theory

3.1. The mean-field regime

In this subsection we focus on the mean-field regime, namely we consider the Hamiltonian in (2.1),

$$
H_N = \sum_{i=1}^{N} \big(-\Delta_{x_i} + V_{\mathrm{ext}}(x_i) \big) + \frac{1}{N-1} \sum_{1 \le i < j \le N} W(x_i - x_j),
$$

with time-independent potentials V_{ext}, W.

From the heuristic discussion in Section 2, we can easily extract two natural conditions which are necessary to justify Bogoliubov's prediction for the excitation spectrum.

- The Hartree minimizer is unique. This is the necessary and sufficient condition to have the complete BEC in (2.16) for low-lying eigenfunctions of H_N; see, e.g., [33, 34, 53].
- The non-degeneracy (2.8) holds true. This condition ensures that the Taylor expansion in (2.6) makes sense, namely the Hessian dominates the error term, and that the Bogoliubov Hamiltonian in (2.19) is bounded from below and diagonalizable; see [20, 48].

In a joint work with M. Lewin, S. Serfaty, and J. P. Solovej [36], we proved that Bogoliubov's prediction is indeed correct under those general conditions on the Hartree minimizer. More precisely, we have the following theorem.

Theorem 3.1 (Validity of Bogoliubov excitation spectrum [36]). *Consider the Hamiltonian H_N in (2.1), where V_{ext} and W satisfy (2.2) and (2.3). Assume that the Hartree minimizer u_0 is unique and non-degenerate. Then for every $j \in \mathbb{N}$, the jth eigenvalue of H_N satisfies*

$$\lim_{N\to\infty} \big(\lambda_j(H_N) - N e_{\mathrm{H}}\big) = \lambda_j(\mathbb{H}_{\mathrm{Bog}}),$$

where the Bogoliubov Hamiltonian is an operator on $\mathcal{F}(\mathfrak{H}_+)$ defined in (2.19).

The result in [36] holds in a more general setting; in particular, it holds in all dimensions and the external potential V_{ext} may vanish at infinity which is relevant to unconfined systems. In the later case, some particles may escape to infinity and we have to add the assumption that any minimizing sequence of the Hartree functional is pre-compact in $L^2(\mathbb{R}^3)$, which is the necessary and sufficient condition for the complete BEC to hold (see [34]).

Our result in [36] was inspired by the pioneer works of Seiringer [55] and Grech and Seiringer [30] who have for the first time derived the Bogoliubov excitation spectrum for a class of trapped bosons in the mean-field model. In [30,55], the interaction potential W is assumed to be bounded and of positive type, namely its Fourier transform satisfies

$$0 \le \widehat{W} \in L^1(\mathbb{R}^3).$$

Under this condition, we have

$$\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \overline{f(x)} f(y) W(x-y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \big|\hat{f}(k)\big|^2 \widehat{W}(k)\,\mathrm{d}k \ge 0. \tag{3.1}$$

Therefore, the uniqueness of the Hartree minimizer is an easy consequence of the convexity of $|u|^2 \mapsto \mathcal{E}_{\mathrm{H}}(u)$ (the convexity of the kinetic part follows from the diamagnetic inequality $|\nabla u(x)| \ge |\nabla |u|(x)|$). Moreover, (3.1) also implies that the operator K with kernel $u_0(x)u_0(y)W(x-y)$ is a positive operator, and hence the non-degeneracy condition (2.8) holds true.

Note that thanks to (2.20), the spectrum of $\mathbb{H}_{\mathrm{Bog}}$ is known explicitly in terms of the spectrum of the one-body operator E_∞ given in (2.11). For the *homogeneous gas* studied in [55], when particles are confined on the torus $[0, L]^3$ with periodic boundary condition and $V_{\mathrm{ext}} = 0$, the eigenvalues of E_∞ are simply given by

$$e_p = \big(|p|^4 + 2|p|^2 \widehat{W}(p)\big)^{1/2}, \quad p \in (2\pi/L)\mathbb{Z}^3 \backslash \{0\}.$$

As already mentioned by Bogoliubov [10], the fact that the elementary excitation e_p behaves linearly for small $|p|$ corresponds to Landau's criterion for superfluidity [32]. More precisely, it implies the wedge-like shape of the joint spectrum of the Hamiltonian momentum, which in particular guarantees that adding a drop with a small velocity will not change the ground state of the system, namely the drop can move without friction. Strictly speaking, the mean-field regime discussed in this subsection corresponds to the choice $L \sim 1$ and $|p|$ is not very small. However, the same picture holds true in the large volume limit $L = L_N \to \infty$; see [21] for rigorous results (the results in [7, 8], up to a suitably scaling argument, are also relevant to the large volume limit).

Ingredients of the proof. Now let us explain the main ideas of the proof in [36]. Our important tool is an *excitation operator* which implements Bogoliubov's c-number substitution. Thanks to the isomorphism of Fock spaces

$$\mathcal{F}\big(L^2(\mathbb{R}^3)\big) = \mathcal{F}\big(\operatorname{Span}(u_0) \oplus \{u_0\}^{\perp}\big) \approx \mathcal{F}\big(\operatorname{Span}(u_0)\big) \otimes_s \mathcal{F}(\mathfrak{H}_+)$$

we can decompose any function $\Psi_N \in \mathfrak{H}^N$ uniquely as

$$\Psi_N = u_0^{\otimes N}\xi_0 + u_0^{\otimes N-1} \otimes_s \xi_1 + u_0^{\otimes N-2} \otimes_s \xi_2 + \cdots + \xi_N$$

with $\xi_k \in \mathfrak{H}_+^k$. Recall that for two functions $\Psi_k \in \mathfrak{H}^k$ and $\Psi_\ell \in \mathfrak{H}^\ell$, we define the symmetric tensor product by

$$\Psi_k \otimes_s \Psi_\ell(x_1, \ldots, x_{k+\ell}) \\ = \frac{1}{\sqrt{k!\ell!(k+\ell)!}} \sum_{\tau \in S_N} \Psi_k(x_{\tau(1)}, \ldots, x_{\tau(k)})\Psi_\ell(x_{\tau(k+1)}, \ldots, x_{\tau(k+\ell)}).$$

As proved in [36], the operator

$$U : \Psi_N \to (\xi_0, \xi_1, \ldots, \xi_N) \tag{3.2}$$

is a unitary transformation from $\mathfrak{H}^N$ to the truncated Fock space

$$\mathcal{F}^{\leq N}(\mathfrak{H}_+) = \mathbf{1}^{\mathcal{N}_+ \leq N}\mathcal{F}(\mathfrak{H}_+),$$

where $\mathcal{N}_+ = \mathrm{d}\Gamma(\mathbf{1}_{\mathfrak{H}_+})$ is the number operator on the excited Fock space $\mathcal{F}(\mathfrak{H}_+)$. The operator U essentially maps $a(u_0)$ and $a^*(u_0)$ to $\sqrt{N - \mathcal{N}_+}$, namely

$$a_x \mapsto \sqrt{N - \mathcal{N}_+}u_0(x) + c_x,$$

where c_x is the annihilation operator on $\mathcal{F}(\mathfrak{H}_+)$. More precisely, we have on $\mathcal{F}^{\leq N}(\mathfrak{H}_+)$

$$UH_NU^* = \mathbf{1}^{\mathcal{N} \leq N}\Big(\sum_{i=0}^{4} \mathcal{L}_i\Big)\mathbf{1}^{\mathcal{N} \leq N}, \tag{3.3}$$

where

$$\begin{aligned}
\mathcal{L}_0 &= Ne_{\mathrm{H}} + \frac{\mathcal{N}_+(\mathcal{N}_+ + 1)}{2N}\Big(\int |u_0|^2\big(W * |u_0|^2\big)\Big),\\
\mathcal{L}_1 &= \sqrt{N - \mathcal{N}_+}\int \big((-\Delta + V_{\mathrm{ext}} + |u_0|^2 * W)u_0\big)(x)c_x \,\mathrm{d}x + \mathrm{h.c.}\\
&\quad + \frac{\mathcal{N}_+\sqrt{N - \mathcal{N}_+}}{N-1}\int \big((|u_0|^2 * W)u_0\big)(x)c_x \,\mathrm{d}x + \mathrm{h.c.},\\
\mathcal{L}_2 &= \int c_x^*(D + K)_x c_x \,\mathrm{d}x + \frac{1 - \mathcal{N}_+}{N-1}\int c_x^*\big(|u_0|^2 * W + K\big)_x c_x \,\mathrm{d}x\\
&\quad + \frac{\sqrt{(N - \mathcal{N}_+)(N - \mathcal{N}_+ - 1)}}{2(N-1)}\iint K(x,y)c_x c_y \,\mathrm{d}x\,\mathrm{d}y + \mathrm{h.c.},\\
\mathcal{L}_3 &= \frac{\sqrt{N - \mathcal{N}_+}}{N-1}\iint W(x-y)\varphi(x)c_y^* c_x c_y \,\mathrm{d}x\,\mathrm{d}y + \mathrm{h.c.},\\
\mathcal{L}_4 &= \frac{1}{N-1}\iint W(x-y)c_x^* c_y^* c_x c_y \,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

By formally taking the limit $N \to \infty$, we obtain immediately the desired convergence

$$UH_N U^* - Ne_{\mathrm{H}} \to \mathbb{H}_{\mathrm{Bog}}. \tag{3.4}$$

Rigorously, we proved in [36, Proposition 5.1] that for every $1 \leq M \leq N$,

$$\pm \mathbf{1}^{\mathcal{N}_+ \leq M}(UH_N U^* - Ne_{\mathrm{H}} - \mathbb{H}_{\mathrm{Bog}})\mathbf{1}^{\mathcal{N}_+ \leq M} \leq C\sqrt{\frac{M}{N}}(\mathbb{H}_{\mathrm{Bog}} + C) \tag{3.5}$$

as quadratic forms on $\mathcal{F}^{\leq M}(\mathfrak{H}_+)$. This justifies the convergence (3.4) in the sectors of *low excitations*, namely $\mathcal{N}_+ \ll N$. The contribution of the sectors of *high excitations*, namely $\mathcal{N}_+ \sim N$, is negligible thanks to the complete BEC (2.16). Using (3.5), we can derive the convergence of quadratic forms in (3.5), which in turns implies the convergence of eigenvalues by the min-max principle.

As a byproduct of our method, we also obtain the information for eigenfunctions.

Theorem 3.2 (Norm approximation for eigenfunctions [36]). *Under the same conditions in Theorem 3.1, the ground state Ψ_N of H_N is simple and satisfies*

$$\lim_{N\to\infty} \|U\Psi_N - \Phi\|_{\mathcal{F}(\mathfrak{H}_+)} = 0, \tag{3.6}$$

where $\Phi \in \mathcal{F}(\mathfrak{H}_+)$ is the unique ground state of the Bogoliubov Hamiltonian $\mathbb{H}_{\mathrm{Bog}}$. A similar convergence holds for the higher eigenfunctions (possibly up to subsequences of $N \to \infty$ in case of degenerate eigenvalues).

The norm approximation (3.6) is much stronger than the complete BEC (2.16). In fact, while (2.16) describes a macroscopic property, (3.6) really contains microscopic information: changing the behavior of a single particle can change the many-body state in norm to the leading order. In particular, (3.6) implies that in the non-interacting case ($W \not\equiv 0$), Ψ_N is *never* close to $u_0^{\otimes N}$ in norm, namely the fluctuations around the Hartree state $u_0^{\otimes N}$ are nontrivial.

3.2. The Gross–Pitaevskii regime

In this subsection, we consider the N-body Hamiltonian

$$H_N = \sum_{i=1}^{N} \big(- \Delta_{x_i} + V_{\mathrm{ext}}(x_i)\big) + \sum_{1\le i<j\le N} N^2 V\big(N(x_i - x_j)\big) \tag{3.7}$$

on $\mathfrak{H}^N = \bigotimes_{\mathrm{sym}}^N L^2(\mathbb{R}^3)$ with time-independent potentials V_{ext}, V. For simplicity, we assume that the external and interaction potentials satisfy

$$0 \le V_{\mathrm{ext}}(x) \le Ce^{C|x|} \text{ for some constant } C > 0, \quad \lim_{|x|\to\infty} V_{\mathrm{ext}}(x) = \infty, \tag{3.8}$$

$$0 \le V \in L^1(\mathbb{R}^3), \quad V \text{ is radially symmetric and compactly supported.} \tag{3.9}$$

In this so-called Gross–Pitaevskii regime, the system is very dilute and the strong correlation between particles at short distances leads to a subtle correction to the leading order which is captured by the *scattering length*

$$8\pi\mathfrak{a}_0 = \inf\left\{ \int_{\mathbb{R}^3} \big(2|\nabla f(x)|^2 + V(x)|f(x)|^2\big)\,\mathrm{d}x, \ \lim_{|x|\to\infty} f(x) = 1\right\}. \tag{3.10}$$

More precisely, the Hartree functional has to be replaced by the *Gross–Pitaevskii functional*

$$\mathcal{E}_{\mathrm{GP}}(u) = \int_{\mathbb{R}^3} \big(|\nabla u(x)|^2 + V_{\mathrm{ext}}(x)|u(x)|^2 + 4\pi\mathfrak{a}_0|u(x)|^4\big)\,\mathrm{d}x. \tag{3.11}$$

Note that by simply restricting to the Hartree states $u^{\otimes N}$ and using $N^3V(N\cdot) \approx \widehat{V}(0)\delta_0$, we would obtain a *wrong* functional with $8\pi\mathfrak{a}_0$ replaced by its first Born approximation $\widehat{V}(0)$. It is not difficult to prove that the Gross–Pitaevskii functional has a unique normalized minimizer φ which is positive and exponentially decay (see [39]).

In [39], Lieb, Seiringer, and Yngvason proved that the ground state energy of H_N in (3.7) satisfies

$$\lim_{N\to\infty} \frac{\lambda_1(H_N)}{N} = \inf_{\|u\|_{L^2(\mathbb{R}^3)}=1} \mathcal{E}_{\mathrm{GP}}(u). \tag{3.12}$$

Later, in [37,38], Lieb and Seiringer proved that if Ψ_N is an approximate ground state, namely $\langle \Psi_N, H_N \Psi_N\rangle = \lambda_1(H_N) + o(N)$, then the complete BEC on the Gross–Pitaevskii minimizer φ holds:

$$\big\langle \Psi_N, a^*(\varphi)a(\varphi)\Psi_N\big\rangle = N + o(N). \tag{3.13}$$

Recently, the BEC with *optimal rate*

$$\big\langle \Psi_N, a^*(\varphi)a(\varphi)\Psi_N\big\rangle = N + O(1) \tag{3.14}$$

was obtained in [6, 9, 31] (the homogeneous case) and [18, 47] (the general trapped case).

Since there are only finitely many excited particles due to (3.14), it is still reasonable to predict the excitation spectrum by Bogoliubov's approximation. A straightforward application of the heuristic arguments in Section 2 predicts that the elementary excitations are eigenvalues of the one-body operator

$$\big(D^{1/2}\big(D + 2\widehat{V}(0)\varphi^2\big)D^{1/2}\big)^{1/2},$$

where D is the mean-field operator associated with the Gross–Pitaevskii equation,

$$D\varphi = 0, \quad D = -\Delta + V_{\mathrm{ext}} + 8\pi\mathfrak{a}_0 - \varepsilon_0.$$

However, as mentioned already by Bogoliubov [10] (which goes back to a remark of Landau), the number $\widehat{V}(0)$ should be replaced by the scattering length $8\pi\mathfrak{a}_0$, similarly to the leading order correction. Therefore, to put Bogoliubov's theory in a good use, after the three steps written in Section 2.2, we need an important modification.

Step 4 (Landau's correction). $\widehat{V}(0)$ should be replaced by $8\pi\mathfrak{a}_0$ everywhere, with $\mathfrak{a}_0$ the scattering length of V.

It is Step 4 that makes the implementation of Bogoliubov's arguments in the Gross–Pitaevskii regime much more challenging than that of the mean-field regime.

In [7], Boccato, Brennecke, Cenatiempo, and Schlein solved this problem for the homogeneous gas. Recently, in a joint work with A. Triay [50], we extended the result for general trapped systems. We have the following theorem.

Theorem 3.3 (Bogoliubov's theory in the Gross–Pitaevskii regime [50]). *Consider the Hamiltonian H_N in (3.7). Let $\lambda_1(H_N)$ be the ground state energy of the Hamiltonian H_N in (3.7). Then the spectrum of $H_N - \lambda_1(H_N)$ below an energy $\Lambda \in [1, N^{1/12}]$ is equal to finite sums of the form*

$$\sum_{i\ge 1} n_i e_i + \mathcal{O}(\Lambda^3 N^{-1/12}), \quad n_i \in \{0, 1, 2, \dots\},$$

where $\{e_i\}_{i=1}^{\infty}$ are the positive eigenvalues of $(D^{1/2}(D + 16\pi\mathfrak{a}_0\varphi^2)D^{1/2})^{1/2}$.

Independently to us, a result similar to Theorem 3.3 was obtained by Brennecke, Schlein, and Schraven in [17]. While our overall approach is similar to that of [7, 17], the detailed implementations are different. In fact, in [50] we introduced several conceptual simplifications and generalizations, which could be helpful for the study of dilute gases in the future. Let us explain some key ideas below.

Ingredients of the proof. Our proof is based on the rigorous approximation

$$T_2^* T_c^* T_1^* U H_N U^* T_1 T_c T_2 \approx \lambda_1(H_N) + \mathrm{d}\Gamma(E_\infty) + o(1)_{N\to\infty} \tag{3.15}$$

on the excited Fock space $\mathcal{F}_+ = \mathcal{F}(\mathfrak{H}_+)$ with $\mathfrak{H}_+ = \{\varphi\}^\perp = QL^2(\mathbb{R}^3)$ with $Q = 1 - |\varphi\rangle\langle\varphi|$.

Here U is the same transformation in (3.2), which factors out the condensation described by the Gross–Pitaevskii minimizer u_0. Consequently, the excited particles are captured by the Hamiltonian in (3.3). Unlike the mean-field regime where $\mathcal{L}_3$ and $\mathcal{L}_4$ are of order $o(1)$, in the Gross–Pitaevskii regime $L_4 \sim N$ and $L_3 \sim O(1)$. Therefore, these terms have to be renormalized by the unitary transformations T_1 and T_c, respectively. After that, we obtain a quadratic Hamiltonian which can be diagonalized by the final unitary transformation T_2.

To define the quadratic transformation T_1, we need to capture the correlation structure of particles. Let $0 \le f \le 1$ be the scattering solution

$$-2\Delta f + Vf = 0 \quad \text{in } \mathbb{R}^3, \qquad \lim_{|x|\to\infty} f(x) = 1. \tag{3.16}$$

We write $\omega = 1 - f$ and for every $0 < \ell \ll 1$ introduce the truncated functions

$$\omega_{\ell,N}(x) = \chi(x/\ell)\omega(Nx), \quad \varepsilon_{\ell,N} = 2\Delta\big(\omega_{\ell,N}(x) - \omega(Nx)\big), \tag{3.17}$$

where $0 \le \chi \le 1$ is a smooth function satisfying $\chi(t) = 1$ if $|x| \le 1/2$ and $\chi(x) = 0$ if $|x| \ge 1$. By choosing T_1 such that

$$T_1^* a^*(g) T_1 = a^*\big(\sqrt{1 + s_1^2}\, g\big) + a(s_1 g), \quad \forall g \in \mathfrak{H}, \tag{3.18}$$

where

$$s_1 = Q^{\otimes 2}\tilde{s}_1 \in \mathfrak{H}_+^2, \quad \tilde{s}_1(x,y) = -N\omega_{\ell,N}(x-y)\varphi(x)\varphi(y),$$

we can replace the short range potential $V(N(x-y))$ in $\mathcal{L}_2$ by the longer range potential $\varepsilon_{\ell,N}(x-y)$. Note that $\varepsilon_{\ell,N}$ is supported in $\{\ell/2 \le |x| \le \ell\}$ and

$$N^3 \int_{\mathbb{R}^3} \varepsilon_{\ell,N} = 8\pi\mathfrak{a}_0. \tag{3.19}$$

When ℓ grows slowly, we are essentially placed in the mean-field regime.

The idea of renormalizing the short-range potential by a Bogoliubov transformation was introduced by Benedikter, de Oliveira, and Schlein [4] to derive the Gross–Pitaevskii dynamics on Fock space. In [16], Brennecke and Schlein adapted the approach in [4] to study the quantum dynamics on $\mathfrak{H}^N$, where they used a generalized Bogoliubov transformation on $\mathcal{F}_+^{\leq N}$ of the form

$$\exp\left(\frac{1}{2}\iint \mathfrak{K}_1(x,y)b_x^* b_y^* \,\mathrm{d}x\,\mathrm{d}y - \text{h.c.}\right) \quad \text{with } b_x = \sqrt{1-\mathcal{N}/N}\,a_x. \tag{3.20}$$

The transformation (3.20) has been also an essential tool in the study of the spectral problem in a series of papers [6–9,17,18]. Our choice of T_1 in (3.18) is different from (3.20) in three aspects.

- First, the operator b_x in (3.20) is not an exact annihilation operator, and hence $\widetilde{T}_1$ only satisfies an approximate form of (3.18). Here our T_1 is a proper Bogoliubov transformation and the exact formula (3.18) simplifies several computations.
- Second, the truncated scattering solution in [4, 16] is defined using Neumann boundary condition on $|x| = \ell N$. Here our choice of $\omega_{\ell,N}$ in (3.17) is simpler and works for a larger class of potentials.
- Third, and most importantly, we take $\ell \ll 1$ instead of $\ell \sim 1$ as in [4,7,16]. Thus T_1 renormalizes $\mathcal{L}_2$ efficiently but and leaves the cubic terms $\mathcal{L}_3$ invariant.

To remove the cubic term $\mathcal{L}_3$, we introduce a cubic transformation of the form

$$T_c = e^S, \quad S = \theta_M \iint k_c(x,y,y')a_x^* a_y^* a_y \,\mathrm{d}x\,\mathrm{d}y - \text{h.c.},$$

where $\theta_M \approx \mathbf{1}(\mathcal{N} \leq M)$ and $k_c(x,y,y')$ is the kernel of the operator $k_c : \mathfrak{H} \to \mathfrak{H}^2$ defined by

$$k_c = Q^{\otimes 2}\tilde{k}_c Q, \quad \tilde{k}_c(x,y,y') = -N^{1/2}\varphi(x)\omega_{\ell,N}(x-y)\delta_{y,y'}$$

with $\tilde{k}_c(x,y,y')$ the kernel of the operator $k_c : \mathfrak{H} \to \mathfrak{H}^2$. The projections $Q : \mathfrak{H} \to \mathfrak{H}_+$ and $Q^{\otimes 2} : \mathfrak{H}^2 \to \mathfrak{H}_+^2$ ensure that $k_c : \mathfrak{H}_+ \to \mathfrak{H}_+^2$, namely the cubic kernel S acts only on excited particles. The cut-off parameter $1 \ll M \ll N$ in θ_M allows us to control the number of excitations. Consequently, we have the simple expansion

$$T_c^* A T_c \approx A - [S,A] + \frac{1}{2}\big[S,[S,A]\big]$$

and the above choice of S comes from the cancelation

$$\mathcal{L}_3 - \big[S, \mathrm{d}\Gamma(-\Delta) + \mathcal{L}_4\big] \approx 0.$$

Here our cubic transformation is slightly simpler than that of [7] since we did not change $\mathcal{L}_3$ in the previous step. The idea of using a cubic generator goes back to the work of Yau and Yin [56] on the Lee–Huang–Yang formula in the thermodynamic limit. The choice $\ell \ll 1$ is again very helpful to separate high and low momenta.

Finally, we end up with the quadratic Hamiltonian

$$\mathrm{d}\Gamma(D) + \frac{1}{2}\int N^3 \varepsilon_{\ell,N}(x-y)\varphi(x)\varphi(y)(2a_x^* a_y + a_x^* a_y^* + a_x a_y)\,\mathrm{d}x\,\mathrm{d}y$$

which can be diagonalized similarly as in the mean-field regime. We find that

$$T_2^* T_c^* T_1^* \mathcal{H} T_1 T_c T_2 \approx \mathrm{const} + \mathrm{d}\Gamma(E), \tag{3.21}$$

where

$$E = \big(D^{1/2}(D+2K)D^{1/2}\big)^{1/2}, \quad K = Q\widetilde{K}Q, \ \widetilde{K}(x,y) = \varphi(x)N^3\varepsilon_{\ell,N}(x-y)\varphi(y).$$

Since $\ell \ll 1$, we have $N^3\varepsilon_{N,\ell} \to 8\pi\delta_0$, which implies that $E \to E_\infty$ in an appropriate sense. This completes the overview of our proof of Theorem 3.3.

4. Further results and open problems

Excitation spectrum. In the mean-field regime, the validity of Bogoliubov's theory for the ground state energy and the excitation spectrum were extended in various directions, including the large volume setting [21], multiple-condensations [49, 54], mixture of Bose gases [41], and higher-order expansions [13, 42, 46, 52]. The intermediate regime between the mean-field and the Gross–Pitaevskii regime was studied in [8]. The regime beyond the Gross–Pitaevskii was studied in [14] (see also [1, 27] for results on the BEC). It is an interesting open problem to extend the results in the Gross–Pitaevskii regime (or beyond) to trapped systems in bounded domains with Neumann or Dirichlet boundary conditions, since this will have interesting implications to systems in the thermodynamic limit.

Quantum dynamics. In the mean-field regime, the method in [36] was developed in [35] to derive the norm approximation for the many-body Schrödinger dynamics. Higher-order expansions in the mean-field regime were also obtained in [12]. The validity of Bogoliubov's theory for the quantum dynamics with singular interaction potentials of the form $N^{3\beta}W(N^\beta x)$ with $0 < \beta < 1$ was obtained in [15, 43–45]. When $\beta = 1$, the Gross–Pitaevskii dynamics was derived in [25, 26], but the justification of Bogoliubov's theory for the dynamics remains open. We refer to the reviews [5, 51] for further discussions on the dynamical problem.

Positive temperatures. As discussed in Section 3, Bogoliubov's theory holds true for eigenvalues belonging to an interval of order 1 above $\lambda_1(H_N)$. This implies the validity of Bogoliubov's theory for the free energy of a temperature of order 1; see, e.g., [36, Theorem 2.3] for an explicit statement. It is an open problem to extend the analysis to higher temperatures. For the homogeneous gas in a unit torus, the critical temperature where we see the BEC phase transition is of order $N^{2/3}$. In this case, the validity of the Gross–Pitaevskii theory has been understood [22], but the validity of Bogoliubov's theory remains unknown.

Thermodynamic limit. In the thermodynamic limit, Bogoliubov's theory is consistent with the Lee–Huang–Yang formula on the ground state energy of dilute Bose gases. In this problem, the leading order behavior is already difficult: the upper bound was proved in 1957 [23] but the lower bound was obtained only some 40 years later [40]. The second order, which requires a correction to Bogoliubov's theory similar to that in the Gross–Pitaevskii regime, was proved recently in [3, 56] (upper bound) and [28, 29] (lower bound). While the second-order lower bound in [29] covers a large class of interaction potentials, including the hard core case, extending this universality to the second-order upper bound remains an open problem. The excitation spectrum seems to be completely out of reach by current techniques; a simple reason is that the existence of the BEC in the thermodynamic limit remains a major open problem in mathematical physics.

Acknowledgments. I would like to thank Mathieu Lewin, Sylvia Serfaty, Jan Philip Solovej, and Arnaud Triay for the collaborations leading to the results reviewed here. I thank Christian Brennecke, Søren Fournais, Christian Hainzl, Marcin Napiórkowski, Benjamin Schlein, and Robert Seiringer for helpful discussions on various aspects of Bogoliubov's theory.

Funding. This research supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC-2111-390814868).

References

[1] A. Adhikari, C. Brennecke, and B. Schlein, Bose–Einstein condensation beyond the Gross–Pitaevskii regime. *Ann. Henri Poincaré* **22** (2021), no. 4, 1163–1233
Zbl 1467.82006 MR 4229531

[2] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, Observation of Bose-Einstein condensation in a dilute atomic vapor. *Science* **269** (1995), no. 5221, 198–201

[3] G. Basti, S. Cenatiempo, and B. Schlein, A new second-order upper bound for the ground state energy of dilute Bose gases. *Forum Math. Sigma* **9** (2021), Paper No. e74
Zbl 07450651 MR 4342115

[4] N. Benedikter, G. de Oliveira, and B. Schlein, Quantitative derivation of the Gross–Pitaevskii equation. *Comm. Pure Appl. Math.* **68** (2015), no. 8, 1399–1482 Zbl 1320.35318 MR 3366749

[5] N. Benedikter, M. Porta, and B. Schlein, *Effective evolution equations from quantum dynamics*. SpringerBriefs Math. Phys. 7, Springer, Cham, 2016 Zbl 1396.81003 MR 3382225

[6] C. Boccato, C. Brennecke, S. Cenatiempo, and B. Schlein, Complete Bose–Einstein condensation in the Gross–Pitaevskii regime. *Comm. Math. Phys.* **359** (2018), no. 3, 975–1026 Zbl 1391.82004 MR 3784538

[7] C. Boccato, C. Brennecke, S. Cenatiempo, and B. Schlein, Bogoliubov theory in the Gross–Pitaevskii limit. *Acta Math.* **222** (2019), no. 2, 219–335 Zbl 1419.82064 MR 3974476

[8] C. Boccato, C. Brennecke, S. Cenatiempo, and B. Schlein, The excitation spectrum of Bose gases interacting through singular potentials. *J. Eur. Math. Soc. (JEMS)* **22** (2020), no. 7, 2331–2403 Zbl 1448.81319 MR 4107508

[9] C. Boccato, C. Brennecke, S. Cenatiempo, and B. Schlein, Optimal rate for Bose–Einstein condensation in the Gross–Pitaevskii regime. *Comm. Math. Phys.* **376** (2020), no. 2, 1311–1395 Zbl 1439.82004 MR 4103969

[10] N. Bogoliubov, On the theory of superfluidity. *Acad. Sci. USSR. J. Phys.* **11** (1947), 23–32 MR 0022177

[11] S. Bose, Plancks Gesetz und Lichtquantenhypothese. *Z. Phys.* **26** (1924), 178–181 Zbl 51.0732.04

[12] L. Boßmann, N. Pavlović, P. Pickl, and A. Soffer, Higher order corrections to the mean-field description of the dynamics of interacting Bosons. *J. Stat. Phys.* **178** (2020), no. 6, 1362–1396 Zbl 1439.82029 MR 4081233

[13] L. Boßmann, S. Petrat, and R. Seiringer, Asymptotic expansion of low-energy excitations for weakly interacting bosons. *Forum Math. Sigma* **9** (2021), Paper No. e28 Zbl 1460.81123 MR 4239621

[14] C. Brennecke, M. Caporaletti, and B. Schlein, Excitation spectrum for Bose gases beyond the Gross–Pitaevskii regime. 2021, arXiv:2104.13003

[15] C. Brennecke, P. T. Nam, M. Napiórkowski, and B. Schlein, Fluctuations of N-particle quantum dynamics around the nonlinear Schrödinger equation. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **36** (2019), no. 5, 1201–1235 Zbl 1419.81042 MR 3985542

[16] C. Brennecke and B. Schlein, Gross–Pitaevskii dynamics for Bose–Einstein condensates. *Anal. PDE* **12** (2019), no. 6, 1513–1596 Zbl 1414.35184 MR 3921312

[17] C. Brennecke, B. Schlein, and S. Schraven, Bogoliubov theory for trapped bosons in the Gross–Pitaevskii regime. 2021, arXiv:2108.11129

[18] C. Brennecke, B. Schlein, and S. Schraven, Bose–Einstein condensation with optimal rate for trapped bosons in the Gross–Pitaevskii regime. 2021, arXiv:2102.11052

[19] K. B. Davis, M. O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, Bose–Einstein condensation in a gas of sodium atoms. *Phys. Rev. Lett.* **75** (1995), 3969–3973

[20] J. Dereziński, Bosonic quadratic Hamiltonians. *J. Math. Phys.* **58** (2017), no. 12, 121101 Zbl 1380.81500 MR 3739173

[21] J. Dereziński and M. Napiórkowski, Excitation spectrum of interacting bosons in the mean-field infinite-volume limit. *Ann. Henri Poincaré* **15** (2014), no. 12, 2409–2439 Zbl 1305.82042 MR 3272826

[22] A. Deuchert and R. Seiringer, Gross–Pitaevskii limit of a homogeneous Bose gas at positive temperature. *Arch. Ration. Mech. Anal.* **236** (2020), no. 3, 1217–1271 Zbl 1439.82006 MR 4076065

[23] F. J. Dyson, Ground-state energy of a hard-sphere gas. *Phys. Rev.* **106** (1957), 20–26 Zbl 0077.23503

[24] A. Einstein, Quantentheorie des einatomigen idealen Gases. *Sitzungsber. Preuß. Akad. Wiss.* **1924** (1924), 261–267

[25] L. Erdős, B. Schlein, and H.-T. Yau, Rigorous derivation of the Gross–Pitaevskii equation with a large interaction potential. *J. Amer. Math. Soc.* **22** (2009), no. 4, 1099–1156 Zbl 1207.82031 MR 2525781

[26] L. Erdős, B. Schlein, and H.-T. Yau, Derivation of the Gross–Pitaevskii equation for the dynamics of Bose–Einstein condensate. *Ann. of Math. (2)* **172** (2010), no. 1, 291–370 Zbl 1204.82028 MR 2680421

[27] S. Fournais, Length scales for BEC in the dilute Bose gas. In *Partial Differential Equations, Spectral Theory, and Mathematical Physics*, pp. 115–133, EMS Ser. Congr. Rep., EMS Press, Zürich, 2021 Zbl 07438889 MR 4331810

[28] S. Fournais and J. P. Solovej, The energy of dilute Bose gases. *Ann. of Math. (2)* **192** (2020), no. 3, 893–976 Zbl 1455.81050 MR 4172623

[29] S. Fournais and J. P. Solovej, The energy of dilute Bose gases II: The general case. 2021, arXiv:2108.12022

[30] P. Grech and R. Seiringer, The excitation spectrum for weakly interacting bosons in a trap. *Comm. Math. Phys.* **322** (2013), no. 2, 559–591 Zbl 1273.82007 MR 3077925

[31] C. Hainzl, Another proof of BEC in the GP-limit. *J. Math. Phys.* **62** (2021), no. 5, Paper No. 051901 Zbl 1467.82090 MR 4256681

[32] L. Landau, Theory of the superfluidity of helium II. *Phys. Rev.* **60** (1941), 356–358 Zbl 0027.18505

[33] M. Lewin, Mean-field limit of Bose systems: rigorous results. *Proceedings from the International Congress of Mathematical Physics at Santiago de Chile, July 2015.* arXiv:1510.04407

[34] M. Lewin, P. T. Nam, and N. Rougerie, Derivation of Hartree's theory for generic mean-field Bose systems. *Adv. Math.* **254** (2014), 570–621 Zbl 1316.81095 MR 3161107

[35] M. Lewin, P. T. Nam, and B. Schlein, Fluctuations around Hartree states in the mean-field regime. *Amer. J. Math.* **137** (2015), no. 6, 1613–1650 Zbl 1329.81430 MR 3432269

[36] M. Lewin, P. T. Nam, S. Serfaty, and J. P. Solovej, Bogoliubov spectrum of interacting Bose gases. *Comm. Pure Appl. Math.* **68** (2015), no. 3, 413–471 Zbl 1318.82030 MR 3310520

[37] E. H. Lieb and R. Seiringer, Proof of Bose-Einstein condensation for dilute trapped gases. *Phys. Rev. Lett.* **88** (2002), 170409

[38] E. H. Lieb and R. Seiringer, Derivation of the Gross–Pitaevskii equation for rotating Bose gases. *Comm. Math. Phys.* **264** (2006), no. 2, 505–537 Zbl 1233.82004 MR 2215615

[39] E. H. Lieb, R. Seiringer, and J. Yngvason, Bosons in a trap: A rigorous derivation of the Gross–Pitaevskii energy functional. *Phys. Rev. A* **61** (2000), 043602

[40] E. H. Lieb and J. Yngvason, Ground state energy of the low density Bose gas. *Phys. Rev. Lett.* **80** (1998), 2504–2507

[41] A. Michelangeli, P. T. Nam, and A. Olgiati, Ground state energy of mixture of Bose gases. *Rev. Math. Phys.* **31** (2019), no. 2, 1950005 Zbl 1418.81104 MR 3917408

[42] P. T. Nam, Binding energy of homogeneous Bose gases. *Lett. Math. Phys.* **108** (2018), no. 1, 141–159 Zbl 1387.81395 MR 3740432

[43] P. T. Nam and M. Napiórkowski, Bogoliubov correction to the mean-field dynamics of interacting bosons. *Adv. Theor. Math. Phys.* **21** (2017), no. 3, 683–738 Zbl 1382.82032 MR 3695801

[44] P. T. Nam and M. Napiórkowski, A note on the validity of Bogoliubov correction to mean-field dynamics. *J. Math. Pures Appl. (9)* **108** (2017), no. 5, 662–688 Zbl 1376.35016 MR 3711470

[45] P. T. Nam and M. Napiórkowski, Norm approximation for many-body quantum dynamics: focusing case in low dimensions. *Adv. Math.* **350** (2019), 547–587 Zbl 1416.81059 MR 3948169

[46] P. T. Nam and M. Napiórkowski, Two-term expansion of the ground state one-body density matrix of a mean-field Bose gas. *Calc. Var. Partial Differential Equations* **60** (2021), no. 3, Paper No. 99 Zbl 1462.81218 MR 4249877

[47] P. T. Nam, M. Napiórkowski, J. Ricaud, and A. Triay, Optimal rate of condensation for trapped bosons in the Gross–Pitaevskii regime. *Anal. PDE*, to appear

[48] P. T. Nam, M. Napiórkowski, and J. P. Solovej, Diagonalization of bosonic quadratic Hamiltonians by Bogoliubov transformations. *J. Funct. Anal.* **270** (2016), no. 11, 4340–4368 Zbl 1338.81430 MR 3484973

[49] P. T. Nam and R. Seiringer, Collective excitations of Bose gases in the mean-field regime. *Arch. Ration. Mech. Anal.* **215** (2015), no. 2, 381–417 Zbl 1317.35213 MR 3294406

[50] P. T. Nam and A. Triay, Bogoliubov excitation spectrum of trapped Bose gases in the Gross–Pitaevskii regime. 2021, arXiv:2106.11949

[51] M. Napiórkowski, Dynamics of interacting bosons: a compact review. 2021, arXiv: 2101.04594

[52] A. Pizzo, Bose particles in a box I. A convergent expansion of the ground state of a three-modes Bogoliubov Hamiltonian. 2015, arXiv:1511.07022

[53] N. Rougerie, Scaling limits of bosonic ground states, from many-body to non-linear Schrödinger. *EMS Surv. Math. Sci.* **7** (2020), no. 2, 253–408 Zbl 1472.35319 MR 4261667

[54] N. Rougerie and D. Spehner, Interacting bosons in a double-well potential: localization regime. *Comm. Math. Phys.* **361** (2018), no. 2, 737–786 Zbl 1397.82060 MR 3828897

[55] R. Seiringer, The excitation spectrum for weakly interacting bosons. *Comm. Math. Phys.* **306** (2011), no. 2, 565–578 Zbl 1226.82039 MR 2824481

[56] H.-T. Yau and J. Yin, The second order upper bound for the ground energy of a Bose gas. *J. Stat. Phys.* **136** (2009), no. 3, 453–503 Zbl 1200.82002 MR 2529681

Phan Thành Nam
Department of Mathematics, LMU Munich, Theresienstrasse 39, 80333 Munich, Germany;
nam@math.lmu.de

DOI 10.4171/8ECM/27

From branching singularities in minimal surfaces to non-smoothness points in ice-water interfaces

Joaquim Serra

Abstract. We review some recent developments on the regularity theory of two classical free boundary problems: the obstacle problem and Stefan's problem.

We emphasize the similarities and differences between these recent results (for the obstacle problem and Stefan's problem) and the regularity theory of integer rectifiable area-minimizing currents (and related problems) developed during the XXth century.

1. Introduction

The aim of this note is to review some recent developments on the regularity theory of two classical free boundary problems: the obstacle problem and Stefan's problem.

We believe that these new developments are better understood and appreciated when one can recognize in them strong parallelisms, and yet crucial differences, with the regularity theories of Plateau's problem, Signorini's problem, and *Almgren's problem*. (Throughout the note, we will use the non-standard (but convenient) keyword *Almgren's problem* to refer to the analog of Plateau's problem in context of integer rectifiable area-minimizing currents of codimension 2 or higher, which was studied by Almgren in his famous work [4].) Consequently, we provide, in addition to a rather complete background on the former two problems, a (partial) historical overview of the latter three, focusing on their connections and analogies with the obstacle problem and Stefan's problem.

Finally, we describe the methods and results in recent works [25–27, 29] concerned with the fine structure of the singular sets in the obstacle problem and Stefan's problem.

2. Five "classical" problems

We begin by presenting—in chronological order of their first appearance—the five problems that will be discussed throughout the note: Plateau's problem (1760s),

2020 Mathematics Subject Classification. 35R35.
Keywords. Stefan's problem, obstacle problem, free boundary.

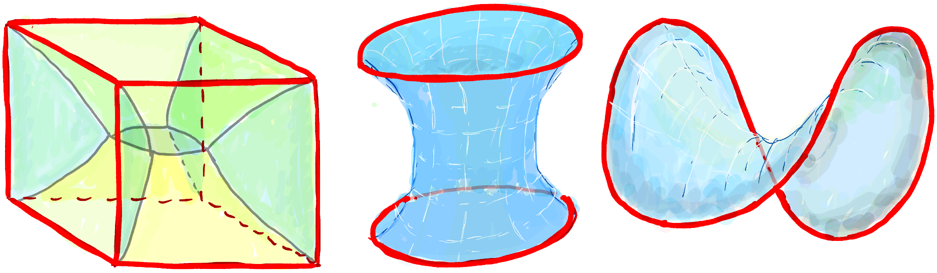


Figure 1. Soap films spanning the red “curves”.

Stefan’s problem (1890s), Signorini’s problem (1950s), the obstacle problem (1960s), and Almgren’s problem (1970s).

2.1. Plateau’s problem: The elegant shapes of soap films

Given a curve in $\mathbb{R}^3$, can one find a surface with minimal area having it as boundary? Raised by Joseph-Louis Lagrange in the 1760s, this problem is one of the most classical and influential ones in the calculus of variations and geometry. It is named after the Belgian physicist, J. Plateau (1801–1883), who experimentally investigated the (physical) geometric laws of soap films and bubbles. By the effect of surface tension, soap films are natural examples of area-minimizing surfaces.

By a well-known classical computation going back to Lagrange, if a piece of an area-minimizing surface is smooth, then its mean curvature (sum of the principal curvatures) must be identically zero.

A difficulty of Plateau’s problem is that area-minimizing “surfaces” may not be surfaces in the classical sense of differential geometry. For instance, physical soap films can take the shapes sketched in Figure 1, and while the center and right ones are smooth surfaces, the soap film on the left is not smooth (or not even locally homeomorphic to a planar disc!) near some of its points.

Between the 1930s and the 1970s, several well-known analysts and geometers, including Almgren, De Giorgi, Douglas, Federer, Fleming, Radó, Reifenberg, and Taylor, among others, yielded outstanding contributions to Plateau’s problem, which shaped its modern theory; see for instance [1, 3, 4, 14, 15, 18, 20, 21, 28, 40–42, 53]. They addressed the following fundamental questions:

(i) Which mathematical objects, that are “surfaces” in some sense, allow for a rigorous solution of the area minimization problem?

(ii) Are such minimizers smooth, possibly outside of a certain singular set?

(iii) What can be said about the singular set? (e.g., is it lower dimensional?)

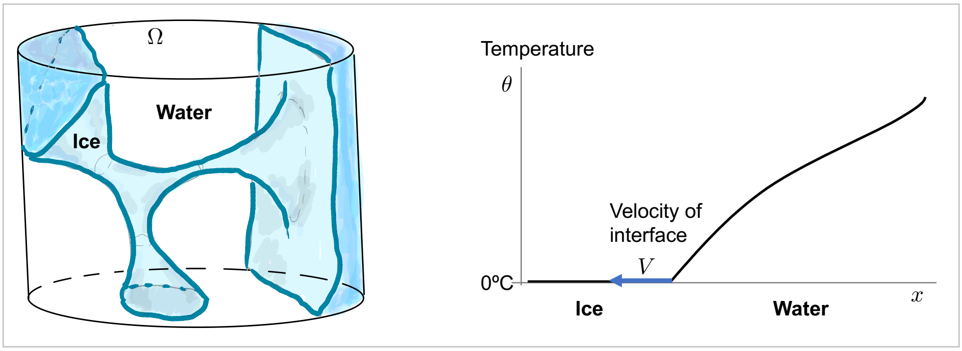


Figure 2. Stefan's problem: ice melting in water.

Thanks to intensive efforts during the XXth century, the answers to these questions are today well understood. In Section 3, we will review some key aspects of the regularity theory of Plateau's problem, i.e., the answers to questions (ii) and (iii). Some of them will have clear parallelisms in the other problems we will discuss.

2.2. Stefan's problem: Ice melting in water

Dating back to the XIXth century, Stefan's problem aims to describe the temperature distribution in a homogeneous medium undergoing a phase change, typically a body of ice at zero degrees centigrade submerged in water. The problem is named after Josef Stefan, a Slovenian physicist who introduced it around 1890; see [52].

The most classical formulation of Stefan's problem (see e.g. [19, 24]) is as follows: let $\Omega \subset \mathbb{R}^3$ be some bounded domain. For concreteness, let us think that Ω is a "cylindric water tank" as drawn in Figure 2. We denote by $\theta = \theta(x,t)$ the temperature of the water at the point $x \in \Omega$ at time $t \in \mathbb{R}^+ := [0,+\infty)$. We assume that $\theta \geq 0$ in $\Omega \times \mathbb{R}^+$. The (nonnegative) temperature at the boundary of the tank is given, and we assume that $\theta = 0$ at $t = 0$.

The set $\{(x,t) \in \Omega \times \mathbb{R}^+ : \theta(x,t) > 0\}$, denoted for brevity by $\{\theta > 0\}$, represents the water while its complement, denoted by $\{\theta = 0\}$, represents the ice. The temperature θ satisfies the heat equation

$$\partial_t \theta - \Delta\theta = 0 \quad \text{in the region } \{\theta > 0\},$$

while in the complement θ is simply zero.

Determining the time-evolving domain $\{\theta > 0\}$ in which the heat equation holds is part of the problem. Equivalently, one must determine where the ice-water interface $\partial\{\theta > 0\}$, also called the *free boundary*, is. For it, an additional equation—so-called

Stefan's condition[1]—is needed:

$$\partial_t \theta = |\nabla_x \theta|^2 \quad \text{on } \partial\{\theta > 0\}. \tag{2.1}$$

It is not difficult to see that, in the previous setting, the ice $\{\theta = 0\}$ must shrink over time. More precisely, if at some point of the tank there is liquid water at some given time, then the same point remains occupied by liquid at all future times.

The relevant regularity questions for Stefan's problem are as follows:

(i) Is the problem well posed?[2]

(ii) Is the free boundary smooth, or may it have singularities?

(iii) If there are singularities, how often may they occur in space and time?

We will discuss the answers to these questions (which remained completely open until the 1970s!) later on in this note.

2.3. Signorini's problem (1950s)

Raised in 1959, Signorini's problem [50] consists in finding the (elastic) equilibrium configuration of an elastic body, resting on a rigid frictionless horizontal plane and subject only to its mass forces.

The difficulty of the problem lies on the fact that one needs to determine which points on the bottom surface of the body will be in contact with the plane (and what is the deformation at the points which are not in contact).[3]

In a very linearized situation, Signorini's problem is reduced to following a minimization problem in the half space $U := \mathbb{R}^3 \cap \{x_3 \geq 0\}$,

$$\min \left\{ \int_U |\nabla u|^2 \, dx \text{ among } u : U \to \mathbb{R} \text{ satisfying } u(x_1, x_2, 0) \geq g(x_1, x_2), \right.$$
$$\left. \lim_{x \to \infty} u(x) = 0 \right\}, \tag{2.2}$$

[1] This extra relation comes from two considerations. First, the normal velocity of the interphase, V, is proportional to the amount of heat absorbed by it (and used to melt the ice). In turn, this flow of heat "entering" the interphase is, by Fourier's law, proportional to the gradient of temperature. Hence, we have $|V| = C|\nabla\theta|$. Second, since $\theta = 0$ on the moving interphase, we obtain that, on it, V and $\nabla\theta$ are parallel and $(\partial_t + V \cdot \nabla)\theta = 0$. Combining the two previous equations and choosing the physical units to make $C = 1$, we obtain Stefan's condition.

[2] In the sense of Hadamard, i.e., given initial and boundary conditions, is there a unique solution which depends continuously on the given data?

[3] If one knew, for instance, that all points are in contact, then the initial and final position of all the boundary points of the body would be obviously determined, and resolving the body's deformation would be much simpler!

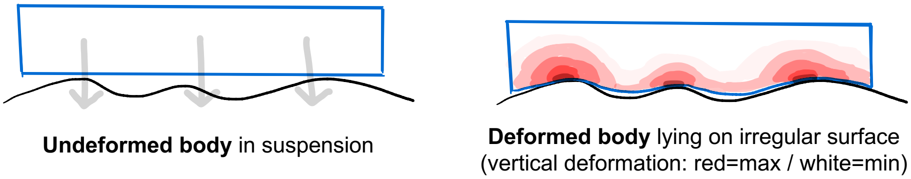

Undeformed body in suspension
Deformed body lying on irregular surface
(vertical deformation: red=max / white=min)

Figure 3. Signorini's problem: an elastic body lying on a surface.

where $g : \mathbb{R}^2 \to \mathbb{R}$ is a smooth prescribed function satisfying $\limsup_{x\to\infty} g < 0$. This problem is often called the *thin obstacle problem* and has other applications as well, such as in the modeling of semipermeable membranes. See [5, 13, 22, 38] and references therein for more information on this problem.

The model (2.2) can be used when the bottom surface of the (undeformed) elastic solid is a small perturbation of a horizontal plane. In order to derive it more easily, let us consider the following variant of the problem: Assume that the bottom surface of the (undeformed) elastic solid is (exactly) a horizontal plane and that, instead, the rigid surface on which the body will rest is a small perturbation of a horizontal plane. The horizontal surface is then described as $\{x_3 = \varepsilon g(x_1, x_2)\}$, where $g : \mathbb{R}^2 \to \mathbb{R}$ is some bounded function—which we assume to be smooth—and where $\varepsilon > 0$ is small. This situation is depicted in Figure 3. Let us suppose for simplicity that $g = -1 + \bar{g}$, where $\bar{g} : \mathbb{R} \to [0, 1]$ is smooth and compactly supported.

The undeformed body "suspended in air" corresponds to $U := \mathbb{R}^3 \cap \{x_3 \geq 0\}$. As we let it rest on the rigid surface, it experiences a deformation. For ε small, horizontal deformations may be neglected, and we can think that the displacements are only vertical. More precisely, there is a function $u : U \to \mathbb{R}$ and a constant $c \in (0, 1)$ such that the point of the solid before occupying the position $(x_1, x_2, x_3) \in U$ in the suspended configuration, now occupies the position $(x_1, x_2, x_3 + \varepsilon u)$ in the resting configuration. We are considering for simplicity the "boundary condition at infinity" $\lim_{x\to\infty} u = -c \in (-1, 0)$, but it would not be difficult to consider other (more realistic) boundary conditions by modifying (2.2) accordingly.

The elastic energy of the deformed body is proportional (at leading order in ε) to $\int_{\mathbb{R}^2} |\nabla u|^2$. Hence, up to replacing u and g by $u - c$ and $g - c$, we obtain (2.2).

Now, although the minimization problem (2.2) leads to a nonlinear Euler–Lagrange equation, the fact that the (convex) Dirichlet energy is minimized inside the convex set $\{u \in H_0^1(U) : u(\cdot, \cdot, 0) \geq g\}$ confers it a very nice mathematical structure. The study of Singorini's problem was the starting point for the study of other similar convex constrained minimization problems with free boundaries, initiating in the 1960s the field of *variational inequalities*.

2.4. The obstacle problem (1960s)

Conceived as a paradigmatic variational inequality, the obstacle problem originates in the papers [8,30,32,35]. The initial motivation of the problem (which gives its name) concerned Plateau's problem with an obstacle. Namely, given a concave function $\psi : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^2$ is a convex smooth domain, and some boundary values $h : \partial\Omega \to \mathbb{R}$ satisfying $h \geq \psi|_{\partial\Omega}$, find a surface with minimal area among all graphs lying above the *obstacle* ψ and spanning the curve $\{(x, h(x)) : x \in \partial\Omega\}$. In other words,

$$\min \left\{ \int_\Omega \sqrt{1 + |\nabla v|^2}\, dx \text{ among } v \geq \psi \text{ satisfying } v|_{\partial\Omega} = h \right\}. \tag{2.3}$$

Determining where the surface will be in contact with the obstacle is part of the problem.

The obstacle problem is actually the "small perturbation version" of (2.3), namely, the same minimization problem where $\sqrt{1 + |\nabla v|^2}$ is replaced by $\frac{1}{2}|\nabla v|^2$. Computing its first variation with respect to nonnegative perturbations, one finds that a minimizer u must satisfy the variational inequality

$$\int_\Omega \nabla v \cdot \nabla \xi \, dx \geq 0 \text{ in } \Omega, \quad \text{for all } \xi \in C_c^\infty(\Omega) \text{ such that } \xi \geq 0. \tag{2.4}$$

Using this one can show that v is lower-semicontinuous, and hence the set $\{x \in \Omega : v(x) > \psi\}$, denoted by $\{v > \psi\}$ for brevity, is open. Since inside the set $\{v > \psi\}$ the solution v can be slightly perturbed in both the upwards and downwards directions, its minimality yields

$$\int_\Omega \nabla v \cdot \nabla \eta \, dx = 0 \text{ in } \Omega, \quad \text{for all } \eta \in C_c^\infty\big(\{v > \psi\}\big). \tag{2.5}$$

Considering the new function $u := v - \psi$ and integrating by parts in (2.4)-(2.5), we obtain the PDE

$$\begin{cases} \Delta u = \max(0, -\Delta\psi), \\ u \geq 0, \end{cases} \tag{2.6}$$

which is also called the obstacle problem.

Although the original motivation of the obstacle problem does not seem very deep, much more interesting applications have been found in the last decades. A beautiful one concerns the configuration of a cloud of Coulomb charges (all with the same sign), which are kept together by a confining electric potential. In the asymptotic regime corresponding to a very large number of charges, the potential generated by them solves a problem of the form (2.6) (see for instance [43] or the introduction of [49]). Other well-known applications are the dam problem (fluid filtration) and optimal stopping problems (for Finance and Probability). As we will see, the obstacle problem in the particular case $-\Delta\psi \equiv 1$ is also closely connected with Stefan's problem.

2.5. Almgren's problem (1970s)

After their existence theory was established in the 1960s (see [20, 21]), the question of (partial) regularity for oriented area-minimizing m-dimensional surfaces in $\mathbb{R}^{m+k}$ (more precisely integer rectifiable area-minimizing m-currents), in codimension $k \geq 2$ was a very natural one. For the sake of brevity, throughout this note we will use the keyword *Almgren's problem* to refer to this problem. It is a convenient and probably fair name for the problem, since Almgren anticipated its mathematical significance and studied it in depth during the last two decades of his life. His complete resolution was published in a famous 950-page posthumous paper [4].

The details of Almgren's proof are so intricate that its correctness was rather a myth until De Lellis and Spadaro deciphered its key ideas and bridged them with shorter (and clearer) arguments in a series of recent papers (see [17] and references therein). In Section 5, we will describe (very roughly) some ideas from this monumental proof, since they have clear parallels in our recent results for the Stefan problem and the obstacle problem.

One aspect that makes codimension $k \geq 2$ area-minimizing surfaces particularly delicate is the phenomenon of branching. As it was already known before the 1970s (by a classical result of Wirtinger and Federer; see [16, Section 1.2] for details), holomorphic "curves" are area-minimizing 2-surfaces in $\mathbb{R}^4$. For example, we can consider $S := \{(x_3 + ix_4)^2 = (x_1 + ix_2)^3\} \subset \mathbb{C}^2 \cong \mathbb{R}^4$. Note that S is not smooth at 0: it has a delicate type of singularity called *branching singularity*. While—as we will see in Section 3—singularities in soap-film-like area-minimizing surfaces in $\mathbb{R}^3$ (or in integer rectifiable codimension 1 currents in $\mathbb{R}^n$ for all n) are always of conical type, zooming in infinitely at a branching singularity, we always obtain a plane, just as in smoothness points. However, near branching singularities, the surface is really a multiple-valued graph over the tangent plane. As we will see, this feature makes the analysis of the problem in codimension $k \geq 2$ much harder than in the case $k = 1$.

3. Classical regularity theory for Plateau's problem (1960s)

In Section 2.1, we stated the three main questions (i)-(ii)-(iii) associated to Plateau's problem. Similar questions apply to all the problems considered here. Now, in the case of Plateau's problem, existence question (i) is very challenging, and the multiple (all valid) answers to it obtained during the XXth century were celebrated breakthroughs (see references in Section 2.1). However, the discussion of (i) does not reveal any parallelism between Plateau's problem and the obstacle problem or Stefan's problem. For this reason (and also because it would take too much space), we will not discuss (i) here and will focus on the regularity part: questions (ii) and (iii).

Although we do not discuss (i), it is perfectly possible to intuitively understand most of the main ideas in the regularity theory for Plateau's problem without giving a completely rigorous definition of "area-minimizing surface". For our purposes, it is enough to think of physical soap films.

In the rest of the section, Γ will denote some prescribed (reasonably regular) contour. We postulate the existence of an "area-minimizing surface" (a physical soap film) spanning Γ, which we denote by S. We review next some of the main ingredients of the regularity theory for such S.

3.1. Minimal surface equation (1760s)

As found by Lagrange in the 1760s, smooth pieces of an area-minimizing surface must have zero mean curvature. As a consequence, if a piece of the surface can be represented by a C^1 graph, then it solves a uniformly elliptic equation with continuous coefficients. Then, linear methods in elliptic PDE (Schauder estimates) can be used to show that the piece of surface must be analytic.

3.2. De Giorgi's "flatness implies smoothness" principle (1961)

One of the most fundamental results for area-minimizing surfaces is the following theorem of De Giorgi [14] (see also [31]). We give a slightly modified version of statement (not involving the excess) due to Savin [46].

Theorem 1 ([14]). *There exists $\varepsilon_\circ > 0$ dimensional such that the following holds. Assume that S has minimal perimeter inside B_1 (i.e., the curve Γ which the soap film S spans does not intersect B_1) such that $S \cap B_1 \subset \{|x_n| \leq \varepsilon_\circ\}$. Then, ∂S is an analytic graph in $B_{1/2}$.*

It will become clear in the next subsections that this theorem is a fundamental pillar of the theory. Let us now recall the heuristic behind its proof: let $B_1' \subset \mathbb{R}^2$ be the unit ball and suppose that $S = \{x_3 = \varepsilon g(x_1, x_2)\}$ with $\varepsilon > 0$ tiny and $g : B_1' \to \mathbb{R}$ bounded. Then the area of S is given by

$$\int_{B_1'} \sqrt{1 + \varepsilon^2 |\nabla g|^2}\, dx_1\, dx_2 = \pi + \varepsilon^2 \int_{B_1'} \frac{|\nabla g|^2}{2}\, dx_1\, dx_2 + O(\varepsilon^4).$$

Hence, for $\varepsilon \downarrow 0$, the fact that $x_3 = \varepsilon g(x_1, x_2)$ has minimal area should imply that g (which is nothing but the x_3 coordinate on S, as a function of x_1, x_2, and divided by ε) must be, approximately, a minimizer of $\int_{B_1'} \frac{|\nabla g|^2}{2}\, dx_1\, dx_2$. In other words, g is approximately harmonic.[4] As a consequence (of this happening at every scale and near every point of S), the smoothness of the limiting harmonic functions as $\varepsilon \downarrow 0$ is inherited by S, which can be shown to be a $C^{1,\alpha}$-graph. The minimal surface equation then implies its analyticity.

3.3. Fleming's monotonicity formula and tangent cones (1962)

A very useful consequence of the minimality of S is the so-called "monotonicity formula". Fix $x \in S$. For $r > 0$, let $B_r(x) \subset \mathbb{R}^3$ denote the Euclidean ball of radius r centered at x. Given $r > 0$ such that $B_r(x) \cap \Gamma = \varnothing$ (recall that Γ is the contour spanned by S), let us consider the dimensionless quantity

$$a_x(r; S) := \frac{1}{r^2} \operatorname{Area}\big(S \cap B_r(x)\big). \tag{3.1}$$

Then, $a_x(r; S)$ is monotone nondecreasing in r (this was first shown in [21]).

To prove the monotonicity formula (at $x = 0$), one compares the area of S in $B_r(0)$ with the area of "competitors" S_t obtained glueing the rescaled surface tS for some $t \in (0, 1)$ inside $B_{tr}(0)$ with a "conical interpolation" $\{x \in \mathbb{R}^n : \frac{rx}{|x|} \in S\}$ in the annulus $B_r \setminus B_{tr}$ (note that in this way S_t coincides with S on ∂B_r).

One can show that $a_x(r; S)$ is constant between $r = 0$ and $r = R$ if and only if S is conical inside B_R, that is, if $t(S \cap B_R) = S \cap B_{tR}$ for all $t \in (0, 1)$.

The previous observation gives essential information on area-minimizing surfaces: they must have conical structure around each point "when looked at the microscope". More precisely, let us consider the "zoomed-in" (around x) surfaces $S^{x,r} := \frac{1}{r}(S - x)$ for $r > 0$. For any fixed $R > 0$, $a_0(R; S^{x,r}) = a_x(Rr; S)$ is monotone increasing (in r) and converges to the constant $a_x(0^+, S)$ as $r \downarrow 0$. Hence we have $0 \le a_0(R; S^{x,r}) - a_0(0^+; S^{x,r}) \downarrow 0$ as $r \downarrow 0$. As a consequence, one can prove that the surface $S^{x,r}$ must be closer and closer to some cone inside any fixed ball B_R, as $r \downarrow 0$. This crucial property was first noticed in [28].

3.4. The classification of minimal cones: Taylor, Almgren, and Simons

By the discussion in the previous subsection, for any given $x \in S$, the "zoomed-in" surface $S^{x,r} \cap B_1$ is arbitrarily close to some area-minimizing cone $\mathscr{C}$, provided that we take r small enough (possibly depending on x). This leads us to the question: what are the possible area-minimizing cones $\mathscr{C}$? The answer depends on the type of objects which we want to admit as "surfaces". As proven in [53], in the case of "soap-film-like minimal surfaces" (Reifenberg [41] or Almgren [3]), there are exactly three

[4] The actual proof of this kind of statement is, of course, more complicated than that: to start with S does not need to be a graph, so first, one must suitably approximate it by graphs, and then one needs to understand how to transfer the regularity of harmonic functions to S. But this gives a good enough idea on how the proof works.

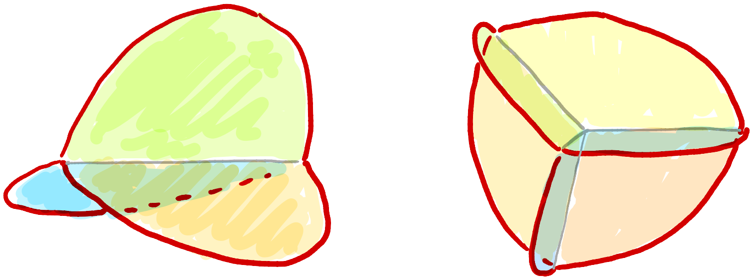


Figure 4. Possible singularities in soap-film-like minimal surfaces: Y type (left) and tetrahedron type (right).

possibilities: a plane, three half-planes meeting in Y shape with angles of 120°, or the cone generated by the edges of a regular tetrahedron centered at 0 (see Figure 4).

An easy computation shows that

$$a_0(r;\mathscr{C}) \equiv \begin{cases} \ell_1 : \pi \approx 3.1 & \text{if } \mathscr{C} \text{ is a plane,} \\ \ell_2 := 3\pi/2 \approx 4.7 & \text{if } \mathscr{C} \text{ is of } Y \text{ type,} \\ \ell_3 := 3\arccos(-1/3) \approx 5.7 & \text{if } \mathscr{C} \text{ is of tetrahedron type.} \end{cases} \tag{3.2}$$

Hence, thanks to Fleming's monotonicity formula, for every point x in a soap film S (in $\mathbb{R}^3$), zoomed-in surface $S^{x,r}$ (r tiny) must be close to one of the previous three possible cones. Moreover, the type of cone is determined by the value of $a_x(0^+; S)$, which necessarily belongs to $\{\ell_1, \ell_2, \ell_3\}$, as (3.2).

Based on our experience when observing physical soap films, we would expect that the zoomed-in surfaces should look like a plane around "most points", but still one important idea is still needed to show this (see next subsection). Still, we can already start to devise the power of De Giorgi's theorem and Fleming's monotonicity formula combined. They imply that for any given $x \in S$, if there exists $r > 0$ such that $a_x(r; S) < \ell_2$, then S will be analytic in some neighborhood of x.[5] Such points x are called *regular points*. All other points are called *singular points*.

Let us close the subsection with an important remark: if instead of soap-film-like minimal surfaces we had considered *boundaries of sets of minimal perimeter* (resp. *integer rectifiable area-minimizing* 2*-currents*) in $\mathbb{R}^3$, then the only possible minimal cones would have been the planes. In particular, De Giorgi's theorem implies that such notions of area-minimizing surfaces in $\mathbb{R}^3$ are analytic unconditionally.

[5] Indeed, since a_x is monotone and $a_x(0^+, S)$ must take one of the three values in (3.2), the assumption $a_x(r, S) < \ell_2$ implies that the value at 0^+ can only be π. Hence for small enough scales, $S^{x,r}$ will be arbitrarily close to a plane, and then De Giorgi's theorem implies its analyticity.

On the other hand, the same strategy—described here for surfaces in $\mathbb{R}^3$—works for hypersurfaces in $\mathbb{R}^n$. In that case, Almgren [2] for $n = 4$ and Simons [51] for $5 \le n \le 7$ proved that if $\mathscr{C}$ is an area-minimizing (hyper-)cone in $\mathbb{R}^n$ and $\mathscr{C} \cap \mathbb{S}^{n-1}$ is smooth, then $\mathscr{C}$ must be a hyperplane. This classification result is important because one can deduce from it that *boundaries of sets of minimal perimeter (resp. integer rectifiable area-minimizing* $(n-1)$*-currents) are analytic in dimensions* $n \le 7$. This dimension 7 is sharp since Simons's cone $\{x_1^2 + x_2^2 + x_3^2 + x_4^2 = x_5^2 + x_6^2 + x_7^2 + x_8^2\}$ is an example of area-minimizing surface (with respect to the two previous notions) in $\mathbb{R}^8$, as shown in [7].

3.5. Federer's dimension reduction principle and partial regularity theorems

In order to complete our heuristic overview of the classical regularity theory for area-minimizing surfaces in $\mathbb{R}^3$, a last key idea is missing: the dimension reduction principle. The first observation we need to make is that the map $\mathfrak{m} : S \to \{\ell_1, \ell_2, \ell_3\}$ defined as

$$\mathfrak{m}(x) := a_x(0^+, S) = \inf\{a_x(r; S) : r > 0\}$$

will be upper-semicontinuous, since it is an infimum of continuous functions. As a consequence, the set of tetrahedron type singular points $X_3 := \{\mathfrak{m} = \ell_3\}$ is closed.

In order to glimpse how Federer's dimension reduction argument works, let us show that X_3 is discrete. Indeed, assume by contradiction that $x_k \in X_3$ converges to $x_\infty \in \mathbb{R}^n \setminus (\bigcup_k \{x_k\} \cup \Gamma)$. Since X_3 is closed, x_∞ belongs to X_3.

Now given $\varepsilon > 0$ arbitrarily small, we can choose $r_\varepsilon > 0$ (depending on x_∞) such that $0 \le a_{x_\infty}(r_\varepsilon, S) - \ell_3 < \varepsilon/2$. On the other hand, since $a_x(r_\varepsilon, S)$ is continuous in x, there exist $\varrho_\varepsilon \in (0, r_\varepsilon)$ such that $0 \le a_x(r_\varepsilon; S) - \ell_3 < \varepsilon$ for all $x \in X_3 \cap B_{\varrho_\varepsilon}(x_\infty)$. Since $x_k \to x_\infty$, for k sufficiently large, we will have $r_k := |x_k - x_\infty| < \varrho_\varepsilon$.

Let us now zoom in: consider $S_k^* := S^{x_\infty, r_k}$ and define $x_k^* := (x_k - x_\infty)/r_k$. Note that, by definition, x_k^* belongs to $\mathbb{S}^2$. By scaling, we have $a_0(1; S_k^*) = a_{x_\infty}(r_k; S)$ and $a_{x_k^*}(1; S_k^*,) = a_{x_k}(r_k; S)$. Hence, by definition of ϱ_ε,

$$\ell_3 = a_0(0^+; S_k^*) \le a_0(S_k^*, 1) = a_{x_\infty}(r_k; S) < a_{x_\infty}(\varrho_\varepsilon; S) \le \ell_3 - \varepsilon$$

and

$$\ell_3 = a_{x_k^*}(0^+; S_k^*) \le a_{x_k^*}(1; S_k^*) = a_{x_k}(r_k; S) < a_{x_\infty}(\varrho_\varepsilon; S) \le \ell_3 - \varepsilon.$$

Hence choosing ε sufficiently small (and k sufficiently large), we find that

- S_k^* will be arbitrarily close to a cone of tetrahedron type (centered at 0),
- S_k^* will be arbitrarily close to a cone about the point $x_k^* \in \mathbb{S}^2$.

This gives an obvious contradiction since the cone of tetrahedron type is clearly not a cone about any of its points in $\mathbb{S}^2$.

A refined version of the same type of argument allows one to show that singular points of Y type have Hausdorff dimension[6] at most one.[7] This type of argument is often called Federer's dimension reduction and works in several contexts where "zoomed-in objects" have some conical structure. The basic principle can be summarized as follows: if $X \subset \mathbb{R}^n$ is at the same time a cone about 0 and about another point $x^* \neq 0$, then X must be translation invariant in the direction x^* (since $tX = X$ and $t(X - x^*) = X - x^*$ for all $t > 0$ imply that $X - (t-1)x^* = X$).

4. Stefan's problem and the obstacle problem during 1970s–2000s

4.1. Duvaut's transformation

From the XIX century formulation of the Stefan problem explained in Section 2.2, it was not even clear if the problem was well posed. A key development was obtained in 1973 by Duvaut [19], who revealed a hidden convex structure in the problem: recall that θ denotes the temperature and consider the transformation

$$u(x,t) := \int_0^t \theta(x,\tau)\, d\tau.$$

Duvaut showed that the new function

$$u : \Omega \times \mathbb{R}^+ \to \mathbb{R}^+$$

satisfies

$$\begin{aligned} \partial_t u - \Delta u &= -\chi_{\{u>0\}}, \\ u &\geq 0, \\ \partial_t u &\geq 0, \end{aligned} \tag{4.1}$$

where χ_A denotes the characteristic function of the set A.

By the strong maximum principle, if u is Duvaut's transformation of a temperature solving the Stefan problem, then it also satisfies the strict monotonicity property

$$\partial_t u > 0 \quad \text{in } \{u > 0\}. \tag{4.2}$$

This seemingly qualitative property was never used in the regularity theory developed in the 1970s. Still, we state it here because it plays an important role in the recent results.

[6] We recall that a subset $X \subset \mathbb{R}^n$ is said to have Hausdorff dimension $\beta \in [0,n]$ if for all $\beta' > \beta$ and for all $\delta > 0$ there exist countably many balls $B_{r_i} z_i$ covering X such that $\sum_i (r_i)^\beta < \delta$. One can easily check from this definition that the Hausdorff dimension of an m-plane in $\mathbb{R}^n$ is m.

[7] Actually, Y points form analytic curves by the deep results in [34, 53].

Since we can easily recover θ from u by computing its time derivative, we see that (4.1) is an equivalent formulation of the Stefan problem. The new formulation is useful because (4.1) enjoys a convex structure: it is the L^2-*gradient flow*[8] of the convex functional

$$J(u) = \int_{\Omega} \left(\frac{1}{2} |\nabla u|^2 + \max(0, u) \right) dx. \tag{4.3}$$

As a consequence, questions such as the well-posedness of solutions become much simpler in the new formulation.

4.2. Stefan's problem as a parabolic obstacle problem

Let us notice that stationary (constant-in-time) solutions of (4.1) satisfy exactly the obstacle problem (2.6) in particular case $-\Delta\psi \equiv 1$, that is

$$\begin{aligned} \Delta u &= \chi_{\{u>0\}}, \\ u &\geq 0. \end{aligned} \tag{4.4}$$

In this respect, (4.1) is a parabolic version of the obstacle problem (4.4). Solutions to (4.4) are critical points (and hence minimizers, since the functional is convex) of $J(u)$. Note that such constant-in-time solutions of (4.1) are never solutions of the Stefan problem (never arise as Duvaut transforms of some temperature) since, for instance, they do not satisfy (4.2). Still, understanding the regularity of the free boundaries for the obstacle problem (4.4) is a logic first step before dealing with time-dependent solutions.

4.3. Obstructions to regularity of the free boundary: Schaeffer's examples (1977)

To study (4.4), a first thing one might try is to construct some explicit solutions. In most simple cases, the obtained free boundaries are smooth.

However, it is possible to find singular free boundaries even in two dimensions, as done by Schaeffer in [48]. He used complex variables to construct solutions of (4.4) in $\mathbb{R}^2$ in which the free boundary has a cusp represented by the curve (Figure 5, left)

$$x_2 = \pm x_1^{2k+\frac{1}{2}}, \quad 0 \leq x_1 \leq 1,$$

[8] The solution u satisfies, for infinitesimal $\tau > 0$,

$$u(\cdot, t+\tau) = \arg\min \left(J(v) + \frac{1}{2\tau} \|v - u(\cdot, t)\|^2_{L^2(\Omega)} \right) + o(\tau),$$

where the minimum is among functions $v : \Omega \to \mathbb{R}^n$ satisfying the prescribed boundary condition for u at time $t + \tau$, i.e., $v = u(\cdot, t+\tau)$ on $\partial\Omega$.

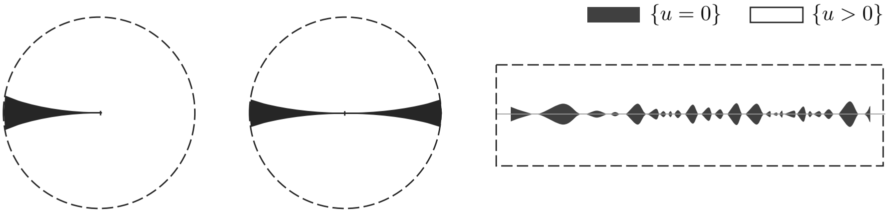


Figure 5. Schaeffer's examples of singular free boundaries.

where $k \in \mathbb{N}$. In this family of examples, the set $\{u > 0\}$ is actually the image of $\{|z| \le 1, \operatorname{Im} z > 0\}$ under the conformal mapping $f(z) = z^2 + i\, z^{4k+1}$, and u satisfies near the origin

$$u(z) \approx \frac{x_2^2}{2} + c_k \operatorname{Im}\big(z^{2k+\frac{3}{2}}\big) + \cdots,$$

where $z = x_1 + ix_2$.

Another family type of singularities (two-sided cusps) was also constructed by Schaeffer (Figure 5, center). In this case, the free boundary is represented by the curves

$$x_2 = \pm\,|x_1|^{2k}, \quad -1 \le x_1 \le 1.$$

In the case of general smooth concave obstacles ψ, Schaeffer noticed that solutions to (2.6) may even have infinitely many cusps (Figure 5, right).

4.4. Caffarelli's breakthrough (1977)

It was not until 1977, with the groundbreaking paper of Caffarelli [9] (and with the paper [33]), that the "modern" regularity theory for (4.1) and (4.4) was initiated. Since, as explained before, (4.4) is a particular case of (4.1)—that of constant-in-time solutions—Caffarelli's results described next apply at the same time to both the obstacle problem and Stefan's problem.

The approach of Caffarelli to the regularity of free boundaries of (4.1)—or of (4.4)—has some rough similarities with the regularity theory of area-minimizing hypersurfaces described in Section 3. In Caffarelli's regularity theory (as in area-minimizing surfaces), *blow-ups* (limiting zoomed-in objects) are central actors. Informally speaking, Caffarelli looks at points on the free boundary through the microscope, and infers "macroscopic properties" of the free boundary from the "microscopic" ones.

For (4.1) the natural scaling of the equation suggests considering, for given $(x_\circ, t_\circ) \in \partial\{u > 0\}$ and $r > 0$,

$$u^{x_\circ, t_\circ, r}(x,t) := \frac{1}{r^2} u(x_\circ + rx, t_\circ + r^2 t).$$

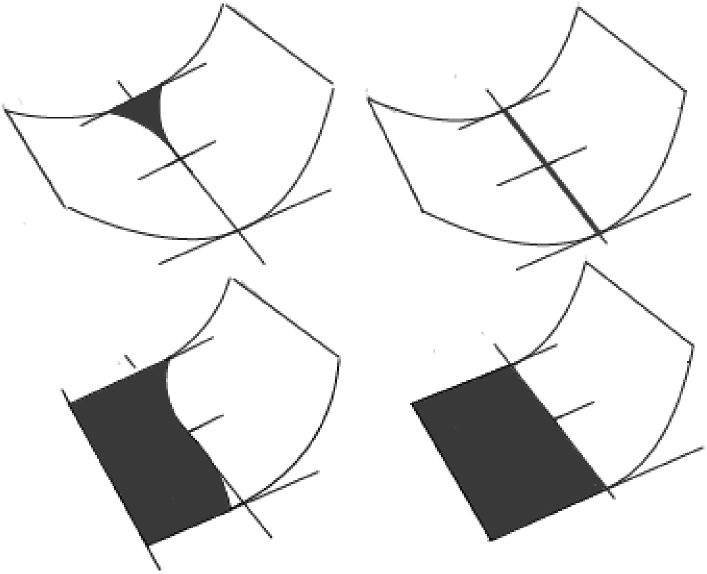


Figure 6. Illustration of the dichotomy in Caffarelli's theorem.

It is easy to see that $u^{x_\circ,t_\circ,r}$ is again a solution of (4.1). *Blow-ups* are defined as accumulation points as $r \downarrow 0$ of $u^{x_\circ,t_\circ,r}$.

The main result from Caffarelli [9], combined with the fundamental analyticity result for C^1 free boundaries of Kinderlehrer and Nirenberg [33], is stated next. (See Figure 6 for an illustration of the result.)

Theorem 2. *Let $\Omega \subset \mathbb{R}^n$ and let $u : \Omega \times (0,T) \to \mathbb{R}$ be a solution of* (4.1)*. For every $(x_\circ, t_\circ)$ belonging to the free boundary $\partial\{u > 0\}$, one of the following two alternatives holds:*

(a) *$u^{x_\circ,t_\circ,r} \to \frac{1}{2}(\max(0, e \cdot x))^2$ as $r \downarrow 0$, for some $e \in \mathbb{S}^{n-1}$; and the free boundary is a (moving) smooth embedded $(n-1)$-surface near $(x_\circ, t_\circ)$;*

(b) *$u^{x_\circ,t_\circ,r_k} \to \frac{1}{2}x \cdot Ax$ for some $r_k \downarrow 0$ and some nonnegative definite matrix A with* $\operatorname{trace}(A) = 1$*; and the free boundary has a cusp-like singularity at $(x_\circ, t_\circ)$.*

Besides the idea of considering blow-ups, the methods used by Caffarelli to prove this result were rather different than those employed in area-minimizing surfaces. For instance, in Caffarelli's approach, a convexity property of blow-ups is crucially used in its classification, and his methods "based on the maximum principle" can be applied to more general non-variational problems, such as fully nonlinear obstacle problems. For a self-contained overview of Caffarelli's 1977 proof, we refer the reader to [23].

4.5. Weiss' epiperimetric inequality approach (1999)

In the paper [54], Weiss introduced a new monotonicity formula for the obstacle problem, which is in many respects analogous to the Federer–Fleming monotonicity formula for area-minimizing surfaces. Given a solution u of the obstacle problem (4.4), and $x_\circ \in \partial\{u > 0\}$, he introduced the *adjusted energy* (recall that the functional J was defined in (4.3))

$$W_{x_\circ}(r, u) = \frac{1}{r^{n+2}} J\big(u; B_r(x_\circ)\big) - \frac{1}{r^{n+3}} \int_{\partial B_r} u.$$

He proved that $W_{x_\circ}(r, u)$ is monotone nondecreasing in r, and constant if and only if u is 2-homogeneous (i.e., $u(tx) = t^2 u(x)$ for all $t > 0$).

Similarly to what happens with area-minimizing surfaces, this monotonicity formula follows from comparing the energy of the solution J of u in B_r with the energy of its natural competitor (defined for given $t \in (0, 1)$)

$$\tilde{u}(x) := \begin{cases} \frac{|x|^2}{r^2} u\big(\frac{rx}{|x|}\big) & x \in B_r \setminus B_{tr}, \\ t^2 u(tx) & x \in B_{tr}. \end{cases}$$

Using Weiss' monotonicity formula, one can show that blow-ups in the obstacle problem must be 2-homogeneous. Similarly, as we explained with $a_{x_\circ}(r, S)$ in the case of Plateau's problem, $W_{x_\circ}(0^+, u)$ can only take two different values: $\frac{|B_1|}{8(n+2)}$ and $\frac{|B_1|}{4(n+2)}$. The lowest possible value defines regular points, while the higher value is attained at singular points.

The paper [54] was the first to introduce methods for the obstacle problem which had a very strong parallelism with those for area-minimizing surfaces (e.g., an "epiperimetric inequality"), reinforcing the connection between the two theories.

4.6. First regularity results on the singular set and open questions

After the results of Caffarelli [9], a natural question was as follows: what else can be said about the singular set?

Besides some first results in two dimensions [11], there was no real progress on this question until 1991, when Sakai [44, 45] obtained a very precise description of singularities for the obstacle problem (4.4) in $\mathbb{R}^2$. He essentially proved that the cusps of Schaeffer (and small analytic perturbations of them) are the only ones which may appear for (4.4) in $\mathbb{R}^2$.

In dimensions $n \geq 3$, where complex analysis is of no use, the first results on the singular set were established again by Caffarelli in 1998 [10] (using the Alt–Caffarelli–Friedmann monotonicity formula) and by Monneau in 2003 [37] (using a new monotonicity formula based on the Weiss one). They established that a solution

u of (4.4) in $\mathbb{R}^n$ ($n \geq 3$) must satisfy at any singular point $x_\circ$,

$$u(x_\circ + x) = \frac{1}{2} x \cdot Ax + o\big(|x|^2\big). \tag{4.5}$$

As a consequence of (4.5) and Whitney's extension theorem, one obtains that the singular set enjoys spatial C^1-regularity, in the sense that they can be covered by $(n-1)$-manifolds of class C^1. Still, this result seems rather weak in the sense that it does not prevent the singular set from being as large as the regular part of the free boundary. In this direction, the following conjecture holds.

Conjecture 3 (Schaeffer [47]). *Generically, solutions of the obstacle problem have smooth free boundaries.*

In other words, the conjecture states that, generically, the free boundary has *no* singular points. Here the word "generically" must be interpreted as "for most boundary values". Until very recently (see Section 6), Conjecture 3 was only known to hold in the plane $\mathbb{R}^2$, a result of Monneau [37].

In the evolutionary case, the "parabolic analog" of Monneau's monotonicity formula [37] for solutions to (4.1) and its consequences were investigated in [6, 36].

5. Almgren's problem and the thin obstacle problem during 1970s–2000s

5.1. Branching singularities of holomorphic curves

As explained in Section 2.5, holomorphic curves such as $S_1 := \{w^2 = z^3\} \subset \mathbb{C}^2 \cong \mathbb{R}^4$, where $w = x_3 + ix_4$ and $z = x_1 + ix_2$, are examples of area-minimizing 2-surfaces in $\mathbb{R}^4$. In the case of S_1, we can write x_3 and x_4 as "two valued functions" of x_1 and x_2: since $\zeta = \sqrt{z^3}$ involves a complex square root, there are two possible values of (x_3, x_4) for each pair (x_1, x_2).

Let us consider two further examples: $S_2 := \{(w - z^2)^2 = z^5\}$ and $S_3 := \{(w - z^2 + z^3)^3 = z^{11}\}$. Both have branching singularities at 0, but they look even more complicated than the case of S_1. In order to understand the singularity of S_2, we need to proceed as follows: we first consider the change of variables $\zeta(z, w) = w - z^2$ and notice that the coordinates (z, ζ) are diffeomorphic to (z, w) near the origin. In the new coordinates, we have $S_2 := \{\zeta^2 = z^5\}$, so we see that the singularity has again two branches (from the complex square root involved in $\zeta = \sqrt{z^5}$). Only after we rectify the coordinates, we can clearly see the structure of the branching singularity of S_2. Something similar happens for S_3. In that case, the new coordinates would be $\zeta(w, z) = w - z^2 + z^3$ and the model singularity $\{\zeta^3 = z^{11}\}$, with three branches from $\sqrt[3]{\cdot}$.

5.2. Almgren's regularity theorem

In [4], Almgren established the following theorem.

Theorem 4. *Let S be an oriented area-minimizing surface*[9] *of dimension $n \geq 2$ in $\mathbb{R}^{m+k}$, where $k \geq 2$.*

Then, S is an analytic submanifold in $\mathbb{R}^{m+k} \setminus \Gamma$, where Γ denotes the boundary[10] *of S, with the exception of a closed set* $\mathrm{Sing}(S)$ *of dimension at most $m-2$ (discrete if $m = 2$).*

The dimensional estimate for the singular set is optimal, as shown by holomorphic curves with branching points.

5.3. Q-valued harmonic functions, frequency formula

In Section 5.1, we saw examples of branching singularities in explicit holomorphic curves. Let us explain next in what sense general oriented area-minimizing surfaces resemble holomorphic curves.

Suppose that $S \subset \mathbb{R}^4$ is any area-minimizing oriented 2-surface[11] and that 0 is a non-smoothness point on it (e.g., an integer rectifiable area-minimizing current). Similarly, as in Section 3.3, $a_0(r; S)$ is monotone nondecreasing and the zoomed-in surfaces $S^{0,r}$ converge towards a cone $\mathcal{C}$, now a 2-surface in $\mathbb{R}^4$. It is not difficult to show that planes are the only possible area-minimizing oriented 2-cones in $\mathbb{R}^4$. The difficulty now is that $\mathcal{C}$ could be a plane with "multiplicity two or higher"; in other words, we could have $a_0(0^+; S) = Q\pi$, for some $Q \geq 2$ (as it happens in branching singularities of holomorphic curves). Note that this cannot happen for codimension 1 surfaces, thanks to De Giorgi's theorem.

What can one do at those "multiplicity points"? Assume first that, up to a rotation, $S^{0,r}$ is close to the plane $\{x_3 = x_4 = 0\}$. If $S^{0,r}$ happened to be (locally near 0) a very flat multiplicity one graph $x_i = \varepsilon f_i(x_3, x_4)$, $i = 1, 2$, then its surface area would be given by an integral of the type

$$\int \sqrt{\left(1+\varepsilon^2|\nabla f_1|^2\right)\left(1+\varepsilon^2|\nabla f_2|^2\right) - \varepsilon^4(\nabla f_1 \cdot \nabla f_2)^2}\, dx_1\, dx_2$$

$$\approx \pi + \frac{\varepsilon^2}{2} \int \left(|\nabla f_1|^2 + |\nabla f_2|^2\right) dx_1\, dx_2.$$

Hence, both f_1 and f_2 would need to be approximate minimizers of the Dirichlet energy! Something similar happens when $\boldsymbol{f} = (f_1, f_2)$ is not a single-valued

[9] Rigorously, assume that S is an integer rectifiable area-minimizing current.

[10] More precisely, Γ is the support of the boundary of the current S.

[11] Integer rectifiable current.

map from $\mathbb{R}^2 \to \mathbb{R}^2$, but a multiple-valued one. More precisely, the pair of functions $\boldsymbol{f}(x_1, x_2)$ do not "return" a point in $\mathbb{R}^2$, but a Q-tuple of them: all the pairs $(x_3/\varepsilon, x_4/\varepsilon)$ for which (x_1, x_2, x_3, x_4) belongs to S. Still in this case, the area would be given by an analogous expression as above. And, as before, the fact that the S is area-minimizing should imply that, as $\varepsilon \downarrow 0$, the multiple-valued functions are approximate minimizers of the Dirichlet energy—appropriately generalized to the context of multiple-valued functions.

The Q-valued Dirichlet minimizers $\boldsymbol{f} : \mathbb{R}^m \to (\mathbb{R}^k)^Q/\!\sim$, where $\sim$ identifies Q-tuples of points in $\mathbb{R}^k$ which are equal up to reordering, are a main object in Almgren's theory. Interesting minimizers such as $x_3 + ix_4 = \sqrt{(x_1 + ix_2)^3}$ have branched structure, where the multiple graphs are "knotted" to one another. In Almgren's theory, the singularities of minimal surfaces are shown to correspond to the singularities of multiple-valued minimizers of the Dirichlet energy, also called *multiple-valued harmonic functions*.

A crucial ingredient in Almgren's theory is the *frequency formula*: if $\boldsymbol{f} : \mathbb{R}^m \to (\mathbb{R}^k)^Q/\!\sim$ $\boldsymbol{f}(x_\circ) = \mathbf{0}$ is Dirichlet-minimizer, then the dimensionless quantity

$$\phi_{x_\circ}(r; \boldsymbol{f}) := \frac{r \int_{B_r(x_\circ)} |\nabla \boldsymbol{f}|^2}{\int_{\partial B_r(x_\circ)} \boldsymbol{f}^2} \tag{5.1}$$

is monotone nondecreasing in r. Moreover, $\phi_{x_\circ}(r; \boldsymbol{f}) \equiv \lambda$ for some $\lambda \geq 0$ ($\phi_{x_\circ}$ is constant in r) if and only if $\boldsymbol{f}(x_\circ + \cdot\,)$ is λ-homogeneous. As a consequence of the frequency formula, whenever $\boldsymbol{f}$ is a multiple-valued harmonic function and $\boldsymbol{f}(x_\circ) = \mathbf{0}$, "blow-up" sequences

$$\boldsymbol{f}^{x_\circ, r_k} := \frac{\boldsymbol{f}(x_\circ + r_k\, \cdot)}{\left(r^{1-m} \int_{\partial B_r} \boldsymbol{f}^2\right)^{1/2}}, \quad r_k \downarrow 0,$$

converge (up to subsequences) towards some homogeneous multiple-valued harmonic function $\boldsymbol{f}^*$.

5.4. Dimension reduction and center manifold

With the frequency formula at hand, we can explain (roughly and naively) some other key ideas in the proof of Theorem 4, which will later have parallels in the obstacle problem and Stefan's problem. Assume that S is a minimal surface (current) of dimension m inside $\mathbb{R}^{m+k}$ that has a singular (or non-smoothness) point at 0. As discussed before, near 0, S will be well approximated by a multiple-valued Dirichlet minimizer $\boldsymbol{f} : \mathbb{R}^m \to (\mathbb{R}^k)^Q/\!\sim$, where $Q \in \mathbb{N}$ is given by $Q = a_0(0^+; S)/|B_1^n|$ (here $|B_1^m|$ denotes the m-dimensional volume of the unit ball of $\mathbb{R}^m$).

Let us only discuss for simplicity the case $m = k = 2$. In that case, we want to show that singular points are isolated. So, assume by contradiction that there was a sequence of singular points $x_k \to 0$ and let $r_k := |x_k|$ be their norms. Consider the blow-up sequence $\boldsymbol{f}^{0,r_k}$, which will converge (up to a subsequence) towards some λ-homogenous (possibly multiple-valued) function $\boldsymbol{f}^*$, where $\lambda = \phi_0(0^+; \boldsymbol{f})$. Now $\boldsymbol{f}^* : \mathbb{R}^2 \to (\mathbb{R}^k)^Q/\sim$ must be of the form $\boldsymbol{f}(r\cos\theta, r\sin\theta) = r^\lambda \boldsymbol{g}(\theta)$, where $\boldsymbol{g}$ is some Q-valued curve. The fact that $\boldsymbol{f}$ is harmonic (Dirichlet minimizer) imposes very strong restrictions on $\boldsymbol{g}$ (e.g., locally each branch $g : \mathbb{R} \to \mathbb{R}^k$ must satisfy the ODE $g'' = \lambda^2 g$). This strong rigidity helps in classifying all possibilities for $\boldsymbol{f}^*$, and one can show that it must be exactly given by a holomorphic curve, e.g.,

$$x_3 + ix_4 = (x_1 + ix_2)^3 \quad \text{or} \quad (x_3 + ix_4)^Q = (x_1 + ix_2)^{Q+1}.$$

Now, if $\boldsymbol{f}^*$ has a (multiplicity Q) branching singularity at 0, then—since we now know that $\boldsymbol{f}^*$ is a homogeneous holomorphic curve—it must be isolated. Hence $\boldsymbol{f}^*$ must be smooth away from 0. This fact—thanks to Allard's version of Theorem 1, which only applies to multiplicity one points $x \in S$ (i.e., $a_x(0^+; S)/\pi = 1$)—implies that S will not have any other singularity in a (sufficiently small) neighborhood of 0.

Still, there is the possibility that—as it happens in the examples S_2 and S_3 given from Section 5.1—, $\boldsymbol{f}^*$ may be a harmonic polynomial. In such cases, the branching singularity will only show itself after we "rectify" $\boldsymbol{f}$, subtracting from it the (single-valued) harmonic polynomial P, which "best fits" $\boldsymbol{f}$ near 0. This idea leads to the notion of *center manifold*: in order to see the branching structure, we must consider the deviation of S, not from the tangent plane, but from the "best fitting" smooth single-valued minimal graph near 0. The frequency function on $\boldsymbol{f} - P$—more precisely $\phi_0(r; \boldsymbol{f} - P)$—is also monotone, and $(\boldsymbol{f} - P)(r_k\,\cdot)$ divided by its L^2 norm on ∂B_1 converges to some new homogeneous blow-up $\boldsymbol{f}^*$. Now, by construction f^* cannot be single-valued, so it must have a branching singularity.

In order to prove the result in higher dimensions, we need an appropriate variant of Federer's dimension reduction principle (previously discussed in the context of area-minimizing surfaces in $\mathbb{R}^3$). The dimension reduction is based on the following simple property: if a function $f : \mathbb{R}^m \to \mathbb{R}^k$ is at the same time λ-homogeneous with respect to 0 and μ-homogeneous with respect to $x_\circ \neq 0$, then necessarily $\lambda = \mu$ and f is translation invariant in the direction $x_\circ$. A similar property holds for multiple-valued functions $\boldsymbol{f}$.

5.5. Almgren's methods applied to Signorini's problem

In [5], the authors devised how to apply the methods introduced by Almgren in the context of area-minimizing currents to Signorini's problem. This leads to a very important progress, as described next.

In order to get rid of superfluous technical details, instead of (2.2), the authors consider the "cleaner" zero obstacle problem: Let $n \geq 2$, and consider $u : B_1 \to \mathbb{R}$

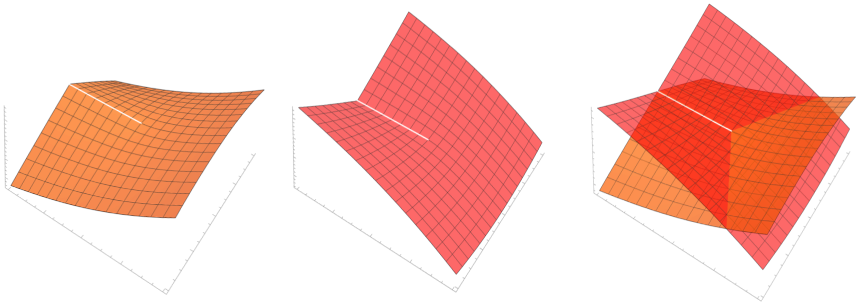


Figure 7. From Signorini's problem to "2-valued harmonic functions".

($B_1 \subset \mathbb{R}^n$ is the unit ball) satisfying

$$\int_{B_1} |\nabla v|^2 \geq \int_{B_1} |\nabla u|^2 \text{ for all } v \in H^1(B_1) \text{ with } v \geq 0 \text{ on } \{x_n = 0\} \text{ and } v = u \text{ on } \partial B_1. \quad (5.2)$$

A key contribution of [5] was to show that functions satisfying (5.2) behave essentially identically to Dirichlet-minimizing 2-valued functions. As a matter of fact, for $n = 2$, explicit examples of minimizers to (5.2) can be obtained, and they are conspicuously related to the examples of branching singularities of holomorphic curves discussed before. For example, a very important explicit solution of (5.2) for $n = 2$ is $u(x_1, x_2) = \operatorname{Re} \sqrt{(x_1 + ix_2)^3}$, where now $\sqrt{\cdot}$ selects only the principal branch. This is clearly related to the branching singularity $(x_3 + ix_4)^2 = (x_1 + ix_2)^3$.

Heuristically, if u is a minimizer of (5.2), then the 2-valued function $(u(x), -u(x))$ can be thought of as "2-valued harmonic function" (see Figure 7).

Among the multiple analogies, the frequency formula $\phi_{x_\circ}(r, u)$—defined exactly as in (5.1) replacing f with u—is also monotone nondecreasing for every point such $x_\circ \in \{x_n = 0\} \cap \{u = 0\}$. The main contribution [5] was to show that if $\lambda := \phi_{x_\circ}(r, u) < 2$ at some *free boundary point* $x_\circ \in \partial\{u > 0\} \cap \{x_n = 0\}$, then either $\lambda \leq 3/2$ or $\lambda \geq 2$. Moreover, they proved that the set of points where the first alternative holds is open and is an $(n-2)$-manifold of class C^1 inside $\{x_n = 0\}$.

6. The singular set in the obstacle problem (2017–2021)

In 2015, more than 30 years after Caffarelli's breakthrough [9] for the obstacle problem, the following important questions remained essentially open in dimensions $n \geq 3$:

- Can we obtain some precise description of singularities in the obstacle problem?
- Is the singular set "small" in some sense? How small?

As we discussed before, satisfactory answers to these questions had been only obtained (through complex variable methods) in dimension $n = 2$ by Sakai [44, 45]. Sakai's methods did not work in higher dimensions, and improving Caffarelli's result required new ideas.

6.1. A finer analysis of the singular set

The first new result in this direction for $n \geq 3$ was established by Colombo, Spolaor, and Velichkov in [12]. By refining the methods of Weiss [54], they proved that at every singular point $x_\circ$, the expansion

$$u(x_\circ + x) = \frac{1}{2} x \cdot Ax + \omega\big(x\big). \tag{6.1}$$

holds with a quantitative logarithmic estimate for the error $|\omega(x)| \leq C|x|^2(\log|x|)^{-\gamma}$, where $\gamma > 0$. Caffarelli in [10] had obtained a qualitative control $|\omega(x)| \leq o(|x|^2)$ using the Alt–Caffarelli–Friedmann monotonicity formula—a different proof of the same qualitative estimate was given later in [37]. Sakai had found in [44,45] the (optimal) rate $|\omega(x)| \leq C|x|^3$ in dimensions $n = 2$. In the proofs of [12], one can glimpse some delicate obstructions to obtaining such a strong result in dimensions $n \geq 3$, although it was not clear if they were only of technical nature (no counterexample was known).

Independently and with different methods, Figalli and the author proved in [27] the following:

Theorem 5 ([27]). *Let u be a solution of the obstacle problem (4.4) in a ball of $\mathbb{R}^n$. For all singular points outside some "anomalous" (relatively open) set of Hausdorff dimension $\leq n-3$, (6.1) holds with $|\omega(x)| \leq C|x|^3$.*

Moreover, there exist examples in $\mathbb{R}^3$ of isolated singular points for which

$$\big|\omega(x)\big| \gg |x|^{2+\varepsilon} \quad \text{as } |x| \to 0 \text{ for all } \varepsilon > 0.$$

The previous theorem suggests that one might be able to give a much more precise description of the solutions than Caffarelli's near "most" singular points. However, not for all of them: the existence, already in $\mathbb{R}^3$, *anomalous* singular points for which $|\omega(x)| \gg |x|^{2+\varepsilon}$ for all $\varepsilon > 0$ is to be kept in mind as a warning of the arduousness of the problem.

The methods introduced in [27] are strongly connected with Almgren's ones for minimal currents. The link between the two (a priori unrelated) problems, found in [27], is as follows. Let u be a solution of the obstacle problem (4.4) in $B_1 \subset \mathbb{R}^n$ with a singular point at 0. In other words, assume that (6.1) holds at $x_\circ = 0$ with $\omega(x) = o(|x|^2)$. We then consider $w(x) = u(x) - \frac{1}{2}x \cdot Ax$. In [27], it was found that (surprisingly!) $\phi_0(r; w)$ is monotone increasing in r, where ϕ_0 is, as before, Almgren's frequency formula. This property allows one to study the so-called *second blow-ups*, namely accumulation of points of the type

$$q(x) = \lim_{r_k \to 0} \frac{w(rx)}{\|w(r_k \cdot)\|_{L^2(\partial B_1)}}.$$

Thanks to the monotonicity of ϕ on w, such second blow-ups q are λ-homogeneous—that is $q(tx) = t^\lambda q(x)$ for all $t \geq 0$—where $\lambda = \phi_0(0^+; w)$. Moreover, in [27], it is found that, outside of an $n-3$ dimensional set of singular points, the second blow-ups have homogeneity $\lambda \geq 3$ and are either harmonic or solutions of the thin obstacle problem (5.2). This allows for a full classification of possible second blow-ups in two dimensions, and in higher dimensions, allows us to perform dimension reduction arguments à la Federer based on the frequency, similarly to Almgren's work for area-minimizing currents in codimensions ≥ 2.

Another insightful result from [27] is that, for all singular points outside some $(n-2)$-dimensional set we have, after rotation, the improved expansion

$$u(x_\circ + x) = \frac{1}{2} x_n^2 + x_n Q(x) + o\big(|x|^3\big), \tag{6.2}$$

where Q is some quadratic polynomial satisfying $\Delta(x_n Q) \equiv 0$. By analogy with Almgren's center manifold, this invites to subtract the polynomial $x_n Q$ in order to investigate higher order expansions (this turned out to be a quite delicate task, and the missing tools in order to perform it were only developed later in [25,29]).

6.2. Generic regularity: Schaeffer's conjecture in low dimensions

Building on the methods of [27], we could recently obtain a positive answer to (Schaeffer's) Conjecture 3 in low dimensions:

Theorem 6 ([25]). *Conjecture 3 holds true in $\mathbb{R}^3$ and $\mathbb{R}^4$.*

More precisely, we can consider 1-parameter monotone (and continuous) families of boundary data $g : \partial\Omega \times (0,1) \to \mathbb{R}_+$, where $\Omega \subset \mathbb{R}^n$ is a bounded smooth domain, satisfying $g(x; \tau') - g(x; \tau) \geq c(\tau' - \tau)$ for all $0 < \tau < \tau' < 1$. We let u^τ be the solution of (4.4) with boundary data $u^\tau = g(\,\cdot\,; \tau)$ on ∂B_1. The"generic regularity" question we want to understand can be phrased as follows: if we choose $\tau \in (0,1)$ randomly (with a uniform distribution), will the free boundary of u^τ be analytic with probability one? We can answer positively this question in dimensions 3 and 4 (the positive answer in two dimensions had already been obtained by Monneau in [37] for $g(x; \tau) = g(x) + \tau$).

Our strategy towards this theorem is reminiscent of *Sard's theorem* in analysis. We aim to prove that the set of "singular values" $\tau \in (0,1)$ has measure zero by improving, at most singular points, the order of approximation of certain polynomial expansions for u^τ. This is a delicate and long proof because the singular set needs to be split into several different subsets and, in each of them, the corresponding set of singular values has measure zero for a different reason.

In order to prove the conjecture in four dimensions, we need to consider the set of all points $x_\circ \in \Omega$, which are singular for some of the solutions u^τ in the family. We then show that, after to removing $(n-2)$-dimensional set, for all the other $x_\circ$, we have an expansion of the type

$$u^\tau(x_\circ + x) = \mathscr{P}(x) + O\big(|x|^5\big),$$

where $\tau = \tau(x_\circ)$ is the value of the parameter for which $x_\circ$ is singular. Here $\mathscr{P}$ is a polynomial of the form (in some orthonormal coordinates depending of $x_\circ$)

$$\mathscr{P}(x) := \frac{1}{2}\left(x_n + \sum_{\alpha=1}^{n-1} \frac{a_\alpha}{2} x_\alpha^2 + \frac{(\sum a_\alpha)}{6} x_n^2 + \sum_{\alpha=1}^{n-1} \left(a_\alpha^2 - \frac{a_\alpha(\sum a_\alpha)}{3}\right)\left(\frac{x_n^3}{12} - \frac{x_\alpha^2 x_n}{2}\right)\right)^2,$$

for some $a_\alpha \geq 0$ ($\alpha = 1, \dots, n-1$). We call $\mathscr{P}$, the "Ansatz", and whose structure is found imposing $\Delta\mathscr{P} = 1 + O(|x|^3)$. In many respects, $\mathscr{P}$ plays an analogous role to Almgren's center manifold: also here the idea is that, only after subtracting a very smooth "tangent object", one is able to see branching-type patterns which can only occur on lower dimensional sets.

We then manage to obtain an approximate monotonicity of (a truncated version of) the frequency function ϕ_0 for the remainder

$$w := u^\tau(x_\circ + \cdot) - \mathscr{P},$$

and perform dimension reduction type of arguments à la Federer–Almgren. However, an interesting feature of the dimension reduction arguments in [25] (which is completely new with respect to Almgren's) is that we need to work not with one single solution but with an increasing family of them (which do not have any other link between them than the monotonicity). And the dimension bounds that we obtain for the union in τ of all "bad points" for the family $\{u^\tau\}_\tau$ are as precise as the estimate one single u^τ.

The existence of solutions with an $(n-3)$-dimensional set of "anomalous points" where the expansion is quadratic, and not better, prevents us from using the same kind of methods for Schaeffer's conjecture in dimensions 5 or higher.

6.3. C^∞ partial regularity

Building on the methods of [25] (and [26]), F. Franceschini and W. Zatoń obtained in [29] the following extremely detailed (and essentially optimal) result:

Theorem 7 ([29]). *Let u be a solution of the obstacle problem* (4.4) *in the unit ball of $\mathbb{R}^n$ and let Σ denote its singular set. There exists a closed set $\Sigma^\infty \subset \Sigma$ such that*

(i) $\dim_{\mathcal{H}}(\Sigma \setminus \Sigma^\infty) \le n-2$ *(countable, if $n=2$);*

(ii) *locally, Σ^∞ is contained in one $(n-1)$-dimensional C^∞ manifold, and at every point $x \in \Sigma^\infty$ the solution u has a polynomial expansion of arbitrarily large order. Moreover, these are consistent from one point to another in the sense of Whitney's extension theorem.*

A key contribution from [29] was to show almost-optimal Lipchitz estimates (in terms of their L^2 norms in a small ball) for the differences $u(x_\circ + \cdot) - \mathcal{P}$, where $x_\circ$ is a singular point, and $\mathcal{P}$ is an Ansatz of arbitrarily large order (nonnegative polynomial satisfying $\Delta\mathcal{P} \approx 1$). Such Lipchitz estimates are needed to prove that Almgren's frequency formula on $w = u - \mathcal{P}$ is monotone. With the previous approach from [25], such estimates had necessarily errors of size $O(|x|^5)$, which was blocking the expansion at order 5. The smarter (and more natural, a posteriori) approach from [29] allows the authors to obtain similar Lipchitz estimates with an error of arbitrarily high order. As a consequence, they obtain a beautiful C^∞ partial regularity result for the singular set: something that seemed inconceivable only a few years ago.

7. The singular set in the Stefan problem (2019–2021)

After Caffarelli's 1977 breakthrough, a main question on the structure of the free boundaries in Stefan's problem remained open: how large may the singular set be? Very simple examples—such as a one-dimensional solution $u(x_3, t)$ for which the ice region is $\{|x_3| \le f(t)\}$ for some f decreasing—show that the singular set in Stefan's problem (in $\mathbb{R}^3$) may be as large as 2-dimensional; at least for some times. The regular part of the free boundary is a moving 2-surface in $\mathbb{R}^3$, so at such "bad" times, the singular set is as large as the regular part! However, in the examples, this may happen only for a very exceptional set of times. This suggests that the singular set should be "smaller" than the set of smooth points as a subset of *spacetime* $\mathbb{R}^3 \times \mathbb{R}$.

In order to measure the dimension of subsets of spacetime in Stefan's problem, it is natural to introduce a Hausdorff dimension associated to the "parabolic scaling" (which leaves the equation invariant). Namely, for a set $E \subset \mathbb{R}^n \times \mathbb{R}$, we write $\dim_{\mathrm{par}}(E) \le \beta$, when for all $\beta' > \beta$, E can be covered by countably many *parabolic cylinders* $B_{r_i}(x_i) \times (t_i - r_i^2, t_i)$, making $\sum_i r_i^{\beta'}$ arbitrarily small. Notice that, if we denote by $\dim_{\mathcal{H}}(E)$ the standard Hausdorff dimension of a set

$$E \subset \mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R},$$

then $\dim_{\mathcal{H}}(E) \le \dim_{\mathrm{par}}(E)$. On the other hand, the time axis has parabolic Hausdorff dimension 2, while it has standard Hausdorff dimension 1.

The only known dimensional bound on the singular set $\Sigma \subset \mathbb{R}^n \times \mathbb{R}$ for solutions to (4.1)-(4.2) in dimensions $n \geq 2$ was the following rather rough estimate: as a consequence of the results in [6, 9], at every singular point $(x_\circ, t_\circ)$, the qualitative expansion

$$u(x_\circ + x, t_\circ + t) = \frac{1}{2} x \cdot Ax + o\big(|x|^2 + |t|\big) \tag{7.1}$$

holds, where $A = A_{x_\circ, t_\circ}$ is a nonnegative definite matrix, satisfying $\mathrm{tr}(A) = 1$, which depends on $(x_\circ, t_\circ)$. As a consequence of (7.1), the set of singular points can be decomposed as $\Sigma = \bigcup_{m=0}^{n-1} \Sigma_m$, where

$$\Sigma_m := \big\{(x_\circ, t_\circ) \in \Sigma : \dim\big(\ker(A_{x_\circ, t_\circ})\big) = m\big\}, \quad m = 0, \dots, n-1.$$

Moreover, for each m, the set $\Sigma_m \cap \{t = t_\circ\}$ can be covered by a C^1 manifold of dimension m. Unfortunately, the previous expansion implies only $C^{1/2}$ regularity in time for the covering manifolds. As shown in [36], (7.1) also implies a (very rough) bound on the parabolic Hausdorff dimension of Σ:

$$\dim_{\mathrm{par}}(\Sigma) \leq n + \frac{1}{2}. \tag{7.2}$$

Since the parabolic dimension of the regular part of the free boundary has dimension $(n-1) + 2 = n + 1$, the previous bound shows that, in some weak sense, the singular set is smaller than the regular one. However, the bound (7.2) does not even rule out the existence of pathological solutions with singular points at every time (not even in two dimensions)!

7.1. Almgren meets Stefan

After the works [25, 27], it was very natural to apply the same kind of arguments to the Stefan's problem (4.1)-(4.2). Given a singular point $(x_\circ, t_\circ)$, let us consider

$$w(x) := u(x_\circ + x, t_\circ + x) - \frac{1}{2} x \cdot Ax.$$

In order to extend our methods from [27] to the parabolic setting, Poon's [39] parabolic version of Almgren's frequency formula plays an important role. Namely, denoting $G(x,t) = (4\pi t)^{-n/2} e^{\frac{|x|^2}{4t}}$ the time-reversed heat kernel, the functional

$$\phi(r, w) := \frac{r^2 \int_{\{t=-r^2\}} |\nabla w|^2 G \, dx}{\int_{\{t=-r^2\}} w^2 G \, dx},$$

can be shown to be monotone in r.[12]

[12] Actually, we need to employ a (new) suitable truncated version of ϕ (which we call ϕ^γ), and its monotonicity can be proved only up to exponentially small errors. But these are technical details.

Thanks to this fact, we can prove that

$$\frac{w(rx, r^2t)}{\|w\|_r} \to q(x,t) \quad \text{as } r \to 0, \tag{7.3}$$

along subsequences and in compact subset of $\{t < 0\}$, where q is a parabolically homogeneous function: namely $q(rx, r^2t) = r^\lambda q(x,t)$ for all $r > 0$, where $\lambda = \phi(0^+, w)$. In (7.3), we denote by $\|w\|_r$ the quantity $(⨍_{\mathcal{C}_r} w^2)^{1/2}$, which measures the "size" of w in the parabolic cylinder $\mathcal{C}_r := B_r \times (-r^2, 0)$.

We then show the following:

(i) If $(x_\circ, t_\circ) \in \Sigma_m$ with $m \le n-2$, then the function q is *always* a *quadratic* caloric polynomial. This means that the expansion (7.1) cannot be improved at any of these points! To obtain an improved dimensional bound on Σ_m, we employ a barrier argument in the spirit of [25] to show that $u > 0$ in $B_r(x_\circ) \times (t_\circ + r^{2-\varepsilon}, \infty)$. In other words, a ball of radius r around one of these singular points will be completely occupied by water after increment of time of size $r^{2-\varepsilon}$. This gives

$$\dim_{\text{par}}(\Sigma_m) \le m, \quad 0 \le m \le n-2.$$

(ii) If $(x_\circ, t_\circ) \in \Sigma_{n-1}$, then q is a homogeneous solution of the parabolic thin obstacle problem. We denote by $\Sigma_{n-1}^{<3}$ the subset at which the homogeneity is less than 3.

(a) If $(0,0) \in \Sigma_{n-1}^{<3}$, we show that $\partial_t q \not\equiv 0$ and that q is *convex* in all directions that are tangential to $\{p_2 = 0\}$. This allows us to perform a dimension reduction that, combined with a barrier argument similar to that in (i), yields

$$\dim_{\text{par}}(\Sigma_{n-1}^{<3}) \le n-2.$$

(b) If $(0,0) \in \Sigma_{n-1} \setminus \Sigma_{n-1}^{<3}$, we show that q is always 3-homogeneous, hence

$$u(x_\circ + \cdot, t_\circ + \cdot) = \frac{1}{2} x \cdot Ax + O\big(|x|^3 + |t|^{3/2}\big). \tag{7.4}$$

This (and a barrier argument similar to the one before) implies that

$$\dim_{\text{par}}(\Sigma_{n-1} \setminus \Sigma_{n-1}^{<3}) \le n-1.$$

Combining these estimates in [26], we obtain the following theorem.

Theorem 8. *The singular set of solutions to* (4.1)-(4.2) *has a parabolic dimension* $n-1$.

Therefore, it is natural to ask ourselves if a similar result holds in the physical space $\mathbb{R}^3$ and, more in general, how often singular points may appear.

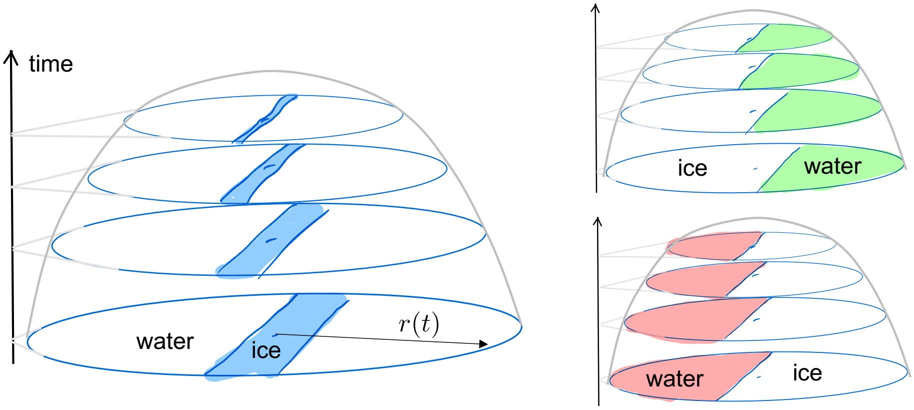


Figure 8. Inside of the shrinking ball $B_{r(t)}(x_\circ)$, the free boundary consists of two fronts, which evolve independently until they meet at time $t_\circ$.

7.2. Cubic expansions and their heuristic interpretation

With a bit of extra work, we can obtain a complete parabolic analog of the main result in [27]: for all singular points $(x_\circ, t_\circ)$ outside of a set of parabolic dimension $n-2$, the following expansion holds[13]

$$u(x_\circ + x, t_\circ + t) = \frac{1}{2}x_n^2 + a|x_n|\left(t + \frac{1}{6}x_n^2\right) + \begin{bmatrix}\text{3-homogeneous caloric}\\ \text{polynomials}\end{bmatrix} + o\big((|x| + |t|^{1/2})^3\big), \tag{7.5}$$

for some $a > 0$. The fact that this coefficient is positive, which turns out to be consequence of (4.2), is crucial.

Indeed, (7.5) implies that, if we look at the free boundary at time $t < t_\circ$ inside a ball of radius $\sqrt{t_\circ - t}$ centered at $x_\circ$, we will see two almost-parallel "independent" fronts which move one towards the other. More precisely, for $t < t_\circ$, we have

$$x_n = \pm 2a\,(t_\circ - t) + o(t_\circ - t) \quad \text{on } \partial\{u > 0\} \cap B_{\sqrt{t_\circ - t}}(x_\circ) \times \{t\}.$$

In this direction, let us (informally) define $r(t)$ as "the largest" radius for which the ice inside $B_\varrho(x_\circ)$ has two connected components for times before t—see Figure 8 (left). The expansion (7.5) actually implies $r(t) \gg \sqrt{t_\circ - t}$, as $t \uparrow t_\circ$.

Now, it is interesting to observe the following: suppose that $r(t)$ happened to stay bounded away from zero as $t \uparrow t_\circ$. Then, inside of some (small) parabolic cylinder $B_\varrho(x_\circ) \times (t_\circ - \varrho^2, t_\circ)$, the "positivity set" $\{u > 0\}$ would consist of exactly two

[13] After choosing an appropriate orthonormal frame depending on $(x_\circ, t_\circ)$.

connected components 1 and 2. We could then define $u^{(i)}$, $i = 1, 2$, as u multiplied by the characteristic function of the component i. Doing so, the two new functions $u^{(i)}$ would both solve (4.1)! Moreover, both functions would have a thick contact set $\{u^{(i)} = 0\}$, so the point $(x_\circ, t_\circ)$ would be regular for the two of them—see Figure 8 (right). Hence the free boundaries of $u^{(i)}$, $i = 1, 2$ (which correspond to the two fronts of u) would be smooth inside $B_\varrho(x_\circ)$ up to the final time $t = t_\circ$. At this final time, $t = t_\circ$, the two fronts $\{x_n = g^{(i)}(x')\}$ would be ordered, smooth, and tangent at least at $x_\circ$. Then, their tangency points in $B_\varrho(x_\circ)$ would necessarily be of one of the following two types.

- Infinite order tangency points of the two functions $g^{(i)}$: near such points the ice would be extremely thin, and hence they should become immediately surrounded by water after $t_\circ$.
- Lower dimensional tangency points: the subset of $g^{(1)} = g^{(2)}$ where the two functions disagree at some finite order k would be automatically contained in a smooth $(n-2)$-dimensional manifold (being contained in the transversal intersection of the graphs of certain $(k-1)$-derivatives of $g^i(x')$).

Of course, the difficulty is that we cannot expect $r(t)$ to be bounded away from zero at typical free boundary points. But it turns out that (with much extra effort) we can improve the bound $r(t) \gg \sqrt{t_\circ - t}$ to $r(t) \geq (t_\circ - t)^{\frac{1}{2+\beta}}$, for some tiny $\beta > 0$ (as $t \uparrow t_\circ$), at "most" singular points. This amounts to proving an expansion like (7.5) but with an error of size $O((|x| + |t|^{1/2})^{3+\beta})$. As we will see, such apparently small improvement is "breaking the parabolic scaling", and will allow us to obtain the same type of conclusions as if $r(t)$ stayed bounded away from zero! But such strong conclusions are not cheap to obtain: in order to improve (7.5) by a tiny positive β, we need to introduce completely new techniques. We need to go beyond Almgren.

7.3. Improving cubic expansions: Life beyond Almgren

Arguably, the most delicate point in [26] is to show that, for all singular points $(x_\circ, t_\circ)$ outside of a set of parabolic dimension $n-2$, the following expansion holds:

$$\begin{aligned} u(x_\circ + x, t_\circ + t) = \frac{1}{2}x_n^2 + a|x_n|\left(t + \frac{1}{6}x_n^2\right) + [\text{3-hom. cal. pol.}] \\ + O\left(\left(|x| + |t|^{1/2}\right)^{3+\beta}\right), \end{aligned} \tag{7.6}$$

for some $\beta > 0$ (which may depend on the point).

Given a singular point $(x_\circ, t_\circ)$ where (7.5) holds, it is natural to consider

$$w(x,t) := u(x_\circ + x, t_\circ + t) - \frac{1}{2}x_n^2 - a|x_n|\left(t + \frac{1}{6}x_n^2\right) - [\text{3-hom. cal. pol.}].$$

Now, one could naively try to show that Almgren frequency is again monotone on such w (this is the first we tried and, as a matter of fact, we thought for a long time that this was the way to go). Unfortunately, since $a > 0$, the cubic term is never a caloric polynomial and the frequency function $\phi(r, w)$ is never (almost) monotone.

In order to improve (7.5), we need a completely new strategy based on barriers, compactness, and certain ad-hoc monotonicity properties, which are much weaker than Almgren's (but which still give some nonempty information).

Our new approach consists in showing, essentially,[14] that

$$\|w\|_{L^\infty(B_r\times(-r^2,0))} \le \omega(r),$$

where ω satisfies the following alternative with $\varepsilon > 0$ arbitrarily small. For all $r > 0$ sufficiently small, we have either

$$\begin{gathered}\Sigma \text{ is } (\varepsilon r)\text{-close to an } (n-2)\text{-plane inside } B_r(x_\circ)\\ \text{for } t \in (t_\circ - r^2, t_\circ) \text{ and } \omega\left(\frac{r}{2}\right) \le \frac{\omega(r)}{2^{3-\varepsilon}};\end{gathered} \tag{7.7}$$

or else, we have

$$\omega\left(\frac{r}{2}\right) \le \frac{\omega(r)}{2^{3+\frac{1}{2}}}. \tag{7.8}$$

In view of the previous alternative, it seems to look at dyadic scales $r = 2^{-i}$ and consider the "upper density" of scales at which (7.7) holds:

$$\vartheta := \limsup_{\ell\to\infty} \frac{\#\left\{i \le \ell : (7.7) \text{ holds at the scale } r = 2^{-i}\right\}}{\ell} \in [0, 1].$$

Now, if $\vartheta = 1$, then as we zoom in around $(x_\circ, t_\circ)$, we see "enough scales" at which the singular set is close to an $(n-2)$-plane to conclude that "Σ is $(n-2)$ dimensional at $(x_\circ, t_\circ)$" (this requires new delicate GMT-type covering arguments). On the other hand, if $\vartheta < 1$, then for a positive $(1-\vartheta)$-proportion of scales, we have (7.8), while for the other scales, we have $\omega(\frac{r}{2}) \le 2^{-3+\varepsilon}$. Taking ε small, we can choose $\beta > 0$ such that $(1-\vartheta)\frac{1}{2} + \vartheta(-\varepsilon) = 3 + \beta$. We then see that

$$\omega(2^{-\ell}) \lesssim 2^{(-3-\frac{1}{2})(1-\theta)\ell}2^{(-3+\varepsilon)\theta\ell}\omega(1) = 2^{-(3+\beta)\ell}\omega(1).$$

This gives (7.6) at such points.

[14] The (over)simplified statement given here is not strictly correct, but it gives a very good approximated idea on how the argument goes. The actual statement is much more involved (see [26, Proposition 11.3]). Although some of the subtleties in the actual statement are important and not mere technicalities, we cannot discuss them here.

7.4. C^∞ partial regularity and optimal dimensional bounds on the singular set

Once we have proven (7.6), we are ready to push the expansion to higher order. For this, we show with a barrier argument that the set $\{u > 0\}$ splits into two separate connected components inside the set $\Omega^\beta := \{|x|^{2+\beta} < -t\}$—here $(x_\circ, t_\circ) = (0, 0)$.

Note that, under the parabolic scaling $(x, t) \to (rx, r^2 t)$, the set Ω^β converges to $\mathbb{R}^n \times (-\infty, 0)$ as $r \to 0$. In other words, we have "broken the parabolic scaling". We then show a C^∞ regularity result (at $(0, 0)$) for the free boundary of solutions of (4.4) in Ω^β which have a "regular point" at $(0, 0)$. Here the difference with respect to the Caffarelli and Kinderlehrer–Nirenberg result is that in our case the domain Ω^β is not a parabolic cylinder: for every time slice space, the equation holds in ball, but its radius goes to zero as $t \uparrow 0$. Nevertheless, we manage to prove a C^∞ regularity which is robust enough to work in our setting. More precisely, to show that if $\bar{u}$ is a solution of the Stefan problem such that $\{\bar{u} = 0\}$ is sufficiently close to $\{x_n \le 0\}$ inside Ω^β, then we have a C^∞ expansion for $\bar{u}$ at $(0, 0)$. We then apply this result to our solution u multiplied by the characteristic functions of each of the two connected components of $\{u > 0\}$ inside Ω^β. In this way, we obtain a C^∞-type regularity for u.

As a corollary of this C^∞ expansion, we are able to prove that, outside an $(n-2)$-dimensional set, if $(x_\circ, t_\circ)$ and (x_1, t_1) are singular points, then

$$|t_\circ - t_1| = o\big(|x_\circ - x_1|^k\big) \quad \text{for every } k \gg 1.$$

This allows to finally establish the following theorem.

Theorem 9 ([26]). *Let $\Omega \subset \mathbb{R}^n$, and let $u \in L^\infty(\Omega \times (0, T))$ solve the Stefan problem (4.1)-(4.2). Then there exists $\Sigma^\infty \subset \Sigma$ (recall that $\Sigma \subset \mathbb{R}^n \times \mathbb{R}$ denotes the singular set) such that*

$$\dim_{\mathrm{par}}(\Sigma \setminus \Sigma^\infty) \le n - 2, \quad \dim_{\mathcal{H}}\left(\big\{t \in (0, T) : \exists\, (x, t) \in \Sigma^\infty\big\}\right) = 0,$$

and $\Sigma^\infty \subset \Omega \times (0, T)$ can be covered by countably many $(n-1)$-dimensional submanifolds in $\mathbb{R}^{n+1}$ of class C^∞.[15]

In a sense, this result says that the singular set can be split into two separate pieces: one is very smooth and extremely rare in time (the set Σ^∞), and one lower dimensional (of parabolic dimension at most $n - 2$).

This is a very precise result. Indeed, it is easy to construct radial examples of solutions to (4.1)-(4.2) for which the singular set contains some $(n-1)$-sphere for countably many times. Such spheres would be covered by the set Σ^∞ in Theorem 9.

[15] Here, the $(n-1)$-submanifolds that cover Σ^∞ are of class C^∞ as subset of $\mathbb{R}^{n+1}$ with the usual Euclidean distance, not with respect to the parabolic distance. So, our statement is much stronger than the previously known results (for instance, [36] proved C^1 regularity of Σ with respect to the parabolic distance, which implies only $C^{1/2}$ regularity in time).

Now, for general solutions, we cannot prove that $\pi_t(\Sigma^\infty)$ is countable as in such examples, but we do prove that it must be a 0-dimensional set (and Hausdorff dimension cannot distinguish between countable and 0-dimensional sets, so the result is sharp in this sense). On the other hand, the complement of Σ^∞ inside Σ instead, is a set of "bad" singular points. These "bad" points do not enjoy a priori any extra spatial regularity, but in exchange, their parabolic dimension is bounded by $n-2$. The fact that points where Caffarelli's quadratic expansion cannot be improved exist (and may be $n-2$ dimensional) can be easily shown by considering any radial solutions in $\mathbb{R}^2$ with a singular point at $(0,0)$.

An important consequence of Theorem 9 is the following very precise bound for the physical case (three spatial dimensions):

Corollary 10 ([26]). *The set of singular times for Stefan's problem in $\mathbb{R}^3$ has Hausdorff dimension at most $1/2$. In particular, it has measure zero.*

Also, Theorem 9 implies that in $\mathbb{R}^2$, the set of singular times for Stefan's problem has zero Hausdorff dimension (prior to our results, it was not even known that in $\mathbb{R}^2$, the set of singular times had measure zero).

In summary, these new results provide a very good picture about how the singular set of the Stefan problem behaves.

Funding. Research supported by Swiss NSF Ambizione Grant PZ00P2 180042 and by the European Research Council (ERC) under Grant Agreement No. 948029.

References

[1] W. K. Allard, On the first variation of a varifold. *Ann. of Math. (2)* **95** (1972), 417–491 Zbl 0252.49028 MR 307015

[2] F. J. Almgren Jr., Some interior regularity theorems for minimal surfaces and an extension of Bernstein's theorem. *Ann. of Math. (2)* **84** (1966), 277–292 Zbl 0146.11905 MR 200816

[3] F. J. Almgren Jr., Existence and regularity almost everywhere of solutions to elliptic variational problems among surfaces of varying topological type and singularity structure. *Ann. of Math. (2)* **87** (1968), 321–391 Zbl 0162.24703 MR 225243

[4] F. J. Almgren Jr., *Almgren's Big Regularity Paper. q-Valued Functions Minimizing Dirichlet's Integral and the Regularity of Area-Minimizing Rectifiable Currents up to Codimension* 2. World Sci. Monogr. Ser. Math. 1, World Scientific Publishing, River Edge, NJ, 2000 Zbl 0985.49001 MR 1777737

[5] I. Athanasopoulos, L. A. Caffarelli, and S. Salsa, The structure of the free boundary for lower dimensional obstacle problems. *Amer. J. Math.* **130** (2008), no. 2, 485–498 Zbl 1185.35339 MR 2405165

[6] A. Blanchet, On the singular set of the parabolic obstacle problem. *J. Differential Equations* **231** (2006), no. 2, 656–672 Zbl 1121.35145 MR 2287901

[7] E. Bombieri, E. De Giorgi, and E. Giusti, Minimal cones and the Bernstein problem. *Invent. Math.* **7** (1969), 243–268 Zbl 0183.25901 MR 250205

[8] H. R. Brezis and G. Stampacchia, Sur la régularité de la solution d'inéquations elliptiques. *Bull. Soc. Math. France* **96** (1968), 153–180 Zbl 0165.45601 MR 239302

[9] L. A. Caffarelli, The regularity of free boundaries in higher dimensions. *Acta Math.* **139** (1977), no. 3-4, 155–184 Zbl 0386.35046 MR 454350

[10] L. A. Caffarelli, The obstacle problem revisited. *J. Fourier Anal. Appl.* **4** (1998), no. 4-5, 383–402 Zbl 0928.49030 MR 1658612

[11] L. A. Caffarelli and N. M. Rivière, Asymptotic behaviour of free boundaries at their singular points. *Ann. of Math. (2)* **106** (1977), no. 2, 309–317 Zbl 0364.35041 MR 463690

[12] M. Colombo, L. Spolaor, and B. Velichkov, A logarithmic epiperimetric inequality for the obstacle problem. *Geom. Funct. Anal.* **28** (2018), no. 4, 1029–1061 Zbl 1428.49042 MR 3820438

[13] D. Danielli and S. Salsa, Obstacle problems involving the fractional Laplacian. In *Recent Developments in Nonlocal Theory*, pp. 81–164, De Gruyter, Berlin, 2018 Zbl 1408.35226 MR 3824211

[14] E. De Giorgi, *Frontiere orientate di misura minima. Seminario di Matematica della Scuola Normale Superiore di Pisa, 1960–61*. Editrice Tecnico Scientifica, Pisa, 1961 MR 0179651

[15] E. De Giorgi, Una estensione del teorema di Bernstein. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)* **19** (1965), 79–85 Zbl 0168.09802 MR 178385

[16] C. De Lellis, Almgren's Q-valued functions revisited. In *Proceedings of the International Congress of Mathematicians. Volume III*, pp. 1910–1933, Hindustan Book Agency, New Delhi, 2010 Zbl 1226.49041 MR 2827872

[17] C. De Lellis and E. Spadaro, Regularity of area minimizing currents III: blow-up. *Ann. of Math. (2)* **183** (2016), no. 2, 577–617 Zbl 1345.49053 MR 3450483

[18] J. Douglas, Solution of the problem of Plateau. *Trans. Amer. Math. Soc.* **33** (1931), no. 1, 263–321 Zbl 0001.14102 MR 1501590

[19] G. Duvaut, Résolution d'un problème de Stefan (fusion d'un bloc de glace à zéro degré). *C. R. Acad. Sci. Paris Sér. A-B* **276** (1973), A1461–A1463 Zbl 0258.35037 MR 328346

[20] H. Federer, The singular sets of area minimizing rectifiable currents with codimension one and of area minimizing flat chains modulo two with arbitrary codimension. *Bull. Amer. Math. Soc.* **76** (1970), 767–771 Zbl 0194.35803 MR 260981

[21] H. Federer and W. H. Fleming, Normal and integral currents. *Ann. of Math. (2)* **72** (1960), 458–520 Zbl 0187.31301 MR 123260

[22] X. Fernández-Real, The thin obstacle problem: a survey. *Publ. Mat.* **66** (2022), no. 1, 3–55 Zbl 07473077 MR 4366205

[23] A. Figalli, Free boundary regularity in obstacle problems. In *Journées équations aux dérivées partielles*, p. Exposé no. 2, 2018

[24] A. Figalli, Regularity of interfaces in phase transitions via obstacle problems—Fields Medal lecture. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures*, pp. 225–247, World Sci. Publ., Hackensack, NJ, 2018 MR 3966728

[25] A. Figalli, X. Ros-Oton, and J. Serra, Generic regularity of free boundaries for the obstacle problem. *Publ. Math. Inst. Hautes Études Sci.* **132** (2020), 181–292 Zbl 1456.35234 MR 4179834

[26] A. Figalli, X. Ros-Oton, and J. Serra, The singular set in the Stefan problem. 2021, arXiv:2103.13379

[27] A. Figalli and J. Serra, On the fine structure of the free boundary for the classical obstacle problem. *Invent. Math.* **215** (2019), no. 1, 311–366 Zbl 1408.35228 MR 3904453

[28] W. H. Fleming, On the oriented Plateau problem. *Rend. Circ. Mat. Palermo (2)* **11** (1962), 69–90 Zbl 0107.31304 MR 157263

[29] F. Franceschini and W. Zatoń, C^∞ partial regularity of the singular set in the obstacle problem. 2021, arXiv:2102.00923

[30] M. Giaquinta and L. Pepe, Esistenza e regolarità per il problema dell'area minima con ostacoli in n variabili. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)* **25** (1971), 481–507 Zbl 0283.49032 MR 305205

[31] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*. Monogr. Math. 80, Birkhäuser, Basel, 1984 Zbl 0545.49018 MR 775682

[32] D. Kinderlehrer, The coincidence set of solutions of certain variational inequalities. *Arch. Rational Mech. Anal.* **40** (1970/71), 231–250 Zbl 0219.49014 MR 271799

[33] D. Kinderlehrer and L. Nirenberg, Regularity in free boundary problems. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **4** (1977), no. 2, 373–391 Zbl 0352.35023 MR 440187

[34] D. Kinderlehrer, L. Nirenberg, and J. Spruck, Regularity in elliptic free boundary problems. *J. Analyse Math.* **34** (1978), 86–119 Zbl 0402.35045 MR 531272

[35] H. Lewy and G. Stampacchia, On the regularity of the solution of a variational inequality. *Comm. Pure Appl. Math.* **22** (1969), 153–188 Zbl 0167.11501 MR 247551

[36] E. Lindgren and R. Monneau, Pointwise regularity of the free boundary for the parabolic obstacle problem. *Calc. Var. Partial Differential Equations* **54** (2015), no. 1, 299–347 Zbl 1327.35455 MR 3385162

[37] R. Monneau, On the number of singularities for the obstacle problem in two dimensions. *J. Geom. Anal.* **13** (2003), no. 2, 359–389 Zbl 1041.35093 MR 1967031

[38] A. Petrosyan, H. Shahgholian, and N. Uraltseva, *Regularity of Free Boundaries in Obstacle-Type Problems*. Grad. Stud. Math. 136, Amer. Math. Soc., Providence, RI, 2012 Zbl 1254.35001 MR 2962060

[39] C.-C. Poon, Unique continuation for parabolic equations. *Comm. Partial Differential Equations* **21** (1996), no. 3-4, 521–539 Zbl 0852.35055 MR 1387458

[40] T. Radó, On Plateau's problem. *Ann. of Math. (2)* **31** (1930), no. 3, 457–469 MR 1502955

[41] E. R. Reifenberg, Solution of the Plateau Problem for m-dimensional surfaces of varying topological type. *Acta Math.* **104** (1960), 1–92 Zbl 0099.08503 MR 114145

[42] E. R. Reifenberg, On the analyticity of minimal surfaces. *Ann. of Math. (2)* **80** (1964), 15–21 Zbl 0151.16702 MR 171198

[43] E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields*. Grundlehren Math. Wiss. 316, Springer, Berlin, 1997 Zbl 0881.31001 MR 1485778

[44] M. Sakai, Regularity of a boundary having a Schwarz function. *Acta Math.* **166** (1991), no. 3-4, 263–297 Zbl 0728.30007 MR 1097025

[45] M. Sakai, Regularity of free boundaries in two dimensions. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **20** (1993), no. 3, 323–339 Zbl 0851.35022 MR 1256071

[46] O. Savin, Minimal surfaces and minimizers of the Ginzburg–Landau energy. In *Symmetry for Elliptic PDEs*, pp. 43–57, Contemp. Math. 528, Amer. Math. Soc., Providence, RI, 2010 Zbl 1225.35227 MR 2759034

[47] D. G. Schaeffer, An example of generic regularity for a non-linear elliptic equation. *Arch. Rational Mech. Anal.* **57** (1975), 134–141 Zbl 0319.35036 MR 387810

[48] D. G. Schaeffer, Some examples of singularities in a free boundary. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **4** (1977), no. 1, 133–144 Zbl 0354.35033 MR 516201

[49] S. Serfaty and J. Serra, Quantitative stability of the free boundary in the obstacle problem. *Anal. PDE* **11** (2018), no. 7, 1803–1839 Zbl 1391.35432 MR 3810473

[50] A. Signorini, Questioni di elasticità non linearizzata e semilinearizzata. *Rend. Mat. e Appl. (5)* **18** (1959), 95–139 Zbl 0091.38006 MR 118021

[51] J. Simons, Minimal varieties in riemannian manifolds. *Ann. of Math. (2)* **88** (1968), 62–105 Zbl 0181.49702 MR 233295

[52] J. Stefan, Über die Theorie der Eisbildung, insbesondere über die Eisbildung im Polarmeere. *Ann. Phys. Chem.* **42** (1891), 269–286

[53] J. E. Taylor, The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces. *Ann. of Math. (2)* **103** (1976), no. 3, 489–539 Zbl 0335.49032 MR 428181

[54] G. S. Weiss, A homogeneity improvement approach to the obstacle problem. *Invent. Math.* **138** (1999), no. 1, 23–50 Zbl 0940.35102 MR 1714335

Joaquim Serra
Department of Mathematics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland;
joaquim.serra@math.ethz.ch

DOI 10.4171/8ECM/12

Elliptic curves and modularity

Jack A. Thorne

Abstract. We survey results and conjectures concerning the modularity of elliptic curves over number fields.

1. Introduction

The modularity conjecture for elliptic curves over $\mathbf{Q}$ was stated with increasing degrees of precision by Taniyama, Shimura, and Weil in the 1950s and 60s. It admits several equivalent formulations, which are discussed in the textbook [17]. The most common asserts that given any elliptic curve E over $\mathbf{Q}$, we can find a newform $f \in S_2(\Gamma_0(N))$ with the property that for all but finitely many prime numbers p, the pth Fourier coefficient $a_p(f)$ in the q-expansion $f(q) = q + \sum_{n \geq 2} a_n(f) q^n$ equals the number

$$a_p(E) = p + 1 - \left|E(\mathbf{F}_p)\right|,$$

which measures the error in the Hasse estimate for the number of points on E modulo p. The newform f is then uniquely determined by E, by the strong multiplicity one theorem for modular forms. Any curve E for which such a newform f exists is said to be modular.

A famous example of a modular elliptic curve is the curve of conductor 11 given by the equation

$$E : y^2 + y = x^3 - x^2.$$

This elliptic curve is modular, with associated newform

$$f(q) = q \prod_{n=1}^{\infty} (1 - q^n)^2 (1 - q^{11n})^2 \in S_2\big(\Gamma_0(11)\big).$$

The modularity conjecture is, on the face of it, a very surprising statement. It is easy to write down elliptic curves over $\mathbf{Q}$; indeed, for any cubic polynomial

$$f(x) = x^3 + ax + b \in \mathbf{Z}[x]$$

2020 Mathematics Subject Classification. Primary 11G05; Secondary 11R39.
Keywords. Elliptic curves, automorphic forms, Langlands program.

of non-zero discriminant, the equation $y^2 = f(x)$ gives an elliptic curve. On the other hand, modular forms begin life as complex analytic objects. Even once admits their algebro-geometric description (as sections of a line bundle on a modular curve, thought of as an algebraic curve over $\mathbf{Q}$), together with the theory of Hecke operators, there is no a priori reason to expect that *every* elliptic curve over $\mathbf{Q}$ should be associated to a newform. Nevertheless, the modularity conjecture was proved for semistable elliptic curves over $\mathbf{Q}$ in 1995 by Wiles and Taylor [40, 45], on the way to proving Fermat's Last Theorem, and finally for all elliptic curves over $\mathbf{Q}$ in 2001 by Breuil, Conrad, Diamond, and Taylor [8].

The modularity conjecture for elliptic curves over $\mathbf{Q}$ can be thought of as a special case of the Langlands program, in a form made precise by Clozel [13]. Newforms give rise to automorphic representations of the adèle group $\mathrm{GL}_2(\mathbf{A}_\mathbf{Q})$. Under Clozel's conjectures, there would be a correspondence between motives of rank n over a number field K (or more concretely, compatible systems of semisimple, n-dimensional representations of the absolute Galois group of K) and automorphic representations of the group $\mathrm{GL}_n(\mathbf{A}_K)$ satisfying a condition that he calls "algebraic." Specialising to elliptic curves, we obtain a precise analogue of the modularity conjecture valid over an arbitrary number field. (We note that such an analogue had already been anticipated, especially in the case of imaginary quadratic fields; cf. [14, 24].)

Our first goal in this article is to state a version of this modularity conjecture for elliptic curves over a general number field K in as down to earth a manner as possible. In particular, our formulation does not use the language of automorphic representations. (This is not original; for example, Taylor's 1994 ICM article [38] contains essentially the same statement that we give here.) Note however that it is not possible to avoid the automorphic theory if one wants to give the most precise statements, or to get to the most important consequences of modularity, such as the analytic continuation of the L-function of an elliptic curve.

We will then continue to discuss some of the many applications of modularity in number theory, beyond the most famous application to Fermat's Last Theorem. It is interesting to note that these range from statements of great theoretical importance (such as the analytic properties of the L-function) to very concrete statements that have no obvious connection to automorphic representations or the Langlands program (such as bounds on the height of solutions to Mordell's equation).

Finally, we will discuss what is known towards the modularity conjecture for elliptic curves over a number field K, beyond the case $K = \mathbf{Q}$. It is natural to break up the discussion depending on whether or not K is totally real (in the sense that each field embedding $K \to \mathbf{C}$ in fact takes values in $\mathbf{R}$). Many of the methods developed to study modularity over $\mathbf{Q}$ translate well to the totally real setting. It is more challenging to study modularity over number fields which are not totally real, but there has been much progress in this direction recently, inspired particularly by applications of Scholze's theory of perfectoid spaces.

2. The modularity conjecture

Let K be a number field, with ring of integers $\mathcal{O}_K$ and absolute Galois group $G_K = \mathrm{Gal}(\overline{K}/K)$ with respect to a fixed choice of algebraic closure $\overline{K}/K$. (More generally, if k is a perfect field, then we will write G_k for the absolute Galois group of k with respect to some fixed choice of algebraic closure.)

Definition 2.1. An elliptic curve over K is a pair (E, ∞), where E is a smooth, projective, connected curve over K and $\infty \in E(K)$ is a marked rational point.

We often take the marked point as given and just say that E is an elliptic curve. Any elliptic curve may be given by a Weierstrass equation

$$y^2 = x^3 + ax + b, \tag{2.1}$$

where $a, b \in \mathcal{O}_K$ and x, y are plane co-ordinates. The closure (in the projective plane $\mathbf{P}^2$) of the affine curve defined by such an equation picks up exactly one extra point at infinity, which is the marked point ∞. The discriminant $\Delta = \Delta(a,b) = -16(4a^3 + 27b^2)$ is non-zero. Conversely, for any pair $(a,b) \in \mathcal{O}_K^2$ such that $\Delta(a,b) \neq 0$, equation (2.1) defines an elliptic curve.

Elliptic curves have a number of important associated structures. The first is the group law: there is a unique way to make any elliptic curve into a commutative algebraic group with identity element $\infty \in E(K)$. The addition law then has a simple characterization: three points P, Q, R sum to ∞ if and only if they are collinear in the Weierstrass embedding (2.1).

The second is the system of reductions modulo v, for v a finite (i.e., non-archimedean) place of the number field K. If the discriminant Δ of a given Weierstrass equation is a v-adic unit, then v is a place of good reduction for the curve E: the reduction modulo v of the Weierstrass equation (2.1) defines an elliptic curve over the residue field $k(v)$ of the completion K_v of K at the place v. This leads to the definition of the quantity

$$a_v(E) = q_v + 1 - \big|E\big(k(v)\big)\big|,$$

where $q_v = |k(v)|$ and $|E(k(v))|$ is the number of points of this reduced curve over the residue field $k(v)$. One can also define a_v at the places where Δ is not a v-adic unit, but this requires the use of a long Weierstrass equation in order to be able to find a model of minimal discriminant at the place v (see [36, Chapter VII]).

The third structure we want to introduce is the compatible system of ℓ-adic Galois representations of E. For each prime number ℓ, the étale cohomology group $H^1_{\text{ét}}(E_{\overline{K}}, \mathbf{Q}_\ell)$ is a 2-dimensional $\mathbf{Q}_\ell$-vector space which receives a continuous action of the absolute Galois group G_K. Fixing a choice of basis, we obtain a continuous representation

$$\rho_{E,\ell} : G_K \to \mathrm{GL}_2(\mathbf{Q}_\ell).$$

(This representation is the same, up to passing to the dual and taking a twist by the cyclotomic character, of the representation afforded by the ℓ-adic Tate module of E.) If v is a finite place of K not dividing ℓ and at which E has good reduction, then the representation $\rho_{E,\ell}$ is unramified at v. By definition, this means that the inertia subgroup I_{K_v} of the decomposition group $G_{K_v} \subset G_K$ acts trivially through $\rho_{E,\ell}$. Moreover, if $\mathrm{Frob}_v \in G_{K_v}/I_{K_v} \cong G_{k(v)}$ denotes the Frobenius element,[1] then we have the equality

$$\mathrm{tr}\, \rho_{E,\ell}(\mathrm{Frob}_v) = a_v(E),$$

a consequence of the Grothendieck–Lefschetz trace formula for the reduction modulo v of the elliptic curve E. We call the collection $(\rho_{E,\ell})_\ell$ of ℓ-adic representations a “compatible system” because these Frobenius traces are independent of ℓ (even though the representations themselves are incomparable, because the topological fields $\mathbf{Q}_\ell$ are pairwise non-isomorphic).

So much for elliptic curves. We next want to introduce the structures “on the automorphic side” that should be matched up with elliptic curves under the modularity conjecture. By analogy with class field theory, which gives a description of the 1-dimensional representations of G_K, these structures should be defined using the “internal arithmetic” of the field K. To write these down, we first need to recall the existence of the adèle ring of K.

Definition 2.2. The finite adèle ring of K is the restricted direct product

$$\mathbf{A}_K^\infty = \prod_{v \text{ finite}}{}' K_v$$

with respect to the valuation rings $\mathcal{O}_{K_v} \subset K_v$. The adèle ring of K is the product $\mathbf{A}_K = \mathbf{A}_K^\infty \times K_\infty$, where $K_\infty = \prod_{v \text{ infinite}} K_v$.

In other words, $\mathbf{A}_K$ is the set of elements $x = (x_v)_v \in \prod_v K_v$ such that for all but finitely many finite places v of K, $x_v \in \mathcal{O}_{K_v}$. The fundamental facts concerning $\mathbf{A}_K$ are that it is a locally compact topological ring, the diagonal embedding $K \to \mathbf{A}_K$ induces the discrete topology on K, and the quotient $\mathbf{A}_K/K$ is compact.

[1] More precisely, the geometric Frobenius element (inverse of the arithmetic Frobenius automorphism $x \mapsto x^{q_v}$ on $\overline{k(v)}$).

Having defined $\mathbf{A}_K$, we can take the $\mathbf{A}_K$-points of any algebraic group over K. In particular, the group $\mathrm{GL}_2(\mathbf{A}_K)$ is then defined. This group can also be realised as the restricted direct product $\prod'_v \mathrm{GL}_2(K_v)$, with respect to the family of open subgroups $\mathrm{GL}_2(\mathcal{O}_{K_v}) \subset \mathrm{GL}_2(K_v)$ for finite places v.

Definition 2.3. Let $\mathfrak{n} \subset \mathcal{O}_K$ be a non-zero ideal. We define the open compact subgroup of $\prod_{v \text{ finite}} \mathrm{GL}_2(\mathcal{O}_{K_v})$:

$$U_1(\mathfrak{n}) = \left\{ \begin{pmatrix} a_v & b_v \\ c_v & d_v \end{pmatrix} \in \prod_{v \text{ finite}} \mathrm{GL}_2(\mathcal{O}_{K_v}) : \forall v, c_v, d_v - 1 \equiv 0 \bmod \mathfrak{n}\mathcal{O}_{K_v} \right\}.$$

If v is an infinite place of K, we let $U_v = \mathrm{SO}_2(\mathbf{R})$ (if $K_v = \mathbf{R}$) or $U_v = \mathrm{U}_2(\mathbf{R})$ (if $K_v \cong \mathbf{C}$). Let $U_\infty = \mathbf{R}_{>0} \cdot \prod_{v|\infty} U_v \subset \mathrm{GL}_2(K_\infty)$. We then define the quotient topological space

$$Y_1(\mathfrak{n}) = \mathrm{GL}_2(K) \backslash \mathrm{GL}_2(\mathbf{A}_K) / U_1(\mathfrak{n}) \times U_\infty.$$

In order to formulate the modularity conjecture, we will look at the singular cohomology groups $H^*(Y_1(\mathfrak{n}), \mathbf{Q})$. These are finite dimensional $\mathbf{Q}$-vector spaces. Indeed, $Y_1(\mathfrak{n})$ can be represented quite concretely, as we now explain. The double quotient $\mathrm{GL}_2(K) \backslash \mathrm{GL}_2(\mathbf{A}_K^\infty) / U_1(\mathfrak{n})$ (where we omit the infinite places) is finite; if $g_1, \dots, g_n \in \mathrm{GL}_2(\mathbf{A}_K^\infty)$ are coset representatives, then $Y_1(\mathfrak{n})$ can itself be identified with the disjoint union of the quotients $\Gamma_i \backslash \mathrm{GL}_2(K_\infty) / U_\infty$, where we define

$$\Gamma_i = \mathrm{GL}_2(K) \cap g_i U_1(\mathfrak{n}) g_i^{-1}$$

(intersection in $\mathrm{GL}_2(\mathbf{A}_K^\infty)$). The groups Γ_i are congruence subgroups of $\mathrm{GL}_2(K)$, which are torsion-free if the ideal $\mathfrak{n}$ is small enough, so these quotients are generalisations of the modular curves arising in the theory of classical modular forms. In fact, if $K = \mathbf{Q}$ and $\mathfrak{n} = (N)$ for a natural number N, then the space $Y_1(\mathfrak{n})$ defined above may be identified with the usual modular curve of level $\Gamma_1(N)$.

The reason for defining $Y_1(\mathfrak{n})$ using the adèle ring is that it makes transparent the definition of Hecke operators, which are necessary to be able to give a precise formulation of the modularity conjecture. The existence of Hecke operators is a consequence of the following observation: if $U \subset \mathrm{GL}_2(\mathbf{A}_K^\infty)$ is any open compact subgroup, let Y_U be the space defined in the same way as $Y_1(\mathfrak{n})$, except with $U_1(\mathfrak{n})$ replaced by U. If $V \subset U$, then there is a natural projection $Y_V \to Y_U$. We can thus form the direct limit

$$\mathcal{A} = \varinjlim_{U} H^*(Y_U, \mathbf{Q}),$$

a representation of $\mathrm{GL}_2(\mathbf{A}_K^\infty)$ which is *smooth*, in the sense that each vector is fixed by some open compact subgroup of $\mathrm{GL}_2(\mathbf{A}_K^\infty)$. Moreover, we can recover $H^*(Y_1(\mathfrak{n}), \mathbf{Q})$

as the space of $U_1(\mathfrak{n})$-invariant vectors of $\mathcal{A}$. General considerations (see e.g. [32, §2.2]) then imply that $H^*(Y_1(\mathfrak{n}), \mathbf{Q})$ has the structure of module for the ring $\mathcal{H}(\mathrm{GL}_2(\mathbf{A}_K^\infty), U_1(\mathfrak{n}))$ of compactly supported $U_1(\mathfrak{n})$-biinvariant functions

$$f : \mathrm{GL}_2(\mathbf{A}_K^\infty) \to \mathbf{Q}.$$

Elements of this ring are what we call Hecke operators.

The most fundamental ones are as follows.

Definition 2.4. Let v be a finite place of K which is prime to $\mathfrak{n}$, and let $\varpi_v \in \mathcal{O}_{K_v}$ be a uniformizer of the valuation ring at v. Then the Hecke operator

$$T_v : H^*\big(Y_1(\mathfrak{n}), \mathbf{Q}\big) \to H^*\big(Y_1(\mathfrak{n}), \mathbf{Q}\big)$$

is the endomorphism induced by the characteristic function $f_v \in \mathcal{H}(\mathrm{GL}_2(\mathbf{A}_K^\infty), U_1(\mathfrak{n}))$ of the double coset $U_1(\mathfrak{n})xU_1(\mathfrak{n})$, where $x = (x_w)_w \in \mathrm{GL}_2(\mathbf{A}_K^\infty)$ is the element with $x_w = 1$ if $w \neq v$ and $x_w = \mathrm{diag}(\varpi_v, 1)$ if $w = v$.

This definition is independent of the choice of uniformizer x_v. The Hecke operator T_v can also be described more concretely as the endomorphism of $H^*(Y_1(\mathfrak{n}), \mathbf{Q})$ induced by a correspondence

$$\begin{tikzcd}
& Y_{U_1(\mathfrak{n}) \cap x U_1(\mathfrak{n}) x^{-1}} \arrow[dl] \arrow[dr] & \\
Y_1(\mathfrak{n}) & & Y_1(\mathfrak{n})
\end{tikzcd}$$

However, its definition is explained most clearly by the local Langlands correspondence for unramified representations of $\mathrm{GL}_2(K_v)$, as we will recall below.

We now have everything we need to state a version of the modularity conjecture.

Conjecture 2.5. *Let E be an elliptic curve over K such that* $\mathrm{End}_K(E) = \mathbf{Z}$*. Then there exists an ideal $\mathfrak{n} \subset \mathcal{O}_K$ and a non-zero class $c_E \in H^*(Y_1(\mathfrak{n}), \mathbf{Q})$ such that for all but finitely many finite places v of K, one has the equality*

$$T_v(c_E) = a_v(E)c_E.$$

Various remarks are in order. The restriction to curves with $\mathrm{End}_K(E) = \mathbf{Z}$ is made because curves with $\mathrm{End}_K(E) \neq \mathbf{Z}$ (in other words, elliptic curves with complex multiplication defined over K) behave differently: their Galois representations $\rho_{E,\ell}$ are abelian and are described by class field theory. We note that the condition $\mathrm{End}_K(E) = \mathbf{Z}$ always holds if K is totally real, for example if $K = \mathbf{Q}$.

Next we ask how this conjecture is related to the more classical conjecture in the case $K = \mathbf{Q}$ referenced in the introduction, which phrases modularity in terms of the modular forms, rather than cohomology classes. The bridge between modular forms and cohomology is in this case given by the Eichler–Shimura isomorphism. This is an isomorphism

$$M_2\big(\Gamma_1(N)\big) \oplus \overline{S_2\big(\Gamma_1(N)\big)} \cong H^1\big(Y_1(N), \mathbf{C}\big)$$

respecting the action of Hecke operators on each side. If p is a prime number not dividing N and f is a newform, then the pth Fourier coefficient of f coincides with the eigenvalue of the Hecke operator T_p on f, which explains how newforms give rise to cohomology classes in $H^1(Y_1(N), \mathbf{C})$ which are eigenvectors for Hecke operators. When the eigenvalues are rational numbers, we can even choose eigenvectors which lie in $H^1(Y_1(N), \mathbf{Q})$.

How is this conjecture related to the formulation of Clozel [13], also referenced in the introduction, which would lead one to associate to each elliptic curve E over K with $\mathrm{End}_K(E) = \mathbf{Z}$ a cuspidal automorphic representation π of $\mathrm{GL}_2(\mathbf{A}_K)$? Such a representation π admits a restricted tensor product decomposition $\pi = \otimes'_v \pi_v$, indexed by the set of places v of the number field K. One can predict the isomorphism class of π_v, as a representation of the group $\mathrm{GL}_2(K_v)$, using the local Langlands correspondence for the group $\mathrm{GL}_2(K_v)$. Let us recall that when v is a finite place, the local Langlands correspondence rec_{K_v} is a bijection between the following two sets of objects:

- the set of isomorphism classes of irreducible smooth representations of $\mathrm{GL}_2(K_v)$ over $\mathbf{C}$,
- the set of isomorphism classes of 2-dimensional Frobenius-semisimple Weil–Deligne representations of the Weil group $W_{K_v} \subset G_{K_v}$ over $\mathbf{C}$.

We can use the local Langlands correspondence to build an irreducible representation $\pi(E)$ of $\mathrm{GL}_2(\mathbf{A}_K^\infty)$ from an elliptic curve E over K, by specifying a Weil–Deligne representation (r_v, N_v) of the group W_{K_v} for each finite place v of K using the local representations $\rho_{E,\ell}|_{W_{K_v}}$. (For an explanation of how to do this, see e.g. [37].) Thus $\pi(E)$ is the restricted tensor product of the local factors $\mathrm{rec}_{K_v}^{-1}(r_v \otimes |\cdot|^{1/2}, N_v)$. In particular, this leads to the following more precise conjecture, which implies Conjecture 2.5.

Conjecture 2.6. *Let E be an elliptic curve over K such that* $\mathrm{End}_K(E) = \mathbf{Z}$, *and let $\pi(E)$ be the irreducible smooth representation of* $\mathrm{GL}_2(\mathbf{A}_K^\infty)$ *associated to E using the local Langlands correspondence. Then there is a* $\mathrm{GL}_2(\mathbf{A}_K^\infty)$*-equivariant injection* $\pi(E) \hookrightarrow \mathcal{A} \otimes_{\mathbf{Q}} \mathbf{C}$.

From this point of view we can explain the importance of the Hecke operators T_v in formulating the modularity conjecture, which is otherwise slightly obscure: if v is a finite place of K, then the local Langlands correspondence restricts to a bijection between the following two sets of objects:

- the set of isomorphism classes of smooth representations of $\mathrm{GL}_2(K_v)$ over $\mathbf{C}$ which are *unramified*, in the sense that the space of $\mathrm{GL}_2(\mathcal{O}_{K_v})$-invariant vectors is non-zero,
- the set of isomorphism classes of 2-dimensional semisimple representations of W_{K_v} which are unramified, in the sense that the inertia group $I_{K_v} \subset W_{K_v}$ acts trivially.

If π_v is an unramified irreducible smooth representation of $\mathrm{GL}_2(K_v)$ and $r \otimes |\cdot|^{1/2} = \mathrm{rec}_{K_v}(\pi_v)$, then the Hecke operator T_v acts by a scalar on the space of $\mathrm{GL}_2(\mathcal{O}_{K_v})$-invariant vectors of π_v which is equal to $\mathrm{tr}\, r(\mathrm{Frob}_v)$. We have already observed that if v is a place of good reduction for the elliptic curve E, then the Grothendieck–Lefschetz trace formula implies the equality $\rho_{E,\ell}(\mathrm{Frob}_v) = a_v(E)$, provided v is prime to ℓ. This explains the essential equality

$$\text{eigenvalue of } T_v = a_v(E)$$

which appears in the statement of Conjecture 2.5.

One can (and should) go further than we do here. For example, is it possible to describe all of the systems of Hecke eigenvalues which appear in $H^*(Y_1(\mathfrak{n}), \mathbf{C})$ in terms of abelian varieties? They cannot all be described in terms of elliptic curves since, for example, there are systems of Hecke eigenvalues which are not all rational numbers and so cannot come from elliptic curves. See [38] for a precise conjectural description in terms of "false generalised elliptic curves."

3. Applications of modularity

We briefly discuss some applications of the modularity conjecture for elliptic curves. Our intent here is not to be exhaustive but rather to give a flavour of some of the many different applications of modularity that exist.

We mention first applications to Fermat's Last Theorem and other Fermat-style problems. Let us recall the strategy to prove Fermat's Last Theorem used in [45]. Let $p \geq 5$ be a prime number, and suppose given a non-trivial solution

$$a^p + b^p = c^p$$

to the Fermat equation in exponent p; thus $a, b, c \in \mathbf{Z}$ are coprime non-zero integers. One associates to such a non-trivial solution the Frey–Hellegoarch elliptic curve

$$E_{a,b,c} : y^2 = x(x - a^p)(x + b^p).$$

After possibly permuting a, b, c (in order to optimise the local behaviour at the prime 2), the minimal discriminant of this elliptic curve over $\mathbf{Q}$ is $2^{-8}(abc)^{2p}$ (see for example the calculation in [35, §4.1]). This implies that the reduction of the p-adic Galois representation $\rho_{E_{a,b,c},p}$ (to be discussed further below) can be ramified only at the prime 2 (and is finite flat at p). The modularity of the curve $E_{a,b,c}$, together with Ribet's level-lowering theorem, then implies the existence of a newform of weight 2 and level $\Gamma_0(2)$, a contradiction.

Variants of this strategy may be employed to study the generalised Fermat equations

$$a^p + b^q = c^r,$$

where $p, q, r \geq 2$ are integers satisfying $1/p + 1/q + 1/r < 1$. Bennett et al. [4] describe a broad range of exponents for which it can be proved using variants of the above modularity-based method that no non-trivial solutions exist. One can also study solutions to these equations in number fields other than $\mathbf{Q}$. Assuming a strengthened version of the modularity conjecture (Conjecture 2.5) for an imaginary quadratic field $K = \mathbf{Q}(\sqrt{-d})$, where $d > 0$ is an even squarefree integer, Şengün and Siksek [34] prove that for all sufficiently large prime numbers p, there are no non-trivial solutions to the Fermat equation in exponent p over $\mathcal{O}_K$. See also [20] for similar (and unconditional) results over real quadratic fields.

These kinds of modular techniques can also be used to get positive (as opposed to non-existence) information about solutions to Diophantine equations. An example is given by the following theorem, taken from the work of von Känel and Matschke [44].

Theorem 3.1. *Let a be a non-zero integer. Then any solution $(x, y) \in \mathbf{Z}^2$ to the equation $y^2 = x^3 + a$ satisfies the estimate*

$$\max\left(\log|x|, \frac{2}{3}\log|y|\right) \leq 1728|a|\big(\log|a| + 4\big).$$

Modularity is also of great importance for studying individual elliptic curves. For example, essentially all known results towards the Birch–Swinnerton-Dyer (BSD) conjecture are restricted to the class of modular elliptic curves. The BSD conjecture concerns the L-function of an elliptic curve over a number field.

Definition 3.2. Let E be an elliptic curve over a number field K. The L-function $L(E, s)$ of E is the function of a complex variable s defined by the Euler product, indexed by finite places v of K:

$$L(E, s) = \prod_{v \text{ bad}} \left(1 - a_v(E)q_v^{-s}\right)^{-1} \prod_{v \text{ good}} \left(1 - a_v(E)q_v^{-s} + q_v^{1-2s}\right)^{-1}.$$

The Hasse estimate implies that this Euler product converges absolutely in the right half-plane $\mathrm{Re}(s) > 3/2$, and defines a complex analytic function there. We then have the following fundamental conjectures.

Conjecture 3.3. *Let E be an elliptic curve over a number field K.*

(1) *(Analytic continuation) The function $L(E, s)$ admits an analytic continuation to the whole complex plane. Defining $\Lambda(E, s) = (2\pi^{-s}\Gamma(s))^{[K:\mathbf{Q}]}L(E, s)$, there is a natural number N and a sign $\varepsilon \in \{\pm 1\}$ such that the functional equation*

$$\Lambda(E, s) = \varepsilon N^{1-s}\Lambda(E, 2-s)$$

holds.

(2) *(Weak BSD) Assuming* (1)*, the order of vanishing of $L(E, s)$ at the point $s = 1$ is equal to the rank r_E of the finitely generated abelian group $E(K)$.*

(3) *(Strong BSD) Assuming* (2)*, one has*

$$\lim_{s\to 1} \frac{L(E, s)}{(s-1)^{r_E}} = P(E)R(E)\big|\,\mathrm{Sha}(E)\big|,$$

where $P(E)$ is a product of local terms, $R(E)$ is the regulator of $E(K)$, and $\mathrm{Sha}(E)$ *is the Tate–Shafarevich group of E. In particular,* $\mathrm{Sha}(E)$ *is finite.*

Here we follow the formulation of the strong BSD conjecture given by Gross [22], to which we refer for the definition of the terms $P(E)$, $R(E)$.

Theorem 3.4. *Let E be an elliptic curve over a number field K. Then*

(1) *if E satisfies Conjecture 2.6, then $L(E, s)$ has an analytic continuation,*

(2) *if E satisfies Conjecture 2.6 and K is totally real, and either $[K : \mathbf{Q}]$ is odd or there exists a finite place v such that the Weil–Deligne representation of W_{K_v} associated to E is indecomposable, then the weak BSD conjecture holds for E provided that the order of vanishing of $L(E, s)$ at the point $s = 1$ is* at most 1.

If E satisfies Conjecture 2.6, then there is a cuspidal automorphic representation π of $\mathrm{GL}_2(\mathbf{A}_K)$ such that $L(E, s) = L(\pi, s)$. In other words, we may identify $L(E, s)$ with an automorphic L-function. The analytic continuation of $L(E, s)$ is then a consequence of the known continuation for such automorphic L-functions [25]. When $K = \mathbf{Q}$ and $L(E, s)$ vanishes to order at most 1, the validity of the weak BSD conjecture follows from the Gross–Zagier formula and work of Kolyvagin [23, 29].

These results were generalised to a general totally real field K by Zhang [46]. It is interesting to note that the Gross–Zagier formula and its generalisations depend on the existence of a modular parameterisation, i.e., a non-constant map from a

Shimura curve defined over K to the elliptic curve E. The existence of such a parameterisation for a curve E satisfying the hypothesis of Theorem 3.4 (2) is a non-trivial consequence of its modularity, in the sense of Conjecture 2.5.

4. Known results

We now discuss what is known towards the modularity conjecture (Conjecture 2.5). First, it is known for elliptic curves over $\mathbf{Q}$ [8, 40, 45].

Theorem 4.1. *Every elliptic curve over* $\mathbf{Q}$ *is modular.*

We review the structure of the proof, which underlies all known generalisations of this theorem. First, we change our point of view slightly by considering the modularity of the Galois representations $\rho_{E,\ell} : G_K \to \mathrm{GL}_2(\mathbf{Q}_\ell)$ associated to an elliptic curve over a number field K. For example, we can make the following definition.

Definition 4.2. Let K be a number field, let ℓ be a prime number, and let $\rho : G_K \to \mathrm{GL}_2(\mathbf{Q}_\ell)$ be a continuous representation. We say that ρ is modular if there exists a non-zero ideal $\mathfrak{n} \subset \mathcal{O}_K$ and a non-zero class $c_\rho \in H^*(Y_1(\mathfrak{n}), \mathbf{Q}_\ell)$ satisfying the following condition: for all but finitely many finite places v of K, $\rho|_{G_{K_v}}$ is unramified, c_ρ is an eigenvector of the Hecke operator T_v, and we have the equality

$$T_v(c_\rho) = \big(\mathrm{tr}\,\rho(\mathrm{Frob}_v)\big)c_\rho.$$

In view of the equality $a_v(E) = \mathrm{tr}\,\rho_{E,\ell}(\mathrm{Frob}_v)$ for prime-to-ℓ places at which E has good reduction, we see that Conjecture 2.5 holds for an elliptic curve E over K if and only if one (or equivalently, all) of its ℓ-adic Galois representations is modular in the above sense.

It is important to note that this notion of modularity is very restrictive. It is believed (and known, in many cases) that any Galois representation which is modular in the above sense must be of weight 2, in the sense defined in [38]. To encompass all (say irreducible 2-dimensional) Galois representations which arise from the étale cohomology of algebraic varieties over K we would need to consider a broader definition of modularity, encompassing all of the algebraic automorphic representations of $\mathrm{GL}_2(\mathbf{A}_K)$ singled out in [13].

We can also define a notion of modularity for representations with coefficients in $\mathbf{F}_\ell$, the field with ℓ elements.

Definition 4.3. Let $\overline{\rho} : G_K \to \mathrm{GL}_2(\mathbf{F}_\ell)$ be a continuous representation. We say that $\overline{\rho}$ is modular if there exists a non-zero ideal $\mathfrak{n} \subset \mathcal{O}_K$ and a non-zero class $c_\rho \in H^*(Y_1(\mathfrak{n}), \mathbf{F}_\ell)$ satisfying the following condition: for all but finitely many finite places v of K, $\rho|_{G_{K_v}}$ is unramified, c_ρ is an eigenvector of the Hecke operator T_v, and we have the equality

$$T_v(c_\rho) = \big(\operatorname{tr} \rho(\mathrm{Frob}_v)\big)c_\rho.$$

Any continuous representation $\rho : G_K \to \mathrm{GL}_2(\mathbf{Q}_\ell)$ may be conjugated to take values in $\mathrm{GL}_2(\mathbf{Z}_\ell)$; reduction modulo ℓ then gives a representation valued in $\mathrm{GL}_2(\mathbf{F}_\ell)$. We write $\overline{\rho} : G_K \to \mathrm{GL}_2(\mathbf{F}_\ell)$ for the semisimplification of this representation, which is (up to isomorphism) independent of any choices. It is easy to prove that if ρ is modular in the sense of Definition 4.2, then $\overline{\rho}$ is modular in the sense of Definition 4.3.

A fundamental idea behind the proof of Theorem 4.1 and its generalisations, first introduced in [45], is that of the modularity lifting theorem, which gives conditions under which one can go in the other direction and "lift" the modularity of the residual representation $\overline{\rho}$ to the characteristic 0 representation ρ. Many such results now exist in the literature, all approximations to the following ideal:

Theorem Schema 4.4. *Let $\rho : G_K \to \mathrm{GL}_2(\mathbf{Q}_\ell)$ be a continuous representation satisfying*

(1) *some global conditions on $\overline{\rho}$, such as the irreducibility of $\overline{\rho}$,*

(2) *some necessary local conditions on ρ, such as that ρ be of weight 2, in the sense of* [38]*,*

(3) *$\overline{\rho}$ is modular.*

Then ρ is modular.

The first such theorem, proved in [40, 45], was sufficient to establish the modularity of semistable elliptic curves over $\mathbf{Q}$ (i.e., those elliptic curves with good or multiplicative reduction everywhere). In order to apply such a theorem, say to prove the modularity of an elliptic curve E, one needs a way to verify the modularity of the residual representation $\overline{\rho}_{E,\ell}$ for some prime ℓ. Wiles was able to do this when $\ell = 3$ and $K = \mathbf{Q}$ by exploiting a few very happy coincidences:

- The homomorphism $\mathrm{GL}_2(\mathbf{Z}_3) \to \mathrm{GL}_2(\mathbf{F}_3)$ given by reduction modulo 3 splits. Consequently, for any elliptic curve E over $\mathbf{Q}$ we can find a representation $\widetilde{\rho} : G_{\mathbf{Q}} \to \mathrm{GL}_2(\mathbf{Z}_3)$ with *finite image* and lifting $\overline{\rho}_{E,3}$.
- The group $\mathrm{GL}_2(\mathbf{F}_3)$ is soluble. The Langlands–Tunnell theorem [43], which gives the automorphy (in the sense of [13]) of 2-dimensional representations of $G_{\mathbf{Q}}$ (or more generally G_K, where K is any number field) with finite soluble image, implies that $\widetilde{\rho}$ may be associated to a weight 1 holomorphic newform.
- There exist plentiful congruences between weight 1 newforms and weight 2 newforms (for example, given by multiplying by a well-chosen weight 1 Eisenstein series). The existence of such congruences is needed to pass from the automorphy

of $\widetilde{\rho}$ to the modularity of $\overline{\rho}$ in our sense (which is also the sense required for application of the modularity lifting theorem in [45]).

Verifying the modularity of $\overline{\rho}_{E,3}$ in this way, Wiles was able to prove the modularity of those semistable elliptic curves over $\mathbf{Q}$ for which $\overline{\rho}_{E,3}$ is irreducible. To take care of those curves for which $\overline{\rho}_{E,3}$ is reducible (or in other words, for which E admits a rational 3-isogeny), he introduced a beautiful trick, the "3-5 switch," exploiting the geometry of modular curves of low level to prove the modularity of $\overline{\rho}_{E,5}$ instead. This suffices since there are no semistable elliptic curves over $\mathbf{Q}$ with a rational 15-isogeny!

4.1. Elliptic curves over totally real fields

The strongest known modularity lifting theorem suitable for applications to the modularity of elliptic curves over totally real number fields K is the following result, taken from [19, Theorem 2].

Theorem 4.5. *Let K be a totally real number field and let E be an elliptic curve over K. Suppose that there exists an odd prime ℓ such that the following conditions are satisfied:*

(1) *$\overline{\rho}_{E,\ell}$ is modular,*

(2) *$\overline{\rho}_{E,\ell}|_{G_{K(\zeta_\ell)}}$ is absolutely irreducible (here ζ_ℓ denotes a primitive ℓth root of unity in the fixed algebraic closure of K).*

Then $\rho_{E,\ell}$ is modular (and hence E itself is modular).

This is very close to optimal! The possibility of proving a theorem like this is based on numerous technical improvements to the methods introduced in [40, 45], which are due to many people. First, one has to understand why it may be reasonable to generalise modularity lifting theorems from the case $K = \mathbf{Q}$ to the case where K is totally real. For a totally real field, the analogues of holomorphic modular forms are Hilbert modular forms. Most of the Galois representations attached to Hilbert modular forms may be constructed and analyzed using Shimura curves and the Jacquet–Langlands correspondence [12], giving a theory quite analogous to the theory of classical modular curves.

Diamond and Fujiwara [16, 21] explained how to generalise the fundamental Taylor–Wiles patching technique introduced in [40] to this context, making it possible to prove the first modularity lifting theorems over totally real fields, and also introducing soluble base change, using [30], as a fundamental tool. At this point the main question was how to impose conditions from ℓ-adic Hodge theory[2] (such as

[2] More normally called p-adic Hodge theory, but we consider ℓ-adic representations in this article.

the above-mentioned weight 2 condition) while still being able to control the Galois deformation theory (in [45] only smooth conditions were considered, in which case computing the tangent space to the deformation functor in terms of Galois cohomology is enough – not so in general). This problem was solved by Kisin [28], who introduced a variant of the Taylor–Wiles method and defined and analysed weight 2 lifting functors using sophisticated results in integral ℓ-adic Hodge theory. Finally, Khare and Wintenberger, on their way to proving Serre's conjecture, introduced an important new technique for constructing liftings of modulo ℓ Galois representations with prescribed properties [27], using modularity lifting theorems and Taylor's potential automorphy technique [39] as an input. This was exploited in a very clever way by Barnet-Lamb, Gee, and Geraghty [3] in order to optimise Kisin's results.

With Theorem 4.5 in hand, we see that for an elliptic curve over a totally real field K to fail to be modular, each of its residual representations must either be degenerate (in the sense that $\overline{\rho}_{E,\ell}|_{G_{K(\zeta_\ell)}}$ is reducible) or must fail to be modular. The coincidences underlying Wiles's proof of the representations $\overline{\rho}_{E,3}$, together with the 3-5 switch, generalise well to the totally real context. Using the geometry of the modular curve $X(7)$, Manoharmayum [31] gave a 3-7 switch argument, making it possible now to prove the following theorem.

Theorem 4.6. *Let E be an elliptic curve over a totally real field K. If $\overline{\rho}_{E,\ell}|_{G_{K(\zeta_\ell)}}$ is absolutely irreducible for any of $\ell = 3$, 5 or 7, then E is modular.*

Using this, Freitas, Le Hung, and Siksek were able to prove the following striking result.

Theorem 4.7. *Let K be a totally real field. Then,*

(1) *there is a finite set $S \subset K$ such that if E is an elliptic curve over K and $j(E) \notin S$, then E is modular,*

(2) *if $[K : \mathbf{Q}] = 2$, then every elliptic curve over K is modular.*

(Here $j(E)$ is the j-invariant, which classifies the $\overline{K}$-isomorphism class of E.) The proof of this theorem is based on the following idea: if E is a non-modular elliptic curve, then, by Theorem 4.6, it must determine a rational point on one of a finite set of modular curves parameterising elliptic curves with some of kind degeneracy of their modulo 3, 5, and 7 Galois representations. (For example, this set would include the curve $X_0(105)$, which parameterises elliptic curves for which each of the modulo 3, 5, and 7 Galois representations is reducible already on G_K.) The first part of Theorem 4.7 is thus a consequence of the observation that each of these modular curves has genus greater than 2, together with Faltings's theorem (i.e., Mordell's conjecture) [18]. The second part, much the harder, is to analyse the points of these modular curves which are defined over real quadratic fields. Similar ideas have been used by

Derickx, Najman, and Siksek to establish also the modularity of elliptic curves over totally real cubic fields [15], and by Box to establish the modularity of elliptic curves over most totally real quartic fields [6].

Here is a “vertical” analogue of Theorem 4.7 (2), proved in [42].

Theorem 4.8. *Let p be a prime, and let $K/\mathbf{Q}$ be a totally real abelian extension, unramified away from the prime p, such that* $\mathrm{Gal}(K/\mathbf{Q})$ *has order a power of p. Then every elliptic curve over K is modular.*

This theorem is again proved by combining modularity lifting theorems and an analysis of rational points on modular curves, although in a different way. The first main ingredient is a new modularity lifting theorem, proved in [41], which removes the assumption that $\rho_{E,\ell}|_{G_{K(\zeta_\ell)}}$ is irreducible. This so-called Taylor–Wiles assumption is used to control certain Galois cohomology groups. The effect of this new theorem is that in proving Theorem 4.8, one needs consider only rational points on the single modular curve $X_0(15)$. This curve has genus 1, so could have infinitely many rational points over a fixed number field (as it does, for example, over $\mathbf{Q}(\sqrt{3})$). However, it turns out that for any field K as in the statement of Theorem 4.8, we in fact have $X_0(15)(K) = X_0(15)(\mathbf{Q})$! Any such field K is contained in the cyclotomic $\mathbf{Z}_p$-extension $\mathbf{Q}_\infty/\mathbf{Q}$, so the natural tool to prove this is Iwasawa theory, and in particular the results of Kato [26].

Looking at Theorems 4.7 and 4.8, it seems reasonable, in principle, to try to prove the modularity of all elliptic curves over any family $\mathcal{F}$ of totally real number fields for which the points of modular curves rational over members of $\mathcal{F}$ can be “organised” in some way. Establishing the modularity of elliptic curves over all totally real fields will require new ideas.

4.2. Elliptic curves over more general number fields

We now consider the modularity of elliptic curves over number fields which are not totally real. Until a few years ago, it was very mysterious how one might hope to prove modularity lifting theorems in this context. First, it is not known in general how to associate Galois representations to Hecke eigenclasses in $H^*(Y_1(\mathfrak{n}), \mathbf{Q}_\ell)$. Indeed, the spaces $Y_1(\mathfrak{n})$ (and their analogues, associated to quaternion algebras over number fields) have no obvious relation to algebraic geometry (for example, when K has a complex place they have no complex structure). Second, even if one could solve this problem, the spaces $Y_1(\mathfrak{n})$ can have non-trivial torsion classes in their cohomology (say with $\mathbf{Z}_\ell$ coefficients) which cannot be described in terms of automorphic representations (see e.g. [5]). Third, the Taylor–Wiles method breaks down because the cohomology groups of $Y_1(\mathfrak{n})$ (again, say, with $\mathbf{Z}_\ell$ coefficients, and now some auxiliary Taylor–Wiles level structure) are not free modules over the group rings of diamond operators that appear in the version of the Taylor–Wiles method developed by Diamond and Fujiwara.

The way forward was explained by Calegari and Geraghty [10]. Assuming a number of conjectures, they explain how to generalise the Taylor–Wiles method and prove modularity lifting theorems over general number fields which can be applied, for example, to prove the modularity of elliptic curves. We will not attempt to formulate these conjectures precisely here but note that their conjectures include the important prescription that there should exist Galois representations associated not just to (algebraic) automorphic representations with complex coefficients, but also to torsion classes in the cohomology of spaces like $Y_1(\mathfrak{n})$. This is a striking enlargement of the Langlands program as outlined in [13]!

To get unconditional results, one still has to establish the conjectures which are taken as a starting point in [10]. Progress towards these conjectures was made first by Scholze, who used his theory of perfectoid spaces to prove the existence of Galois representations attached to Hecke eigenclasses in the groups $H^*(Y_1(\mathfrak{n}), \mathbf{Z}_\ell)$ when K is a CM field, i.e., a totally imaginary quadratic extension of a totally real field [33]. Using the further results of Caraiani and Scholze on the cohomology of noncompact Shimura varieties [11], the 10-author collaboration [1] established enough of the Calegari–Geraghty conjectures to be able to establish unconditional modularity lifting theorems over CM fields. These sufficed to be able to prove, for example, the potential modularity of all elliptic curves E over CM fields K (i.e., the modularity of the base change E_L, for some finite extension L/K depending on E – a result which implies in particular the *meromorphic* continuation to $\mathbf{C}$ of $L(E, s)$).

Separately, Boxer, Calegari, Gee, and Pilloni studied the application of the Calegari–Geraghty method in the context of the coherent cohomology of Siegel type Shimura varieties [7]. The problems faced here are analogous, but different, to those arising out of the singular cohomology of the locally symmetric spaces $Y_1(\mathfrak{n})$. Nevertheless these authors were able to prove unconditional modularity lifting theorems that can be applied to the Galois representations arising from abelian surfaces over totally real fields. As a particular consequence, they are able to prove the potential modularity of elliptic curves over any quadratic extension of a totally real field (not necessarily CM) – the first general results of this kind that can be applied to elliptic curves over non-CM fields. An excellent guide to the path to the results of the last few paragraphs can be found in the survey article [9].

What about modularity (as opposed to potential modularity) of elliptic curves? To prove modularity using modularity lifting theorems, one needs a source of modular residual representations. Unfortunately, one can no longer use Wiles's idea to prove the modularity of representations $\overline{\rho}_{E,3}$ for elliptic curves E when the base field K is not totally real. The reason is that, although the Langlands–Tunnell theorem applies over arbitrary base number fields, there is no known way to construct congruences between the automorphic representations it gives and those automorphic representations which contribute to the cohomology of locally symmetric spaces. A solution to this problem would also allow the construction of the Galois representations associated to algebraic Maass forms, a famously difficult open problem!

Nevertheless, we were able to establish the following theorem in [2].

Theorem 4.9. *Let K be a CM field, and let E be an elliptic curve over K with multiplicative reduction at each place $v|5$ of K. Then $\overline{\rho}_{E,3}$ is modular.*

Corollary 4.10. *Let K be a CM field such that $\zeta_5 \notin K$. Then a positive proportion of elliptic curves over K are modular.*

The proof of Theorem 4.9 is based on the idea of a kind of 2-3 switch: we want to find an auxiliary elliptic curve A such that $\overline{\rho}_{A,3} \cong \overline{\rho}_{E,3}$ and $\overline{\rho}_{A,2}$ extends to a representation of G_{K^+}, where K^+ is the maximal totally real subfield of K. A tricky 2-adic modularity lifting theorem would then imply the modularity of A, hence of $\overline{\rho}_{A,3} \cong \overline{\rho}_{E,3}$. In fact, the existence of such an auxiliary curve A is a delicate matter (partly explained by the fact that the modular curve $X(6)$ has genus 1) and we need to take a more circuitous route, for which we refer to [2].

The local conditions at the 5-adic places in Theorem 4.9 are always satisfied after possibly replacing K by a soluble CM extension. Since we are free to make a soluble base change when establishing the modularity of a given elliptic curve E (by cyclic base change [30]), a sufficiently powerful modularity lifting theorem would, when combined with Theorem 4.9, prove the modularity of most elliptic curves over a given CM field.

The modularity lifting theorems established in [1] apply only to elliptic curves which have either good reduction at each place of K above the fixed prime ℓ, with ℓ unramified in K, or which have good ordinary/multiplicative reduction at each place of K above ℓ. Thus we do not have yet access to theorems such as those proved by Kisin over totally real fields [28], in which an arbitrary amount of ramification is permitted. If such theorems can be established in the future, then it seems reasonable to hope that it will be possible to prove e.g. the modularity of all elliptic curves over imaginary quadratic fields.

Acknowledgements. I am grateful to the anonymous referee for their useful comments.

Funding. The author's work received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 714405).

References

[1] P. B. Allen, F. Calegari, A. Caraiani, T. Gee, D. Helm, B. V. Le Hung, J. Newton, P. Scholze, R. L. Taylor, and J. A. Thorne, Potential automorphy over CM fields. 2018, arXiv:1812.09999

[2] P. B. Allen, C. Khare, and J. A. Thorne, Modularity of $\mathrm{GL}_2(\mathbb{F}_p)$-representations over CM fields. 2019, arXiv:1910.12986

[3] T. Barnet-Lamb, T. Gee, and D. Geraghty, Congruences between Hilbert modular forms: constructing ordinary lifts. *Duke Math. J.* **161** (2012), no. 8, 1521–1580 Zbl 1297.11028 MR 2931274

[4] M. A. Bennett, I. Chen, S. R. Dahmen, and S. Yazdani, Generalized Fermat equations: a miscellany. *Int. J. Number Theory* **11** (2015), no. 1, 1–28 Zbl 1390.11065 MR 3280939

[5] N. Bergeron and A. Venkatesh, The asymptotic growth of torsion homology for arithmetic groups. *J. Inst. Math. Jussieu* **12** (2013), no. 2, 391–447 Zbl 1266.22013 MR 3028790

[6] J. Box, Elliptic curves over totally real quartic fields not containing $\sqrt{5}$ are modular. *Trans. Amer. Math. Soc.* **375** (2022), no. 5, 3129–3172 Zbl 07502495 MR 4402658

[7] G. Boxer, F. Calegari, T. Gee, and V. Pilloni, Abelian surfaces over totally real fields are potentially modular. *Publ. Math. Inst. Hautes Études Sci.* **134** (2021), 153–501 MR 4349242

[8] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, On the modularity of elliptic curves over **Q**: wild 3-adic exercises. *J. Amer. Math. Soc.* **14** (2001), no. 4, 843–939 Zbl 0982.11033 MR 1839918

[9] F. Calegari, Reciprocity in the Langlands program since Fermat's Last Theorem. 2021, arXiv:2109.14145

[10] F. Calegari and D. Geraghty, Modularity lifting beyond the Taylor–Wiles method. *Invent. Math.* **211** (2018), no. 1, 297–433 Zbl 06830049 MR 3742760

[11] A. Caraiani and P. Scholze, On the generic part of the cohomology of non-compact unitary Shimura varieties. 2019, arXiv:1909.01898

[12] H. Carayol, Sur les représentations l-adiques associées aux formes modulaires de Hilbert. *Ann. Sci. École Norm. Sup. (4)* **19** (1986), no. 3, 409–468 Zbl 0616.10025 MR 870690

[13] L. Clozel, Motifs et formes automorphes: applications du principe de fonctorialité. In *Automorphic Forms, Shimura Varieties, and L-Functions, Vol. I (Ann Arbor, MI, 1988)*, pp. 77–159, Perspect. Math. 10, Academic Press, Boston, MA, 1990 Zbl 0705.11029 MR 1044819

[14] J. E. Cremona, Hyperbolic tessellations, modular symbols, and elliptic curves over complex quadratic fields. *Compositio Math.* **51** (1984), no. 3, 275–324 Zbl 0546.14027 MR 743014

[15] M. Derickx, F. Najman, and S. Siksek, Elliptic curves over totally real cubic fields are modular. *Algebra Number Theory* **14** (2020), no. 7, 1791–1800 Zbl 1471.11178 MR 4150250

[16] F. Diamond, The Taylor–Wiles construction and multiplicity one. *Invent. Math.* **128** (1997), no. 2, 379–391 Zbl 0916.11037 MR 1440309

[17] F. Diamond and J. Shurman, *A First Course in Modular Forms*. Grad. Texts in Math. 228, Springer, New York, 2005 Zbl 1062.11022 MR 2112196

[18] G. Faltings, Endlichkeitssätze für abelsche Varietäten über Zahlkörpern. *Invent. Math.* **73** (1983), no. 3, 349–366 Zbl 0588.14026 MR 718935

[19] N. Freitas, B. V. Le Hung, and S. Siksek, Elliptic curves over real quadratic fields are modular. *Invent. Math.* **201** (2015), no. 1, 159–206 Zbl 1397.11086 MR 3359051

[20] N. Freitas and S. Siksek, The asymptotic Fermat's last theorem for five-sixths of real quadratic fields. *Compos. Math.* **151** (2015), no. 8, 1395–1415 Zbl 1391.11065 MR 3383161

[21] K. Fujiwara, Galois deformations and arithmetic geometry of Shimura varieties. In *International Congress of Mathematicians. Vol. II*, pp. 347–371, EMS Press, Zürich, 2006 Zbl 1130.11028 MR 2275601

[22] B. H. Gross, Lectures on the conjecture of Birch and Swinnerton-Dyer. In *Arithmetic of L-Functions*, pp. 169–209, IAS/Park City Math. Ser. 18, Amer. Math. Soc., Providence, RI, 2011 Zbl 1285.11096 MR 2882691

[23] B. H. Gross and D. B. Zagier, Heegner points and derivatives of L-series. *Invent. Math.* **84** (1986), no. 2, 225–320 Zbl 0608.14019 MR 833192

[24] F. Grunewald, H. Helling, and J. Mennicke, SL_2 over complex quadratic number fields. I. *Algebra i Logika* **17** (1978), no. 5, 512–580, 622 Zbl 0483.10024 MR 555260

[25] H. Jacquet and R. P. Langlands, *Automorphic Forms on* $\mathrm{GL}(2)$. Lecture Notes in Math. 114, Springer, Berlin, 1970 Zbl 0236.12010 MR 0401654

[26] K. Kato, p-adic Hodge theory and values of zeta functions of modular forms. Cohomologies p-adiques et applications arithmétiques. III. *Astérisque* **295** (2004), 117–290 Zbl 1142.11336 MR 2104361

[27] C. Khare and J.-P. Wintenberger, On Serre's conjecture for 2-dimensional mod p representations of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. *Ann. of Math. (2)* **169** (2009), no. 1, 229–253 Zbl 1196.11076 MR 2480604

[28] M. Kisin, Moduli of finite flat group schemes, and modularity. *Ann. of Math. (2)* **170** (2009), no. 3, 1085–1180 Zbl 1201.14034 MR 2600871

[29] V. A. Kolyvagin, Finiteness of $E(\mathbf{Q})$ and $\mathrm{CH}(E, \mathbf{Q})$ for a subclass of Weil curves. *Izv. Akad. Nauk SSSR Ser. Mat.* **52** (1988), no. 3, 522–540 Zbl 0662.14017 MR 954295

[30] R. P. Langlands, *Base Change for* $\mathrm{GL}(2)$. Ann. of Math. Stud. 96, Princeton University Press, Princeton, NJ, 1980 Zbl 0444.22007 MR 574808

[31] J. Manoharmayum, On the modularity of certain $\mathrm{GL}_2(\mathbb{F}_7)$ Galois representations. *Math. Res. Lett.* **8** (2001), no. 5-6, 703–712 Zbl 1001.11023 MR 1879814

[32] J. Newton and J. A. Thorne, Torsion Galois representations over CM fields and Hecke algebras in the derived category. *Forum Math. Sigma* **4** (2016), Paper No. e21 Zbl 1404.11072 MR 3528275

[33] P. Scholze, On torsion in the cohomology of locally symmetric varieties. *Ann. of Math. (2)* **182** (2015), no. 3, 945–1066 Zbl 1345.14031 MR 3418533

[34] M. H. Şengün and S. Siksek, On the asymptotic Fermat's last theorem over number fields. *Comment. Math. Helv.* **93** (2018), no. 2, 359–375 Zbl 1430.11045 MR 3811755

[35] J.-P. Serre, Sur les représentations modulaires de degré 2 de Gal$(\overline{\mathbf{Q}}/\mathbf{Q})$. *Duke Math. J.* **54** (1987), no. 1, 179–230 Zbl 0641.10026 MR 885783

[36] J. H. Silverman, *The Arithmetic of Elliptic Curves*. Grad. Texts in Math. 106, Springer, New York, 1986 Zbl 0585.14026 MR 817210

[37] J. Tate, Number theoretic background. In *Automorphic forms, Representations and L-Functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 2*, pp. 3–26, Proc. Sympos. Pure Math. 33, Amer. Math. Soc., Providence, RI, 1979 Zbl 0422.12007 MR 546607

[38] R. Taylor, Representations of Galois groups associated to modular forms. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pp. 435–442, Birkhäuser, Basel, 1995 Zbl 0864.11022 MR 1403943

[39] R. Taylor, On the meromorphic continuation of degree two L-functions. *Doc. Math.* **Extra Vol.** (2006), 729–779 Zbl 1138.11051 MR 2290604

[40] R. Taylor and A. Wiles, Ring-theoretic properties of certain Hecke algebras. *Ann. of Math. (2)* **141** (1995), no. 3, 553–572 Zbl 0823.11030 MR 1333036

[41] J. A. Thorne, Automorphy of some residually dihedral Galois representations. *Math. Ann.* **364** (2016), no. 1-2, 589–648 Zbl 1404.11079 MR 3451399

[42] J. A. Thorne, Elliptic curves over $\mathbb{Q}_\infty$ are modular. *J. Eur. Math. Soc. (JEMS)* **21** (2019), no. 7, 1943–1948 Zbl 1443.11103 MR 3959855

[43] J. Tunnell, Artin's conjecture for representations of octahedral type. *Bull. Amer. Math. Soc. (N.S.)* **5** (1981), no. 2, 173–175 Zbl 0475.12016 MR 621884

[44] R. von Känel and B. Matschke, Solving S-unit, Mordell, Thue, Thue–Mahler and generalized Ramanujan–Nagell equations via Shimura–Taniyama conjecture. 2016, arXiv:1605.06079

[45] A. Wiles, Modular elliptic curves and Fermat's last theorem. *Ann. of Math. (2)* **141** (1995), no. 3, 443–551 Zbl 0823.11029 MR 1333035

[46] S. Zhang, Heights of Heegner points on Shimura curves. *Ann. of Math. (2)* **153** (2001), no. 1, 27–147 Zbl 1036.11029 MR 1826411

Jack A. Thorne
Department of Pure Mathematics and Mathematical Statistics, University of Cambridge,
Wilberforce Road, Cambridge CB3 0WB, UK; thorne@dpmms.cam.ac.uk

Public lectures

DOI 10.4171/8ECM/37

From art and circuit design to geometry and combinatorics

Bojan Mohar

Abstract. These notes provide a detailed insight on the interplay between crossing numbers of graphs and random geodesic drawings, and try to explain a relationship with the main fundamental open questions about crossing numbers of graphs. A very general class of geodesic drawings on the sphere attaining the Hill bound is presented.

1. Introduction

Crossing number minimization in drawings of graphs on surfaces appears in diverse applications across disciplines. It came into mathematical research through problems in modern constructionist art. Crossing number problems have various applications within engineering (e.g., design of large electrical circuits [17]) and in computer science.

Later it became a useful concept in theoretical questions about graph drawing, algorithm design, and robotics, and became an important notion in discrete and computational geometry. The famous crossing lemma made a very surprising impact within pure mathematics, after it was discovered that it gives greatly simplified proofs for various (seemingly unrelated) hard geometric [32] and algebraic problems [29]. On the other hand, the rectilinear crossing number is related to the classical Sylvester four-point problem, which gave motivation for developments of geometric probability theory. We refer to [26, 27] for a more complete overview of this area of mathematics.

These notes are related to the public talk at the 8th ECM in Portorož, Slovenia, where the author presented some of the complex issues related to geodesic drawings of graphs on surfaces and crossing minimization in such drawings, with a special emphasis on random drawings.

The presentation herein includes historical remarks, it overviews the main fundamental open problems about crossing numbers of graphs, and through a generalization of Sylvester's four-point problem gives special emphasis on random geodesic

2020 Mathematics Subject Classification. 05C10.
Keywords. Crossing number, geodesic drawing, graph drawing.

drawings of graphs on surfaces. When dealing with random geodesic drawings on the sphere, we give a very general class of geodesic drawings attaining the conjectured minimum crossing number.

2. Hill conjecture and Hill drawings

English painter Anthony Hill[1] made an extraordinary conjecture in the 1950s that remained unanswered until today despite serious attacks using powerful machinery in trying to resolve his conjecture. Starting with a question underlying some of his painting projects, Hill tried to understand how to draw $\binom{n}{2}$ connections between n objects so that the painting would involve a minimum number of under or over-crossings. This lead to the formal notion of the *crossing number* of a graph, which he introduced in a mathematical paper jointly with Harary [13].

Given a graph G, one can consider its drawing in the plane (or in some other surface), where vertices are represented as distinct points and edges are drawn as rectifiable arcs joining the corresponding points. One may restrict attention to *good drawings*, where we request that any two edges intersect in at most one point, which is either their common endvertex or a (proper) crossing of two arcs, and no three arcs have their pairwise crossings in the same point. The *crossing number of a (good) drawing D* of a graph G is the number of crossings of pairs of edges in D, and the *crossing number of the graph G*, denoted by $\mathrm{cr}(G)$, is the minimum crossing number taken over all good drawings of G in the plane.

Hill found a general drawing for any complete graph K_n of order n that involves precisely

$$H(n)=\frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor\left\lfloor\frac{n-1}{2}\right\rfloor\left\lfloor\frac{n-2}{2}\right\rfloor\left\lfloor\frac{n-3}{2}\right\rfloor=\begin{cases}\frac{1}{64}n(n-2)^2(n-4), & n \text{ is even;}\\ \frac{1}{64}(n-1)^2(n-3)^2, & n \text{ is odd}\end{cases} \tag{2.1}$$

crossings. Based on these drawings and inability of producing any drawing with less than $H(n)$ crossings, Hill conjectured the following.

Conjecture 2.1 (Hill, 1959). *For any complete graph K_n of order n, we have*

$$\mathrm{cr}(K_n) = \frac{1}{4}\left\lfloor\frac{n}{2}\right\rfloor\left\lfloor\frac{n-1}{2}\right\rfloor\left\lfloor\frac{n-2}{2}\right\rfloor\left\lfloor\frac{n-3}{2}\right\rfloor.$$

[1] Anthony Hill (1930–2020) was one of leading modern British painters. The following is an abstract from his obituary in *Guardian*: “Anthony Hill, who has died aged 90, was a singular, but not solitary, figure in the art world. An artist under two names, and a mathematician and writer under more than one alias, he was a member of the constructionist group of geometrical abstract artists that emerged in Britain in the mid-1950s, and was its leading theoretician.”

Hill's drawings of complete graphs are called *cylindrical drawings* because they can be realized on a cylinder in such a way that all vertices lie (evenly split) on the two circles forming the cylinder and no edge crosses those two circles. Soon after these drawings were published in [13], Blažek and Koman [8] found another kind of drawings of complete graphs involving precisely the same number of crossings. Their drawings correspond to 2-*page drawings* in which the vertices are drawn on the boundary of a unit disk in the plane and no edge crosses this boundary, so each edge is drawn entirely inside the disk or entirely outside. It has been proved quite recently that no cylindrical [2] and no 2-page drawing [1] of K_n has fewer than $H(n)$ crossings, thus giving the first real support to the conjecture of Hill.

We will say that a drawing D of the complete graph K_n is a *Hill drawing* if it has precisely $H(n)$ crossings.

No other Hill drawings of complete graphs have been discovered until 2014 when Ábrego et al. [3] described modifications of cylindrical drawings of K_{2n} yielding Hill drawings of K_{2n+1}, K_{2n+2}, and K_{2n+3} that are different in the sense that they are not "shellable". In Section 4.3, we describe a much more general class of Hill drawings that include in particular all known examples of Hill drawings. These drawings have not appeared previously in the mathematical literature.[2]

After 60+ years, Conjecture 2.1 is still widely open. It has been confirmed for every $n \leq 12$ (with K_{11} and K_{12} confirmed in [23]), but it is still unresolved for $n = 13$ and beyond. In fact, the weaker, asymptotic version of Conjecture 2.1 is also open.

Conjecture 2.2 (Asymptotic Hill conjecture).

$$\operatorname{cr}(K_n) = \frac{1}{64}n^4\big(1 - o(1)\big) = \frac{3}{8}\binom{n}{4}\big(1 - o(1)\big).$$

Of course, the main problem is to show the lower bound – that there are no better drawings than those with precisely $H(n)$ crossings. To that effect, there are several exciting new results that were obtained through elaborate analysis of drawings and use of semidefinite programming and Razborov's flag algebra calculus [18, 25].

Theorem 2.3 (Balogh, Lidický, and Salazar [7]). *For every sufficiently large n,*

$$\operatorname{cr}(K_n) \geq 0.985\, H(n).$$

In the same paper [7], the authors also proved that the spherical geodesic crossing number of K_n (see Section 4.3 for the definition) is asymptotically at least $0.996H(n)$.

[2] After the author put a preprint of this construction on the arXiv [20] in 2018, he was informed that almost the same construction was mentioned by Kyncl on `math`*overflow* [16].

3. Turán's brick factory problem

Turán's brick factory problem asks for the minimum number of crossings in a drawing of a complete bipartite graph $K_{m,n}$. During World War II, Turán was forced to work in a brick factory, pushing wagon loads of bricks from kilns to storage sites, and the corresponding rail network with m kilns and n storage barracks was the same as a special drawing of the complete bipartite graph $K_{m,n}$. Crossings of rail tracks made the transport challenging, and Turán, inspired by this situation, asked himself how the rail network might be redesigned to minimize the number of crossings between the railway tracks [33].

Paul Turán discussed his brick factory problem during 1950s in his talks, and Zarankiewicz and Urbanik, who attended some of his presentations, independently found drawings of complete bipartite graphs for which they claimed that they are optimal [34, 35]. Unfortunately, proofs in both published papers were flawed. This was discovered only a couple of years later, and the claimed minimum number of crossings was turned into the following conjecture, which remains widely open even today.

Conjecture 3.1. *For any positive integers m and n, the crossing number* $\mathrm{cr}(K_{m,n})$ *of the complete bipartite graph with* $m + n$ *vertices is equal to*

$$Z(m,n) = \frac{1}{4}\left\lfloor \frac{m}{2} \right\rfloor \left\lfloor \frac{m-1}{2} \right\rfloor \left\lfloor \frac{n}{2} \right\rfloor \left\lfloor \frac{n-1}{2} \right\rfloor .$$

The conjecture has since been confirmed for the cases where one of the parameters is at most 6 and also for $K_{7,7}$ and $K_{7,8}$, but it remains open even for such small graphs as $K_{7,9}$ and $K_{9,9}$.[3]

4. Geodesic drawings

When we consider drawings of graphs in the plane, where each edge is drawn as a straight-line segment, we come to the notion of the rectilinear crossing number $\overline{\mathrm{cr}}(G)$. It turns out that for most small graphs the rectilinear crossing number is equal to the usual crossing number. However, it was discovered early that $\mathrm{cr}(K_8) = 18$ and $\overline{\mathrm{cr}}(K_8) = 19$ and that this difference extends to larger graphs. However, it was open for a long time how large can be the difference $\overline{\mathrm{cr}}(K_n) - \mathrm{cr}(K_n)$. A breakthrough was made in 2004 by Lovász, Vesztergombi, Wagner, and Welzl [19], who proved that the normalized rectilinear crossing number of K_n is strictly greater than the corresponding limit for the usual crossing number. Building on the work in [19], Ábrego, Cetina,

[3] The cases with even parameters were not mentioned since they would follow from odd cases by known parity arguments; see, e.g. [27].

Probability	Shape	Author
$3/4$	Triangle	Cayley and Sylvester
$1/2$		DeMorgan
$2/3$		Wilson
$> 1/2$		Ingleby
$5/8$		(No name given)
$1 - 35/(12\pi^2)$	Disk	Woolhouse
$25/36$	Rectangle	[10]
$2(18 - \sqrt{5})/45$	Regular pentagon	[9]
$683/972$	Regular hexagon	[14, p. 46]

Table 1. Answers to Sylvester's question. The upper part of the table is taken from [24], where the complementary probabilities for non-convex 4-gon are shown.

Fernández-Merchant, Leaños, and Salazar [5] improved the bounds from [19]. They also found today's best upper bounds [4]. Their results are summarized in the following inequalities:

$$0.379972 < \frac{277}{729} \leq \lim_{n\to\infty} \overline{\operatorname{cr}}(K_n)/\binom{n}{4} \leq \frac{83247328}{218791125} < 0.380488. \qquad (4.1)$$

Unlike for the Hill conjecture, we are lacking understanding of the rectilinear crossing number and there is no good evidence about whether the lower or the upper bound in (4.1) is closer to the normalized limit.

The rectilinear crossing number of complete graphs is tightly related to an old problem in geometric probability that was originally proposed by Julius Sylvester in 1864, and which we will discuss next.

4.1. Sylvester's four-point problem

In 1864, Sylvester asked [30] "what is the probability that four randomly chosen points in the plane form a convex 4-gon?". As it turned out, the problem was ill-posed since by 1865, at least six solutions were received, all with different answers (see the first six entries in Table 1). Depending on the method chosen to pick points from the infinite plane, a number of different solutions are possible, and Sylvester concluded [31] that his problem does not admit a determinate solution (see also [24]).

The reason for so many distinct answers was that it was not clear what "randomly chosen points" in the plane would be. Sylvester himself changed the question a year later [31]. The revised four-point problem asks for the probability $q(R)$ that four points chosen at random in a bounded planar region R have a convex hull which is a quadrilateral.

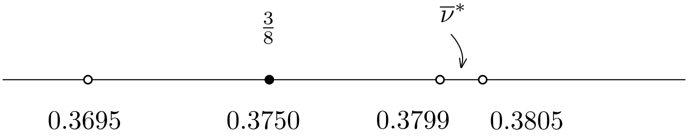


Figure 1. Currently best bounds on the asymptotic values of normalized crossing numbers of large complete graphs: $0.3695 \le \mathrm{cr}(K_n)/\binom{n}{4} \le \frac{3}{8}$ and $0.3799 \le \overline{\mathrm{cr}}(K_n)/\binom{n}{4} \le 0.3805$.

Scheinerman and Wilf linked the Sylvester problem to the rectilinear crossing number [28]. Let $\overline{\nu}^*$ be the limit of $\overline{\mathrm{cr}}(K_n)/\binom{n}{4}$ as $n \to \infty$, and let $q(R)$ be as defined above. Scheinerman and Wilf proved that $\overline{\nu}^* = \inf q(R)$, where the infimum is taken over all open planar sets R whose area is 1 (equivalently, over all unions of finitely many disjoint circles). Let us recall that today's best estimates for the rectilinear crossing number of complete graphs given in (4.1) yield that $0.3799 < \overline{\nu}^* < 0.3805$; see Figure 1.

Note that, in the plane, four points form a convex quadrilateral if and only if the six line segments joining pairs of these points make a crossing. We will use this interpretation in the sequel.

One can pose similar questions when considering randomly chosen points on any surface. Suppose that $\mathbb{S}$ is a compact Riemannian surface and that μ is a probability measure[4] on $\mathbb{S}$. Then we define $q(\mu)$ as the probability that for four μ-randomly chosen points, two of the six geodesics joining pairs of these points cross each other. Then $q(\mu)$ is called the *geometric crossing probability* of μ.

The geometric crossing probabilities are related to the *geodesic crossing number* of the complete graph on $\mathbb{S}$, for which we consider all drawings of the graph in which all edges are drawn as shortest geodesic segments on $\mathbb{S}$. This is not hard to see and we prove it as Lemma 4.1 below. But let us first discuss random drawings.

Let μ be a probability measure on $\mathbb{S}$. We say that μ is *geodesically non-degenerate* if the probability that two μ-random points x, y are distinct and that they are joined with a unique geodesic is equal to 1, and the probability that a third random point lies on this geodesic is 0. If this is the case, then, with probability 1, n randomly selected points define a unique good geodesic drawing of K_n. Such a drawing will be referred to as a *μ-random drawing*.

Lemma 4.1. *Let $\overline{\nu}^*(\mathbb{S})$ be the limit of $\overline{\mathrm{cr}}_{\mathbb{S}}(K_n)/\binom{n}{4}$ (where n tends to infinity) and let $q(\mu)$ be as defined above. Then $\overline{\nu}^*(\mathbb{S}) = \inf q(\mu)$, where the infimum is taken over all geodesically non-degenerate probability, measures μ. The same holds when the infimum is taken over all uniform probability measures whose support is the union of finitely many disjoint disks on $\mathbb{S}$.*

[4] The measure μ has to fulfill some simple non-degeneracy conditions, which will be discussed later.

Proof. Every μ-random drawing D_n of K_n gives an upper bound on the geodesic crossing number of K_n in $\mathbb{S}$. For any four points $a, b, c, d \in \mathbb{S}$, let Q_{abcd} be equal to 1 if two of the geodesics between points a, b, c, d in $\mathbb{S}$ cross each other. Note that $\mathbb{E}(Q_{abcd}) = q(\mu)$ if a, b, c, d are chosen at random with respect to μ, and that

$$\operatorname{cr}(D_n) = \sum \left\{ Q_{abcd} \mid \{a, b, c, d\} \in \binom{V}{4} \right\}.$$

By linearity of expectations, we have

$$\mathbb{E}(\operatorname{cr}(D_n)) = \mathbb{E}\left(\sum Q_{abcd}\right) = \sum \mathbb{E}(Q_{abcd}) = \binom{n}{4} q(\mu).$$

This implies that $\overline{\nu}^*(\mathbb{S}) \le q(\mu)$ for every μ.

To establish equality, let $\delta > 0$ and consider an optimal geodesic drawing D of K_n in $\mathbb{S}$, such that $\operatorname{cr}(D)/\binom{n}{4} - \overline{\nu}^* < \delta$. Let $x_1, \ldots, x_n \in \mathbb{S}$ be the vertices of D. There are $\varepsilon > 0$ and balls $B_1, \ldots, B_n$ centered at these vertices, each of area ε, such that for any choice of points $x_i' \in B_i$ $(1 \le i \le n)$, the geodesic drawing on these points has exactly the same crossings as D.

Let μ_n be the uniform measure on $B_1 \cup \cdots \cup B_n$. We claim that $q(\mu_n)$ is close to $\operatorname{cr}(D)/\binom{n}{4}$. Let us consider four μ_n-random points. With probability at least $1 - O(1/n)$, the four points are in distinct balls B_{i_1}, B_{i_2}, B_{i_3}, B_{i_4} and the four indices i_1, i_2, i_3, i_4 are chosen uniformly at random from $[n]$. Thus, the probability that the geodesics on these four points induces a crossing is at most $(1 - O(1/n))^{-1} \operatorname{cr}(D)/\binom{n}{4}$. This shows that

$$q(\mu_n) \le \big(1 + o(1)\big) \operatorname{cr}(D)/\binom{n}{4} \le \big(1 + o(1)\big)(\overline{\nu}^* + \delta).$$

By letting $\delta \to 0$ and $n \to \infty$, we conclude that

$$\overline{\nu}^* \le \inf_{\mu} q(\mu) \le \lim_{n \to \infty} q(\mu_n) \le \overline{\nu}^*.$$

This completes the proof. ■

4.2. Sylvester's problem on the sphere

Moon [22] proved that the expected number of crossings in random drawings of K_n on the unit sphere in $\mathbb{R}^3$ is asymptotically the same as the conjectured crossing number of K_n. His result can be expressed as follows.

Theorem 4.2 (Moon [22]). *Let μ be the uniform probability distribution on the unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$. Then $q(\mu) = 3/8$.*

Guy, Jenkyns, and Schaer [12] considered the crossing number of K_n on the flat torus (obtained from the unit square by identifying opposite sides). Their computation shows the following.

Theorem 4.3 (Guy, Jenkyns, and Schaer [12]). *Let $\mathbb{T}$ be the flat torus obtained from a rectangle in the plane by identifying opposite sides. Let μ be the uniform probability distribution on $\mathbb{T}$. Then $q(\mu) = 5/18$.*

As noted in [12], the Sylvester crossing probability is the same for every rectangle model of the flat torus. However, they neglected the possibility of other parallelogram representations of the flat torus. Interestingly, they give smaller crossing probabilities.

Theorem 4.4 (Elkies [11]). *Let $\mathbb{T}_\alpha$ be the flat torus obtained from a rhombus with side length 1 and angle α $(0 < \alpha \le \pi/2)$ by identifying opposite sides. If μ_α is the uniform distribution on $\mathbb{T}_\alpha$, then*

$$q(\mu_\alpha) \ge \frac{22}{81}.$$

The smallest value occurs at $\alpha = \pi/3$, where $q(\mu_{\pi/3}) = \frac{22}{81}$.

In [15], Koman bounded the crossing number of K_n in the projective plane:

$$\frac{41}{273}\binom{n}{4} \le \mathrm{cr}_{\mathbb{N}_1}(K_n) \le \frac{39}{128}\binom{n-1}{4}, \tag{4.2}$$

where the left inequality holds only when $n \ge 11$.

Below we give an improvement of Koman's upper bound by using the model of the projective plane as the surface endowed with constant curvature 1 and considering random drawings.

Let $\mathbb{P}^2$ be the projective plane obtained from the unit sphere $\mathbb{S}^2$ by identifying all antipodal pairs of points. This defines the projective plane as a surface of constant curvature 1. Its total area is one half of the area of the unit sphere, $A(\mathbb{P}^2) = 2\pi$. The geodesics in $\mathbb{P}^2$ are the *great semicircles*, each of which has length equal to π.

Theorem 4.5. *The uniform distribution μ on $\mathbb{P}^2$ has crossing probability $q(\mu) = 3\pi^{-2}$. Consequently,*

$$\overline{\mathrm{cr}}_{\mathbb{P}^2}(K_n) \le 3\pi^{-2}\binom{n}{4}.$$

Proof. Let us consider two random points in $\mathbb{P}^2$ and let ℓ denote the length of the geodesic joining them. We claim that $\mathbb{E}(\ell) = 1$. To see this, we may assume that the first point is the North pole of $\mathbb{S}^2$. Then $\ell = \alpha$, where $0 \le \alpha \le \pi/2$ is the angle between the lines through the origin in $\mathbb{R}^3$ and the two points. Now,

$$\mathbb{E}(\ell) = \iint_S \alpha\, dS = \int_0^{2\pi}\int_0^{\pi/2} \alpha \sin\alpha\, d\alpha\, ds = 2\pi \int_0^{\pi/2} \alpha \sin\alpha\, d\alpha = 1.$$

Next, we consider the conditional probability of a crossing of two random segments S_1, S_2, conditioned on their lengths ℓ_1 and ℓ_2. The two great semicircles containing S_1 and S_2 cross each other at a point p. The segments are positioned randomly on these two segments, so the probability that they both contain p (which is the only way they would cross) is equal to $(\ell_1/\pi) \cdot (\ell_2/\pi)$. Thus the conditional probability of the event X that S_1 and S_2 cross is

$$\Pr[X \mid \ell_1, \ell_2] = \frac{\ell_1}{\pi} \cdot \frac{\ell_2}{\pi}.$$

Since ℓ_1 and ℓ_2 are independent and $\mathbb{E}(\ell_1) = \mathbb{E}(\ell_2) = 1$, we get

$$\mathbb{E}(X) = \pi^{-2}\,\mathbb{E}(\ell_1\ell_2) = \pi^{-2}\,\mathbb{E}(\ell_1)\,\mathbb{E}(\ell_2) = \pi^{-2}.$$

Finally, we have that $q(\mu) = 3\,\mathbb{E}(X) = 3\pi^{-2}$. ■

Elkies [11] realized that $3\pi^{-2} < 39/128$ and concluded that the bound of Theorem 4.5 asymptotically beats Koman's upper bound (4.2). In comparison with the Hill conjecture, it was conjectured in [11] that the bound of the theorem is best possible. However, Arroyo, McQuillan, Richter, Salazar, and Sullivan [6] recently found better drawings of complete graphs in the projective plane. Their drawings can also be approximated with random drawings, but the probability measure is not uniform.

Theorem 4.6 ([6]). $\overline{\mathrm{cr}}_{\mathbb{P}^2}(K_n) < 0.3024\binom{n}{4}$.

Note that $0.3024 < 3\pi^{-2} \approx 0.304$.

4.3. Antipodal drawings on the sphere

Let $\mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$. For any two points $p, q \in \mathbb{S}^2$, consider the great circle through p and q (the great circle is unique unless q is antipodal to p in which case there are many). The shorter of the two segments on this circle from p to q is called a *geodesic arc* (or just a *geodesic*). Any geodesic arc joining two antipodal points in $\mathbb{S}^2$ is a half of a great circle and will be referred to as a *half-circle*.

A *geodesic drawing* of a graph G on $\mathbb{S}^2$ is a drawing in which all edges are drawn as geodesic arcs. We define the *geodesic crossing number* of the graph G on the sphere as the minimum number of crossings of edges of G in a geodesic drawing of the graph, and denote it by $\mathrm{cr}_{\mathbb{S}^2}(G)$.

A set P of points in $\mathbb{S}^2$ is *in general position* if no three points in P lie on a common great circle in the sphere.

Let $k \geq 3$ be a positive integer and let $n = 2k$. The graph M_n obtained from the complete graph K_n by removing edges of a perfect matching in K_n is isomorphic to the complete k-partite graph $K_{2,2,\ldots,2}$ with k parts of size 2 each. The edge-set of this graph consists of $\binom{k}{2}$ 4-cycles, each of which joins two parts of size 2 and is called a *basic* 4*-cycle* in M_n.

We will consider some special drawings of M_n. Let P be a set of k points in general position in $\mathbb{S}^2$. Let $\mathcal{S}$ be obtained from P by adding, for each $p \in P$, its antipodal point $\bar{p}$ into $\mathcal{S}$. The geodesic drawing of M_n on these points, where each antipodal pair represents a pair of nonadjacent vertices in M_n, is said to be an *antipodal geodesic drawing of* M_n. We will denote by $D_n(P)$ the antipodal drawing of M_n determined by P.

Lemma 4.7 ([20]). *For every $k \geq 3$, every antipodal drawing $D_n(P)$ of M_n has precisely $\frac{1}{4}k(k-1)(k-2)(k-3)$ crossings, and by adding any geodesic half-circle between a pair of antipodal points $p, \bar{p}$ $(p \in P)$, we obtain precisely $\frac{1}{2}(k-1)(k-2)$ additional crossings.*

Proof. Note that every pair of points $p, q \in P$ together with their antipodes $\bar{p}$, $\bar{q}$ determines a great circle Q_{pq} that consists of four edges forming the basic 4-cycle between $\{p, \bar{p}\}$ and $\{q, \bar{q}\}$. Any two such great circles Q_{pq} and Q_{rs} cross twice and make two crossings if $\{p, q\} \cap \{r, s\} = \emptyset$. If $|\{p, q\} \cap \{r, s\}| = 1$, then they do not cross. Thus, the edges in each Q_{pq} participate in precisely $2\binom{k-2}{2} = (k-2)(k-3)$ crossings. By summing up these numbers over all $\binom{k}{2}$ possibilities for the pair $\{p, q\}$, we count each crossing twice, so

$$\mathrm{cr}(D_n) = \frac{1}{2}\binom{k}{2}(k-2)(k-3) = \frac{1}{4}k(k-1)(k-2)(k-3).$$

By adding any great circle through two antipodal points $p, \bar{p}$, $p \in P$, we separate $k-1$ of the points in $P \cup \overline{P}$ from their antipodal pairs. There are precisely $(k-1)(k-2)$ edges joining them. Because of the antipodal symmetry of the drawing D_n, precisely half of these edges cross each half-circle. Thus, each half-circle is crossed $\frac{1}{2}(k-1)(k-2)$ many times. ■

We say that a set P of points in $\mathbb{S}^2$ has *strength* s if there is a choice of half-circles joining each point in P with its antipodal point $\bar{p}$ such that these half-circles cross each other s times.

Corollary 4.8 ([20]). *If a set P of k points in general position on $\mathbb{S}^2$ has strength s, then the drawing $D_n(P)$ can be extended to a geodesic drawing of the complete graph K_n with $H(n) + s$ crossings.*

Proof. We extend the drawing D_n by adding half-circles joining the antipodal pairs $p, \bar{p}$ for $p \in P'$ so that these half-circles make s crossings among each other. By Lemma 4.7, the number of crossings is $\frac{1}{4}k(k-1)(k-2)(k-3) + \frac{1}{2}(k-1)(k-2)|P| + s$, which is equal to $H(n) + s$. ■

It is easy to see that there are many sets of strength 0. They give rise to antipodal Hill drawings.

Corollary 4.9. *Let $P \subset \mathbb{S}^2$ be a set of k points in general position in $\mathbb{S}^2$, whose strength is* 0. *Then the geodesic drawing $D_n(P)$ $(n = 2k)$ can be extended to a Hill drawing of K_n. This drawing has the following additional properties:*

(a) *the drawing is antipodally symmetric except for the drawing of the half-circles joining antipodal pairs;*

(b) *for every vertex v of K_n, the edges incident with v participate in precisely $\frac{1}{16}(n-2)^2(n-4)$ crossings;*

(c) *by deleting any point from $P \cup \overline{P}$, we obtain a drawing of K_{n-1} with precisely $H(n-1)$ crossings;*

(d) *by adding any new point (in general position with respect to P) and adding geodesics from that point to $P \cup \overline{P}$, we obtain a geodesic drawing of K_{n+1} with precisely $H(n+1)$ crossings.*

Proof. Statements (a)–(c) are easy observations and their proof is left for the reader. To prove (d), let $Q = P \cup \{q\}$, where q is the added point. Consider the corresponding drawing of K_{n+2} for $\widehat{Q}$. Note that Q may no longer have strength 0, but since P has strength 0, there is a drawing where the only half-circle intersecting other half-circles is the half-circle joining q and $\bar{q}$. All these added crossings disappear after removing $\bar{q}$, and thus by (b), the extended drawing of K_{n+1} has $H(n+2) - \frac{1}{16}n^2(n-2) = H(n+1)$ crossings. ■

The last corollary implies that $\mathrm{cr}_{\mathbb{S}^2}(K_n) \leq H(n)$ for every positive integer n. This result is surprising in two ways. Firstly, it is known that the rectilinear crossing number (geodesic version in the Euclidean plane) of complete graphs is strictly larger than the usual crossing number. So, assuming the Hill conjecture, it is surprising that the geodesic crossing number in the sphere is not different. Secondly, the abundance of obtained Hill drawings is also quite unexpected.

4.4. Moon's result revisited

A probability measure μ on the sphere $\mathbb{S}^2$ is *non-degenerate* if $\mu(C) = 0$ for each great circle C. This is equivalent to saying that the probability that n μ-random points on the sphere lie in general position is equal to 1 (with probability 1, they are all distinct and no three are on the same great circle). Further, we say that μ is *antipodally symmetric* if for any μ-measurable set $A \subseteq \mathbb{S}^2$, its antipodal set $\overline{A}$ has the same measure, $\mu(\overline{A}) = \mu(A)$.

As mentioned before (see Theorem 4.2), Moon proved that random geodesic drawings of complete graphs on the sphere have asymptotically about the same number of crossings as the conjectured best drawings. Corollary 4.8 gives a simple explanation of this phenomenon. Indeed, in a forthcoming work [21] the following result with several interesting consequences is derived.

Theorem 4.10 (Mohar and Wesolek [21]). *Let μ be a non-degenerate antipodally symmetric probability distribution on the unit sphere $\mathbb{S}^2$. Then a μ-random set of n points on $\mathbb{S}^2$ joined by geodesics gives rise to a drawing D_n of the complete graph K_n such that* $\operatorname{cr}(D_n)/H(n) = 1 + o(1)$ *asymptotically almost surely.*

References

[1] B. M. Ábrego, O. Aichholzer, S. Fernández-Merchant, P. Ramos, and G. Salazar, The 2-page crossing number of K_n. *Discrete Comput. Geom.* **49** (2013), no. 4, 747–777 Zbl 1269.05078 MR 3068573

[2] B. M. Ábrego, O. Aichholzer, S. Fernández-Merchant, P. Ramos, and G. Salazar, Shellable drawings and the cylindrical crossing number of K_n. *Discrete Comput. Geom.* **52** (2014), no. 4, 743–753 Zbl 1306.05166 MR 3279547

[3] B. M. Ábrego, O. Aichholzer, S. Fernández-Merchant, P. Ramos, and B. Vogtenhuber, Non-shellable drawings of K_n with few crossings. In *Proc. of the 26th Canadian Conference on Computational Geometry (CCCG 2014)*, Halifax, Nova Scotia, Canada, 2014

[4] B. M. Ábrego, M. Cetina, S. Fernández-Merchant, J. Leaños, and G. Salazar, 3-symmetric and 3-decomposable geometric drawings of K_n. *Discrete Appl. Math.* **158** (2010), no. 12, 1240–1458 Zbl 1228.05214 MR 2652001

[5] B. M. Ábrego, M. Cetina, S. Fernández-Merchant, J. Leaños, and G. Salazar, On $\leq k$-edges, crossings, and halving lines of geometric drawings of K_n. *Discrete Comput. Geom.* **48** (2012), no. 1, 192–215 Zbl 1247.52010 MR 2917207

[6] A. Arroyo, D. McQuillan, R. B. Richter, G. Salazar, and M. Sullivan, Drawings of complete graphs in the projective plane. *J. Graph Theory* **97** (2021), no. 3, 426–440 MR 4313189

[7] J. Balogh, B. Lidický, and G. Salazar, Closing in on Hill's conjecture. *SIAM J. Discrete Math.* **33** (2019), no. 3, 1261–1276 Zbl 1419.05050 MR 3982073

[8] J. Blažek and M. Koman, A minimal problem concerning complete plane graphs. In *Theory of Graphs and its Applications (Proc. Sympos. Smolenice, 1963)*, pp. 113–117, Publ. House Czech. Acad. Sci., Prague, 1964 Zbl 0161.20601 MR 0174042

[9] H. T. Croft, K. J. Falconer, and R. K. Guy, *Unsolved Problems in Geometry. Unsolved Problems in Intuitive Mathematics, II*. Probl. Books in Math., Springer, New York, 1991 Zbl 0748.52001 MR 1107516

[10] R. Deltheil, *Probabilités géométriques*. Traité du calcul des probabilités et de ses applications, Gauthier-Villars, Paris, 1926

[11] N. D. Elkies, Crossing numbers of complete graphs. In *The Mathematics of Various Entertaining Subjects. Vol. 2*, pp. 218–249, Princeton Univ. Press, Princeton, NJ, 2017 MR 3701434

[12] R. K. Guy, T. Jenkyns, and J. Schaer, The toroidal crossing number of the complete graph. *J. Combinatorial Theory* **4** (1968), 376–390 Zbl 0172.48804 MR 220630

[13] F. Harary and A. Hill, On the number of crossings in a complete graph. *Proc. Edinburgh Math. Soc. (2)* **13** (1962/63), 333–338 Zbl 0118.18902 MR 163299

[14] M. G. Kendall and P. A. P. Moran, *Geometrical Probability*. Griffin's Statistical Monographs & Courses 10, Hafner Publishing, New York, 1963 Zbl 0105.35002 MR 0174068

[15] M. Koman, On the crossing numbers of graphs. *Acta Univ. Carolin. Math. Phys.* **10** (1969), no. 1–2, 9–46 Zbl 0256.05103 MR 288049

[16] J. Kyncl, Drawings of complete graphs with $Z(n)$ crossings. https://mathoverflow.net/questions/128878/drawings-of-complete-graphs-with-zn-crossings. Accessed 2020-9-3

[17] F. T. Leighton, New lower bound techniques for VLSI. *Math. Systems Theory* **17** (1984), no. 1, 47–70 Zbl 0488.94048 MR 738751

[18] L. Lovász, *Large Networks and Graph Limits*. Amer. Math. Soc. Colloq. Publ. 60, Amer. Math. Soc., Providence, RI, 2012 Zbl 1292.05001 MR 3012035

[19] L. Lovász, K. Vesztergombi, U. Wagner, and E. Welzl, Convex quadrilaterals and k-sets. In *Towards a Theory of Geometric Graphs*, pp. 139–148, Contemp. Math. 342, Amer. Math. Soc., Providence, RI, 2004 Zbl 1071.05028 MR 2065260

[20] B. Mohar, On a conjecture by Anthony Hill. 2020, arXiv:2009.03418

[21] B. Mohar and A. Wesolek, Random geodesic drawings on the sphere. In preparation

[22] J. W. Moon, On the distribution of crossings in random complete graphs. *J. Soc. Indust. Appl. Math.* **13** (1965), 506–510 Zbl 0132.40305 MR 179106

[23] S. Pan and R. B. Richter, The crossing number of K_{11} is 100. *J. Graph Theory* **56** (2007), no. 2, 128–134 Zbl 1128.05018 MR 2350621

[24] R. E. Pfiefer, The historical development of J. J. Sylvester's four point problem. *Math. Mag.* **62** (1989), no. 5, 309–317 Zbl 0705.52005 MR 1031429

[25] A. A. Razborov, Flag algebras. *J. Symbolic Logic* **72** (2007), no. 4, 1239–1282 Zbl 1146.03013 MR 2371204

[26] M. Schaefer, The graph crossing number and its variants: a survey. *Electron. J. Combin.* **DS21** (2013), Dynamic Surveys, 90 pp. Zbl 1267.05180 MR 4336223

[27] M. Schaefer, *Crossing Numbers of Graphs*. Discrete Math. Appl. (Boca Raton), CRC Press, Boca Raton, FL, 2018 Zbl 1388.05005 MR 3751397

[28] E. R. Scheinerman and H. S. Wilf, The rectilinear crossing number of a complete graph and Sylvester's "four point problem" of geometric probability. *Amer. Math. Monthly* **101** (1994), no. 10, 939–943 Zbl 0834.05022 MR 1304316

[29] R. Schwartz, J. Solymosi, and F. de Zeeuw, Simultaneous arithmetic progressions on algebraic curves. *Int. J. Number Theory* **7** (2011), no. 4, 921–931 Zbl 1231.11120 MR 2812643

[30] J. J. Sylvester, Question 1491. *The Educational Times (London)* (1864)

[31] J. J. Sylvester, On a special class of questions on the theory of probabilities. *Birmingham British Assoc. Rept.* (1865), 8–9

[32] L. A. Székely, Crossing numbers and hard Erdős problems in discrete geometry. *Combin. Probab. Comput.* **6** (1997), no. 3, 353–358 Zbl 0882.52007 MR 1464571

[33] P. Turán, A note of welcome. *J. Graph Theory* **1** (1977), 7–9

[34] K. Urbaník, Solution du problème posé par P. Turán. *Colloq. Math.* **3** (1955), 200–201

[35] K. Zarankiewicz, On a problem of P. Turan concerning graphs. *Fund. Math.* **41** (1954), 137–145 Zbl 0055.41605 MR 63641

Bojan Mohar
Department of Mathematics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada;
mohar@sfu.ca

DOI 10.4171/8ECM/28

European mathematics: A history in stamps

Robin Wilson

Abstract. It is surprising how many hundreds of postage stamps from around the world have featured mathematics and its history. For the 8ECM meeting in Portorož I was invited to present a public lecture on those stamps that are related to European mathematics, and I illustrated it with more than 200 examples. I also designed a stamp exhibition for the Congress, also including over 200 mathematical stamps with historical commentary. In this article I present a selection of the stamps that I selected for this lecture and exhibition. The treatment is historical and is presented chronologically.

1. Greek mathematics

European mathematics is often taken to begin with Ancient Greece, although its origins can be traced back further. From around 600 BC, the subject flourished throughout the eastern Mediterranean where the Greeks developed deductive logical reasoning and proof – the hallmark of their work, especially in geometry.

An early Greek mathematician, around 600 BC, was *Thales of Miletus* (Figure 1(a)), who predicted a solar eclipse and showed how to cause electricity in feathers by rubbing them with a stone. In geometry he reportedly proved that a circle is bisected by any diameter and that the base angles of an isosceles triangle are equal.

Another semi-legendary figure is *Pythagoras of Samos* (Figure 1(b)), who formed a School in Crotona to further the study of mathematics and science. Supposedly believing that "All is number", the Pythagoreans emphasised the "mathematical arts" of arithmetic, geometry, astronomy, and music, later known as the "quadrivium". Several stamps feature the well-known *Pythagorean theorem* for a right-angled triangle (Figure 1(c)), that the areas of the squares on its two smallest sides add up to the area of the square on its largest side – or in algebraic form (which the Greeks did not use), $a^2 + b^2 = c^2$. We do not know who first proved this, but its connection with right-angled triangles had already been known many years earlier, in Mesopotamia and elsewhere.

2020 Mathematics Subject Classification. 97A30.
Keywords. European mathematics, history of mathematics, postage stamps.

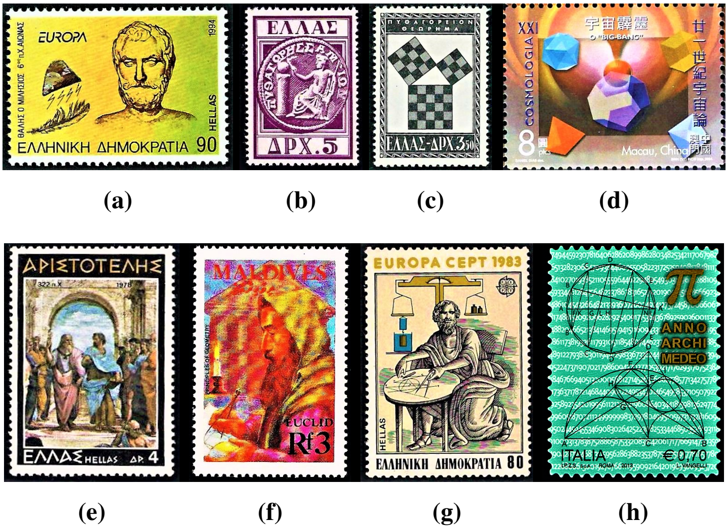


Figure 1. *Greek mathematics*
(a) Thales, (b) Pythagoras, (c) Pythagorean theorem, (d) Platonic solids,
(e) Plato’s Academy, (f) Euclid, (g) Archimedes, (h) Archimedes’ geometry

The scene then moved to Athens, which became the most important intellectual centre in Greece. In 387 BC, the philosopher Plato founded a school in the suburb of Athens named “Academy”, and Plato’s Academy became the focal point for mathematical study. Convinced that mathematical training was essential for his ideal citizens, Plato emphasised the quadrivium subjects of the Pythagoreans and discussed the five regular (or “Platonic”) solids (Figure 1(d)), while his pupil Aristotle formalised deductive reasoning. In Raphael’s fresco *The School of Athens*, Plato and Aristotle are shown on the steps of the Academy (Figure 1(e)).

Around 300 BC, following the military successes of Alexander the Great, mathematical activity moved to Alexandria in the Egyptian part of the Greek world. The first important mathematician there was *Euclid* (Figure 1(f)) who is mainly remembered for his *Elements*, the most widely read and influential mathematical work of all time. A model of deductive reasoning, it presented plane and solid geometry, arithmetic, and number theory, by building them up from a small number of axioms to a great hierarchy of results that he derived in a logical and systematic order.

One of the greatest of mathematicians, around 250 BC, was *Archimedes* of Syracuse, now in Sicily (Figure 1(g)). In geometry he investigated spheres and cylinders and compared the surface areas and volumes of sections of these; he also listed the thirteen “Archimedean” (or semi-regular) solids, and found estimates for π by considering polygons that approximate a circle (Figure 1(h)). In mechanics he found the law of moments for a balance and invented the Archimedean screw for raising water. In statics he stated Archimedes' principle on the weight of an object immersed in water, but no contemporary evidence exists for the well-known story that he used his principle to test the purity of a gold crown or that he jumped out of his bath and ran naked through the streets celebrating his discovery.

2. Early European mathematics

We now turn briefly to the Islamic world from AD 750 onwards. United by their new religion, and with Baghdad lying on the east-west trade routes, their scholars developed Greek writings from the west and Hindu writings from India. Some of our present terminology dates from this period: the word "algorithm" (a step-by-step procedure for solving a problem) comes from al-Khwārizmī, a Persian mathematician whose influential book on arithmetic introduced the Indian decimal place-value system to the Islamic world. He also wrote a book on solving equations, *Kitāb al-jabr wal-muqābala* (Calculation by Completion and Balancing), whose title gives us the word "algebra".

The Islamic world developed in all directions, and by the year 1000 it had spread across the top of Africa and up into southern Europe through Spain and Italy. Córdoba became the scientific capital of Europe, while Islamic decorative art and architecture spread through southern Spain and Portugal (Figure 2(a)) and included the magnificent geometrical arches in the mosque at Córdoba, and the tilings in Granada's Alhambra.

Meanwhile, in Europe, the period from 500 to 1000 had become known as the "Dark Ages". Much of the legacy from the ancient world was forgotten, and the general level of culture was low. Revival of interest began with the French scholar *Gerbert of Aurillac* (Figure 2(b)) who trained in Catalonia and introduced the Hindu–Arabic numerals to Christian Europe, using an abacus that he designed for the purpose. A major figure in the Church, he was crowned Pope in 999.

The Hindu-Arabic numbers were also popularised by Leonardo of Pisa, or *Fibonacci* (Figure 2(c)) in his *Liber Abbaci* (Book of Calculation) of 1202. This famous work contains many problems from arithmetic and algebra, such as his *rabbits problem* (Figure 2(d)) that leads to the *Fibonacci sequence* of numbers, $1, 2, 3, 5, 8, 13, \ldots,$ where each successive number is the sum of the previous two. These numbers also arise in the arrangements of seeds in sunflowers and pine cones.

Another notable figure was the Catalan mystic *Ramon Lhull* (Figure 2(e)), who believed that *all* knowledge could be obtained by combining God's "divine attributes", such as power, wisdom, and goodness. His combinatorial ideas spread through Europe, later influencing such figures as Mersenne and Leibniz.

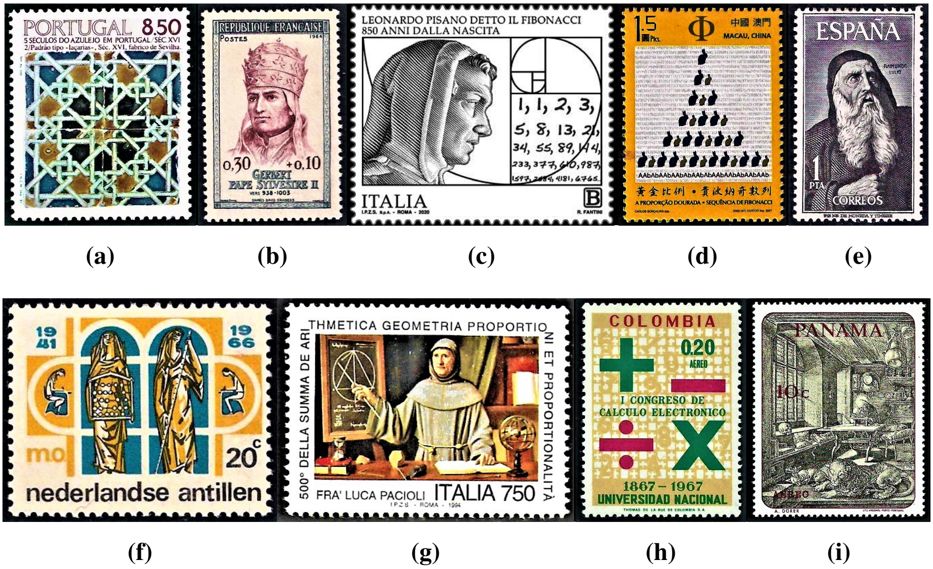

PORTUGAL 8,50
REPUBLIQUE FRANÇAISE
GERBERT PAPE SYLVESTRE II
LEONARDO PISANO DETTO IL FIBONACCI 850 ANNI DALLA NASCITA
1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144,
ITALIA
Φ
MACAU, CHINA
ESPAÑA
CORREOS
(a)
(b)
(c)
(d)
(e)
nederlandse antillen
20c
500° DELLA SUMMA DE ARITHMETICA GEOMETRIA PROPORTIONI ET PROPORTIONALITÀ
FRA' LUCA PACIOLI ITALIA 750
COLOMBIA 0.20
I CONGRESO DE CALCULO ELECTRONICO
1867 – 1967 UNIVERSIDAD NACIONAL
PANAMA
10c
(f)
(g)
(h)
(i)

Figure 2. *Early European mathematics*
(a) Tiling pattern, (b) Gerbert, (c) Fibonacci, (d) rabbits problem, (e) Llull,
(f) arithmetic and geometry, (g) Pacioli, (h) arithmetic symbols, (i) Dürer engraving

The Middle Ages renaissance in learning was largely due to three developments: the establishment of universities, the translation of Arabic texts into Latin, and the invention of printing. The first European university was in Bologna, founded in 1088, with Paris and Oxford following soon after. For hundreds of years, the curriculum was based on the Greek *quadrivium* (Figure 2(f)).

The invention of printing around 1440 enabled mathematical works to become widely available for the first time. At first, these were in Latin for the scholar, but gradually vernacular works appeared at prices accessible to all. These included texts in arithmetic, algebra, and geometry, and practical works on the mathematics of commerce. Important among these vernacular works was *Luca Pacioli*'s *Summa* in Italian (Figure 2(g)), a 600-page compilation of contemporary mathematics that included the first account of double-entry bookkeeping. Printing also led to a standardisation of *mathematical notation*: the symbols + and − first appeared in a German arithmetic text of 1489, but × and ÷ were not used until much later (Figure 2(h)).

It was around this time that painters learnt how to give visual depth through geometrical perspective. Two of these were Brunelleschi who designed the dome of Florence Cathedral, and his friend Alberti who presented mathematical rules for perspective and insisted that "the first duty of a painter is to know geometry". Piero della Francesca wrote books on perspective that included polyhedron woodcuts by

Leonardo da Vinci who warned: “Let no one who is not a mathematician read my work”. Another famous artist was Albrecht Dürer, who learnt perspective in Italy and introduced it into Germany. His engraving *St Jerome in His Study* shows his use of perspective (Figure 2(i)).

3. The age of exploration

The Renaissance period also coincided with many great sea voyages and explorations. In Portugal, Prince Henry the Navigator devoted his wealth and energies to maritime exploration, while Vasco da Gama became the first European to reach the west coast of India. Other well-known explorers included the Italian Christopher Columbus and Ferdinand Magellan of Portugal.

Such explorers needed accurate maps, and attempts to represent the spherical earth on a flat surface led to various types of map projection for use by navigators. Most notable was that of *Gerard Mercator* (Figure 3(a)) who projected the globe onto a vertical cylinder and adjusted the scale so that the lines of latitude and longitude appeared straight, as did fixed compass directions. Another early European to apply mathematical techniques to cartography was *Pedro Nunes* (Figure 3(b)), royal cosmographer and the leading figure in Portuguese nautical science.

For navigating at sea, *astrolabes* were used to determine latitude by measurements of the altitudes of heavenly bodies such as the sun or pole star (Figure 3(c)); other instruments included quadrants (in the shape of a quarter-circle, or 90°) and *sextants* (a sixth of a circle, or 60°) (Figure 3(d)). To measure an object’s altitude, you viewed it along the top edge of the instrument, and the position of a movable rod on the rim gave the reading.

The 16th century was also important for astronomy, which was completely transformed when *Nicolaus Copernicus* replaced the Greek earth-centred planetary system by one with the sun at the centre and the earth as just one of the planets in circular orbits around it; his book *De Revolutionibus Orbium Coelestium* (On the Revolutions of the Heavenly Spheres) was published in 1543 (Figure 3(e)). The Copernican system aroused much controversy, bringing its supporters into conflict with the church which placed the earth at the centre of creation.

Before the invention of the telescope, the greatest observer of the heavens was the Danish astronomer Tycho Brahe, who designed instruments of unequalled accuracy and measured over 700 stars. His assistant *Johannes Kepler* (Figure 3(f)) is remembered for his laws of planetary motion. From Tycho’s extensive observations, he proposed *elliptical* orbits for the planets, with the sun at one focus, and introduced the word “focus” into mathematics. Kepler also rotated curves around an axis and found the volumes of many solids of revolution by summing thin discs, foreshadowing the integral calculus of some years later.

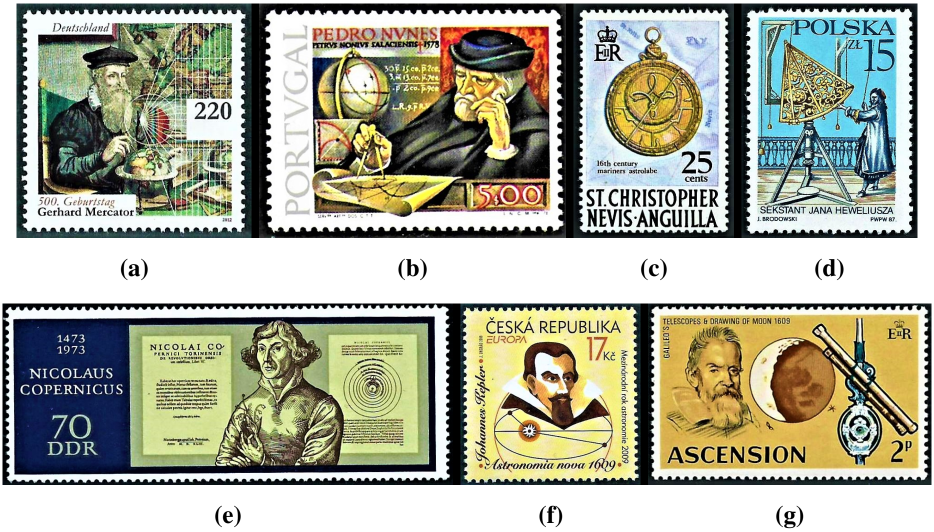


Figure 3. *The age of exploration*
(a) Mercator, (b) Nunes, (c) mariner's astrolabe, (d) sextant,
(e) Copernicus, (f) Kepler, (g) Galileo

Another Copernican supporter was *Galileo Galilei* (Figure 3(g)), who made extensive use of the telescope, drawing our moon's surface and discovering the moons of Jupiter and Saturn. His mechanics book *Discorsi e Dimostrazioni Matematiche Intorno a Due Nuove Scienze* (Discourses and Mathematical Demonstrations Relating to Two New Sciences) investigated uniform and accelerated motion and explained why the path of a projectile must be a parabola.

4. The 17th century

A major difficulty of the time, particularly for navigators and astronomers, was numerical calculation. In 1614 John Napier of Scotland introduced his "logarithms", designed to replace lengthy multiplications and divisions by easier additions and subtractions. These soon led to practical instruments based on a logarithmic scale, such as the *slide rule* (Figure 4(a)); dating from around 1630, they were used for over 300 years until pocket calculators appeared in the 1970s. The Slovenian mathematician *Jurij Vega* also published a celebrated compendium of logarithms, as well as 7-figure and 10-figure tables that ran to many editions (Figure 4(b)), and calculated π to 140 decimal places.

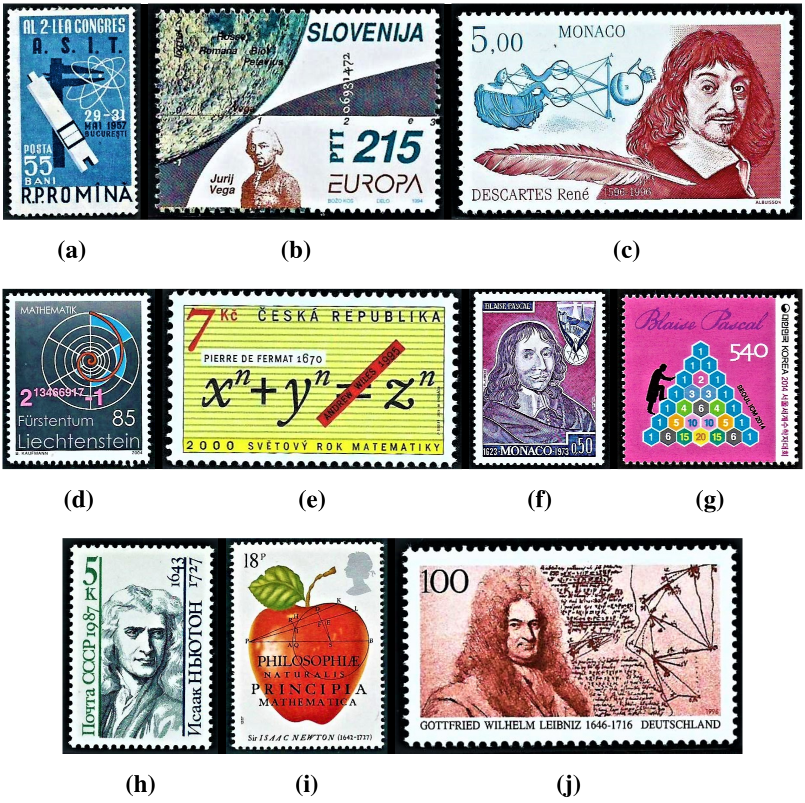


Figure 4. *The 17th century*
(a) Slide rule, (b) Vega's logarithms, (c) Descartes,
(d) Mersenne prime, (e) Fermat's last theorem, (f) Pascal, (g) Pascal's triangle,
(h) Newton, (i) Newton's *Principia*, (j) Leibniz

Meanwhile, in France, *René Descartes* (Figure 4(c)) solved an ancient problem of Pappus who had asked for the locus of a point that moved in a specified way relative to certain fixed lines. To solve this, Descartes named two lengths x and y and calculated every other length in terms of them, obtaining a quadratic expression (a conic) as the required path. In this way he introduced algebraic methods into geometry (a development that would continue over the next 100 years), but not the "Cartesian coordinates" that are named after him.

Marin Mersenne was a minimite friar living just outside Paris, who made great advances in the mathematical theory of sound and who is mainly remembered for listing prime numbers of the form $2^n - 1$, such as 3, 7, and 31. Fifty-one of these *Mersenne primes* are now known, and Figure 4(d) exhibits the largest Mersenne prime that had been discovered up to 2004.

Pierre de Fermat is mainly remembered for analytic geometry and number theory. In particular, he famously claimed to have proved *Fermat's last theorem*, that the equation $x^n + y^n = z^n$ has no non-zero integer solutions when $n > 2$ (Figure 4(e)). This was eventually proved by *Andrew Wiles* in 1995, as indicated on the stamp by the bar across the equals sign.

Blaise Pascal (Figure 4(f)) showed an early interest in mathematics – when only 16 he discovered his "hexagon theorem" about six points on a conic. One of the earliest to explore the theory of probability, he is also remembered for "Pascal's principle" in hydrodynamics, *Pascal's triangle* of binomial coefficients (Figure 4(g)), and for an early calculating machine, operated by cogged wheels, that could add and subtract.

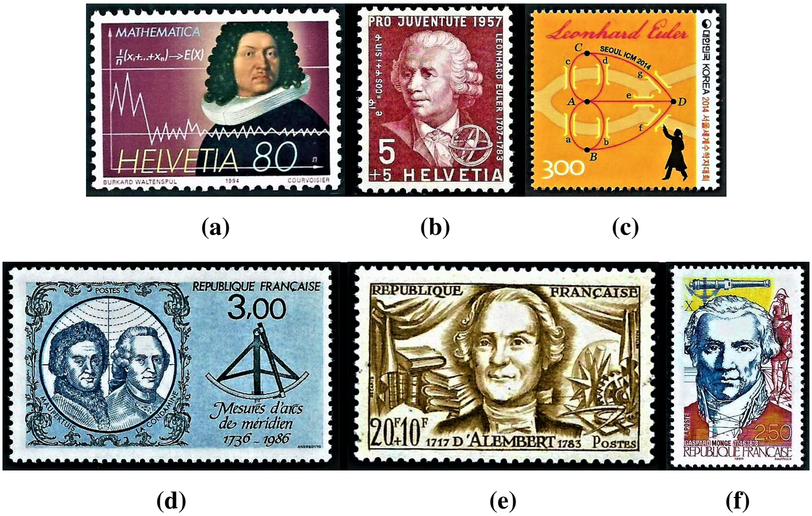
In England, *Isaac Newton* (Figure 4(h)) was born in 1642, and at Cambridge University he was appointed Lucasian Professor of Mathematics, a post later held by Stephen Hawking. Together with Leibniz (but independently) he recognised the inverse relationship between differentiation and integration, the two branches of the calculus.

The story of Newton and the apple is well known. Seeing it fall, he realised that the gravitational force that pulled it to earth was the same as the force that keeps the moon orbiting the earth and the earth orbiting the sun – and claimed that this motion was governed by a "universal law of gravitation", where the force of attraction between two objects varies inversely as the square of the distance between them. In his *Philosophiae Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy) of 1687, Newton used this law to explain Kepler's laws of elliptical planetary motion, and to account for cometary orbits, the variation of tides, and much else besides (Figure 4(i)).

Newton justly claimed priority for the calculus, but it was *Gottfried Leibniz* who was the first to publish it (Figure 4(j)). But his calculus was different from Newton's, being based on geometry rather than on velocity and motion. Also, his notation was more versatile than Newton's: his "D" for differentiation and his integral sign, which are still used today, were introduced within just three weeks of each other in the autumn of 1675.

5. The 18th century

The Bernoulli family included several distinguished Swiss mathematicians. In his book of 1713 on the "Art of Conjecturing", *Jakob Bernoulli* presented his *law of large numbers* (Figure 5(a)). With his brother Johann, he was the first to develop Leibniz's calculus, introducing the word "integral" and applying calculus to such curves as cycloids and spirals.



(a) (b) (c)

(d) (e) (f)

Figure 5. *The 18th century*
(a) Bernoulli, (b) Euler, (c) Königsberg bridges problem,
(d) geodetic missions, (e) d'Alembert, (f) Monge

Leonhard Euler also grew up in Switzerland, but spent his working life at the scientific academies of St. Petersburg and Berlin. The most prolific mathematician of all time, he contributed to almost every branch of mathematics and physics, from number theory and the calculus to mechanics, astronomy, and optics. Euler introduced the notations e for exponential, f for a function, i for $\sqrt{-1}$, and Σ for summation, and linked the exponential and trigonometric functions via his equation $e^{i\varphi} = \cos\varphi + i \sin\varphi$ as shown on the stamp (Figure 5(b)). In 1735 he solved the *Königsberg bridges problem* of deciding whether one can cross the seven bridges of the city visiting no bridge twice, but he never drew the associated graph that is often attributed to him (Figure 5(c)).

Newton had predicted that the earth's rotation causes a flattening at the poles, whereas an alternative theory of Descartes claimed that it is elongated. In the 1730s *geodetic missions* went to Peru (led by Charles-Marie de la Condamine) and Lapland (led by Pierre Louis de Maupertuis) to measure the swing of a pendulum and ascertain who was correct (Figure 5(d)). These missions confirmed Newton's view: the earth is flattened at the poles.

In France a leading Enlightenment figure was Jean d'Alembert (Figure 5(e)), who attempted to put the calculus on a firm basis by formalising the idea of a limit. He also derived the wave equation that describes the motion of a vibrating string, and in later years wrote many mathematical and scientific articles for Denis Diderot's *Encyclopédie* (Encyclopedia).

Napoleon Bonaparte's rise to power in France led to important developments in mathematics. Napoleon himself was interested in the subject – there's even a "Napoleon's theorem" – and his close friend, the geometer *Gaspard Monge* (Figure 5(f)), while investigating fortress gun emplacements, developed improved methods for projecting 3-dimensional objects onto a plane; this became known as "descriptive geometry".

6. The 19th century

An important consequence of the French Revolution was the founding in Paris of the École Polytechnique, where the finest mathematicians of the day – Monge, Lagrange, Laplace and Cauchy – taught students who were destined to serve in both military and civilian capacities.

Joseph-Louis Lagrange (Figure 6(a)) wrote on mechanics, functions, and number theory; and proved that every positive integer can be written as the sum of four squares. *Pierre-Simon Laplace* (Figure 6(b)) is remembered for the Laplace transform of a function and Laplace's equation in physics and wrote a monumental five-volume treatise on celestial mechanics that earned him the title of "the Newton of France". Shortly after the French Revolution, a commission was set up to standardise weights and measures and introduce a metric system; led by Lagrange, its members included Laplace and Monge.

Work in analysis continued with *Augustin-Louis Cauchy* (Figure 6(c)). The calculus was still on shaky foundations, but Cauchy rescued it with formal treatments of limits and continuity, while also developing complex analysis. Meanwhile, in Prague, *Bernard Bolzano* (Figure 6(d)) had formalised the idea of continuity before Cauchy, proving the "intermediate value theorem" that a continuous function takes every value between its greatest and least values.

In algebra a major breakthrough in 1826 occurred when the Norwegian *Niels Abel* (Figure 6(e)) solved a long-standing problem. Although there were general formulas for solving polynomial equations of degrees 2, 3, and 4, none was known for those of higher degrees. Abel showed that no such formulas can exist. Abel's work was continued by *Évariste Galois* (Figure 6(f)), who explained in algebraic terms exactly *which* equations can be solved. Galois had a short and turbulent life, being sent to jail for political activities and dying tragically in a duel at the age of 20, having sat up the previous night summarising all his mathematical achievements for posterity.

The Irishman William Rowan Hamilton was a child prodigy who discovered an error in Laplace's writings while a teenager and was appointed Astronomer Royal of Ireland when he was still a student. He made important advances in mechanics and geometrical optics, and while attempting to generalise the complex numbers

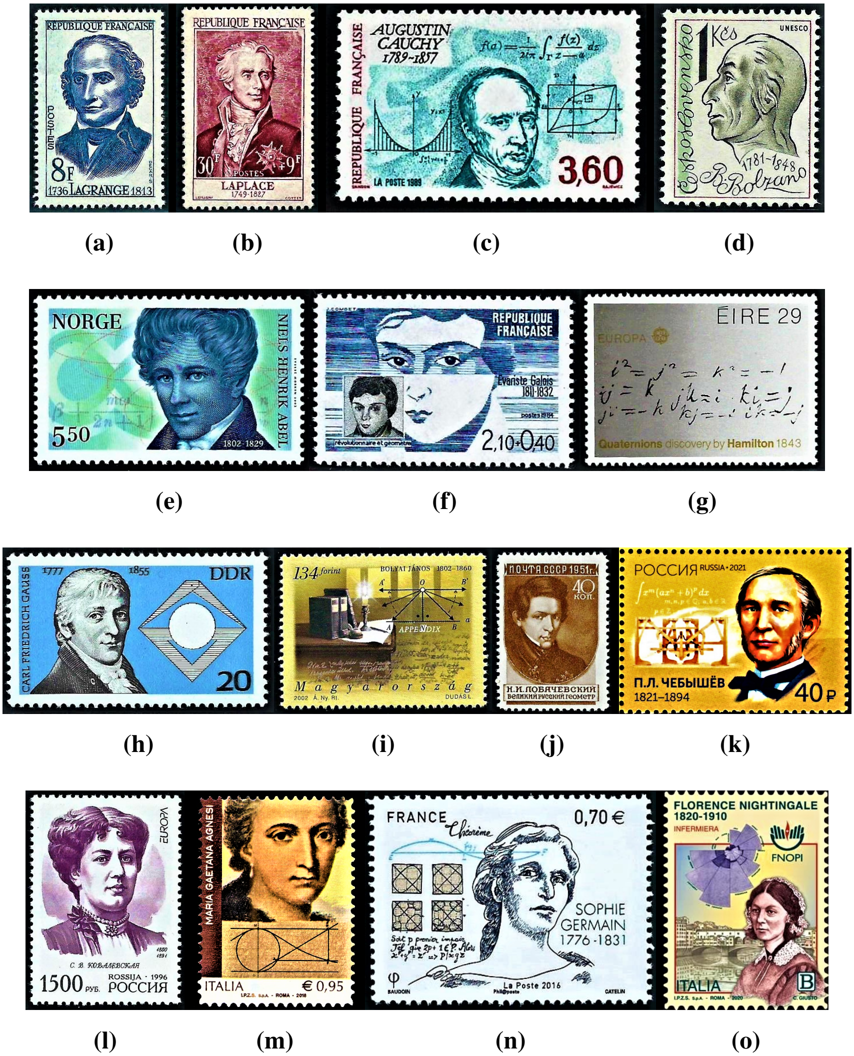


Figure 6. *The 19th century*
(a) Lagrange, (b) Laplace, (c) Cauchy, (d) Bolzano,
(e) Abel, (f) Galois, (g) Hamilton's quaternions,
(h) Gauss and polygon, (i) Bolyai's geometry, (j) Lobachevsky, (k) Chebyshev,
(l) Kovalevskaya, (m) Agnesi, (n) Germain, (o) Nightingale

discovered the *quaternions*, a non-commutative system that involves *three* interconnected square roots of -1 (i, j, and k), as shown in Figure 6(g).

Meanwhile, in Germany, *Carl Friedrich Gauss* worked in many areas, from complex numbers (the "Gaussian number plane") to statistics (the "Gaussian distribution"). One of the greatest mathematicians of all time, he also discovered which regular polygons can be drawn by straight-edge and compasses alone – these include triangles and pentagons, and also a regular polygon with 17 sides (Figure 6(h)).

In the early 19th century, there were important developments in geometry. Euclid's *Elements* opens with five "postulates" – four are straightforward, but the fifth is more complicated and seemed to be provable from the others. One version of it was the "parallel postulate": "given a line L and any point O not on it, there is a *unique* line through O, parallel to L". For two millennia, mathematicians tried to deduce this from the other postulates, but they were unsuccessful because there are geometries that satisfy the first four postulates but not the fifth: these have *infinitely many* lines through O parallel to L (Figure 6(i)). Described around 1830 by *Nikolai Lobachevsky* of Russia (Figure 6(j)) and János Bolyai of Hungary, they forced mathematicians to ask: "Which geometry corresponds to the world we live in?" – our familiar Euclidean geometry or a non-Euclidean one? (The familiar spherical geometry of our globe is not a true geometry in the Euclidean sense, as two lines, or great circles, meet in more than one point.)

In Russia, *Pafnuty Chebyshev* (Figure 6(k)) investigated orthogonal functions ("Chebyshev polynomials"), probability ("Chebyshev's inequality"), and prime numbers. *Sofia Kovalevskaya* (Figure 6(l)) contributed to mathematical analysis and partial differential equations and won a coveted prize from the French Academy of Sciences for a memoir on the rotation of bodies; barred by her gender from studying in Russia, she later became the first female professor in Stockholm.

Other women mathematicians who have featured on stamps include *Maria Gaetana Agnesi* (Figure 6(m)), who published an early book on the calculus and after whom the cubic curve known as the "witch of Agnesi" is named, and *Sophie Germain* (Figure 6(n)), whose pioneering work on prime numbers and Fermat's last theorem greatly impressed Gauss; she also made important contributions to the theory of elasticity. *Florence Nightingale* saved many lives through her sanitary improvements in Crimean War hospitals; an accomplished statistician, she analysed Crimean mortality data and displayed them using her "polar diagrams", as depicted in Figure 6(o).

7. The 20th century

It was in the 20th century that mathematicians created the subject as we now know it. What follows is a brief selection.

Henri Poincaré (Figure 7(a)) worked on the still-unsolved "three-body problem" of determining the simultaneous motion of the sun, earth, and moon. A gifted populariser of mathematics, he also developed algebraic topology, differential equations, celestial mechanics, and much else. The range of *David Hilbert* was also immense – from number theory, "Hilbert space", and a space-filling curve (Figure 7(b)) to potential theory and the theory of gases. In 1900 he gave a celebrated lecture at the International Congress of Mathematicians in Paris, posing 23 mathematical problems that set the agenda for research over the coming century.

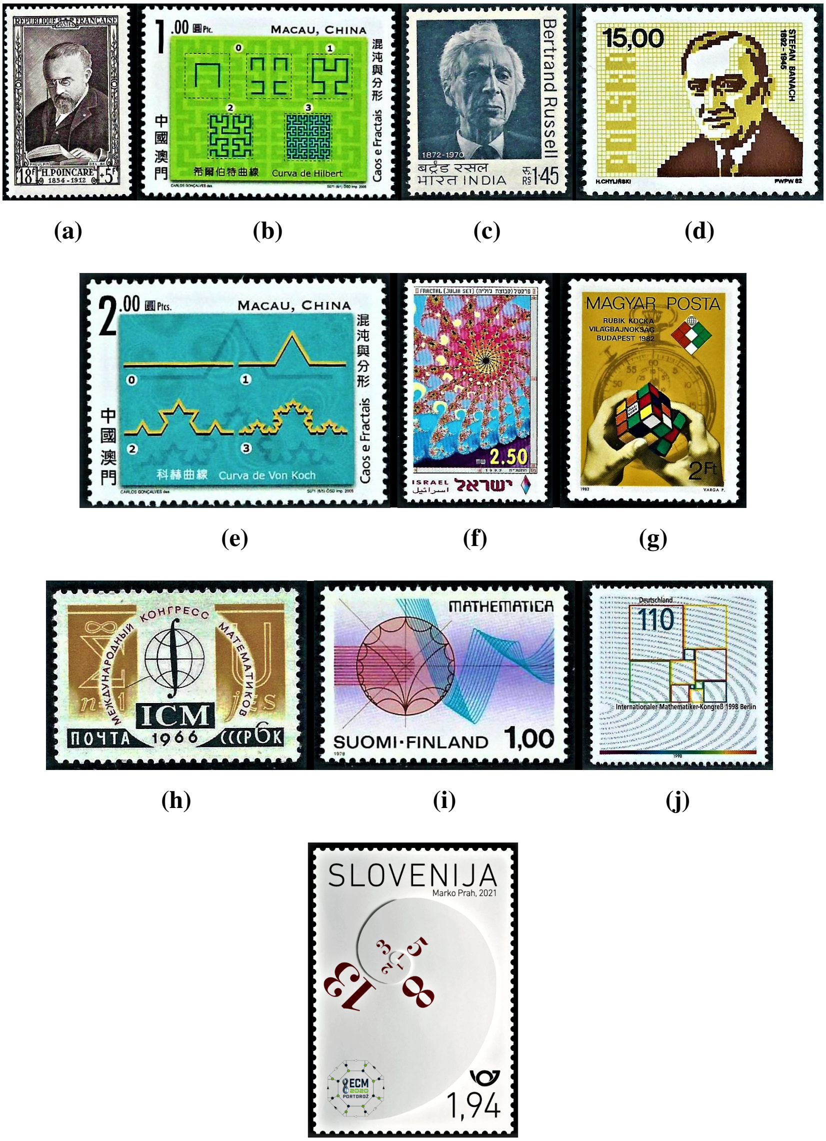
In England, *Bertrand Russell* (Figure 7(c)) made fundamental contributions to mathematical logic, such as "Russell's paradox", and with A. N. Whitehead wrote a three-volume *Principia Mathematica* on the foundations of mathematics, while in Poland *Stefan Banach* (Figure 7(d)) helped to create modern functional analysis and develop links between topology and algebra: the term *Banach space* is named after him.

Fractal patterns are "self-similar", in that they reproduce themselves for ever when magnified or reduced, such as von Koch's snowflake curve (Figure 7(e)), which has infinite length but encloses a finite area. Figure 7(f) features a *Julia set*, a fractal pattern that arises from iterating a quadratic formula.

For something more light-hearted, Figure 7(g) shows *Rubik's cube*, whose faces can be rotated to yield over 10^{19} different patterns; the object is to restore the original colours. In the early 1980s, when the craze was at its peak, over 100 million cubes were sold.

Mathematics continues to advance at an ever-increasing rate, and since 1897 the *International Congresses of Mathematicians* have been held regularly around the world, at which thousands of mathematicians gather to learn about the most recent developments in their subject. Several of these gatherings have been commemorated on stamps – those from Europe include *Moscow* in 1966 (Figure 7(h)), *Helsinki* in 1978 (Figure 7(i)), and *Berlin* in 1998 (Figure 7(j)). As for the European Congresses of Mathematics, only two stamps have been issued: in 1996 for the second congress in Hungary, and recently for the eighth one, 8ECM, at Portorož, showing the Fibonacci sequence (Figure 7(k)).

Postage stamps provide an attractive vehicle for presenting mathematics and its development. This brief account has shown how the subject has been shaped by factors ranging from scientific and geographical developments and trade to education. Crucial to this story have been the attempts to solve a wide range of theoretical and practical problems, as well as the subject's internal logic by which it has progressed to increasingly greater abstractness.



(a) (b) (c) (d)

(e) (f) (g)

(h) (i) (j)

(k)

Figure 7. *The 20th century*
(a) Poincaré, (b) Hilbert curve, (c) Russell, (d) Banach,
(e) von Koch curve, (f) Julia set, (g) Rubik's cube,
(h) 1966 ICM Moscow, (i) 1978 ICM Helsinki, (j) 2008 ICM Berlin,
(k) 2021 8ECM Portorož

See [5, 6] for further information about mathematical stamps, and [1–4] for accounts of the history of mathematics. Many mathematical stamps are featured on the website www.mathematicalstamps.eu

Acknowledgements. I should like to thank Tomaž Pisanski for inviting me to give this public lecture and to design the associated mathematical stamps exhibition at Portorož. I am also very grateful to Matjaž Krnc for assistance with this exhibition, and also for his help with the production of this article.

References

[1] J. Barrow-Green, J. Gray, and R. Wilson, *The History of Mathematics: A Source-Based Approach: Volume 1*. AMS/MAA Textbooks 45, MAA Press/American Mathematical Society, Providence, RI, 2019 Zbl 1426.01001

[2] J. Barrow-Green, J. Gray, and R. Wilson, *The History of Mathematics: A Source-Based Approach: Volume 2*. AMS/MAA Textbooks 61, MAA Press/American Mathematical Society, Providence, RI, 2022 Zbl 07350589

[3] V. J. Katz, *A History of Mathematics: An Introduction*. Harper Collins College Publishers, New York, 1993 Zbl 1065.01501 MR 1200894

[4] D. J. Struik, *A Concise History of Mathematics*. 4th edn., Dover Publications, New York, 1987 Zbl 0645.01003 MR 919604

[5] R. Wilson, Stamp corner. *Math. Intelligencer* **6–44** (1984–2022)

[6] R. Wilson, *Stamping Through Mathematics*. Springer, New York, 2001 Zbl 0984.00007 MR 1836397

Robin Wilson
Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK;
robin.wilson@open.ac.uk

Minisymposia keynote lectures

DOI 10.4171/8ECM/19

Detecting arrays for effects of single factors

Charles J. Colbourn and Violet R. Syrotiuk

Abstract. Determining correctness and performance for complex engineered systems necessitates testing the system to determine how its behavior is impacted by factors and interactions among them. Of particular concern is to determine which settings of single factors (main effects) impact the behavior significantly. Detecting arrays for main effects are test suites that ensure that the impact of each main effect is witnessed even in the presence of d or fewer other significant main effects. Separation in detecting arrays dictates the presence of at least a specified number of such witnesses. A new parameter, corroboration, enables the fusion of levels while maintaining the presence of witnesses. Detecting arrays for main effects, having various values for the separation and corroboration, are constructed using error-correcting codes and separating hash families. The techniques are shown to yield explicit constructions with few tests for large numbers of factors.

1. Introduction

Combinatorial testing [33,45] addresses the design and analysis of test suites in order to evaluate correctness (and, more generally, performance) of complex engineered systems. We first introduce some basic definitions. There are k factors $F_1, \dots, F_k$. Each factor F_i has a set $S_i = \{v_{i1}, \dots, v_{is_i}\}$ of s_i possible *levels* (or *values* or *options*). A *test* is an assignment of a level from $v_{i1}, \dots, v_{is_i}$ to F_i, for each $1 \leq i \leq k$. The execution of a test yields a measurement of a *response*. When $\{i_1, \dots, i_t\} \subseteq \{1, \dots, k\}$ and $\sigma_{i_j} \in S_{i_j}$, the set $\{(i_j, \sigma_{i_j}) : 1 \leq j \leq t\}$ is a *t-way interaction*. The value of t is the *strength* of the interaction. A *main effect* is a 1-way interaction. A test on k factors *covers* $\binom{k}{t}$ t-way interactions. A *test suite* is a collection of tests. A test suite is typically represented as an $N \times k$ array $A = (\sigma_{i,j})$ in which $\sigma_{i,j} \in S_j$ when $1 \leq i \leq N$ and $1 \leq j \leq k$. The *size* of the test suite is N and its *type* is $(s_1, \dots, s_k)$. Tests correspond to rows of A, and factors correspond to its columns.

2020 Mathematics Subject Classification. Primary 05B40; Secondary 05B15, 68R05, 62K99.
Keywords. Combinatorial testing, detecting array, covering array, orthogonal array, error-correcting code.

When the response of interest can depend on one or more interactions, each having strength at most t, a test suite must cover each interaction in at least one row (test). To make this precise, let $A = (a_{i,j})$ be a test suite of size N and type $(s_1, \dots, s_k)$. Let $T = \{(i_j, \sigma_{i_j}) : 1 \leq j \leq t\}$ be a t-way interaction. Then $\rho_A(T)$ denotes the set $\{r : a_{r,i_j} = \sigma_{i_j}, 1 \leq j \leq t\}$ of rows of A in which the interaction is covered. A t-way interaction T must have $|\rho_A(T)| \geq 1$ in order to impact the response. For a set $\mathcal{T}$ of interactions, $\rho_A(\mathcal{T}) = \bigcup_{T \in \mathcal{T}} \rho_A(T)$.

When used in practical testing applications, as in [1,24,50], further requirements arise. First, if we suppose that some set $\mathcal{T}$ of interactions are those that significantly impact the response, yet there is another interaction $T \notin \mathcal{T}$ for which $\rho_A(T) \subseteq \bigcup_{S \in \mathcal{T}} \rho_A(S)$, the responses are inadequate to determine whether or not T impacts the response significantly. This requirement was explored in [17], and later in [19,40]. Secondly, one or more tests may fail to execute correctly, and yield no response or yield outlier responses. To mitigate this, Seidel et al. [51] impose stronger "separation" requirements on the test suite.

Extending definitions in [17,19,51], we formally define the test suites with which we are concerned. Let A be a test suite of size N and type $(s_1, \dots, s_k)$. Let $\mathcal{I}_t$ be the set of all t-way interactions for A. Our objective is to identify the set $\mathcal{T} \subseteq \mathcal{I}_t$ of interactions that have significant impact on the response. In so doing, we assume that at most d interactions impact the response. Without limiting d, it can happen that no test suite of type $(s_1, \dots, s_k)$ exists for any value of N [40].

An $N \times k$ array A of type $(s_1, \dots, s_k)$ is $(\bar{d}, t, \delta)$-*locating* if $|\rho_A(\mathcal{R}) \cap \rho_A(\mathcal{T})| < \delta \Leftrightarrow \mathcal{R} = \mathcal{T}$ whenever $\mathcal{R}, \mathcal{T} \subseteq \mathcal{I}_t$, $|\mathcal{R}| \leq d$, and $|\mathcal{T}| \leq d$. In this paper, we enforce a condition that is stronger [17]. An $N \times k$ array A of type $(s_1, \dots, s_k)$ is (d, t, δ)-*detecting* if $|\rho_A(T) \setminus \rho_A(\mathcal{T})| < \delta \Leftrightarrow T \in \mathcal{T}$ whenever $\mathcal{T} \subseteq \mathcal{I}_t$, and $|\mathcal{T}| = d$. To record all of the parameters, we use the notation $\mathsf{DA}_\delta(N; d, t, k, (s_1, \dots, s_k))$. To emphasize that different factors may have different numbers of levels, this is called a *mixed* detecting array. When all factors have the same number, v, of levels, the array is *uniform* and the notation is simplified to $\mathsf{DA}_\delta(N; d, t, k, v)$. The parameter δ is the *separation* of the detecting array [51], and the definition in [17] is recovered by setting $\delta = 1$. Rows in $\rho_A(T) \setminus \rho_A(\mathcal{T})$ are *witnesses* for T that are not masked by interactions in $\mathcal{T}$. A separation of δ necessitates δ witnesses, ensuring that fewer than δ missed or incorrect measurements cannot result in an interaction's impact being lost.

Setting $d = 0$ in the definition, $\mathcal{T} = \emptyset$, and $\rho_A(\emptyset) = \emptyset$. Then a $(0, t, \delta)$-detecting array is an array in which each t-way interaction is covered in at least δ rows. This leads to a standard class of testing arrays: a *covering array* $\mathsf{CA}_\delta(N; t, k, (s_1, \dots, s_k))$ is equivalent to a $\mathsf{DA}_\delta(N; 0, t, k, (s_1, \dots, s_k))$. The simpler notation $\mathsf{CA}_\delta(N; t, k, v)$ is employed when it is uniform.

An *orthogonal array* $\mathsf{OA}_\delta(N; t, k, v)$, A, enforces the stronger condition that for every t-way interaction T, we have that $|\rho_A(T)| = \delta$. Orthogonal arrays are the subject of a vast literature [28], in part because of their applications in experimental design and error-correcting codes. Covering arrays have also been much more extensively studied [13, 33, 45] than detecting arrays and their variants; they are usually defined only in the case when $\delta = 1$, and in a more direct manner than by exploiting the equivalence with certain detecting arrays. Often constructions of covering arrays focus on the uniform cases. In part this is because a $\mathsf{CA}_\delta(N; t, k, (s_1, \dots, s_{i-1}, s_i - 1, s_{i+1}, \dots, s_k))$ can be obtained from a $\mathsf{CA}_\delta(N; t, k, (s_1, \dots, s_{i-1}, s_i, s_{i+1}, \dots, s_k))$ by making any two levels of the ith factor identical. This operation is *fusion* (see, e.g., [15]).

Supporting fusion for detecting arrays motivates the definition of a further parameter [20]. When applied to detecting arrays with $d \geq 1$, fusion may reduce the number of witnesses. Increasing the separation cannot overcome this problem, unless the number of *distinct* witnesses increases.

Let A be an $N \times k$ array. Let $T = \{(i_j, \sigma_{i_j}) : 1 \leq j \leq t\}$ be a t-way interaction for A. Let $C = \{c_i : 1 \leq i \leq d\}$ be a set of d column indices of A with $\{i_1, \dots, i_t\} \cap \{c_1, \dots, c_d\} = \emptyset$. Define a set system on the ground set $\{(c, f) : c \in C, f \in S_c\}$ by

$$\mathcal{S}_{A,T,C} = \big\{\{(c_1, v_1), \dots, (c_d, v_d)\} : T \cup \{(c_1, v_1), \dots, (c_d, v_d)\} \text{ is covered in } A\big\}.$$

Lemma 1.1. *An array A is (d, t, δ)-detecting if and only if for every t-way interaction T and every set C of d disjoint columns, every subset X of the ground set of $\mathcal{S}_{A,T,C}$ that is disjoint from fewer than δ sets in $\mathcal{S}_{A,T,C}$ satisfies $|X| > d$.*

Proof. First suppose that for some t-way interaction $T = \{(i_j, \sigma_{i_j}) : 1 \leq j \leq t\}$ and some set $C = \{c_i : 1 \leq i \leq d\}$ of d disjoint columns, in the set system $\mathcal{S}_{A,T,C}$ there is a set of elements $X = \{(c_1, v_1), \dots, (c_d, v_d)\}$ for which fewer than δ sets in the set system contain no element of X. Define $T_i = \{(i_j, \sigma_{i_j}) : 1 \leq j \leq t - 1\} \cup \{(c_i, v_i)\}$. Set $\mathcal{T} = \{T_1, \dots, T_d\}$. Then $T \notin \mathcal{T}$ but $|\rho_A(T) \setminus \rho_A(\mathcal{T})| < \delta$, so A is not (d, t, δ)-detecting.

In the other direction, suppose that A is not (d, t, δ)-detecting, and consider a set $\mathcal{T} = \{T_1, \dots, T_d\}$ of d t-way interactions and a t-way interaction T for which $T \notin \mathcal{T}$ but $|\rho_A(T) \setminus \rho_A(\mathcal{T})| < \delta$. Without loss of generality, there is no interaction $T' \in \mathcal{T}$ for which T and T' share a factor set to different levels in each and so, because $T \neq T'$, T' contains a factor not appearing in T. For each $T_i \in \mathcal{T}$, let c_i be a factor in T_i that is not in T, and suppose that $(c_i, v_i) \in T_i$ for $1 \leq i \leq d$. Then the set $X = \{(c_i, v_i) : 1 \leq i \leq d\}$, when removed from $\mathcal{S}_{A,T,C}$, leaves fewer than δ sets. ■

Lemma 1.1 implies that a (d, t, δ)-detecting array must cover each t-way interaction at least $d + \delta$ times; indeed when $d \geq 1$, for each t-way interaction T and every column c not appearing in T, interaction T must be covered in at least $d + 1$

rows containing distinct levels in column c. In particular, a necessary condition for a $\mathsf{DA}_\delta(N; d, t, k, (s_1, \ldots, s_k))$ to exist is that $d < \min(s_i : 1 \le i \le k)$ (see also [17]).

These considerations lead to the parameter of interest. For array A, with t-way interaction T and set C of d disjoint columns, suppose that, in $\mathcal{S}_{A,T,C}$, for each column in C one performs fewer than s fusions of elements within those arising from that column. Further suppose that, no matter how these fusions are done, the resulting set system has the property that every subset X of the ground set of $\mathcal{S}_{A,T,C}$ that is disjoint from fewer than δ sets in $\mathcal{S}_{A,T,C}$ satisfies $|X| \ge d + 1$. Then (T, C) has *corroboration* s in A. When every choice of (T, C) has corroboration (at least) s in a $\mathsf{DA}_\delta(N; d, t, k, (s_1, \ldots, s_k))$, it has *corroboration* s. We extend the notation as $\mathsf{DA}_\delta(N; d, t, k, (s_1, \ldots, s_k), s)$ to include corroboration s as a parameter.

In this paper, we focus on detecting arrays for single factors, or main effects. In Section 2, we briefly summarize what is known about the construction of detecting arrays. In Section 3, we define and construct certain arrays, perfect and separating hash families, which are subsequently used to construct detecting arrays with different values of separation and corroboration. In Section 4, we unify a number of constructions for detecting arrays that employ hash families by providing a general column replacement method, and present the small detecting arrays needed. In Section 5, we examine the consequences of applying the general construction.

2. Covering arrays and Sperner partition systems

As observed in [17], one method to construct detecting arrays is to use covering arrays of higher strength. The following records consequences for separation and corroboration.

Lemma 2.1. *A* $\mathsf{CA}_\lambda(N; t, k, v)$ *is*

(1) *a* $\mathsf{DA}_\delta(N; d, t-d, k, v, 1)$ *with* $\delta = \lambda(v-d)v^{d-1}$, *and*

(2) *a* $\mathsf{DA}_\delta(N; d, t-d, k, v, v-d)$ *with* $\delta = \lambda(d+1)^{d-1}$

whenever $1 \le d < \min(t, v)$.

Proof. Let A be a $\mathsf{CA}_\lambda(N; t, k, v)$. Let d satisfy $1 \le d < \min(t, v)$. Let T be a $(t-d)$-way interaction, and let C be a set of d columns not appearing in T. Using the parameters of the covering array, $\mathcal{S}_{A,T,C}$ contains at least λv^d sets, and each element appears in at least λv^{d-1} of them. Suppose that d elements of $\mathcal{S}_{A,T,C}$ are removed, and further suppose that the numbers of elements deleted for the d factors are $e_1, \ldots, e_d$ (so that $d = \sum_{i=1}^{d} e_i$). Then the number of remaining sets is $\lambda \prod_{i=1}^{d}(v - e_i)$, which is minimized at $\delta = \lambda(v-d)v^{d-1}$. This establishes the first statement. For the second, performing at most $v - d - 1$ fusions within each factor of $\mathcal{S}_{A,T,C}$ and then deleting at most d elements leaves at least $\delta = \lambda(d+1)^{d-1}$ sets by a similar argument. ■

The effective construction of detecting arrays is well motivated by practical testing applications, in which the need for higher separation to mitigate the effects of outlier responses, and higher corroboration to support fusion of levels, arise. Despite this, other than the construction from covering arrays of higher strength, few constructions are available. In [60], uniform $(1, t)$-detecting arrays with separation 1, corroboration 1, and few factors are studied. This was extended in [53, 55] to (d, t)-detecting arrays, and further to mixed detecting arrays in [54]. Each of these focusses on the determination of a lower bound on the number of rows in terms of d, t, and v, and the determination of cases in which this bound can be met. For $d + t \geq 2$, however, the number of rows must grow at least logarithmically in k, because every two columns must be distinct. Hence the study of arrays meeting bounds that are independent of k necessarily considers only small values of k. In addition, none of these addresses separation or corroboration.

For larger values of k, algorithmic methods are developed in [51]. The algorithms include randomized methods based on the Stein–Lovász–Johnson framework [31, 37, 57], and derandomized algorithms using conditional expectations (as in [10, 11]); randomized methods based on the Lovász local lemma [3, 25] and derandomizations using Moser–Tardos resampling [44] (as in [16]). Although these methods produce $(1, t)$-mixed detecting arrays for a variety of separation values, they have not been applied for $d > 1$ or to increase the corroboration. Extensions to larger d for locating arrays are considered in [35].

When $t = 1$, one is considering detecting arrays for main effects. A *Sperner family* is a family of subsets of some ground set such that no set in the family is a subset of any other. Meagher, Moura, and Stevens [42] introduced Sperner partition systems as a natural variant of Sperner families. An (n, v)-*Sperner partition system* is a collection of partitions of some n-set, each into v nonempty classes, such that no class of any partition is a subset of a class of any other. In [36, 42], the largest number of classes in an (n, v)-Sperner partition system is determined exactly for infinitely many values of n for each v. In [12, 26], lower and upper bounds are established for all n and each v. As noted there, given an (n, v)-Sperner partition system with k partitions, if we number the elements using $\{1, \ldots, n\}$ and number the sets in each partition with $\{1, \ldots, v\}$, we can form an $n \times k$ array in which cell (r, c) contains the set number to which element r belongs in partition c. This array is a $\mathsf{DA}_1(n; 1, 1, k, v, 1)$, and indeed every such DA arises in this way. Even when $d = t = s = \delta = 1$, the largest value of k as a function of n is not known precisely. Therefore, it is natural to seek useful bounds and effective algorithms for larger values of the parameters.

3. Perfect and separating hash families

3.1. Separating hash families: Definitions

An $(N;k,v)$-*hash family* is an $N \times k$ array on v symbols. Colbourn and Torres-Jiménez [23] relax the requirement that each row have the same number of symbols. An $N \times k$ array is a *heterogeneous hash family*, or $\mathsf{HHF}(N;k,(v_1,\dots,v_N))$, when the ith row contains (at most) v_i symbols for $1 \le i \le N$.

An $(N;k,v,\{w_1,w_2,\dots,w_t\})$-*separating hash family* of *index* λ is an $(N;k,v)$-hash family A that satisfies the property: for any $C_1,C_2,\dots,C_t \subseteq \{1,2,\dots,k\}$ such that $|C_1| = w_1, |C_2| = w_2, \dots, |C_t| = w_t$, and $C_i \cap C_j = \emptyset$ for every $i \neq j$, whenever $c \in C_i, c' \in C_j$, and $i \neq j$, different symbols appear in columns c and c' in each of at least λ rows. The notation $\mathsf{SHF}_\lambda(N;k,v,\{w_1,w_2,\dots,w_t\})$ is used. See, for example, [2, 48, 56]; and see [5] for the similar notion of "partially hashing." The notation $\mathsf{SHHF}_\lambda(N;k,(v_1,\dots,v_N),\{w_1,w_2,\dots,w_t\})$ is used for heterogeneous arrays. We remark that an $\mathsf{SHF}_1(N;k,v,\{1,d\})$ is a *frameproof code* (see, for example, [56,59]), a type of *strong* separating hash family [47,48].

When $w_1 = \cdots = w_k = 1$, we recover a more widely studied class of arrays. A *perfect hash family* $\mathsf{PHF}_\lambda(N;k,v,t)$ is an $(N;k,v)$-hash family, in which in every $N \times t$ subarray, at least λ rows each consisting of distinct symbols. Mehlhorn [43] introduced perfect hash families, and they have subsequently found many applications in combinatorial constructions [58]. The definition for PHF extends naturally to perfect *heterogeneous* hash families; we use the notation $\mathsf{PHHF}_\lambda(N;k,(v_1,\dots,v_N),t)$.

We employ a further extension that incorporates two types of symbols, as proposed in [14]. Let $\Sigma_v = \{0,\dots,v-1\}$. An $\mathsf{SHHF}_\lambda(N;k,(v_1,\dots,v_N),\{1,d^\circ\})$ is an $N \times k$ array for which

(1) the jth row contains symbols from $\Sigma_{v_j} \cup \{\circ\}$;

(2) for every $C_1, C_2 \subseteq \{1,2,\dots,k\}$ with $|C_1| = 1$, $|C_2| = d$, and $C_1 \cap C_2 = \emptyset$, there are λ rows, indexed by $\{\rho_1,\dots,\rho_\lambda\}$, so that for each ρ_j, the set S of symbols appearing in columns of C_2 in row ρ_j is a subset of $\Sigma_{v_{\rho_j}} \cup \{\circ\}$, and the symbol in the column of C_1 in row ρ_j belongs to $\Sigma_{v_{\rho_j}} \setminus S$.

When the array is homogeneous, the notation $\mathsf{SHF}_\lambda(N;k,v,\{1,d^\circ\})$ is used. Every $\mathsf{SHF}_\lambda(N;k,v,\{1,d\})$ is an $\mathsf{SHF}_\lambda(N;k,v,\{1,d^\circ\})$, and by treating $\circ$ as a symbol like the rest, every $\mathsf{SHF}_\lambda(N;k,v,\{1,d^\circ\})$ is an $\mathsf{SHF}_\lambda(N;k,v+1,\{1,d\})$.

3.2. Separating hash families: Some constructions

Existence of SHFs is well studied for $\delta = 1$ (see [52] and references therein), but these appear not to have been studied when $\delta > 1$. We employ a number of standard ideas to construct SHHFs from other SHHFs in the following lemma.

Lemma 3.1. *Suppose that an* $\mathsf{SHHF}_\delta(N;k,\{v_1,\dots,v_N\},\{1,d^\circ\})$ *exists, in which some symbol in the* j*th row appears in* c *columns. Then*

1. *an* $\mathsf{SHHF}_\delta(N;k,\{v_1,\dots,v_{j-1},v_j+1,v_{j+1},\dots,v_N\},\{1,d^\circ\})$ *exists;*
2. *when* $\delta > 1$, *an* $\mathsf{SHHF}_{\delta-1}(N-1;k,\{v_1,\dots,v_{j-1},v_{j+1},\dots,v_N\},\{1,d^\circ\})$ *exists;*
3. *when* $c = k$, *an* $\mathsf{SHHF}_{\delta-1}(N-1;k,\{v_1,\dots,v_{j-1},v_{j+1},\dots,v_N\},\{1,d^\circ\})$ *exists;*
4. *an* $\mathsf{SHHF}_\delta(N;k-c,\{v_1,\dots,v_{j-1},v_j-1,v_{j+1},\dots,v_N\},\{1,d^\circ\})$ *exists;*
5. *an* $\mathsf{SHHF}_\delta(N-1;c,\{v_1,\dots,v_{j-1},v_{j+1},\dots,v_N\},\{1,d^\circ\})$ *exists;*
6. *an* $\mathsf{SHHF}_\delta(N;k+1,\{v_1+1,\dots,v_j+1,\dots,v_N+1\},\{1,d^\circ\})$ *exists;*
7. *if an* $\mathsf{SHHF}_{\delta'}(M;k,\{w_1,\dots,w_M\},\{1,d^\circ\})$ *also exists, then*
   an $\mathsf{SHHF}_{\delta+\delta'}(N+M;k,\{v_1,\dots,v_N,w_1,\dots,w_M\},\{1,d^\circ\})$ *exists.*

Proof. Let A be the stated SHHF. Then (1) holds because permitting an additional symbol in row j does not require its use. Deleting any row of A can reduce its index by at most one, so (2) holds. When $c = k$ (and in particular when $v_j = 1$), row j accomplishes no separations in A, so (3) holds. To establish (4), choose a symbol that occurs c times in row j, and delete all columns containing that symbol in row j. For (5), choose a symbol that occurs c times in row j, and delete all columns containing any other symbol in row j; then apply (3). For (6), add a column to A that, in each row, contains a symbol not appearing in A. For (7), vertically juxtapose the two arrays. ■

Stinson, Wei, and Chen [59] use an expurgation technique to establish lower bounds on k for which an $\mathsf{SHF}_1(N;k,v,\{1,d\})$ exists. One consequence of their results is the following.

Theorem 3.2 ([59]). *An* $\mathsf{SHF}_1(N;k,v,\{1,2\})$ *exists for* $k = \lceil \frac{1}{2}(\frac{v^2}{2v-1})^{\frac{N}{2}} \rceil$.

Unfortunately, Theorem 3.2 does not provide competitive lower bounds on the achievable numbers of columns in our applications. We therefore develop a number of other constructions.

Lemma 3.3. *An* $\mathsf{SHF}_1(N+1;N\cdot v^N,v,\{1,1^\circ\})$ *exists whenever* $N \geq 1$ *and* $v \geq 1$.

Proof. Form an $N \times v^N$ array A consisting of all distinct column vectors from Σ_v^N. For $0 \leq i \leq N$, form an $(N+1) \times v^N$ array A_i by inserting a row consisting entirely of $\circ$ after row i when $1 \leq i \leq N$, or before row 1 when $i = 0$. Horizontally juxtapose $A_0,\dots,A_N$ to form B, the $\mathsf{SHF}_1(N+1;N\cdot v^N,v,\{1,1^\circ\})$. The verification is routine, as follows. Consider two distinct columns γ and c of B. When γ and c are from the same A_i, the two columns disagree in at least one row because such a row appears in A. On the other hand, when γ is in A_i and c is in A_j with $j \neq i$, row $j+1$ of the resulting array contains $\circ$ in column c, but contains an element of Σ_v in column γ, so the desired separation is ensured. ■

We consider cases with "few" rows next.

Lemma 3.4. *Let $d \geq 2$, $\delta \geq 1$, and $d > \alpha \geq 1$. Then*

$$k \leq k_{max} = \max\left(v_1, \dots, v_{d+\delta-\alpha}, \left\lfloor \frac{1}{\delta} \sum_{i=1}^{d+\delta-\alpha} (v_i - 1) \right\rfloor\right)$$

whenever an $\mathsf{SHHF}_\delta(d+\delta-\alpha; k, (v_1, \dots, v_{d+\delta-\alpha}), \{1, d\})$ *exists.*

Proof. Let A be an $\mathsf{SHHF}_\delta(d+\delta-\alpha; k, (v_1, \dots, v_{d+\delta-\alpha}), \{1, d\})$. An entry in A is a *private* entry if it contains the only occurrence of a symbol in its row. If some row contains only private entries, then $k \leq \max(v_1, \dots, v_{d+\delta-\alpha})$. If some column c were to contain $d+1-\alpha$ entries that are not private, for each of $d+1-\alpha$ such rows choose a column that contains the same symbol as in column c. Let X be the set of at most $d+1-\alpha$ columns so chosen. There could be at most $\delta-1$ rows separating c from X, which cannot arise. Consequently, every column of A contains at least δ private entries, and at most $d-\alpha$ that are not private. Row i employs v_i symbols and hence contains at least $k - v_i + 1$ entries that are not private. Hence $(d-\alpha)k \geq \sum_{i=1}^{d+\delta-\alpha}(k - v_i + 1)$. Hence $\sum_{i=1}^{d+\delta-\alpha}(v_i - 1) \geq \delta k$ and the bound follows. ■

When $\delta = 1$, Blackburn [6] establishes that an $\mathsf{SHF}_1(N; k, v, \{1, d\})$ can exist only when $k \leq dv^{\lceil \frac{N}{d} \rceil} - d$. To establish this, partition the N rows into d classes; then when the largest class has r rows in it, amalgamate all rows in the class into a single row on v^r symbols. He employs a version of Lemma 3.4, using $\delta = 1$ and not exploiting heterogeneity, to obtain the upper bound on k already mentioned. Our heterogeneous bound underlies an improvement in the upper bound in some situations. Unfortunately, although the amalgamation strategy cannot reduce a separation $\delta \geq 2$ to zero, it can nonetheless reduce it to 1. Hence Lemma 3.4 does not lead to an effective upper bound on k as a function of N when $\delta > 1$.

For certain SHHFs, this bound can be met by generalizing a well-known construction for perfect hash families [6, 41, 61]; indeed, it can be extended to employ $\circ$ symbols.

Lemma 3.5. *Let $\delta \geq 1$ and $d > \alpha \geq 1$. Let $v_1 = \cdots = v_\delta \geq v_{\delta+1} \geq \cdots \geq v_{d+\delta-\alpha}$. Then an* $\mathsf{SHHF}_\delta(d+\delta-\alpha; \max(v_\delta, \lfloor \frac{1}{\delta} \sum_{j=1}^{d+\delta-\alpha} v_j \rfloor), \{v_1, \dots, v_{d+\delta-\alpha}\}, \{1, d^\circ\})$ *exists.*

Proof. If $k = v_\delta$, form δ rows that contain only private entries, and adjoin $d-\alpha$ arbitrary rows to produce the SHHF. Henceforth, we suppose that $k > v_\delta = v_1$. In a

$(d+\delta-\alpha)\times k$ array, place $\circ$ entries so that (1) each of the k columns contains exactly $d-\alpha$ $\circ$ entries, and (2) for $1\le j\le d+\delta-\alpha$, row j contains at least $k-v_j$ $\circ$ entries. When this can be done, in each row fill the remaining entries with distinct symbols. Then no matter how $C_1=\{\gamma\}$ is chosen, there are δ rows in which γ contains a private entry, so the array is an $\mathsf{SHHF}_\delta(d+\delta-\alpha;k,\{v_1,\dots,v_{d+\delta-\alpha}\},\{1,d^\circ\})$.

We next determine the values of k for which this is possible. Because each column contains $d-\alpha$ entries equal to $\circ$, the array contains $k(d-\alpha)$ entries equal to $\circ$. On the other hand, row j contains at least $k-v_j$ entries equal to $\circ$. Hence $\sum_{j=1}^{d+\delta-\alpha}(k-v_j)\le k(d-\alpha)$, leading to the stated bound on k. It remains to ensure that the $\circ$ entries can be placed to meet the row and column constraints simultaneously; this follows from classical work on {0, 1}-matrices with fixed row and column sums ([9], for example). ■

For larger numbers of rows, the elementary constructions of Lemma 3.1 are useful, but they typically decrease the number of columns or the index. Hence further constructions are needed. One addition method follows.

Lemma 3.6. *If an* $\mathsf{SHF}(N;k,v,\{1,d^\circ\};\delta)$ *and an* $\mathsf{SHF}(N';k',v,\{1,d^\circ\};\delta)$ *both exist, then an* $\mathsf{SHF}(N+N';k+k',v,\{1,d^\circ\};\delta)$ *exists.*

Proof. Let A be an $\mathsf{SHF}(N;k,v,\{1,d^\circ\};\delta)$ and let B be an $\mathsf{SHF}(N';k',v,\{1,d^\circ\};\delta)$. Let $E_{n\times\kappa}$ denote an $n\times\kappa$ array in which every entry is $\circ$. Then the array $\left(\begin{smallmatrix} A & E_{N\times k'} \\ E_{N'\times k} & B \end{smallmatrix}\right)$ is an $\mathsf{SHF}(N+N';k+k',v,\{1,d^\circ\};\delta)$. The verification is routine. ■

To yield a larger increase in the number of columns, we also employ a *composition* [4] or *column replacement* [13] method.

Theorem 3.7. *Suppose that an* $\mathsf{SHHF}_\delta(N;k,\{k_1,\dots,k_N\},\{1,d\})$ *exists, Further suppose that an* $\mathsf{SHHF}_\beta(M_i;k_i,\{v_{i1},\dots,v_{iM_i}\},\{1,d^\circ\})$ *exists for each* $1\le i\le N$. *Then an* $\mathsf{SHHF}_{\delta\beta}(\sum_{i=1}^N M_i;k,\{v_{ij}:1\le i\le N,1\le j\le M_i\},\{1,d^\circ\})$ *exists.*

Proof. Let A be the $\mathsf{SHHF}_\delta(N;k,\{k_1,\dots,k_N\},\{1,d\})$. Form an arbitrary bijection between the k_i *symbols* permitted in row i of A and the k_i *columns* of B_i, the $\mathsf{SHHF}_\beta(M_i;k_i,\{v_{i1},\dots,v_{iM_i}\},\{1,d^\circ\})$. Replace each symbol of A by its associated column in B_i to form an array D_i. Vertically juxtapose $D_1,\dots,D_N$ to form D. Then D has $\sum_{i=1}^N M_i$ rows and k columns; the largest numbers of symbols that can appear in its rows is given by $\{v_{ij}:1\le i\le N,1\le j\le M_i\}$. Now consider an arbitrary column γ and a set C_2 of d columns of D not containing γ. Column γ is separated from C_2 in δ rows of A, say $\rho_1,\dots,\rho_\delta$. Consider a particular such row, ρ_j. Suppose that A contains symbol ν in column γ; let S be the set of symbols appearing in row ρ_j in columns of C_2. Then S does not contain ν, and column ν is separated from all columns in S in β rows of B_{ρ_j} that do not contain $\circ$ in column ν. But then in D_{ρ_j}, there are β rows in which column γ does not contain $\circ$, and the symbol in column γ is not the same as in any column of C_2. This establishes that D is the desired SHHF. ■

Restricting to SHFs, an $\mathsf{SHF}_\delta(N; k, \kappa, \{1, d\})$ and an $\mathsf{SHF}_\beta(M; \kappa, v, \{1, d^\circ\})$ yield an $\mathsf{SHF}_{\delta\beta}(NM; k, v, \{1, d^\circ\})$. Theorem 3.7 is particularly useful because the array in which replacements are made is a separating, rather than a perfect, hash family. It is also effective because the construction of v-ary SHHFs can employ SHHFs with much larger alphabets.

3.3. Separating hash families: Codes for $d = 1$

Here we consider the case when $d = 1$, i.e., the case of perfect hash families. A v-ary code C of *length* N is a subset of Σ_v^n. (See [38, 39] for definitions in coding theory.) Each $c \in C$ is a *codeword*. The *size* of C is the number $|C|$ of codewords, and its *minimum distance* is the smallest Hamming distance between any two distinct codewords. A v-ary code with length N, size k, and minimum distance δ is an $(N, k, \delta)_v$ *code*. Let $A_v(N, d)$ denote the maximum size of an $(N, k, \delta)_v$ code. Treating columns of a $\mathsf{PHF}_\delta(N; k, v, 2)$ as codewords, one obtains an $(N, k, \delta)_v$ code; the converse also holds. More generally, we have the following lemma.

Lemma 3.8. *An* $\mathsf{SHF}_\delta(N; k, v, \{1, d\})$ *exists whenever* $A_v(N, N - \lfloor \frac{N-\delta}{d} \rfloor) \geq k$.

Proof. Treat codewords of an $(N, k, N - \lfloor \frac{N-\delta}{d} \rfloor)_v$ code as columns of an $N \times k$ array on v symbols. Consider any set $C_1 = \{\gamma\}$ of one column, and any disjoint set C_2 of d columns. For each $\gamma' \in C_2$, there can be at most $\lfloor \frac{N-\delta}{d} \rfloor$ rows in which γ and γ' share a symbol. Then there can be at most $N - \delta$ rows in which γ shares a symbol with one or more columns of C_2, and hence at least δ rows in which γ shares a symbol with no column of C_2. This establishes the result. ■

For $\delta \geq 3$, the existence question for such codes is far from settled, particularly when $v > 2$. In Section 5, we use constructions of $\mathsf{SHF}_\lambda(N; k, 5, \{1, d\})$s, $\mathsf{SHF}_\lambda(N; k, 6, \{1, d\})$s, and $\mathsf{SHF}_\lambda(N; k, 5, \{1, d^\circ\})$s. Therefore, for $d = 1$, in Table 1 we provide lower bounds on the number of columns achieved for these parameters.

To justify these entries, first consider the cases when $\delta \in \{1, 2\}$. For type $\{1, 1\}$ with $v \in \{5, 6\}$, the values are exact and arise from the easy observations that all distinct column vectors form a code of distance 1, while all column vectors whose total weight is $0 \pmod{v}$ form a code of distance 2. Considering the case of type $\{1, 1^\circ\}$ for $v = 5$ with $\delta \in \{1, 2\}$, only one example is given for an $\mathsf{SHF}_\delta(N; k, 5, \{1, 1^\circ\})$ having more columns than the $\mathsf{SHF}_\delta(N; k, 5, \{1, 1\})$, namely the $\mathsf{SHF}_1(6; 18750, 5, \{1, 1^\circ\})$. This SHF was initially found by computation, but Lemma 3.3 provides an easier construction.

	$\delta = 1$			$\delta = 2$			$\delta = 3$			$\delta = 4$		
x	$(5, 1)$	$(5, 1^\circ)$	$(6, 1)$	$(5, 1)$	$(5, 1^\circ)$	$(6, 1)$	$(5, 1)$	$(5, 1^\circ)$	$(6, 1)$	$(5, 1)$	$(5, 1^\circ)$	$(6, 1)$
1	25	25	36	25	25	36	25	25	34	25	25	30
2	125	125	216	125	125	216	125	125	159	125	125	146
3	625	625	1296	625	625	1296	625	625	953	263	263	819
4	3125	3125	7776	3125	3125	7776	1597	1597	5718	1225	1225	4914
5	15625	18750	46656	15625	15625	46656	7985	7985	34278	4375	4375	22719
6										17500	17500	28320

Table 1. Lower bounds on k for $\mathsf{SHF}_\delta(x + \delta; k, v, \{1, d\})$ for $v \in \{5, 6\}$ and $d \in \{1, 1^\circ\}$.

Now let us turn to $\delta \geq 3$. Most research effort for the construction of codes has concentrated on "small" values of v. Indeed, for $v = 6$ there appears to have been no systematic effort to construct 6-ary codes. However, for $v = 5$, the situation is quite different. For large values of N, typically one resorts to using linear codes, as tabulated in [27]. In addition, Bogdanova and Östergård [7] tabulate lower bounds on $A_5(N, d)$ for $N \leq 11$ obtained by standard code constructions, by computation, and certain explicit constructions [32]. Some entries were subsequently improved upon in [34]. In particular, Laaksonen and Östergård [34] show that $A_5(8, 5) \geq 165$ and $A_5(9, 5) \geq 725$; in the repository of codes associated with their paper they provide explicit solutions to establish that $A_5(8, 5) \geq 257$ and $A_5(9, 5) \geq 857$. In [7], it is shown that $A_5(7, 5) \geq 53$ and $A_5(8, 6) \geq 45$. We improve these two bounds next, obtained via computations using `cliquer` [46].

Lemma 3.9. $A_5(7, 5) \geq 57$ *and* $A_5(8, 6) \geq 50$.

Proof. We write codewords omitting the commas and parentheses. Consider the nine codewords $C_7 = \{1111111, 1242342, 1200224, 1324443, 1333020, 1420300, 1432233, 1043404, 1004032\}$. When $a_0 \cdots a_{N-1}$ is a codeword, any vector $b_0 \cdots b_{N-1}$ with $b_{i+s \bmod N} = a_i$ for some s is a cyclic shift of the codeword. The 57 distinct cyclic shifts of the codewords in C_7 form a $(7, 57, 5)_5$ code. In the same manner, the 50 cyclic shifts of $\{11214402, 11023313, 12001200, 13441344, 14330040, 10322424, 22030434, 23232323\}$ form an $(8, 50, 6)_5$ code. ■

Known codes provide powerful constructions for $\mathsf{SHF}_\delta(N; k, 5, \{1, 1\})$s, which in turn yield lower bounds on the number of columns in $\mathsf{SHF}_\delta(N; k, 5, \{1, 1^\circ\})$s and $\mathsf{SHF}_\delta(N; k, 6, \{1, 1\})$s. We failed to find any cases with $\delta \geq 3$ in which an $\mathsf{SHF}_\delta(N; k, 5, \{1, 1^\circ\})$ has more columns than an $\mathsf{SHF}_\delta(N; k, 5, \{1, 1\})$, although we expect that this can happen for larger sizes.

We have no tables of 6-ary codes. Lemma 3.1 (6) constructs 6-ary codes from 5-ary ones, to which the remaining constructions of the lemma can be applied. In addition, we adapted the "replace-one-column–random extension" randomized algorithm from [16] in order to construct SHFs of index δ. We do not describe the method here, noting only that it is a heuristic technique that is not expected to produce optimal sizes. We also report some 6-ary codes, again found using `cliquer` [46].

Lemma 3.10. *For* $v = 6$, $A_6(13, 10) \geq 78$, $A_6(12, 9) \geq 108$, $A_6(11, 8) \geq 132$, $A_6(10, 7) \geq 186$, $A_6(9, 6) \geq 258$, $A_6(16, 12) \geq 96$, $A_6(15, 11) \geq 180$, $A_6(14, 10) \geq 546$, *and* $A_6(11, 7) \geq 660$.

Proof. Form the following sets of codewords:

N, d	Starter codewords
13,10	0001105451315
12,9	000000000000, 002322454401, 012345012345, 013415321203
11,8†	00205141502, 00541020145
10,7	0000143153, 0023442314, 0140504324, 0303030303
9,6	000214121, 000300342, 004135045, 004343154, 010232454, 025403241, 042042042
16,12†	0000414354453414
15,11†	000125304403521, 000450523325054
14,10†	00043515451534, 00103445544301, 00134204402431, 00214412003553, 00452540045254, 01014503530541, 01025245354252
11,7†	00015524122, 00133100525, 00150303051, 00224151422, 00314010413, 00342404243, 00501242105, 01035134235

When a † is shown, adjoin the codeword $a_{N-1} \cdots a_0$ whenever $a_0 \cdots a_{N-1}$ is a codeword. Form all distinct cyclic shifts (as in the proof of Lemma 3.9). Then develop each codeword under the additive action of $\mathbb{Z}_6$. ∎

3.4. Separating hash families: $d = 2$

For separations $\{1, 2\}$ and $\{1, 2^\circ\}$, we report results for $v \in \{5, 6\}$ in Table 2. To produce Table 2, we apply Lemma 3.8 to the $(n, k, \delta)_5$ and $(n, k, \delta)_6$ codes described earlier. We also employ certain linear codes, noting that a linear code with parameters $[n, k, \delta]_q$ yields an $(n, q^k, \delta)_q$ code [38, 39]. In particular, we employ linear codes with parameters $[6, 3, 4]_5$, $[10, 3, 7]_5$, $[12, 4, 8]_5$, $[15, 5, 9]_5$, $[18, 5, 11]_5$, $[15, 6, 8]_5$,

	$\delta = 1$			$\delta = 2$			$\delta = 3$			$\delta = 4$		
x	$(5, 2)$	$(5, 2^{\circ})$	$(6, 2)$	$(5, 2)$	$(5, 2^{\circ})$	$(6, 2)$	$(5, 2)$	$(5, 2^{\circ})$	$(6, 2)$	$(5, 2)$	$(5, 2^{\circ})$	$(6, 2)$
1	8	10	10	6	7	7		6			6	
2	25	25	36	25	25	34	25	25	30	25	25	27
3	34	46	51	26	36	39		26	32			28
4	125	125	162	125	125	146	57	57	132	50	50	126
5		170	202			152		62				
6	263	274	819	174	174	702	135	135	258	125	125	186
7	352	356		177	179							
8	857	868	2106	625	625	1944	625	625	660	625	625	626
9		900	2592			2106						
10	3125	3125	3564	780	780	2592			864			
11			5346	1000	1125	3564		730	1296			864
12		3645	8423	3125	3125	5184	3125	3125	3126		730	1296
13	5000	6562	14124		3645	7776					1125	1944
14	15625	15625	44172	4096	6562	11664		3645	4374	3125	3125	3126
15				5000	10125	17496	4096	6562	6562		3645	4374
16				15625	18225	42984	15625	15625	15626	15625	15625	15626

Table 2. Lower bounds on k for $\mathsf{SHF}_{\delta}(x + \delta; k, v, \{1, d\})$ for $v \in \{5, 6\}$ and $d \in \{2, 2^{\circ}\}$.

$[20, 6, 12]_5$ [27]. Lemmas 3.5 and 3.1 are applied. In order to apply Theorem 3.7, we employ SHFs on larger alphabets. The primary source of these is the following standard construction of orthogonal arrays.

Theorem 3.11. *Let q be a prime power, where $q \geq s \geq 2$, and $d \geq 1$. Then an* $\mathsf{OA}_1(q^s; s, q + 1, q)$ *exists. Hence, in addition, for $1 \leq \delta \leq (q + 1) - (s - 1)d$, an* $\mathsf{SHF}_{\delta}((s - 1)d + \delta; q^s, q, \{1, d\})$ *exists.*

Proof. The construction of the $\mathsf{OA}_1(q^s; s, q + 1, q)$ is very well known [28, 38, 39], but we repeat it here. To form the orthogonal array A, index the q^s rows by the q^s polynomials with coefficients in $\mathbb{F}_q$ of degree less than s. Index the columns by elements of $\mathbb{F}_q \cup \{\infty\}$. In a row indexed by the polynomial $f(x) = \sum_{i=0}^{s-1} a_i x^i$, and column indexed by r, place the entry $f(r)$ when $r \in \mathbb{F}_q$, or the entry a_{s-1} when $r = \infty$. This is the required orthogonal array. To form the SHF, B, first select $R \subseteq \mathbb{F}_q \cup \{\infty\}$ with $|R| = (s - 1)d + \delta$, and let A_R be the array obtained from A by including only columns whose indices are in R. Transpose A_R to form B. Then two columns can agree in at most $s - 1$ rows, and so one column disagrees with each of d other columns in at least δ rows, so we have the desired SHF. ■

Despite applying each of the constructions discussed thus far, for many parameter sets the bound obtained is weak. We also employ a heuristic computational method using random extension (as in [16]) to establish lower bounds on k for these situations. We expect that many or most of the entries can be increased, particularly when $v = 6$.

3.5. Separating hash families: $d = 3$

For separations $\{1, 3\}$ and $\{1, 3^\circ\}$, we report results for $v \in \{5, 6\}$ in Table 3. We follow the same strategy as when $d = 2$. In this case, we use linear codes with parameters $[11, 3, 8]_5$, $[22, 4, 16]_5$, $[25, 5, 17]_5$, $[29, 5, 20]_5$, $[34, 5, 24]_5$, $[16, 4, 12]_8$, and $[16, 4, 12]_9$ [27]. When the number of symbols is not a prime power, we also apply the natural extension of Theorem 3.11 to transversal designs and incomplete transversal designs (see [13], for example). A $(6, 225, 5)_{15}$ code results from the existence of four mutually orthogonal latin squares of order 15 [49], and a $(6, 98, 5)_{10}$ code results from four mutually incomplete orthogonal latin squares of order 10 with a hole of size 2 [8].

It is worthwhile to remark that in order to produce an $\mathsf{SHF}_3(N; 5^6, 5, \{1, 3\})$ one could use an $[N, 6, \Delta]_5$ linear code with $3\Delta - 2N \geq 3$. According to [27], the smallest N for which such a linear code is known has $N = 39$. Nevertheless, an $\mathsf{SHF}_3(9; 25^3, 25, \{1, 3\})$ and an $\mathsf{SHF}_1(4; 25, 5, \{1, 3\})$ both exist by Theorem 3.11, and hence an $\mathsf{SHF}_3(36; 5^6, 5, \{1, 3\})$ exists by Theorem 3.7.

4. Constructing detecting arrays from hash families

Now we return to detecting arrays. Perhaps the easiest connection with hash families is the following.

Lemma 4.1. *If an* $\mathsf{SHF}_1(N; k, v, \{1, 1\})$ *exists, a* $\mathsf{DA}_1(v(N + 1); 1, 1, k, v, 1)$ *exists.*

Proof. Form the $\mathsf{SHF}_1(N; k, v, \{1, 1\})$ on symbols Σ_v, append a row consisting of all 0s, and apply the action of the cyclic group $\mathbb{Z}_v$. To verify that this works, consider a column γ and a symbol σ, and let R be the rows in which σ appears in column γ. Let $\gamma' \neq \gamma$. Some row in R in the orbit of the all-0 row contains σ in column γ'. A different row of R contains a symbol not equal to σ in column γ', from the orbit of a row of the SHF in which columns γ and γ' contain different symbols. ■

In [20], a second approach is explored, there called *h-inflation*, which uses an $\mathsf{SHF}_\delta(N; k, v + 1, \{1, d\})$ to make a detecting array on v symbols. In [21], yet another approach is developed for general d and general t; when $t = 1$ it employs an array that is equivalent to an $\mathsf{SHF}_\delta(N; k, v, \{1, d^\circ\})$. Rather than reviewing each approach, we develop a common generalization of all three, in the case that $t = 1$. Later we revisit these constructions.

	$\delta = 1$			$\delta = 2$			$\delta = 3$			$\delta = 4$		
x	$(5,3)$	$(5,3^\circ)$	$(6,3)$	$(5,3)$	$(5,3^\circ)$	$(6,3)$	$(5,3)$	$(5,3^\circ)$	$(6,3)$	$(5,3)$	$(5,3^\circ)$	$(6,3)$
1	8	10	10	6	7	7						
2	12	15	15	8	10	10	6	8	8	6	7	7
3	25	25	34	25	25	30	25	25	27	15	15	19
4	26	36	37		28	32			28		16	
5	29	54	54		41	41			29		20	20
6	57	67	132	50	50	126	41	41	72	25	25	64
7	64	98									27	
8		103		64	98			50			31	
9	125	125	186	125	125	132	65	98	108	65	65	78
10		150	216			144					82	82
11	200	225	324		130	216	80		144			108
12	320	405	552	200	225	552	125	125	216	125	125	192
13	512	730	730	320	405		200	225	324			216
14		760	864	512	730	730	320	405	486	200	225	324
15	625	830	1296				512	730	730	320	405	486
16		1125	1944		760	760				512	730	730
17	807	2025	2916		830	1130						
18	1270	3645	4374	625	910	1866	625	760	1033	625		
19	2032	6562	6562		2578	3092		1186	1701			
20		6567	6567	704	5141	5141		2345	2813		760	760
21		6833	6833	968	5651	5651		4700	4700		830	841
22		7515	7515	1332	6214	6214	704	5151	5151		1043	1043
23	2179	8265	8265		6833	6833	968	5655	5655		1304	1304
24	3125	9091	9091	1600	7515	7515	1332	6214	6214	704	1663	1663
25	4096	10000	10240	2560	8265	8265	1600	6833	6833	1000	2134	2134
26	5632	13000	15360	4096	9091	9091	2560	7515	7515	1600	2749	2916
27	15625	16900	32770		10000	10000	4096	8265	8265	2560	3645	4374
28		22926		5632	13000	13000		9091	9091	4096	6562	6562
29		34386	34386	7744	16900	16900		10000	10000			
30				15625	21970	32770	5632	13000	13000		7696	7696
31					32250		7744	16900	16900		10000	10000
32							10648	21970	21970		13000	13000
33							15625	30255	32770	4516	16900	16900
34										6753	21970	21970
35										10118	28563	28563
36										15625		32770

Table 3. Lower bounds on k for $\mathsf{SHF}_\delta(x + \delta; k, v, \{1, d\})$ for $v \in \{5, 6\}$ and $d \in \{3, 3^\circ\}$.

To begin, we extend the notion of detecting arrays for single factors to permit a column in which not all symbols need appear. A $\mathsf{DA}_\delta(N; d, 1, k^\circ, v)$ is an $N \times (k+1)$ array A in which the columns contain symbols from Σ_v, where the first k columns are indexed by Σ_k and the last is indexed by $\circ$, so that when $T = \{\gamma\}$ with $0 \le \gamma < k$, $|\rho_A(T) \setminus \rho_A(\mathcal{T})| < \delta \Leftrightarrow T \in \mathcal{T}$ whenever $\mathcal{T} \subseteq \mathcal{I}_1$, and $|\mathcal{T}| = d$. In plain English, we require that each of the first k columns be separated from any set of d other columns (possibly including the last) at least δ times, *but no such requirement is placed on the last column.*

Evidently, every $\mathsf{DA}_\delta(N; d, 1, k+1, v)$ is a $\mathsf{DA}_\delta(N; d, 1, k^\circ, v)$; moreover, by deleting column $k+1$, every $\mathsf{DA}_\delta(N; d, 1, k^\circ, v)$ yields a $\mathsf{DA}_\delta(N; d, 1, k, v)$. Incorporating corroboration as a parameter parallels the definitions provided at the outset.

The general construction that we use for detecting arrays for single factor effects follows.

Construction 4.2. *Suppose that there exist*

(1) *an* $\mathsf{SHHF}_\delta(N; k, (\ell_1, \ldots, \ell_N), \{1, d^\circ\})$, *and*

(2) *a* $\mathsf{DA}_\beta(M_j; d, 1, \ell_j^\circ, v, s)$ *for each* $1 \le j \le N$.

Then a $\mathsf{DA}_{\beta\delta}(\sum_{j=1}^N M_j; d, 1, k, v, s)$ *exists.*

Proof. Let the symbols of the $\mathsf{SHHF}_\delta(N; k, (\ell_1, \ldots, \ell_N), \{1, d^\circ\})$, A, in row j be $\Sigma_{\ell_j} \cup \{\circ\}$. For $1 \le j \le N$, let B_j be a $\mathsf{DA}_\beta(M_j; d, 1, \ell_j^\circ, v, s)$, in which the first ℓ_j columns are indexed by Σ_{ℓ_j}, and the last by $\circ$. (There is a natural bijection between the symbols in row j of A and the column indices of B_j.) Replace each symbol of row j of A by the corresponding column of B_j to form an $M_j \times k$ array, E_j, on v symbols, for $1 \le j \le N$. Then vertically juxtapose $E_1, \ldots, E_N$ to form E. For any column γ of E and any set C_2 of d disjoint columns of A, there are (at least) δ rows $\{\rho_1, \ldots, \rho_\delta\}$ in which column γ contains a non-$\circ$ symbol ψ, and the columns of C_2 contain symbols S with $\psi \notin S$. Let $\nu \in \Sigma_v$. For each $1 \le j \le \delta$, in E_{ρ_j} there is a set R of $r \ge d + \beta$ rows in which the column ψ (arising from symbol ψ of row ρ_j of A) contains ν so that no selection $\mathcal{T}$ of d (column,value) pairs within the columns of S have $|\rho_{E_{\rho_j}}(R) \setminus \rho_{E_{\rho_j}}(\mathcal{T})| < \beta$. This establishes the desired separation. Corroboration is limited by the number of *distinct* witnesses; each of $E_{\rho_1}, \ldots, E_{\rho_\delta}$ ensures corroboration s individually, but each may employ the same witnesses. Hence the corroboration of E is (at least) s. ■

Using an $\mathsf{SHHF}_\delta(N; k, \kappa, \{1, d\})$ and a $\mathsf{DA}_\beta(M; d, 1, \kappa, v, s)$, the variant of Construction 4.2 enables one to use ingredient arrays not involving $\circ$.

Although we have already explored constructing the $\mathsf{SHF}_\delta(N; k, \kappa, \{1, d^\circ\})$, the effective application of Construction 4.2 requires that we establish the existence of suitable $\mathsf{DA}_\delta(M; d, 1, \kappa^\circ, v, s)$s, at least for small values of κ. We resort to one basic construction using orthogonal arrays.

Lemma 4.3. *Suppose that an* $\mathsf{OA}_1(q^2; 2, q+1, q)$ *exists. Let* $d \geq 1$*,* $\delta \geq 1$*, and* $s \geq 1$ *satisfy* $sd + \delta \leq q$*. Then a* $\mathsf{DA}_\delta((sd+\delta)q; 1, d, q^\circ, q, s)$ *exists.*

Proof. Let R be the $\mathsf{OA}_1(q^2; 2, q+1, q)$. Set $\phi = sd + \delta$. Choose any set L_ϕ of ϕ symbols in the last column, and delete all rows from R that contain a symbol not in L_ϕ to form an array S having ϕq rows. In the last column, each of ϕ symbols appears q times; in each of the remaining columns of S, every symbol appears precisely ϕ times. Let $T = \{(\gamma, \nu)\}$ with $0 \leq \gamma < q$ and $\nu \in \Sigma_q$. Consider the ϕ rows in $\rho_S(T)$; these rows agree in column γ *but in no other column*. No matter how fewer than s fusions are performed in at most d columns to form array A (so that ν remains a symbol in column γ), it follows that $|\rho_A(T)| \geq \phi - (s-1)d$, and moreover that $\rho_A(T)$ contains a set of at least $\phi - (s-1)d$ rows that mutually disagree on all other columns. Because $\phi - (s-1)d \geq d + \delta$, S is a $\mathsf{DA}_\delta((sd+\delta)q; 1, d, q^\circ, q, s)$. ■

Lemma 4.3 produces arrays that need not contain all q symbols in the final column. In these cases, we obtain a $\mathsf{DA}_\delta((sd+\delta)q; 1, d, q^\circ, q, s)$ but not a $\mathsf{DA}_\delta((sd+\delta)q; 1, d, q+1, q, s)$. To obtain various DAs on $q+1$ symbols, we employ definitions and results from finite projective geometry. (For relevant background, see [29,30].) In the projective plane $PG(2, q)$, an (n, r)-*arc* is a set A of n points with at most r on a line; the largest n for which there is an (n, r)-arc in $PG(2, q)$ is denoted by $m_r(2, q)$. A t-*blocking set* in $PG(2, q)$ is a set B of points so that every line meets B in at least t points. Whenever A is an (m, n)-arc in $PG(2, q)$, $B = PG(2, q) \setminus A$ is a $(q+1-n)$-blocking set of size $q^2 + q + 1 - m$.

Lemma 4.4. *Let* q *be a prime power. Let* $d \geq 1$*,* $\delta \geq 1$*, and* $s \geq 1$ *satisfy* $sd + \delta \leq q$*. Then a* $\mathsf{DA}_\delta(q^2 - m_{q-sd-\delta}(2, q); 1, d, q+1, q, s)$ *exists.*

Proof. Use an $(m, q - sd - \delta)$-arc in $PG(2, q)$ with $m = m_{q-sd-\delta}(2, q)$ to form an $(sd + \delta + 1)$-blocking set of size $q^2 + q + 1 - m$. The dual blocking set (i.e., the configuration obtained from the blocking set by interchanging points and lines) is a set of $q^2 + q + 1 - m$ lines so that every point belongs to at least $sd + \delta + 1$. Delete any point and the $q + 1$ lines through it to form a set of (at least) $q^2 - m$ lines so that every remaining point belongs to at least $sd + \delta$. Use the $q + 1$ deleted lines, omitting the deleted point, to form the columns of the DA.

Let $T = \{(\gamma, \nu)\}$ with $0 \leq \gamma < q + 1$ and $\nu \in \Sigma_q$. Then $\rho(T)$ contains $sd + \delta$ rows that agree in column γ but in no other column. So similar arguments to those used in the proof of Lemma 4.3 show that the array is in fact a DA with the required parameters. ■

When $q = sd + \delta$, Lemma 4.4 yields precisely q^2 rows, the entire orthogonal array. Exact values for $m_r(2, q)$ are not known in general and form the focus of much research. For our running examples with $q = 5$, however, exact values are known:

$m_0(2,5) = 0, m_1(2,5) = 1, m_2(2,5) = 6, m_3(2,5) = 11, m_4(2,5) = 16, m_5(2,5) = 25$, and $m_6(2,5) = 31$ [29, Table 25].

Lemmas 4.3 and 4.4, together with Construction 4.2, unify earlier constructions, as follows. The h-inflation developed in [20] is equivalent to applying Construction 4.2 using the DAs from Lemma 4.4 along with an $\mathsf{SHF}_\delta(N; k, q+1, \{1, d\})$. The method of [21] is equivalent *when restricted to* $t = 1$ to applying Construction 4.2 using the DAs from Lemma 4.3 along with an $\mathsf{SHF}_\delta(N; k, q, \{1, d^\circ\})$. Instead by removing the last column of the DA from Lemma 4.3 and using an $\mathsf{SHF}_\delta(N; k, q, \{1, d\})$, we recover a construction in the same vein as Lemma 4.1. However, there are important differences. First, Lemma 4.1 needs no assumption that v is a prime power. More importantly, where the application of Construction 4.2 requires a DA_1 (having at least $2q$ rows), Lemma 4.1 instead modifies the SHF by adding an all-0 row, so that instead of the DA_1 we can employ only q rows. To reconcile this apparent discrepancy, form the $\mathsf{DA}_1(2q; 1, 1, q, q, 1)$ so that it contains all q constant rows (rows in which all symbols are the same). Apply Construction 4.2 using an $\mathsf{SHF}_1(N; k, q, \{1, d\})$ to form an $2Nq \times q$ array E. The manner of construction ensures that each of the constant rows appears (at least) N times in E. Because these are only useful when $\mathcal{T}$ contains no main effects using the symbol used in T, $N - 1$ copies of each of these constant rows are unnecessary and can be deleted. This recovers Lemma 4.1 (when q is a prime power), and indeed leads to a useful generalization of Construction 4.2.

By choosing symbol 0 to be in L_ϕ, Lemma 4.3 produces a $\mathsf{DA}_\delta((sd + \delta)q; 1, d, q, q, s)$ having q constant rows. Using this, we provide a construction (stated for the homogeneous case) for corroboration $s = 1$.

Construction 4.5. *If an* $\mathsf{SHF}_\delta(N; k, \kappa, \{1, d\})$ *and a* $\mathsf{DA}_\beta(M; d, 1, \kappa, v, 1)$ *having* v *constant rows exist, a* $\mathsf{DA}_{\beta\delta}(NM - (N - \beta\delta)v; d, 1, k, v, 1)$ *exists.*

Each replacement of columns of the DA into a row of the SHF yields (at least) v constant rows. Then the verification follows that of Construction 4.2, after deleting all but $\beta\delta$ copies of each constant row. We leave the details to the reader. Instead, we explore a powerful construction employing the detecting arrays of Lemma 4.3, restricting to a prime power number of symbols. A row in $\Sigma_v^{\kappa+1}$ is *nearly constant* if each of the first κ entries contains the same symbol, and the last entry is 0.

Construction 4.6. *Let* q *be a prime power. If an* $\mathsf{SHF}_\delta(N; k, q, \{1, d^\circ\})$ *exists, then when* $d + \beta \leq q$*, a* $\mathsf{DA}_{\beta\delta}((N(d + \beta - 1) + \delta)q; d, 1, k, q, 1)$ *exists.*

Proof. Let the symbols of the $\mathsf{SHF}_\delta(N; k, q, \{1, d^\circ\})$, A, be $\Sigma_q \cup \{\circ\}$. Form a $\mathsf{DA}_\beta((d+\beta)q; d, 1, q^\circ, q, 1)$, B, using Lemma 4.3 choosing $L_\phi = \{0, \ldots, d+\beta-1\}$. Let the first q columns of the $\mathsf{DA}_\beta(M; d, 1, q^\circ, q, 1)$ be indexed by Σ_q, and index the $(q+1)$st by $\circ$. The q rows containing 0 in the last column are nearly constant. Remove them from B to form a $(d + \beta - 1)q \times (q + 1)$ array B'. Replace each symbol of row j of A by the corresponding column of B' to form an $(d+\beta-1)q \times k$ array, E_j, on q symbols, for $1 \leq j \leq N$. Form a $q \times k$ array D_1 consisting of each *constant* row. For $1 \leq i \leq \delta$, let D_i be a copy of D. Then vertically juxtapose $E_1, \ldots, E_N$ and $D_1, \ldots, D_\delta$ to form F.

To verify that F is the desired $\mathsf{DA}_{\beta\delta}((N(d+\beta-1)+\delta)q; d, 1, k, q, 1)$, consider $T = \{(\gamma, \nu)\}$. It is necessary and sufficient that in every column $g \neq \gamma$, and every set X of d symbols, at least $\beta\delta$ rows of F contain ν in column γ but contain no symbol of X in column g. To establish this for a specific T and g, partition the N rows of A into classes, as follows:

(1) A_1 contains the τ_1 rows in which columns γ and g contain distinct symbols from Σ_q;

(2) A_2 contains the τ_2 rows in which column γ contains a symbol from Σ_q, and column g contains $\circ$;

(3) A_3 contains the τ_3 rows in which column g contains a symbol from Σ_q, and column γ contains $\circ$;

(4) A_4 contains the τ_4 rows in which columns γ and g contain the same symbol from Σ_q;

(5) A_5 contains the τ_5 rows in which columns γ and g both contain $\circ$.

The separation requirements of the SHF ensure that $\tau_1 + \tau_2 \geq \delta$ in order to separate column γ from column g, and that $\tau_1 + \tau_3 \geq \delta$ in order to separate column g from column γ.

Next we define disjoint classes of rows of F as follows.

(1) For $1 \leq j \leq \min(\delta, \tau_1)$, F_j contains $(d+\beta)q$ rows of F consisting of
- the $(d+\beta-1)q$ arising from the jth row of A_1, and
- the q rows of D_j.

(2) For $1 \leq j \leq \delta - \tau_1$, G_j contains $(2d+2\beta-1)q$ rows of F consisting of
- the $(d+\beta-1)q$ arising from the jth row of A_2,
- the $(d+\beta-1)q$ arising from the jth row of A_3, and
- the q rows of $D_{j+\tau_1}$.

It suffices to check that in each of $F_1, \ldots, F_{\tau_1}$ and each of $G_1, \ldots, G_{\delta-\tau_1}$, at least $d+\beta$ rows have ν in column γ and distinct symbols in column g. Let us check F_j. In the $(d+\beta-1)q$ rows arising from the column replacement of B' into the jth row of A_1, each ν in column γ appears in exactly $d+\beta-1$ rows, with a different symbol in column g in each. Because nearly constant rows have been deleted to form B', none of these $d+\beta-1$ symbols is ν, so the row from D_j that is constant equal to ν provides the final symbol in column g. Initially, we proceed in the same manner for G_j. In the $(d+\beta-1)q$ rows arising from the column replacement of B' into the jth row of A_2, each ν in column γ appears in exactly $d+\beta-1$ rows; in these rows, column g contains each of $\{1,\dots,d+\beta-1\}$. When $\nu \not\in \{1,\dots,d+\beta-1\}$, the all-$\nu$ row from $D_{j+\tau_1}$ provides the final row needed. When $\nu \in \{1,\dots,d+\beta-1\}$, consider the $(d+\beta-1)q$ rows arising from the column replacement of B' into the row chosen from A_3 (these rows have not been considered before). In these rows, every $\nu \in \{1,\dots,d+\beta-1\}$ appears in column γ with *every* symbol of Σ_q in column g. This completes the verification. ■

When the DA from Lemma 4.3 is used, Construction 4.6 improves on Construction 4.5. By imposing the further condition on the SHF that the class A_4 of rows not be empty, one could eliminate some of the constant rows added. Even without this restriction, Construction 4.6 may include more constant rows than are needed in certain cases. We do not pursue this further here.

5. Consequences

Now we consider some consequences of Constructions 4.2 and 4.6 using detecting arrays from Lemmas 4.3 and 4.4, along with the SHFs tabulated in Tables 1, 2, and 3, and SHHFs produced from these by Lemma 3.1 (4). Of course we do not attempt to list all of the detecting arrays generated; instead we compare different approaches. Our interest is in constructing detecting arrays for complex engineered systems of moderate to large sizes. In Table 4, we report upper bounds on the number N of rows in a $\mathsf{DA}_\delta(N; 1, d, k, 5)$ (with corroboration 1) for various values of d, k, and δ.

In constructing Table 4, we apply Lemma 3.1 (7) and the observation that a DA need not have more rows than a DA having larger index but the same parameters otherwise.

The effectiveness of the methods employed in producing detecting arrays for single factor effects with many columns enables us to produce such arrays for larger systems. Although we do not expect that the arrays found have the fewest possible rows in general, it is striking how few rows suffice for large numbers of columns.

Comparing the results from Constructions 4.2 and 4.6 in Table 4, one finds that Construction 4.6 almost always yields the fewest rows. Perhaps this is no surprise, because Construction 4.6 typically succeeds in eliminating many rows using the nearly constant rows of the detecting array ingredient. Despite this, Construction 4.2 often remains competitive, because it uses hash families in which the unusual $\circ$ symbol does not appear, and which can be heterogeneous. Indeed Construction 4.2 can lead to the better result, as we illustrate next. Using an $\mathsf{SHF}_2(8; 126, 6, \{1, 3\})$ and $\mathsf{DA}_2(25; 1, 3, 6, 5)$, Construction 4.2 yields a $\mathsf{DA}_4(200; 1, 3, 100, 5)$. Using an

		$k = 10$		$k = 100$		$k = 1000$		$k = 10000$	
d	δ	4.6,4.3	4.2,4.4	4.6,4.3	4.2,4.4	4.6,4.3	4.2,4.4	4.6,4.3	4.2,4.4
1	1	15	20	20	30	30	50	35	60
1	2	25	30	30	40	40	60	45	70
1	3	35	40	40	50	50	70	60	90
1	4	40	45	50	60	60	80	70	100
1	8	70	75	80	90	100	120	120	125
2	1	25	38	55	75	115	165	155	225
2	2	35	48	70	90	150	186	190	270
2	3	45	50	100	120	165	225	205	285
2	4	70	80	100	120	200	225	220	300
2	8	110	120	170	188	290	360	320	400
3	1	35	48	155	164	290	380	425	560
3	2	45	50	175	175	325	400	490	640
3	3	90	100	230	200	360	500	555	720
3	4	90	100	230	200	410	500	620	775
3	8	160	175	400	400	540	650	820	950

Table 4. Upper bounds on N for a $\mathsf{DA}_\delta(N; 1, d, k, 5)$. Two upper bounds are given for each. The first employs Construction 4.6 using Lemma 4.3, and SHFs of type $\{1, d^\circ\}$ with $v = 5$. The second employs Construction 4.2 using Lemma 4.4, and SHHFs of type $\{1, d\}$ with five or six symbols per row whose existence is implied by the SHF tables.

$\mathsf{SHF}_2(11; 125, 5, \{1, 3^\circ\})$ and $\mathsf{DA}_2(25; 1, 3, 5^\circ, 5)$, Construction 4.6 yields a $\mathsf{DA}_4(230; 1, 3, 100, 5)$. The hash family with 6 symbols is enough smaller than that with five symbols in addition to $\circ$ that the usual advantage of exploiting nearly constant rows is overcome. Naturally finding hash families with fewer rows might impact such comparisons. Although we do not believe that the hash families here have the fewest rows (or the most columns), we do believe that Construction 4.2 can, in certain cases, yield fewer rows than Construction 4.6.

Potential improvements in the sizes of the detecting arrays could result from finding better SHFs and SHHFs. They could also arise from a more detailed analysis of the redundant rows produced by Constructions 4.2 and 4.6; for this purpose, a post-optimization strategy from [18] may prove useful computationally, but we have not employed that here. In this paper, we applied the constructions only to DAs from Lemmas 4.3 and 4.4, which have v or $v + 1$ columns. Naturally, the same constructions can be applied to DAs having more columns (permitting the hash families to have more symbols and hence fewer rows). We expect that such an extension would be effective, given a larger collection of detecting arrays to use as ingredients.

Of most importance from the standpoint of applications is that the column replacement techniques and associated ingredients developed underlie effective and efficient methods to produce detecting arrays for the effects of single factors so that specified values of separation and corroboration can be achieved. Finally, many of the techniques developed here extend in a natural manner to detecting t-way interactions, not just the effects of single factors [21, 22].

Acknowledgments. Thanks to Yasmeen Akhtar, Randy Compton, Erin Lanus, and Stephen Seidel for helpful discussions. Special thanks to an anonymous referee for improvements in the presentation.

Funding. This research was supported by NSF grant #1813729, and by the Software Test & Analysis Techniques for Automated Software Test program by OPNAV N-84, U.S. Navy.

References

[1] A. N. Aldaco, C. J. Colbourn, and V. R. Syrotiuk, Locating arrays: A new experimental design for screening complex engineered systems. *SIGOPS Oper. Syst. Rev.* **49** (2015), no. 1, 31–40

[2] N. Alon, G. Cohen, M. Krivelevich, and S. Litsyn, Generalized hashing and parent-identifying codes. *J. Combin. Theory Ser. A* **104** (2003), no. 1, 207–215 Zbl 1036.94015 MR 2018429

[3] N. Alon and J. H. Spencer, *The Probabilistic Method. With an appendix on the life and work of Paul Erdős*. 3rd edn., Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Hoboken, NJ, 2008 Zbl 1148.05001 MR 2437651

[4] M. Atici, S. S. Magliveras, D. R. Stinson, and W.-D. Wei, Some recursive constructions for perfect hash families. *J. Combin. Des.* **4** (1996), no. 5, 353–363 Zbl 0914.68087 MR 1402122

[5] A. Barg, G. Cohen, S. Encheva, G. Kabatiansky, and G. Zémor, A hypergraph approach to the identifying parent property: the case of multiple parents. *SIAM J. Discrete Math.* **14** (2001), no. 3, 423–431 Zbl 1011.94014 MR 1857594

[6] S. R. Blackburn, Frameproof codes. *SIAM J. Discrete Math.* **16** (2003), no. 3, 499–510 Zbl 1041.68063 MR 2002175

[7] G. T. Bogdanova and P. R. J. Östergård, Bounds on codes over an alphabet of five elements. *Discrete Math.* **240** (2001), no. 1-3, 13–19 Zbl 1005.94032 MR 1855043

[8] A. E. Brouwer, Four MOLS of order 10 with a hole of order 2. *J. Statist. Plann. Inference* **10** (1984), no. 2, 203–205 Zbl 0553.05022 MR 760405

[9] R. A. Brualdi, Matrices of zeros and ones with fixed row and column sum vectors. *Linear Algebra Appl.* **33** (1980), 159–231 Zbl 0448.05047 MR 585770

[10] R. C. Bryce and C. J. Colbourn, The density algorithm for pairwise interaction testing. *Software Testing, Verification, and Reliability* **17** (2007), 159–182

[11] R. C. Bryce and C. J. Colbourn, A density-based greedy algorithm for higher strength covering arrays. *Software Testing, Verification, and Reliability* **19** (2009), 37–53

[12] Y. Chang, C. J. Colbourn, A. Gowty, D. Horsley, and J. Zhou, New bounds on the maximum size of Sperner partition systems. *European J. Combin.* **90** (2020), 103165, 18 Zbl 1458.05028 MR 4125527

[13] C. J. Colbourn, Covering arrays and hash families. In *Information Security, Coding Theory and Related Combinatorics*, pp. 99–135, NATO Sci. Peace Secur. Ser. D Inf. Commun. Secur. 29, IOS, Amsterdam, 2011 Zbl 1341.68134 MR 2963127

[14] C. J. Colbourn, D. Horsley, and V. R. Syrotiuk, A hierarchical framework for recovery in compressive sensing. *Discrete Appl. Math.* **236** (2018), 96–107 Zbl 1431.94014 MR 3739778

[15] C. J. Colbourn, G. Kéri, P. P. R. Soriano, and J.-C. Schlage-Puchta, Covering and radius-covering arrays: constructions and classification. *Discrete Appl. Math.* **158** (2010), no. 11, 1158–1180 Zbl 1231.05033 MR 2629893

[16] C. J. Colbourn, E. Lanus, and K. Sarkar, Asymptotic and constructive methods for covering perfect hash families and covering arrays. *Des. Codes Cryptogr.* **86** (2018), no. 4, 907–937 Zbl 1383.05045 MR 3770276

[17] C. J. Colbourn and D. W. McClary, Locating and detecting arrays for interaction faults. *J. Comb. Optim.* **15** (2008), no. 1, 17–48 Zbl 1149.90090 MR 2375213

[18] C. J. Colbourn and P. Nayeri, Randomized post-optimization for t-restrictions. In *Information Theory, Combinatorics, and Search Theory*, pp. 597–608, Lecture Notes in Comput. Sci. 7777, Springer, Heidelberg, 2013 Zbl 1309.05034 MR 3076131

[19] C. J. Colbourn and V. R. Syrotiuk, On a combinatorial framework for fault characterization. *Math. Comput. Sci.* **12** (2018), no. 4, 429–451 Zbl 1433.68268 MR 3870157

[20] C. J. Colbourn and V. R. Syrotiuk, Detecting arrays for main effects. In *Algebraic Informatics*, pp. 112–123, Lecture Notes in Comput. Sci. 11545, Springer, Cham, 2019 Zbl 1434.68340 MR 3976191

[21] C. J. Colbourn and V. R. Syrotiuk, Covering strong separating hash families. In *Finite Fields and Their Applications*, pp. 189–198, De Gruyter Proc. Math., De Gruyter, Berlin, 2020 Zbl 1466.05019 MR 4204971

[22] C. J. Colbourn and V. R. Syrotiuk, There must be fifty ways to miss a cover. In *50 Years of Combinatorics, Graph Theory, and Computing*, pp. 319–333, Discrete Math. Appl. (Boca Raton), CRC Press, Boca Raton, FL, 2020 Zbl 1451.05190 MR 4368178

[23] C. J. Colbourn and J. Torres-Jimenez, Heterogeneous hash families and covering arrays. In *Error-Correcting Codes, Finite Geometries and Cryptography*, pp. 3–15, Contemp. Math. 523, Amer. Math. Soc., Providence, RI, 2010 Zbl 1226.05061 MR 2766009

[24] R. Compton, M. T. Mehari, C. J. Colbourn, E. De Poorter, and V. R. Syrotiuk, Screening interacting factors in a wireless network testbed using locating arrays. In *IEEE INFOCOM International Workshop on Computer and Networking Experimental Research Using Testbeds (CNERT)*, IEEE Press, 2016

[25] P. Erdős and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions. In *Infinite and Finite Sets (Colloq., Keszthely, 1973; Dedicated to P. Erdős on His 60th Birthday), Vol. II*, pp. 609–627, North-Holland, Amsterdam, 1975 Zbl 0315.05117 MR 0382050

[26] A. Gowty and D. Horsley, More constructions for Sperner partition systems. *J. Combin. Des.* **29** (2021), no. 9, 579–606 MR 4284176

[27] M. Grassl, Bounds on the minimum distance of linear codes and quantum codes. Available at http://www.codetables.de. Accessed on 2019-08-30

[28] A. S. Hedayat, N. J. A. Sloane, and J. Stufken, *Orthogonal Arrays*. Springer Ser. Statist., Springer, New York, 1999 Zbl 0935.05001 MR 1693498

[29] J. W. P. Hirschfeld and L. Storme, The packing problem in statistics, coding theory and finite projective spaces. *J. Statist. Plann. Inference* **72** (1998), no. 1-2, 355–380 Zbl 0958.51013 MR 1655203

[30] J. W. P. Hirschfeld and L. Storme, The packing problem in statistics, coding theory and finite projective spaces: update 2001. In *Finite Geometries*, pp. 201–246, Dev. Math. 3, Kluwer Acad. Publ., Dordrecht, 2001 Zbl 1025.51012 MR 2061806

[31] D. S. Johnson, Approximation algorithms for combinatorial problems. *J. Comput. System Sci.* **9** (1974), 256–278 Zbl 0296.65036 MR 449012

[32] J. G. Kalbfleisch, R. G. Stanton, and J. D. Horton, On covering sets and error-correcting codes. *J. Combinatorial Theory Ser. A* **11** (1971), 233–250 Zbl 0181.22401 MR 290860

[33] D. R. Kuhn, R. Kacker, and Y. Lei, *Introduction to Combinatorial Testing*. CRC Press, Boca Raton, FL, 2013 Zbl 1272.68004

[34] A. Laaksonen and P. R. J. Östergård, New lower bounds on error-correcting ternary, quaternary and quinary codes. In *Coding Theory and Applications*, pp. 228–237, Lecture Notes in Comput. Sci. 10495, Springer, Cham, 2017 Zbl 1429.94090 MR 3705120

[35] E. Lanus, C. J. Colbourn, and D. C. Montgomery, Partitioned search with column resampling for locating array construction. In *2019 IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pp. 214–223, IEEE Press, 2019

[36] P. C. Li and K. Meagher, Sperner partition systems. *J. Combin. Des.* **21** (2013), no. 7, 267–279 Zbl 1269.05107 MR 3055150

[37] L. Lovász, On the ratio of optimal integral and fractional covers. *Discrete Math.* **13** (1975), no. 4, 383–390 Zbl 0323.05127 MR 384578

[38] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes. I*. North-Holland Mathematical Library 16, North-Holland Publishing, Amsterdam, 1977 Zbl 0369.94008 MR 0465509

[39] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes. II*. North-Holland Mathematical Library 16, North-Holland Publishing, Amsterdam, 1977 Zbl 0369.94008 MR 0465510

[40] C. Martínez, L. Moura, D. Panario, and B. Stevens, Locating errors using ELAs, covering arrays, and adaptive testing algorithms. *SIAM J. Discrete Math.* **23** (2009/10), no. 4, 1776–1799 Zbl 1210.94126 MR 2570203

[41] S. Martirosyan and T. V. Trung, On t-covering arrays. *Des. Codes Cryptogr.* **32** (2004), no. 1-3, 323–339 Zbl 1046.05020 MR 2072336

[42] K. Meagher, L. Moura, and B. Stevens, A Sperner-type theorem for set-partition systems. *Electron. J. Combin.* **12** (2005), Note 20, 6 Zbl 1077.05097 MR 2180805

[43] K. Mehlhorn, *Data structures and algorithms. 1: Sorting and searching*. EATCS Monographs on Theoretical Computer Science, Springer, Berlin, 1984 Zbl 0556.68001 MR 756413

[44] R. A. Moser and G. Tardos, A constructive proof of the general Lovász local lemma. *J. ACM* **57** (2010), no. 2, Art. 11, 15 Zbl 1300.60024 MR 2606086

[45] C. Nie and H. Leung, A survey of combinatorial testing. *ACM Comput. Surv.* **43** (2011), no. 2, 29 Zbl 1293.68080

[46] S. Niskanen and P. R. J. Östergård, Cliquer user's guide, version 1.0. Tech. Rep. T48, Communications Laboratory, Helsinki University of Technology, Espoo, Finland, 2003

[47] X. Niu and H. Cao, Constructions and bounds for separating hash families. *Discrete Math.* **341** (2018), no. 9, 2627–2638 Zbl 1422.94033 MR 3828774

[48] P. Sarkar and D. R. Stinson, Frameproof and IPP codes. In *Progress in Cryptology—INDOCRYPT 2001 (Chennai)*, pp. 117–126, Lecture Notes in Comput. Sci. 2247, Springer, Berlin, 2001 Zbl 1011.94547 MR 1934490

[49] P. J. Schellenberg, G. H. J. Van Rees, and S. A. Vanstone, Four pairwise orthogonal Latin squares of order 15. *Ars Combin.* **6** (1978), 141–150 Zbl 0433.05014 MR 526098

[50] S. A. Seidel, M. T. Mehari, C. J. Colbourn, E. De Poorter, I. Moerman, and V. R. Syrotiuk, Analysis of large-scale experimental data from wireless networks. In *IEEE INFOCOM International Workshop on Computer and Networking Experimental Research Using Testbeds (CNERT)*, pp. 535–540, IEEE, 2018

[51] S. A. Seidel, K. Sarkar, C. J. Colbourn, and V. R. Syrotiuk, Separating interaction effects using locating and detecting arrays. In *International Workshop on Combinatorial Algorithms*, pp. 349–360, Springer, Cham, 2018 Zbl 06932716

[52] C. Shangguan, X. Wang, G. Ge, and Y. Miao, New bounds for frameproof codes. *IEEE Trans. Inform. Theory* **63** (2017), no. 11, 7247–7252 Zbl 1390.94884 MR 3724426

[53] C. Shi, Y. Tang, and J. Yin, The equivalence between optimal detecting arrays and super-simple OAs. *Des. Codes Cryptogr.* **62** (2012), no. 2, 131–142 Zbl 1283.05045 MR 2886266

[54] C. Shi, Y. Tang, and J. Yin, Optimum mixed level detecting arrays. *Ann. Statist.* **42** (2014), no. 4, 1546–1563 Zbl 1297.62177 MR 3262460

[55] C. Shi and C. M. Wang, Optimum detecting arrays for independent interaction faults. *Acta Math. Sin. (Engl. Ser.)* **32** (2016), no. 2, 199–212 Zbl 1331.05046 MR 3441302

[56] J. N. Staddon, D. R. Stinson, and R. Wei, Combinatorial properties of frameproof and traceability codes. *IEEE Trans. Inform. Theory* **47** (2001), no. 3, 1042–1049 Zbl 1001.94032 MR 1829330

[57] S. K. Stein, Two combinatorial covering theorems. *J. Combinatorial Theory Ser. A* **16** (1974), 391–397 Zbl 0287.05002 MR 340062

[58] D. R. Stinson, T. van Trung, and R. Wei, Secure frameproof codes, key distribution patterns, group testing algorithms and related structures. *J. Statist. Plann. Inference* **86** (2000), no. 2, 595–617 Zbl 1054.94013 MR 1768292

[59] D. R. Stinson, R. Wei, and K. Chen, On generalized separating hash families. *J. Combin. Theory Ser. A* **115** (2008), no. 1, 105–120 Zbl 1131.68070 MR 2378859

[60] Y. Tang and J. X. Yin, Detecting arrays and their optimality. *Acta Math. Sin. (Engl. Ser.)* **27** (2011), no. 12, 2309–2318 Zbl 1260.05022 MR 2853789

[61] R. A. Walker II and C. J. Colbourn, Perfect Hash families: constructions and existence. *J. Math. Cryptol.* **1** (2007), no. 2, 125–150 Zbl 1128.05012 MR 2345113

Charles J. Colbourn
School of Computing and Augmented Intelligence, Arizona State University,
P.O. Box 878809, Tempe, AZ 85287, USA; colbourn@asu.edu

Violet R. Syrotiuk
School of Computing and Augmented Intelligence, Arizona State University,
P.O. Box 878809, Tempe, AZ 85287, USA; syrotiuk@asu.edu

DOI 10.4171/8ECM/39

Digital collections of examples in mathematical sciences

James H. Davenport

Abstract. Some aspects of computer algebra (notably computation group theory and computational number theory) have some good databases of examples, typically of the form “all the X up to size n”. But most of the others, especially on the polynomial side, are lacking such, despite the utility they have demonstrated in the related fields of SAT and SMT solving. We claim that the field would be enhanced by such community-maintained databases, rather than each author hand-selecting a few, which are often too large or error-prone to print, and therefore difficult for subsequent authors to reproduce.

1. Introduction

Mathematicians have long had useful collections, either of systematic data or examples. One of the oldest known such is the cuneiform tablet known as Plimpton 322, which dates back to roughly 1800BC; see [23, pp. 172–176], or a more detailed treatment in [42, 50]. This use of systematic tables of data spawned the development on logarithmic, trigonometric, and nautical tables: Babbage’s difference engine was intended to mechanise the production of such tables. But there were also tables of purely mathematical interest: the author recalls using an 1839 table of logarithms and what are now known as Zech logarithms [59] (but in fact they go back at least to [41]), i.e., tables of the function $\log x \mapsto \log(1+x)$, at least over $\mathbf{R}$; Jacobi’s table [34] was modulo p^n for all the prime powers $p^n < 1000$.

1.1. Data citation

Citation and referencing is an important point of modern scholarship—Harvard-style referencing is generally attributed to [43], and the history of *Science Citation Index* is described in [29]. It is well understood, and practically all research students, and many undergraduates, get lessons in article citation practices.

2020 Mathematics Subject Classification. Primary 00A35; Secondary 12-04, 20-04.
Keywords. Benchmarking, citation, OpenMath.

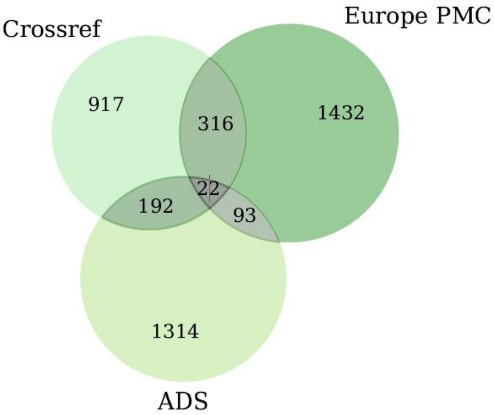


Figure 1. Overlaps between data citation harvesters [56, Figure 5].

Despite the success of article citation, data citation is a mess in practice [56]: only 1.16% of dataset DOIs in Zenodo are cited[1] (and 98.5% of these are self-citations). It is still a subject of some uncertainty: [36, 46] and significant changes are still being proposed [25]. Worse, perhaps, it is poorly harvested; see Figure 1. Assuming independence and looking at the overlap statistics, we can estimate that there are between 4,000 to 20,000 datasets waiting to be cited. In such circumstances, de facto people cite a paper if they can find one.

2. Pure mathematics

2.1. On-line Encyclopedia of Integer Sequences

This database [52] can be said to have "colonised the high ground" in mathematics: mathematicians from all sub-disciplines use it. It has evolved from a private enterprise, for a long time at http://www.research.att.com/∼njas/sequences, to a system maintained by a foundation, and now at https://oeis.org/. The recommended citation is "OEIS Foundation Inc. (2022), The On-Line Encyclopedia of Integer Sequences, published electronically at https://oeis.org, [date]", but the author had originally to search the website to find it!

2.2. Group theory

The classification of finite simple groups, as well as being a *tour de force* in mathematics, also means that we have a complete database here. In most other areas, we have to be content with "small" databases.

[1] In contrast, 60% of papers in Natural Science and Engineering *had* a citation in the next two years [39, 49].

An example of this is the transitive groups acting on n points, where various authors have contributed: [17] ($n \leq 11$); [51] ($n = 12$); [16] ($n = 14, 15$); [32] ($n = 16$); [33] ($17 \leq n \leq 31$); [18] ($n = 32$). These are available in the computer algebra system GAP (and MAGMA), except that (for reasons of space) $n = 32$ is not in the default build for GAP.

These are really great resources (if that is what you want), but how does one cite this resource: "[55, `transgrp` library]"?

There are several other libraries such as primitive groups. But it could be argued that (finite) group theory is "easy": for a given n, there are a finite number and we "just" have to list them.

2.3. L-functions and modular forms

The L-functions and modular forms database, known as LMFDB and hosted at lmfdb.org is a third example of mathematical databases. The recommended citation, "The LMFDB Collaboration, The L-functions and modular forms database, http://www.lmfdb.org, 2021" is directly linked from the home page, which is a good model to follow.

Computation in this area had a long history, from [9] and [54] to the current database, which is the work of a significant number of people. The early computations gave rise to the Birch–Swinnerton-Dyer conjectures [10], now a Clay Millennium Prize topic. The current computations are in active use by mathematicians; see Poonen's remarks in [27].

3. SAT and SMT solving

3.1. SAT solving

SAT solving is normally seen as solving a Boolean expression written in conjunctive normal form (CNF).

The 3-SAT problem is as follows: given a 3-literals/clause CNF satisfiability problem,

$$\underbrace{(l_{1,1} \vee l_{1,2} \vee l_{1,3})}_{\text{Clause 1}} \wedge (l_{2,1} \vee l_{2,2} \vee l_{2,3}) \wedge \cdots \wedge (l_{N,1} \vee l_{N,2} \vee l_{N,3}), \tag{1}$$

where $l_{i,j} \in \{x_1, \overline{x_1}, x_2, \overline{x_2}, \ldots\}$; is it satisfiable? In other words, is there an assignment of $\{T, F\}$ to the x_i such that all the clauses are *simultaneously* true.

3-SAT is the quintessential NP-complete problem [24]. 2-SAT is polynomial, and k-SAT for $k > 3$ is polynomial-transformable into 3-SAT. In practice, we deal with SAT—i.e., no limitations on the length of the clauses and no requirement that all clauses have the same length.

Let n be the number of i such that x_i (and/or $\overline{x_i}$) actually occur. Typically n is of a similar size to N.

Despite the problem class being NP-complete, nearly all examples are easy (e.g., SAT-solving has been routinely used in the German car industry for over twenty years [38]): either easily solved (SAT) or easily proved insoluble (UNSAT). For random problems there seems to be a distinct phase transition between the two [2,3,30], with the hard problems typically lying on the boundary.

This means that constructing difficult examples is itself difficult, and a topical research area [5,53].

SAT solving has many applications, so we want effective solvers for "real" problems, not just "random" ones. This gives us the fundamental question: what does this mean?

3.2. SAT contests

These are described at http://www.satcompetition.org. They have been run since 2002. In the early years, there were distinct tracks for industrial/handmade/random problems; this has been abandoned.

The methodology is that the organisers accept submissions (from contestants[2] and others), then produce a list of problems (in DIMACS, a standard format), set a time (and memory) limit, and see how many of the problems the submitted systems can solve on the contest hardware.

SAT is easy to certify (the solver just produces a list of values of the x_i). Verifying UNSAT is much harder, but since 2013 the contest has required proofs of UNSAT for the UNSAT track, and since 2020 in all tracks, in DRAT: a specified format (some of these proofs have been $> 100\,\mathrm{GB}$).

The general feeling is that these contests have really pushed the development of SAT solvers, roughly speaking ×2/year. For comparison, Linear Programming has done ×1.8 over a greater timeline and with more rigorous dcoumentation [11].

3.3. SMT: Life beyond SAT

Consider a theory T, with variables y_j, and various Boolean-valued statements in T of the form $F_i(y_1, \ldots, y_n)$, and a CNF $\mathcal{L}$ in the form of (1) with $F_i(y_1, \ldots, y_n)$ rather than just x_i. In principle, T can be anything: those currently supported[3] are given in Figure 2.

[2] In 2020, contestants were required to submit at least 20 problems, as well as a solver.

[3] By the SMT-LIB standard [6], which also says " New logics are added to the standard opportunistically, once enough benchmarks are available".

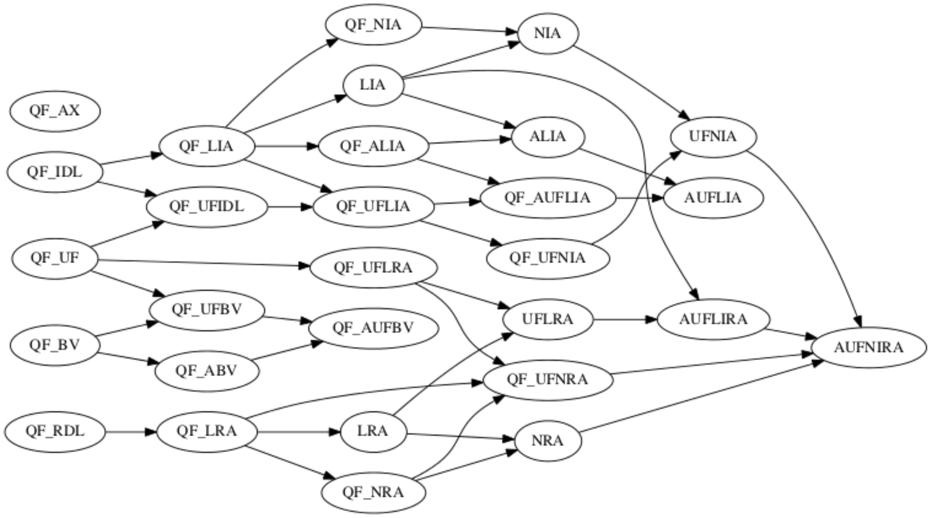


Figure 2. Available logics (March 2022) https://smtlib.cs.uiowa.edu/logics.shtml.

For example `QF_NRA` is the quantifier-free theory of nonlinear real arithmetic, and `QF_LRA` (linear real arithmetic) is included in this. Both `QF_NRA` and `QF_UFLRA` (uninterpreted functions and linear real arithmetic) are included in `QF_UFNRA`.

Then the SAT/UNSAT question is similar: do there exist values of y_i such that $\mathcal{L}$ is true (SAT), or can we state that no such exist (UNSAT), and the community runs SMT competitions (https://smt-comp.github.io/2022/). There is a separate track for each theory T, as the problems will be different. Within each, the problems are subdivided as industrial/crafted/random.

The SMT-LIB format [6] provides a standard input format. The question of proving UNSAT is in general unsolved (but see [37] for one particular theory T).

There has been substantial progress in SMT-solving over the years, possibly similar to SAT, and probably also spurred by the contests.

4. Computer algebra: Where are we?

Obviously, group theory and others are parts of computer algebra: What about the rest of computer algebra?

In general, the problems of computer algebra have a bad worst-case complexity, and we want effective solvers for "real" problems, not just "random" ones. The question, as in SAT and SMT, is "what does this mean?".

But there are also various logistical challenges.

(1) *Format:* there is no widely accepted common standard. We do have OpenMath [15], but it is not as widely supported as we would like.

(2) *Contests:* there are currently none. Could SIGSAM organise them?

(3) *Problem sets:* there are essentially no independent ones. Each author chooses his own.

(4) *Archive:* not really.

We now consider various specific problems.

4.1. Polynomial GCD

This problem is NP-hard (for sparse polynomials, even univariate) [26, 48]. Even for dense polynomials, it can be challenging for multivariates. There is no standard database: one has to trawl previous papers (and often need to ask the authors, as the polynomials were too big to print in the paper). Verification is a challenge: one can check that the result is *a common divisor*, but verifying *greatest* is still NP-hard [48].

4.2. Polynomial factorisation

This is known to be polynomial time for dense encodings [40], even though their exponent is large, and much work has gone into better algorithms; e.g. [1]. Presumably it is NP-hard for sparse encodings, though the author does not know of an explicit proof. There is no standard database: one has to trawl previous papers (and often needs to ask the authors, as the polynomials were too big to print in the paper).

Verification is a challenge: one can check that the result is *a factorisation*, but checking completeness (i.e., that these factors are irreducible) seems to be as hard as the original problem in the worst cases.

It is worth noting that, with probability 1, a random dense polynomial is irreducible (and easily proved so by the Musser test [47]), so the question "what are the *interesting* problems?" is vital.

4.3. Gröbner bases

The computation of Gröbner bases has many applications, from engineering to cryptography. But this has doubly exponential (with respect to n, the number of variables) worst-case complexity [45], even for a prime ideal [20]. If we take n "random" equations in n variables, they will satisfy the conditions for the Shape Lemma [7] and have $D \leq n^n$ solutions, so a Gröbner base in a purely lexicographical order will look like

$$\{p_1(x_1), x_2 - p_2(x_1), x_3 - p_3(x_1), \ldots, x_n - p_n(x_1)\}, \tag{2}$$

where p_1 is a polynomial of degree D in x_1 and the other p_i are polynomials of degree at most $D - 1$ in x_1. Experience shows that the coefficients of the p_i will generally be large (theoretically, they can be D times as long as the input coefficients). Conversely, if we have $n + 1$ equations, there are generally no solutions and the Gröbner base is $\{1\}$, much shorter than (2).

The good news from the point of view of this paper is that there is a collection [8], but it is very old (1996), so most of the examples are trivial with today's hardware and software, and completely static. Worse, some of the examples are only available in PDF.

There is always a Gröbner base (no concept of UNSAT as such) but it is not clear what a useful certificate of "G is a Gröbner base for input L" might mean in general (but see [4]). If $G = \{g_1, \ldots, g_M\}$ is a Gröbner base of $F = \{f_1, \ldots, f_N\}$, then a general certificate would consist of three components:

(1) a proof that G is a Gröbner base, which would mean that every S-polynomial $S(g_i, g_j)$ reduces to 0 under G, which is easily checked;

(2) a proof that $(F) \subseteq (G)$, which could be a set of $\lambda_{i,j}$ such that every $f_i = \sum \lambda_{i,j} g_j$;

(3) a proof that $(G) \subseteq (F)$, which could be a set of $\mu_{i,j}$ such that every $g_i = \sum \mu_{i,j} f_j$.

However, the $\lambda_{i,j}$ and $\mu_{i,j}$ might be (and generally are) extremely large.

4.4. Real algebraic geometry

Again, the problem of describing the decomposition of $\mathbf{R}^n$ sign-invariant for a set S of polynomials f_i in n variables has doubly exponential (with respect to n) worst-case complexity [14]. However, unlike Gröbner bases, it seems that this is the "typical" complexity, though the author knows no formal statement of this. For a given problem, the complexity can vary greatly: [14, Theorem 7] is an example of a polynomial p in $3n+4$ variables such that *any* cylindrical algebraic decomposition (CAD), with respect to one order, of $\mathbf{R}^{3n+4}$ sign-invariant for p has $O\left(2^{2^n}\right)$ cells, but with respect to another order has 3 cells:

$$\begin{aligned} p := x^{n+1}&\left(\left(y_{n-1} - \frac{1}{2}\right)^2 + (x_{n-1} - z_n)^2\right)\left((y_{n-1} - z_n)^2 + (x_{n-1} - x_n)^2\right) \\ &+ \sum_{i=1}^{n-1} x^{i+1}\left((y_{i-1} - y_i)^2 + (x_{i-1} - z_i)^2\right)\left((y_{i-1} - z_i)^2 + (x_{i-1} - x_i)^2\right) \\ &+ x\left((y_0 - 2x_0)^2 + \left(\alpha^2 + \left(x_0 - \frac{1}{2}\right)\right)^2\right) \\ &\quad \times \left((y_0 - 2 + 2x_0)^2 + \left(\alpha^2 + \left(x_0 - \frac{1}{2}\right)\right)^2\right) + a. \end{aligned}$$

The bad order (eliminating x, then $y_0, \alpha, x_0, z_1, y_1, z_1, \ldots, x_n, a$) needs $O(2^{2^n})$ (Maple: 141 when $n = 0$) cells. Any order eliminating a first says that R^{3n+3} is undecomposed, and the only question is $p = 0$, which is linear in a, and we get three cells: $p < 0$, $p = 0$, and $p > 0$.

However, if we replace a by a^3, the topology is essentially the same, but the discriminant is no longer trivial, and the "good" order now generates 213 cells in Maple, rather than three.

There is a collection [58], not quite as old as [8] (2014 was the last update), but still completely static. The DEWCAD project [12] might update this, but there are still issues of long-term conservation. The format has learned from [8] and each example is available in text, Maple input, and QEPCAD.

If we are just looking at computing a CAD, which we might wish to do for motion planning purposes [57], there is no concept of UNSAT, and the question of certificates of correctness is essentially unsolved. Attempts to produce a formally verified CAD algorithm have also so far been unsuccessful [21].

However, CAD was invented [22] for the purpose of quantifier elimination, i.e., converting $Q_k x_k Q_{k+1} x_{k+1} \cdots Q_n x_n \Phi(f_i)$, where $Q_i \in \{\exists, \forall\}$ and Φ is a Boolean combination of equalities and inequalities in the f_i, into $\Psi(g_1, \ldots, g_{n'})$, where Ψ is a Boolean combination of equalities and inequalities in the g_i, polynomials in $x_1, \ldots, x_{k-1}$, and if the statement is fully quantified, the result is a Boolean. A common case, particularly in program verification, is the fully existential case (all Q_i are $\exists$), where Φ is "something has gone wrong", and we want to show that this cannot happen. Then SAT is easy (exhibit values of x_i such that Φ is true, but UNSAT is much harder to certify. See [37] for some steps in this direction.

4.5. Integration

The computational complexity of integration (i.e., given a formula f in a class $\mathcal{L}$, is there a formula $g \in \mathcal{L}$, or in an agreed extension of $\mathcal{L}$, such that $g' = f$) is essentially unknown (but integration certainly involves GCD, factorisation, etc.). When $\mathcal{L}$ includes algebraic functions, difficult questions of algebraic geometry arise (see [28, as corrected in [44]]), and there is no known bound on the complexity of these.

"Paper" mathematics produced large databases of integrals (e.g. [31]), but these are (at best) in PDF, and the way they are commonly printed makes it extremely hard to recover semantics from the layout. Probably the best current database is described in [35]. But these databases are almost entirely of successful (SAT in our notation) examples, and there is almost no collection of UNSAT ($\nexists g \in \mathcal{L} : g' = f$) examples. Algorithm-based software (e.g. [28]) has an internal proof of UNSAT, but I know of no software that can exhibit it. That proof is typically very reliant on the underlying mathematics.

A new question here is the "niceness" of the output in the SAT case. Jeffrey and Rich [35] give the example of

$$\int \frac{5x^4}{(1+x)^6}\,\mathrm{d}x = \frac{x^5}{(1+x)^5}, \tag{3}$$

where Maple's answer is

$$\frac{-10}{(1+x)^3} + \frac{5}{(1+x)^4} - \frac{5}{(1+x)} - \frac{1}{(1+x)^5} + \frac{10}{(1+x)^2}. \tag{4}$$

Note that (4) is not just an ugly form of the right-hand side of (3): the two differ by 1, which is a legitimate constant of integration.

While some element of "niceness" is probably beyond automation, "simplicity" in the sense of [19], essentially minimal Kolmogorov complexity, is probably a good proxy, and could be automatically judged (at least in principle: there are probably some messy system-dependent issues in practice).

5. Conclusions

(1) The field of computer algebra really ought to invest in the sort of contests that have stimulated the SAT and SMT worlds.

(2) This requires much larger databases of "relevant" problems than we currently have, and they need to be properly curated.

   + The technology of collaborative working, e.g. wikis or GitHub, has greatly advanced since the days of [8], which should make collaborative construction of example sets easier, and would also help with the preservation challenge.

   – Although OpenMath is in principle a suitable system-neutral notation that could be the standard input (and output) format, such a use would challenge OpenMath implementations. This would be a good development, though.

(3) This would allow much better benchmarking practices; see the description in [13].

(4) There are significant challenges in providing "certificates", not just of UNSAT in the case of integration, but elsewhere in algebra. For example, asserting $g = \gcd(f_1, f_2)$ involves, not just the claim that g divides f_1 and f_2, but also that f_1/g, f_2/g are relatively prime, which may be much harder to demonstrate.

Acknowledgements. The author is grateful to Dr. Uncu for his comments on drafts, and to the organisers of the MIDAS session at the 8th European Congress of Mathematicians for prompting these reflections.

Funding. This work was partially supported by EPSRC Grant EP/T015713/1.

References

[1] J. Abbott, V. Shoup, and P. Zimmermann, Factorization in $\mathbb{Z}[x]$: the searching phase. In *ISSAC 2000*, edited by C. Traverso, pp. 1–7, ACM, New York, 2000 Zbl 1326.68339

[2] D. Achlioptas and C. Moore, Random k-SAT: two moments suffice to cross a sharp threshold. *SIAM J. Comput.* **36** (2006), no. 3, 740–762 Zbl 1120.68096 MR 2263010

[3] D. Achlioptas and Y. Peres, The threshold for random k-SAT is $2^k \log 2 - O(k)$. *J. Amer. Math. Soc.* **17** (2004), no. 4, 947–973 Zbl 1093.68075 MR 2083472

[4] E. A. Arnold, Modular algorithms for computing Gröbner bases. *J. Symbolic Comput.* **35** (2003), no. 4, 403–419 Zbl 1046.13018 MR 1976575

[5] T. Balyo and L. Chrpa, Using algorithm configuration tools to generate hard SAT benchmarks. In *The Eleventh International Symposium on Combinatorial Search (SoCS 2018)*, pp. 133–137, 2018

[6] C. Barrett, P. Fontaine, and C. Tinelli, The SMT-LIB standard: Version 2.6. 2021, http://smtlib.cs.uiowa.edu/papers/smt-lib-reference-v2.6-r2021-05-12.pdf

[7] E. Becker, M. G. Marinari, T. Mora, and C. Traverso, The shape of the Shape Lemma. In *ISSAC 1994*, pp. 129–133, ACM, Baltimore, MD, 1994 Zbl 0925.13006

[8] D. Bini and B. Mourrain, Polynomial test suite. 1996, http://www-sop.inria.fr/saga/POL/

[9] B. J. Birch and H. P. F. Swinnerton-Dyer, Notes on elliptic curves. I. *J. Reine Angew. Math.* **212** (1963), 7–25 Zbl 0118.27601 MR 146143

[10] B. J. Birch and H. P. F. Swinnerton-Dyer, Notes on elliptic curves. II. *J. Reine Angew. Math.* **218** (1965), 79–108 Zbl 0147.02506 MR 179168

[11] R. Bixby, Computational progress in linear and mixed integer programming. 2015, presentation at ICIAM 2015

[12] R. Bradford, J. H. Davenport, M. England, A. Sadeghimanesh, and A. Uncu, The DEWCAD Project: pushing back the Doubly Exponential Wall of Cylindrical Algebraic Decomposition. *ACM Commun. Comput. Algebra* **55** (2021), no. 3, 107–111 MR 4363371

[13] M. Brain, J. Davenport, and A. Griggio, Benchmarking solvers, SAT-style. In *SC^2 2017 Satisfiability Checking and Symbolic Computation CEUR Workshop 1974*, pp. 1–15, 2017

[14] C. W. Brown and J. H. Davenport, The complexity of quantifier elimination and cylindrical algebraic decomposition. In *ISSAC 2007*, edited by C. Brown, pp. 54–60, ACM, New York, 2007 Zbl 1190.68028 MR 2396184

[15] S. Buswell et al., The OpenMath Standard 2.0 Revision 1. 2017, http://www.openmath.org

[16] G. Butler, The transitive groups of degree fourteen and fifteen. *J. Symbolic Comput.* **16** (1993), no. 5, 413–422 Zbl 0813.20003 MR 1271082

[17] G. Butler and J. McKay, The transitive groups of degree up to eleven. *Comm. Algebra* **11** (1983), no. 8, 863–911 Zbl 0518.20003 MR 695893

[18] J. J. Cannon and D. F. Holt, The transitive permutation groups of degree 32. *Experiment. Math.* **17** (2008), no. 3, 307–314 Zbl 1175.20004 MR 2455702

[19] J. Carette, Understanding expression simplification. In *ISSAC 2004*, pp. 72–79, ACM, New York, 2004 Zbl 1134.68596 MR 2126927

[20] A. L. Chistov, Double-exponential lower bound for the degree of any system of generators of a polynomial prime ideal. *St. Petersburg Math. J.* **20** (2009), no. 6, 983–1001 Zbl 1206.13031

[21] C. Cohen and A. Mahboubi, A formal quantifier elimination for algebraically closed fields. In *CICM 2010*, edited by S. Autexier et al., pp. 189–203, Springer, Berlin, 2010 Zbl 1286.68394

[22] G. E. Collins, Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Automata Theory and Formal Languages (Second GI Conf., Kaiserslautern, 1975)*, pp. 134–183, Lecture Notes in Comput. Sci. 33, Springer, Berlin, 1975 Zbl 0318.02051 MR 0403962

[23] J. H. Conway and R. K. Guy, *The Book of Numbers*. Copernicus, New York, 1996 Zbl 0866.00001 MR 1411676

[24] S. Cook, *On the minimum computation time of functions*. Ph.D. thesis, Department of Mathematics, Harvard University, 1966

[25] M. Daquino et al., The OpenCitations data model. In *The 19th International Semantic Web Conference (ISWC 2020)*, pp. 447–463, 2020

[26] J. Davenport and J. Carette, The sparsity challenges. In *SYNASC 2009*, edited by S. Watt et al., pp. 3–7, 2010

[27] J. Davenport, B. Poonen, J. Maynard, H. Helfgott, P. Tiep, and L. Cruz-Filipe, Machine-assisted proofs. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary Lectures*, pp. 1085–1110, World Sci. Publ., Hackensack, NJ, 2018 Zbl 1452.68262 MR 3966753

[28] J. H. Davenport, *On the Integration of Algebraic Functions*. Lecture Notes in Comput. Sci. 102, Springer, Berlin, 1981 Zbl 0471.14009 MR 617377

[29] E. Garfield, The evolution of the Science Citation Index. *Int. Microbiol.* **10** (2007), 65–69

[30] I. Gent and T. Walsh, The SAT phase transition. In *ECAI 1994*, edited by A. Cohn, pp. 105–109, John Wiley, New York, 1994

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. 7th edn., Academic Press, Amsterdam, 2007 Zbl 1208.65001 MR 2360010

[32] A. Hulpke, *Konstruktion transitiver Permutationsgruppen*. Ph.D. thesis, RWTH Aachen, 1996

[33] A. Hulpke, Constructing transitive permutation groups. *J. Symbolic Comput.* **39** (2005), no. 1, 1–30 Zbl 1131.20003 MR 2168238

[34] C. Jacobi, *Canon arithmeticus, sive tabulae quibus exhibentur pro singulis numeris primis vel primorum potestatibus infra 1000 numeri ad datos indices et indices ad datos numeros pertinentes*. Typis Academicis, Berolini, 1839

[35] D. J. Jeffrey and A. D. Rich, Reducing expression size using rule-based integration. In *CICM 2010*, edited by S. Autexier et al., pp. 234–246, Springer, Berlin, 2010 Zbl 1286.68517

[36] J. Kratz and C. Strasser, Data publication consensus and controversies [version 3]. *F1000Research* **3** (2014), Article No. 94

[37] G. Kremer, E. Ábrahám, M. England, and J. H. Davenport, On the implementation of cylindrical algebraic coverings for satisfiability modulo theories solving. In *SYNASC 2021*, pp. 37–39, 2021

[38] W. Küchlin and C. Sinz, Proving consistency assertions for automotive product data management. *J. Autom. Reasoning* **24** (2000), no. 1–2, 145–163 Zbl 0968.68042

[39] V. Larivière, Y. Gingras, and E. Archambault, The decline in the concentration of citations, 1900–2007. *J. Amer. Soc. Info. Sci. Technol.* **60** (2009), no. 4, 858–862

[40] A. K. Lenstra, H. W. Lenstra Jr., and L. Lovász, Factoring polynomials with rational coefficients. *Math. Ann.* **261** (1982), no. 4, 515–534 Zbl 0488.12001 MR 682664

[41] G. Leonelli, *Supplément logarithmique. Théorie des logarithmes additionels et diductifs*. Brossier, Bordeaux, 1803

[42] D. F. Mansfield, Plimpton 322: a study of rectangles. *Found. Sci.* **26** (2021), no. 4, 977–1005 Zbl 07554371 MR 4334265

[43] E. Mark, Maturation, fecundation, and segmentation of Limax campestris, Binney. *Bulletin of the Museum of Comparative Zoology at Harvard College* **6** (1881), no. 12, 173–625

[44] D. Masser and U. Zannier, Torsion points, Pell's equation, and integration in elementary terms. *Acta Math.* **225** (2020), no. 2, 227–313 Zbl 1470.11163 MR 4205408

[45] E. W. Mayr and S. Ritscher, Dimension-dependent bounds for Gröbner bases of polynomial ideals. *J. Symbolic Comput.* **49** (2013), 78–94 Zbl 1258.13032 MR 2997841

[46] H. Mooney and M. Newton, The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication* **1** (2012), no. 1, Article No. eP1035

[47] D. R. Musser, On the efficiency of a polynomial irreducibility test. *J. Assoc. Comput. Mach.* **25** (1978), no. 2, 271–282 Zbl 0372.68014 MR 488309

[48] D. A. Plaisted, New NP-hard and NP-complete polynomial and integer divisibility problems. *Theoret. Comput. Sci.* **31** (1984), no. 1-2, 125–138 Zbl 0572.68027 MR 752098

[49] D. Remler, Are 90% of academic papers really never cited? Reviewing the literature on academic citations. 2014, http://blogs.lse.ac.uk/impactofsocialsciences/2014/04/23/academic-papers-citation-rates-remler/

[50] E. Robson, Neither Sherlock Holmes nor Babylon: a reassessment of Plimpton 322. *Historia Math.* **28** (2001), no. 3, 167–206 Zbl 0991.01001 MR 1849797

[51] G. F. Royle, The transitive groups of degree twelve. *J. Symbolic Comput.* **4** (1987), no. 2, 255–268 Zbl 0683.20002 MR 922391

[52] N. J. A. Sloane, The On-Line Encyclopedia of Integer Sequences. *Notices Amer. Math. Soc.* **50** (2003), no. 8, 912–915 Zbl 1044.11108 MR 1992789

[53] I. Spence, Weakening cardinality constraints creates harder satisfiability benchmarks. *ACM J. Exp. Algorithmics* **20** (2015), Article No. 1.4 MR 3353196

[54] H. Swinnerton-Dyer et al., Numerical tables on elliptic curves. In *Modular Functions of One Variable IV*, pp. 74–144, Lecture Notes in Math. 476, Springer, Berlin, 1975 Zbl 1214.11006

[55] The GAP Group, GAP—Groups, algorithms, and programming, version 4.11.1. 2021, https://www.gap-system.org

[56] S. van de Sandt et al., Practice meets principle: Tracking software and data citations to Zenodo DOIs. 2019, arXiv:1911.00295

[57] D. Wilson, J. Davenport, M. England, and R. Bradford, A "Piano Movers" Problem reformulated. In *SYNASC 2013*, pp. 53–60, 2013

[58] D. J. Wilson, R. J. Bradford, and J. H. Davenport, A repository for CAD examples. *ACM Commun. Comput. Algebra* **46** (2012), no. 3, 67–69 Zbl 1322.68294

[59] J. Zech, *Tafeln der Additions- und Subtractions-Logarithmen*. Weidmannsche Buchhandlung, Berlin, 1849

James H. Davenport
Department of Computer Science, University of Bath, Bath BA2 7AY, UK;
j.h.davenport@bath.ac.uk

DOI 10.4171/8ECM/18

Closed G_2-structures on compact quotients of Lie groups

Anna Fino and Alberto Raffero

Abstract. G_2-structures defined by a closed non-degenerate 3-form constitute the starting point in various known and potentially effective methods to obtain holonomy G_2-metrics on compact 7-manifolds. Albeit linear, the closed condition is quite restrictive, and no general results on the existence of closed G_2-structures on compact 7-manifolds are currently known. In this paper, we review some results regarding compact locally homogeneous spaces admitting invariant closed G_2-structures. In particular, we consider the case of compact quotients of simply connected Lie groups by discrete subgroups.

1. Introduction

A G_2-structure is a special type of G-structure that occurs on certain 7-dimensional smooth manifolds. More precisely, it is a reduction of the structure group of the frame bundle of a 7-manifold M from the general linear group $GL(7,\mathbb{R})$ to the compact exceptional Lie group G_2. The existence of a G_2-structure on M is equivalent to the orientability of M and the existence of a spin structure on it, namely to the vanishing of the the first and second Stiefel–Whitney classes of M.

Since every 7-manifold admitting G_2-structures is spin, it also admits almost contact structures. The interplay between the existence of special types of G_2-structures and of contact structures has been recently investigated in [2, 13, 26].

The existence of a G_2-structure on M can also be described in terms of differential forms. Indeed, it is characterized by the existence of a 3-form $\varphi \in \Omega^3(M)$ with pointwise stabilizer isomorphic to G_2. This is also equivalent to requiring that φ is *non-degenerate*; namely that at each point p of M one has that

$$\iota_X\varphi \wedge \iota_X\varphi \wedge \varphi \neq 0,$$

for every non-zero tangent vector $X \in T_pM$, where ι_X denotes the contraction by X. Every such 3-form φ gives rise to a Riemannian metric g_φ and to an orientation on M. More precisely, g_φ and the corresponding Riemannian volume form dV_φ are

2020 Mathematics Subject Classification. Primary 53C10; Secondary 53C30.
Keywords. Closed G_2-structure, locally homogeneous space, Lie algebra, lattice, Laplacian soliton, exact G_2-structure.

related to φ as follows:

$$g_\varphi(X,Y)dV_\varphi = \frac{1}{6}\iota_X\varphi \wedge \iota_Y\varphi \wedge \varphi.$$

Moreover, at each point p of M, the 3-form φ can be written as

$$\varphi = e^{127} + e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245},$$

where $(e^1, \dots, e^7)$ is a g_φ-orthonormal basis of the cotangent space T_p^*M, and e^{ijk} denotes the wedge product $e^i \wedge e^j \wedge e^k$.

Let $*$ be the Hodge star operator determined by g_φ and the orientation, and let ∇ be the Levi-Civita connection of g_φ. By [19], the 3-form φ is parallel with respect to ∇ if and only if it is closed and co-closed; namely $d\varphi = 0$ and $d * \varphi = 0$. In this case, the G_2-structure is said to be *parallel* or *torsion-free*, its intrinsic torsion vanishes identically, the Riemannian metric g_φ is Ricci-flat (see also [4]), and $\mathrm{Hol}(g_\varphi)$ is isomorphic to a subgroup of G_2. Notice that the conditions $\nabla\varphi = 0$ and $d * \varphi = 0$ are both non-linear in φ, as both ∇ and $*$ depend on g_φ, which is determined by φ.

The existence of Riemannian metrics with holonomy equal to G_2 was first proved by Bryant in [7], where some non-compact examples of Riemannian 7-manifolds with holonomy G_2 were given. The first complete (but still non-compact) examples were obtained by Bryant and Salamon in 1989 [9], and the first compact examples were constructed by Joyce in 1994 [35, 36]. Further compact examples admitting holonomy G_2 metrics were obtained in [12, 34, 38, 39].

A G_2-structure defined by a non-degenerate 3-form φ satisfying the linear condition $d\varphi = 0$ is said to be *closed* or *calibrated*, since φ defines a calibration on M, namely $\varphi|_\xi \leq \mathrm{vol}_\xi$, for every oriented tangent 3-plane ξ (cf. [30]). The codifferential of a closed G_2-structure φ is given by

$$d * \varphi = \tau \wedge \varphi,$$

for a unique 2-form τ belonging to the irreducible 14-dimensional space $\Lambda^2_{14} \cong \mathfrak{g}_2$. This 2-form is usually called the *torsion form* of the closed G_2-structure φ, and it satisfies the identities $\tau \wedge \varphi = - * \tau$ and $\tau \wedge *\varphi = 0$. Note that $\tau = d^*\varphi$, and therefore $d^*\tau = 0$. As a consequence, $d\tau = \Delta_\varphi\varphi$, where $\Delta_\varphi = dd^* + d^*d$ denotes the Hodge Laplacian of g_φ.

By [8], the scalar curvature of the metric g_φ induced by a closed G_2-structure is given by

$$\mathrm{Scal}(g_\varphi) = -\frac{1}{2}|\tau|^2,$$

and so it is non-positive. Notice that this is not a restrictive condition on compact manifolds.

By [56], a compact homogeneous 7-manifold cannot admit any invariant closed non-parallel G_2-structure. On the other hand, there exist many examples of compact locally homogeneous 7-manifolds admitting *invariant* G_2-structures of this type; see for instance [3, 8, 11, 14, 15, 23, 37, 50]. All these examples are compact quotients of simply connected Lie groups by co-compact discrete subgroups (lattices). Further examples of compact manifolds admitting closed non-parallel G_2-structures are given in [16, 51] and they are obtained resolving the singularities of 7-orbifolds.

In Section 2, we review known examples of compact locally homogeneous spaces admitting invariant closed G_2-structures and known classification results for Lie algebras admitting closed G_2-structures. A classification is currently available for 7-dimensional Lie algebras that are non-solvable [23] and for those having a non-trivial center [11, 26]. The classification of solvable Lie algebras with a trivial center admitting closed G_2-structures is still missing.

A geometric flow evolving closed G_2-structures was introduced by Bryant in [8]. Self-similar solutions to this flow correspond to the so-called *Laplacian solitons*, namely to closed G_2-structures φ satisfying the condition $\Delta_\varphi\varphi = \lambda\varphi + \mathcal{L}_X\varphi$, for some real constant λ and some vector field X on M, where $\mathcal{L}_X\varphi$ denotes the Lie derivative of φ with respect to X. In Section 3, after reviewing general properties of the Laplacian flow and of Laplacian solitons, we present some recent results obtained in [26], where left-invariant Laplacian solitons on Lie groups with a non-trivial center were considered.

A Laplacian soliton φ is called *expanding* if $\lambda > 0$. In this case, the G_2-form φ has to be *exact*, i.e., $\varphi = d\alpha$, for some 2-form α on M. By [42, 44], a non-parallel Laplacian soliton on a compact 7-manifold must be expanding with $\mathcal{L}_X\varphi \neq 0$.

Currently, it is still not known whether exact G_2-structures may occur on compact 7-manifolds. In Section 4, we review the results of [18, 22, 28], where this problem was considered in the case when the compact 7-manifold M is the quotient of a 7-dimensional simply connected Lie group G by a co-compact discrete subgroup $\Gamma \subset$ G, and the exact G_2-structure on M is induced by a left-invariant one on G. In [18,28], it was shown that there are no examples of this type whenever the group G satisfies suitable extra assumptions. In the recent joint work with L. Martín Merchán [22], we extended the previous results, showing that every compact manifold $M = \Gamma\backslash$G as above does not admit any exact G_2-structure which is induced by a left-invariant one on G.

2. Compact locally homogeneous examples and classification results for Lie algebras

Let M be a 7-manifold endowed with a G_2-structure φ and consider its automorphism group

$$\mathrm{Aut}(M,\varphi) := \left\{ f \in \mathrm{Diff}(M) \mid f^*\varphi = \varphi \right\}.$$

Note that $\mathrm{Aut}(M, \varphi)$ is a closed Lie subgroup of the full isometry group $\mathrm{Isom}(M, g_\varphi)$ of the Riemannian manifold (M, g_φ).

When M is compact, $\mathrm{Aut}(M, \varphi)$ is compact, too, and its Lie algebra is given by

$$\mathfrak{aut}(M, \varphi) = \left\{X \in \mathfrak{X}(M) \mid \mathcal{L}_X \varphi = 0\right\}.$$

In particular, every $X \in \mathfrak{aut}(M, \varphi)$ is a Killing vector field for the metric g_φ; namely $\mathcal{L}_X g_\varphi = 0$.

When φ is parallel, g_φ is Ricci-flat, and it follows from the Bochner–Weitzenböck technique that every Killing vector field must be parallel with respect to the Levi-Civita connection of g_φ. Consequently, the Lie algebra $\mathfrak{aut}(M, \varphi)$ is abelian. Moreover, its possible dimensions are $0, 1, 3$ or 7, depending on $\mathrm{Hol}^0(g_\varphi)$ being equal to G_2, $\mathrm{SU}(3)$, $\mathrm{SU}(2)$ or $\{1\}$, respectively.

If the G_2-structure φ is closed and non-parallel, namely $\tau = d^*\varphi \neq 0$, then for every $X \in \mathfrak{aut}(M, \varphi)$ the closed 2-form $\iota_X \varphi$ is Δ_φ-harmonic, since $*(\iota_X \varphi) = \frac{1}{2} \iota_X \varphi \wedge \varphi$ is also closed. There is then an injective map

$$X \in \mathfrak{aut}(M, \varphi) \mapsto \iota_X \varphi \in \mathcal{H}^2(M),$$

and thus $\dim \mathfrak{aut}(M, \varphi) \leq b_2(M)$, where $b_2(M) = \dim \mathcal{H}^2(M) = \dim H^2_{\mathrm{dR}}(M)$ is the second Betti number of M. Moreover, it is possible to prove the following.

Theorem 2.1 ([56]). *Let M be a compact 7-manifold with a closed non-parallel* G_2*-structure φ. Then, $\mathfrak{aut}(M, \varphi)$ is abelian and its dimension is at most* 6.

Therefore, the identity component of $\mathrm{Aut}(M, \varphi)$ is a compact abelian Lie group whose dimension is bounded above by $\min\{6, b_2(M)\}$. As a consequence, a compact 7-manifold M with a closed non-parallel G_2-structure φ cannot be homogeneous; namely neither $\mathrm{Aut}(M, \varphi)$ nor a subgroup thereof can act transitively on M. In contrast to this last result, it is possible to construct non-compact homogeneous examples; see for instance [55].

The first example of compact 7-manifold M admitting closed G_2-structures but not admitting any parallel G_2-structure was constructed by Fernández in [14]. In this example, $M = \Gamma \backslash N$ is a compact nilmanifold; i.e., the compact quotient of a 7-dimensional simply connected nilpotent Lie group N by a co-compact discrete subgroup (lattice) Γ. Moreover, the closed G_2-structure φ on $\Gamma \backslash N$ considered in [14] is induced by a left-invariant one on the Lie group N. In particular, the pair $(\Gamma \backslash N, \varphi)$ is a locally homogeneous space that is not globally homogeneous, as the transitive action of N on $\Gamma \backslash N$ does not preserve the 3-form φ. In other words, N is not a subgroup of $\mathrm{Aut}(\Gamma \backslash N, \varphi)$.

Remark. By Malcev's criterion [49], a nilpotent Lie group admits lattices if and only if its Lie algebra admits a basis with rational structure constants.

We now consider the following problem.

Problem 2.2. Study the existence of invariant closed G$_2$-structures on compact 7-manifolds of the form $\Gamma\backslash\mathrm{G}$, where G is a 7-dimensional simply connected Lie group and $\Gamma \subset \mathrm{G}$ is a co-compact discrete subgroup.

We recall that a G$_2$-structure on $\Gamma\backslash\mathrm{G}$ is said to be *invariant* if it is induced by a left-invariant one on the Lie group G. Therefore, an invariant closed G$_2$-structure on $\Gamma\backslash\mathrm{G}$ is completely determined by a G$_2$-structure φ on the Lie algebra $\mathfrak{g}$ of G which is closed with respect to the Chevalley–Eilenberg differential d of $\mathfrak{g}$.

A 3-form φ on a 7-dimensional Lie algebra $\mathfrak{g}$ defines a G$_2$-structure if and only if the symmetric bilinear map

$$b_\varphi : \mathfrak{g}\times\mathfrak{g} \to \Lambda^7\mathfrak{g}^*, \quad b_\varphi(v,w) = \frac{1}{6}\iota_v\varphi\wedge\iota_w\varphi\wedge\varphi$$

satisfies the condition $\det(b_\varphi)^{1/9} \neq 0 \in \Lambda^7\mathfrak{g}^*$ and the symmetric bilinear form

$$g_\varphi := \det(b_\varphi)^{-1/9} b_\varphi : \mathfrak{g}\times\mathfrak{g} \to \mathbb{R}$$

is positive definite; see e.g. [32]. In particular, for any choice of orientation on $\mathfrak{g}$, the map

$$b_\varphi : \mathfrak{g}\times\mathfrak{g} \to \Lambda^7\mathfrak{g}^* \cong \mathbb{R}$$

has to be positive or negative definite.

By [52], a simply connected Lie group G admits lattices only if its Lie algebra $\mathfrak{g}$ is unimodular; i.e., $\mathrm{tr}(\mathrm{ad}_X) = 0$, for every $X \in \mathfrak{g}$.

In the sequel, the structure equations of an n-dimensional Lie algebra with respect to a basis of covectors $(e^1,\dots,e^n)$ of $\mathfrak{g}^*$ will be specified by the n-tuple $(de^1,\dots,de^n)$. Moreover, we will use the shortening $e^{ijk\cdots}$ to denote the wedge product of covectors $e^i\wedge e^j\wedge e^k\wedge\cdots$.

In [23], we classified all unimodular non-solvable Lie algebras admitting closed G$_2$-structures, up to isomorphism, obtaining the following result.

Theorem 2.3 ([23]). *A unimodular non-solvable Lie group* G *admits left-invariant closed* G$_2$*-structures if and only if its Lie algebra* $\mathfrak{g}$ *is isomorphic to one of the following:*

$$\mathfrak{q}_1 = \left(-e^{23}, -2e^{12}, 2e^{13}, 0, -e^{45}, \frac{1}{2}e^{46} - e^{47}, \frac{1}{2}e^{47}\right),$$

$$\mathfrak{q}_2 = \left(-e^{23}, -2e^{12}, 2e^{13}, 0, -e^{45}, -\mu e^{46}, (1+\mu)e^{47}\right), \quad -1 < \mu \leq -\frac{1}{2},$$

$$\mathfrak{q}_3 = \left(-e^{23}, -2e^{12}, 2e^{13}, 0, -\mu e^{45}, \frac{\mu}{2}e^{46} - e^{47}, e^{46} + \frac{\mu}{2}e^{47}\right), \quad \mu > 0,$$

$$\mathfrak{q}_4 = (-e^{23}, -2e^{12}, 2e^{13}, -e^{14} - e^{25} - e^{47}, e^{15} - e^{34} - e^{57}, 2e^{67}, 0).$$

The first three Lie algebras in the previous list decompose as a product of the form $\mathfrak{sl}(2,\mathbb{R}) \oplus \mathfrak{r}$, where the radical $\mathfrak{r}$ is unimodular and centerless. The Lie algebra $\mathfrak{q}_4$ is indecomposable and its Levi decomposition is given by $\mathfrak{q}_4 \cong \mathfrak{sl}(2,\mathbb{R}) \ltimes \mathfrak{r}$, where $\mathfrak{r} \cong \mathbb{R} \ltimes \mathbb{R}^3$.

As a consequence of the previous result, a unimodular Lie algebra with a non-trivial center admitting closed G_2-structures must be solvable.

It is well known that every nilpotent Lie algebra is unimodular and has a non-trivial center. Nilpotent Lie algebras admitting closed G_2-structures were considered in [11], where the following classification result was obtained.

Theorem 2.4 ([11]). *A 7-dimensional nilpotent Lie algebra admits closed* G_2*-structures if and only if it is isomorphic to one of the following:*

$$\begin{aligned}
\mathfrak{n}_1 &= (0,0,0,0,0,0,0),\\
\mathfrak{n}_2 &= (0,0,0,0,e^{12},e^{13},0),\\
\mathfrak{n}_3 &= (0,0,0,e^{12},e^{13},e^{23},0),\\
\mathfrak{n}_4 &= (0,0,e^{12},0,0,e^{13}+e^{24},e^{15}),\\
\mathfrak{n}_5 &= (0,0,e^{12},0,0,e^{13},e^{14}+e^{25}),\\
\mathfrak{n}_6 &= (0,0,0,e^{12},e^{13},e^{14},e^{15}),\\
\mathfrak{n}_7 &= (0,0,0,e^{12},e^{13},e^{14}+e^{23},e^{15}),\\
\mathfrak{n}_8 &= (0,0,e^{12},e^{13},e^{23},e^{15}+e^{24},e^{16}+e^{34}),\\
\mathfrak{n}_9 &= (0,0,e^{12},e^{13},e^{23},e^{15}+e^{24},e^{16}+e^{34}+e^{25}),\\
\mathfrak{n}_{10} &= (0,0,e^{12},0,e^{13}+e^{24},e^{14},e^{46}+e^{34}+e^{15}+e^{23}),\\
\mathfrak{n}_{11} &= (0,0,e^{12},0,e^{13},e^{24}+e^{23},e^{25}+e^{34}+e^{15}+e^{16}-3e^{26}),\\
\mathfrak{n}_{12} &= (0,0,0,e^{12},e^{23},-e^{13},2e^{26}-2e^{34}-2e^{16}+2e^{25}).
\end{aligned}$$

In [26], we dealt with the more general case of unimodular solvable non-nilpotent Lie algebras with a non-trivial center admitting closed G_2-structures. There, we obtained a characterization that is based on the following observation. Let W be a 7-dimensional vector space endowed with a G_2-structure φ. Choosing a non-zero vector $z \in W$ and a complementary vector subspace $V \subset W$ so that $W \cong V \oplus \mathbb{R}z$, one can write

$$\varphi = \tilde{\omega} \wedge \theta + \rho,$$

where $\theta \in W^*$ is the dual of z, $\tilde{\omega} \in \Lambda^2 V^*$, and $\rho \in \Lambda^3 V^*$. The 3-form φ defines a G_2-structure on W if and only if it is definite; namely for each non-zero vector $w \in W$ the contraction $\iota_w \varphi$ has rank six. Moreover, the 3-form φ on W is definite if and only if the 3-form ρ on V is definite; i.e., for each non-zero vector $v \in V$ the contraction

$\iota_v\rho$ has rank four, and $\widetilde{\omega}$ is a taming form for the complex structure J induced by ρ and one of the two orientations of V; namely $\widetilde{\omega}(v, Jv) > 0$ for every non-zero vector $v \in V$.

Using this property, in [26] we proved that a Lie algebra $\mathfrak{g}$ with a non-trivial center endowed with a closed G_2-structure φ must be the central extension of a 6-dimensional Lie algebra $\mathfrak{h}$ by means of a closed 2-form $\omega_0 \in \Lambda^2\mathfrak{h}^*$; namely $\mathfrak{g} = \mathfrak{h} \oplus \mathbb{R}z$ and its Lie bracket is given by

$$[z, \mathfrak{h}] = 0, \quad [x, y] = -\omega_0(x, y)z + [x, y]_{\mathfrak{h}}, \quad \forall x, y \in \mathfrak{h}.$$

Moreover, $\varphi = \widetilde{\omega} \wedge \theta + \rho$, where θ is a 1-form on $\mathfrak{g}$ satisfying the condition $d\theta = \omega_0$, ρ is a definite 3-form on $\mathfrak{h}$ such that $d\rho = -\omega_0 \wedge \widetilde{\omega}$, and $\widetilde{\omega}$ is a symplectic form on $\mathfrak{h}$ that tames the almost complex structure induced by ρ and a suitable orientation. If the 2-form $\widetilde{\omega}$ is symplectic, the 1-form θ is a contact form on $\mathfrak{g}$ and $(\mathfrak{g}, \theta)$ is the *contactization* of $(\mathfrak{h}, \widetilde{\omega})$; see [1]. In this last case, the Lie algebra $\mathfrak{g}$ admits both a closed G_2-structure and a contact structure. This is reminiscent of the Boothby–Wang construction in [5].

As a first consequence of this characterization, we determined all isomorphism classes of nilpotent Lie algebras admitting closed G_2-structures that arise as the contactization of a 6-dimensional symplectic nilpotent Lie algebra $(\mathfrak{h}, \omega_0)$, showing that any such Lie algebra must be isomorphic to one of the following Lie algebras: $\mathfrak{n}_9$, $\mathfrak{n}_{10}$, $\mathfrak{n}_{11}$, $\mathfrak{n}_{12}$.

Then, we proved that there exist eleven unimodular solvable non-nilpotent Lie algebras with a non-trivial center admitting closed G_2-structures, up to isomorphism, achieving in this way the classification of all isomorphism classes of unimodular Lie algebras with a non-trivial center admitting closed G_2-structures.

Theorem 2.5 ([26]). *Let $\mathfrak{g}$ be a 7-dimensional unimodular solvable non-nilpotent Lie algebra with a non-trivial center. Then, $\mathfrak{g}$ admits closed G_2-structures if and only if it is isomorphic to one of the following:*

$$\begin{aligned}
\mathfrak{s}_1 &= (e^{23}, -e^{36}, e^{26}, e^{26} - e^{56}, e^{36} + e^{46}, 0, 0),\\
\mathfrak{s}_2 &= (e^{16} + e^{35}, -e^{26} + e^{45}, e^{36}, -e^{46}, 0, 0, 0),\\
\mathfrak{s}_3 &= (-e^{16} + e^{25}, -e^{15} - e^{26}, e^{36} - e^{45}, e^{35} + e^{46}, 0, 0, 0),\\
\mathfrak{s}_4 &= (0, -e^{13}, -e^{12}, 0, -e^{46}, -e^{45}, 0),\\
\mathfrak{s}_5 &= (e^{15}, -e^{25}, -e^{35}, e^{45}, 0, 0, 0),\\
\mathfrak{s}_6 &= (\alpha e^{15} + e^{25}, -e^{15} + \alpha e^{25}, -\alpha e^{35} + e^{45}, -e^{35} - \alpha e^{45}, 0, 0, 0), \quad \alpha > 0,\\
\mathfrak{s}_7 &= (e^{25}, -e^{15}, e^{45}, -e^{35}, 0, 0, 0),\\
\mathfrak{s}_8 &= (e^{16} + e^{35}, -e^{26} + e^{45}, e^{36}, -e^{46}, 0, 0, e^{34}),
\end{aligned}$$

$$\mathfrak{s}_9 = (-e^{26} + e^{35}, e^{16} + e^{45}, -e^{46}, e^{36}, 0, 0, e^{34}),$$
$$\mathfrak{s}_{10} = (e^{23}, -e^{36}, e^{26}, e^{26} - e^{56}, e^{36} + e^{46}, 0, 2e^{16} + e^{25} - e^{34} + \sqrt{3}e^{24} + \sqrt{3}e^{35}),$$
$$\mathfrak{s}_{11} = (e^{23}, -e^{36}, e^{26}, e^{26} - e^{56}, e^{36} + e^{46}, 0, 2e^{16} + e^{25} - e^{34} - \sqrt{3}e^{24} - \sqrt{3}e^{35}).$$

In particular, $\mathfrak{g}$ is the contactization of a symplectic Lie algebra if and only if it is isomorphic either to $\mathfrak{s}_{10}$ or to $\mathfrak{s}_{11}$.

By the characterization above, we know that $\mathfrak{g}$ has to be the central extension of a unimodular symplectic Lie algebra $\mathfrak{h}$ endowed with a closed (possibly non-degenerate) 2-form ω_0 and a suitable pair of forms $(\tilde{\omega}, \rho)$. Such an extension is determined by any representative in the cohomology class $[\omega_0] \in H^2(\mathfrak{h})$, and the proof of the theorem follows after an inspection of all 6-dimensional unimodular symplectic Lie algebras that exist up to isomorphism (cf. [20, 47]).

As far as we know, the following problem remains open.

Problem 2.6. Classify all 7-dimensional solvable Lie algebras with a trivial center admitting closed G_2-structures, up to isomorphism.

3. Laplacian solitons

A special class of closed G_2-structures that has attracted a lot of attention in recent years is given by *Laplacian solitons*. These G_2-structures are closely related to the self-similar solutions to the *Laplacian flow* for closed G_2-structures, a geometric flow that was introduced by Bryant in [8] as a tool to potentially deform a closed G_2-structure towards a parallel one.

Definition 3.1 ([8]). Let φ_0 be a closed G_2-structure on a 7-manifold M. The *Laplacian flow* starting at φ_0 is the initial value problem

$$\begin{cases} \partial_t \varphi(t) = \Delta_{\varphi(t)} \varphi(t), \\ d\varphi(t) = 0, \\ \varphi(0) = \varphi_0, \end{cases}$$

where $\Delta_{\varphi(t)}$ is the Hodge Laplacian of $g_{\varphi(t)}$.

The stationary points of the Laplacian flow are parallel G_2-structures, even on non-compact manifolds (see [43] for the explicit computation in the non-compact case). If $\varphi(t)$ is a family of closed G_2-structures solving the Laplacian flow, then $\varphi(t) \in [\varphi_0] \in H^3_{\mathrm{dR}}(M)$; namely the de Rham cohomology class $[\varphi(t)]$ is constant in t. Moreover, the evolution equation of the metric $g_{\varphi(t)}$ induced by $\varphi(t)$ coincides with the Ricci flow of $g_{\varphi(t)}$ up to lower order terms; namely

$$\partial_t g_{\varphi(t)} = -2\,\mathrm{Ric}(g_{\varphi(t)}) + \text{l.o.t.}$$

Remark. On a compact manifold M, the Laplacian flow is the gradient flow of Hitchin's volume functional

$$\mathcal{V} : \varphi \in [\varphi_0] \mapsto \int_M \varphi \wedge *\varphi.$$

This functional is monotonically increasing along the flow, its critical points are parallel G$_2$-structures, and they are strict local maxima. See [6,43] and the arXiv version of [31] for more details.

The short-time existence and uniqueness of the solution to the Laplacian flow on a compact manifold were proved by Bryant and Xu in [6].

Theorem 3.2 ([6]). *Let M be a compact 7-manifold with a closed G$_2$-structure φ_0. Then, the Laplacian flow starting at φ_0 has a unique solution defined for short time $t \in [0, \varepsilon)$, with ε depending on φ_0.*

The geometric and analytic properties of the Laplacian flow have been deeply investigated by Lotay and Wei in [44–46], and further results are available in [10, 21, 57]. Moreover, various lower-dimensional reductions of the flow were studied in [21, 24, 27, 40]. Explicit examples of solutions to the flow are also known; see for instance [17, 24, 41] for examples on simply connected Lie groups with left-invariant closed G$_2$-structures, and [33] for a cohomogeneity one example on the 7-torus.

A closed G$_2$-structure φ on a 7-manifold M is said to be a *Laplacian soliton* if it satisfies the equation

$$\Delta_\varphi \varphi = \lambda\varphi + \mathcal{L}_X \varphi,$$

for some real constant λ and some vector field X on M. These G$_2$-structures give rise to self-similar solutions to the Laplacian flow, namely to solutions of the form $\varphi(t) = \sigma(t) f_t^* \varphi$, where $\sigma(t)$ is a real-valued function of t, and $f_t \in \mathrm{Diff}(M)$. Laplacian solitons are expected to model finite time singularities of the Laplacian flow; see [43] for more details.

Depending on the sign of λ, one can introduce the following definitions.

Definition 3.3. A Laplacian soliton φ is called *shrinking* if $\lambda < 0$, *steady* if $\lambda = 0$ and *expanding* if $\lambda > 0$.

Some restrictions to the existence of a Laplacian soliton on a compact manifold are known.

Theorem 3.4 ([42, 44]). *On a compact 7-manifold, a non-parallel Laplacian soliton φ must satisfy the equation $\Delta_\varphi \varphi = \lambda\varphi + \mathcal{L}_X \varphi$, with $\lambda > 0$ and $\mathcal{L}_X \varphi \neq 0$. Moreover, the only steady Laplacian solitons are given by parallel G$_2$-structures.*

Thus, a non-parallel Laplacian soliton on a compact manifold must be expanding. The following problem is still open.

Problem 3.5. Do there exist expanding Laplacian solitons on compact manifolds?

The non-compact setting is less restrictive, and various homogeneous examples of steady, shrinking, and expanding solitons are known [3, 24, 25, 41, 53, 54]. More recently, complete inhomogeneous examples of steady and shrinking solitons were obtained in [3, 27]. These examples are of gradient type; i.e., X is a gradient vector field.

By [41], any left-invariant Laplacian soliton φ on a Lie group G is *semi-algebraic*; i.e., the vector field X is defined by a 1-parameter group of automorphisms induced by a derivation D of the Lie algebra $\mathfrak{g}$. Some results on semi-algebraic solitons on unimodular Lie algebras with a non-trivial center have been recently obtained in [26]. For instance, under a natural assumption on the derivation D, it is possible to relate the constant λ to a certain eigenvalue of D and to the norm of the torsion form τ of the semi-algebraic soliton φ. Moreover, the following result can be proved.

Theorem 3.6 ([26]). *Let $\mathfrak{g}$ be a unimodular Lie algebra with a non-trivial center $\mathfrak{z}(\mathfrak{g})$ admitting a semi-algebraic soliton φ. Then the following conditions hold:*

(1) *if $\mathfrak{g}$ is the contactization of a symplectic Lie algebra, then $\lambda = |\tau|^2$ and thus φ must be expanding;*

(2) *if $\dim \mathfrak{z}(\mathfrak{g}) = 2$, then $\mathfrak{g}$ has to be isomorphic to one of the following Lie algebras: $\mathfrak{n}_1$, $\mathfrak{n}_2$, $\mathfrak{n}_3$, $\mathfrak{n}_4$, $\mathfrak{n}_5$, $\mathfrak{n}_6$, $\mathfrak{n}_7$, $\mathfrak{s}_5$, $\mathfrak{s}_6$, $\mathfrak{s}_7$.*

If $\dim \mathfrak{z}(\mathfrak{g}) = 1$, some non-existence results for semi-algebraic solitons on certain Lie algebras are also known [26], but a general result is still missing.

Remark. All known examples of Lie algebras admitting shrinking or steady Laplacian solitons have a trivial center. It would be interesting to establish whether the existence of these types of solitons forces the Lie algebra to be centerless.

4. Exact G_2-structures

An expanding Laplacian soliton φ is an *exact* G_2-structure. Indeed, since φ is closed and $\Delta_\varphi \varphi = d\tau$, the condition $\Delta_\varphi \varphi = \lambda\varphi + \mathcal{L}_X \varphi$ can be rewritten as follows:

$$\varphi = d\left(\frac{1}{\lambda}(\tau - \iota_X \varphi)\right).$$

In the literature, all known examples of compact 7-manifolds M admitting closed G_2-structures, but not admitting parallel G_2-structures, have $b_1(M) > 0$ and $b_3(M) > 0$; see [11, 14–16, 50, 51]. A longstanding open question concerns the existence of closed G_2-structures on compact 7-manifolds with $b_3(M) = 0$, such as the 7-sphere. Notice that, in this case, any closed G_2-structure would be defined by an exact 3-form. A natural question is then the following.

Problem 4.1. Does there exist a compact 7-manifold admitting exact G$_2$-structures?

In this section, we consider this problem in the case when the manifold is the compact quotient of a simply connected unimodular Lie group G by a lattice.

The following example constructed in [18] shows that exact G$_2$-structures occur on unimodular Lie algebras.

Example 4.2. Let $\mathfrak{s}$ be the 7-dimensional unimodular solvable Lie algebra with structure equations

$$\begin{aligned}
&de^1 = -2e^{17}, \quad de^2 = -4e^{27}, \quad de^3 = \frac{9}{2}e^{37},\\
&de^4 = \frac{5}{2}e^{47} - e^{13}, \quad de^5 = \frac{1}{2}e^{57} - 6e^{37} - e^{14} - e^{23},\\
&de^6 = -\frac{3}{2}e^{67} - 6e^{47} + 3e^{13} + e^{15} + e^{24}, \quad de^7 = 0.
\end{aligned}$$

This Lie algebra is a semidirect product of the form $\mathfrak{s} = \mathbb{R} \ltimes \mathfrak{n}$, where $\mathfrak{n}$ is a codimension one 4-step nilpotent ideal, and it satisfies the conditions $b_2(\mathfrak{s}) = 0 = b_3(\mathfrak{s})$. Moreover, $\mathfrak{s}$ admits the exact G$_2$-structure

$$\begin{aligned}
\varphi &= e^{127} + e^{347} + e^{567} + e^{135} - e^{146} - e^{236} - e^{245}\\
&= d\left(\frac{1}{6}e^{12} + \frac{23}{7}e^{34} + 2e^{36} - 2e^{45} + e^{56}\right).
\end{aligned}$$

Consequently, the simply connected solvable Lie group S with Lie algebra $\mathfrak{s}$ is endowed with a left-invariant exact G$_2$-structure obtained from φ via left multiplication.

As we already recalled, a Lie group G admitting lattices must be unimodular. In the case of solvable Lie groups, a stronger necessary condition for the existence of lattices is known; namely the group must be strongly unimodular (cf. [29, Prop. 3.3]). We recall the definition here.

Definition 4.3 ([29]). A solvable Lie group G with Lie algebra $\mathfrak{g}$ and nilradical $\mathfrak{n}$ is said to be *strongly unimodular* if $\mathrm{tr}(\mathrm{ad}_X)|_{\mathfrak{n}^i/\mathfrak{n}^{i+1}} = 0$, for every $X \in \mathfrak{g}$, where $\mathfrak{n}^0 = \mathfrak{n}$, and $\mathfrak{n}^i = [\mathfrak{n}, \mathfrak{n}^{i-1}]$, $i \geq 1$, is the ith term in the descending central series of $\mathfrak{n}$.

For instance, the simply connected solvable Lie group S in Example 4.2 is unimodular but not strongly unimodular, so it does not admit any compact quotient by a lattice.

In [18], we showed that a strongly unimodular $(2,3)$-trivial Lie algebra $\mathfrak{g}$, namely with $b_2(\mathfrak{g}) = b_3(\mathfrak{g}) = 0$, does not admit any exact G$_2$-structure. Therefore, there are no compact examples of the form $\Gamma\backslash$G admitting invariant exact G$_2$-structures whenever the Lie algebra of G is $(2,3)$-trivial. To prove this result, we used the property that a $(2,3)$-trivial Lie algebra $\mathfrak{g}$ is solvable and $\mathfrak{g} = \mathbb{R} \ltimes \mathfrak{n}$, with $\mathfrak{n}$ a codimension one nilpotent ideal (see [48]), and we classified all 7-dimensional strongly unimodular $(2,3)$-trivial Lie algebras.

One can then investigate what happens if either $b_3(\mathfrak{g}) = 0$ and $b_2(\mathfrak{g}) \neq 0$ or if no conditions on the Betti numbers of $\mathfrak{g}$ are imposed. A first partial answer to this problem was given in [28].

Theorem 4.4 ([28]). *If the Lie algebra $\mathfrak{g}$ of* G *has a codimension one nilpotent ideal, then any compact quotient* $\Gamma\backslash\mathrm{G}$ *does not admit any invariant exact* G_2*-structure. If in addition* G *is completely solvable, namely* ad_X *has only real eigenvalues for every* $X \in \mathfrak{g}$, *then* $\Gamma\backslash\mathrm{G}$ *does not have any exact* G_2*-structure at all.*

In [22], we investigated the existence of invariant exact G_2-structures on compact quotients of Lie groups without introducing any extra assumption on the Lie algebra $\mathfrak{g}$, and we proved the following result.

Theorem 4.5 ([22]). *A potential compact 7-manifold M with an exact* G_2*-structure φ cannot be of the form $M = \Gamma\backslash\mathrm{G}$, where* G *is a 7-dimensional simply connected Lie group,* $\Gamma \subset \mathrm{G}$ *is a lattice, and the exact* G_2*-structure φ on M is invariant.*

To prove this result, we focused on 7-dimensional unimodular Lie algebras $\mathfrak{g}$ and we studied the non-solvable case and the solvable case separately. By Theorem 2.3, there are four non-solvable unimodular Lie algebras admitting closed G_2-structures, up to isomorphism. The first three Lie algebras are decomposable, and by a direct computation we showed that b_φ is never definite for every exact 3-form φ on each one of them. The remaining Lie algebra $\mathfrak{q}_4$ is indecomposable, and for this we proved that the corresponding simply connected Lie group does not admit any lattice. In the solvable case, $\mathfrak{g}$ has a codimension one unimodular ideal $\mathfrak{s}$, and the existence of a G_2-structure φ on $\mathfrak{g}$ allows one to consider the g_φ-orthogonal decomposition $\mathfrak{g} = \mathfrak{s} \oplus \mathbb{R}$, where $\mathbb{R}$ denotes the orthogonal complement of $\mathfrak{s}$. As a Lie algebra, $\mathfrak{g}$ is then a semidirect product of the form $\mathfrak{g} = \mathfrak{s} \rtimes_D \mathbb{R}$, for some derivation D of $\mathfrak{s}$. Moreover, the G_2-structure φ on $\mathfrak{g}$ can be written as follows:

$$\varphi = \omega \wedge \eta + \rho,$$

where $\eta := z^\flat$ is the metric dual of a unit vector $z \in \mathbb{R}$, and the pair (ω, ρ) defines an SU(3)-structure on $\mathfrak{s}$. By imposing that φ is an exact non-degenerate 3-form and using that $\mathfrak{g}$ has to be strongly unimodular, one sees that no examples can be found also in the solvable case.

Acknowledgments. The first author would like to thank Mahir Bilen Can, Sergey Grigorian, and Sema Salur for the invitation to give a seminar in the Minisymposium "Geometries Defined by Differential Forms" at 8ECM.

Funding. The authors were partly supported by GNSAGA of INdAM and by the project PRIN 2017 "Real and Complex Manifolds: Topology, Geometry and Holomorphic Dynamics."

References

[1] D. V. Alekseevskiĭ, Contact homogeneous spaces. *Funktsional. Anal. i Prilozhen.* **24** (1990), no. 4, 74–75 Zbl 0721.53042 MR 1092805

[2] M. F. Arikan, H. Cho, and S. Salur, Existence of compatible contact structures on G_2-manifolds. *Asian J. Math.* **17** (2013), no. 2, 321–333 Zbl 1337.53064 MR 3078933

[3] G. Ball, Quadratic closed G_2-structures. 2020, arXiv:2006.14155

[4] E. Bonan, Sur des variétés riemanniennes à groupe d'holonomie G_2 ou spin (7). *C. R. Acad. Sci. Paris Sér. A-B* **262** (1966), A127–A129 Zbl 0134.39402 MR 196668

[5] W. M. Boothby and H. C. Wang, On contact manifolds. *Ann. of Math. (2)* **68** (1958), 721–734 Zbl 0084.39204 MR 112160

[6] R. Bryant and F. Xu, Laplacian flow for closed G_2-structures: short time behavior. 2011, arXiv:1101.2004

[7] R. L. Bryant, Metrics with exceptional holonomy. *Ann. of Math. (2)* **126** (1987), no. 3, 525–576 Zbl 0637.53042 MR 916718

[8] R. L. Bryant, Some remarks on G_2-structures. In *Proceedings of Gökova Geometry-Topology Conference 2005*, pp. 75–109, Gökova Geometry/Topology Conference (GGT), Gökova, 2006 Zbl 1115.53018 MR 2282011

[9] R. L. Bryant and S. M. Salamon, On the construction of some complete metrics with exceptional holonomy. *Duke Math. J.* **58** (1989), no. 3, 829–850 Zbl 0681.53021 MR 1016448

[10] G. Chen, Shi-type estimates and finite-time singularities of flows of G_2 structures. *Q. J. Math.* **69** (2018), no. 3, 779–797 Zbl 1425.53033 MR 3859207

[11] D. Conti and M. Fernández, Nilmanifolds with a calibrated G_2-structure. *Differential Geom. Appl.* **29** (2011), no. 4, 493–506 Zbl 1222.53059 MR 2811660

[12] A. Corti, M. Haskins, J. Nordström, and T. Pacini, G_2-manifolds and associative submanifolds via semi-Fano 3-folds. *Duke Math. J.* **164** (2015), no. 10, 1971–2092 Zbl 1343.53044 MR 3369307

[13] X. de la Ossa, M. Larfors, and M. Magill, Almost contact structures on manifolds with a G_2 structure. 2021, arXiv:2101.12605

[14] M. Fernández, An example of a compact calibrated manifold associated with the exceptional Lie group G_2. *J. Differential Geom.* **26** (1987), no. 2, 367–370 Zbl 0604.53013 MR 906398

[15] M. Fernández, A family of compact solvable G_2-calibrated manifolds. *Tohoku Math. J. (2)* **39** (1987), no. 2, 287–289 Zbl 0609.53011 MR 887944

[16] M. Fernández, A. Fino, A. Kovalev, and V. Muñoz, A compact G_2-calibrated manifold with first Betti number $b_1 = 1$. *Adv. Math.* **381** (2021), Paper No. 107623 Zbl 1472.53061 MR 4206793

[17] M. Fernández, A. Fino, and V. Manero, Laplacian flow of closed G_2-structures inducing nilsolitons. *J. Geom. Anal.* **26** (2016), no. 3, 1808–1837 Zbl 1344.53041 MR 3511459

[18] M. Fernández, A. Fino, and A. Raffero, Exact G_2-structures on unimodular Lie algebras. *Monatsh. Math.* **193** (2020), no. 1, 47–60 Zbl 1455.53049 MR 4127433

[19] M. Fernández and A. Gray, Riemannian manifolds with structure group G_2. *Ann. Mat. Pura Appl. (4)* **132** (1982), 19–45 (1983) Zbl 0524.53023 MR 696037

[20] M. Fernández, V. Manero, A. Otal, and L. Ugarte, Symplectic half-flat solvmanifolds. *Ann. Global Anal. Geom.* **43** (2013), no. 4, 367–383 Zbl 1266.53035 MR 3038540

[21] J. Fine and C. Yao, Hypersymplectic 4-manifolds, the G_2-Laplacian flow, and extension assuming bounded scalar curvature. *Duke Math. J.* **167** (2018), no. 18, 3533–3589 Zbl 07009771 MR 3881202

[22] A. Fino, L. Martín-Merchán, and A. Raffero, Exact G_2-structures on compact quotients of Lie groups. 2021, arXiv:2108.11664

[23] A. Fino and A. Raffero, Closed G_2-structures on non-solvable Lie groups. *Rev. Mat. Complut.* **32** (2019), no. 3, 837–851 Zbl 1428.53032 MR 3995432

[24] A. Fino and A. Raffero, Closed warped G_2-structures evolving under the Laplacian flow. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **20** (2020), no. 1, 315–348 Zbl 1451.53127 MR 4088743

[25] A. Fino and A. Raffero, Remarks on homogeneous solitons of the G_2-Laplacian flow. *C. R. Math. Acad. Sci. Paris* **358** (2020), no. 4, 401–406 Zbl 1475.53107 MR 4134249

[26] A. Fino, A. Raffero, and F. Salvatore, Closed G_2-structures on unimodular Lie algebras with non-trivial center. *Transform. Groups* (2022), DOI 10.1007/s00031-021-09683-8

[27] U. Fowdar, S^1-invariant Laplacian flow. *J. Geom. Anal.* **32** (2022), no. 1, Paper No. 17 Zbl 07446091 MR 4349461

[28] M. Freibert and S. Salamon, Closed G_2-eigenforms and exact G_2-structures. *Rev. Mat. Iberoam.* **38** (2022), no. 6, 1827–1866 Zbl 07628545 MR 4516173

[29] H. Garland, On the cohomology of lattices in solvable Lie groups. *Ann. of Math. (2)* **84** (1966), 175–196 Zbl 0142.26702 MR 207916

[30] R. Harvey and H. B. Lawson, Jr., Calibrated geometries. *Acta Math.* **148** (1982), 47–157 Zbl 0584.53021 MR 666108

[31] N. Hitchin, The geometry of three-forms in six dimensions. *J. Differential Geom.* **55** (2000), no. 3, 547–576 Zbl 1036.53042 MR 1863733

[32] N. Hitchin, Stable forms and special metrics. In *Global Differential Geometry: The Mathematical Legacy of Alfred Gray (Bilbao, 2000)*, pp. 70–89, Contemp. Math. 288, Amer. Math. Soc., Providence, RI, 2001 Zbl 1004.53034 MR 1871001

[33] H. Huang, Y. Wang, and C. Yao, Cohomogeneity-one G_2-Laplacian flow on the 7-torus. *J. Lond. Math. Soc. (2)* **98** (2018), no. 2, 349–368 Zbl 1403.53056 MR 3873112

[34] D. Joyce and S. Karigiannis, A new construction of compact torsion-free G_2-manifolds by gluing families of Eguchi-Hanson spaces. *J. Differential Geom.* **117** (2021), no. 2, 255–343 Zbl 1464.53067 MR 4214342

[35] D. D. Joyce, Compact Riemannian 7-manifolds with holonomy G_2. I. *J. Differential Geom.* **43** (1996), no. 2, 291–328 Zbl 0861.53022 MR 1424428

[36] D. D. Joyce, Compact Riemannian 7-manifolds with holonomy G_2. II. *J. Differential Geom.* **43** (1996), no. 2, 329–375 Zbl 0861.53023 MR 1424428

[37] I. Kath and J. Lauret, A new example of a compact ERP G_2-structure. *Bull. Lond. Math. Soc.* **53** (2021), no. 6, 1692–1710 MR 4375925

[38] A. Kovalev, Twisted connected sums and special Riemannian holonomy. *J. Reine Angew. Math.* **565** (2003), 125–160 Zbl 1043.53041 MR 2024648

[39] A. Kovalev and N.-H. Lee, $K3$ surfaces with non-symplectic involution and compact irreducible G_2-manifolds. *Math. Proc. Cambridge Philos. Soc.* **151** (2011), no. 2, 193–218 Zbl 1228.53064 MR 2823130

[40] B. Lambert and J. D. Lotay, Spacelike mean curvature flow. *J. Geom. Anal.* **31** (2021), no. 2, 1291–1359 Zbl 1465.53081 MR 4215263

[41] J. Lauret, Laplacian flow of homogeneous G_2-structures and its solitons. *Proc. Lond. Math. Soc. (3)* **114** (2017), no. 3, 527–560 Zbl 1380.53074 MR 3653239

[42] C. Lin, Laplacian solitons and symmetry in G_2-geometry. *J. Geom. Phys.* **64** (2013), 111–119 Zbl 1259.53066 MR 3004019

[43] J. D. Lotay, Geometric flows of G_2 structures. In *Lectures and Surveys on* G_2*-Manifolds and Related Topics*, pp. 113–140, Fields Inst. Commun. 84, Springer, New York, 2020 Zbl 1447.53086 MR 4295856

[44] J. D. Lotay and Y. Wei, Laplacian flow for closed G_2 structures: Shi-type estimates, uniqueness and compactness. *Geom. Funct. Anal.* **27** (2017), no. 1, 165–233 Zbl 1378.53078 MR 3613456

[45] J. D. Lotay and Y. Wei, Laplacian flow for closed G_2 structures: real analyticity. *Comm. Anal. Geom.* **27** (2019), no. 1, 73–109 Zbl 1421.53032 MR 3951021

[46] J. D. Lotay and Y. Wei, Stability of torsion-free G_2 structures along the Laplacian flow. *J. Differential Geom.* **111** (2019), no. 3, 495–526 Zbl 07036514 MR 3934598

[47] M. Macrì, Cohomological properties of unimodular six dimensional solvable Lie algebras. *Differential Geom. Appl.* **31** (2013), no. 1, 112–129 Zbl 1263.53045 MR 3010082

[48] T. B. Madsen and A. Swann, Multi-moment maps. *Adv. Math.* **229** (2012), no. 4, 2287–2309 Zbl 1238.53017 MR 2880222

[49] A. I. Malcev, On a class of homogeneous spaces. *Amer. Math. Soc. Translation* **1951** (1951), no. 39, 33 Zbl 0034.01701 MR 0039734

[50] V. Manero, Compact solvmanifolds with calibrated and cocalibrated G_2-structures. *Manuscripta Math.* **162** (2020), no. 3-4, 315–339 Zbl 1441.53017 MR 4109489

[51] L. Martín-Merchán, A compact non-formal closed G_2 manifold with $b_1 = 1$. 2020, *Math. Nachr.* (to appear); arXiv:2005.04924

[52] J. Milnor, Curvatures of left invariant metrics on Lie groups. *Advances in Math.* **21** (1976), no. 3, 293–329 Zbl 0341.53030 MR 425012

[53] M. Nicolini, Laplacian solitons on nilpotent Lie groups. *Bull. Belg. Math. Soc. Simon Stevin* **25** (2018), no. 2, 183–196 Zbl 1393.37088 MR 3819121

[54] M. Nicolini, New examples of shrinking Laplacian solitons. *Q. J. Math.* **73** (2022), no. 1, 239–259 MR 4395079

[55] F. Podestà and A. Raffero, Closed G_2-structures with a transitive reductive group of automorphisms. 2019, *Asian J. Math.* (to appear); arXiv:1911.13052

[56] F. Podestà and A. Raffero, On the automorphism group of a closed G_2-structure. *Q. J. Math.* **70** (2019), no. 1, 195–200 Zbl 1414.53019 MR 3927848

[57] F. Xu and R. Ye, Existence, convergence and limit map of the Laplacian flow. 2009, arXiv:0912.0074

Anna Fino
Dipartimento di Matematica, Università degli Studi di Torino, Via Carlo Alberto 10,
10123 Torino, Italy; and Department of Mathematics and Statistics, Florida International
University, Miami, FL 33199, USA; annamaria.fino@unito.it, afino@fiu.edu

Alberto Raffero
Dipartimento di Matematica, Università degli Studi di Torino, Via Carlo Alberto 10,
10123 Torino, Italy; alberto.raffero@unito.it

DOI 10.4171/8ECM/40

Computational/algorithmic thinking in school mathematics

Djordje M. Kadijevich

Abstract. As a result of the globalization and internationalization of the mathematics curriculum, there is, for example, a rapidly developing interest in including computational/algorithmic thinking (CT/AT) in mathematics education. After briefly presenting an emerging educational context regarding the application of CT, this contribution first examines critical issues of CT/AT concerning the notion of CT/AT, the state of CT/AT-oriented educational research, and the integration of CT/AT in the school mathematics curriculum. Then, it presents how CT might be cultivated through data practice and, to this end, data modeling using interactive displays is applied. The contribution ends with a summary of the issues examined and implications for research and practice. This contribution is an extended version of a keynote talk delivered at the symposium "Mathematics in Education."

1. Introduction

Today, technology is increasingly used in all areas of work and life. To practice problem solving with technology successfully, apart from applying disciplinary reasoning (i.e., reasoning applied in the particular discipline such as mathematics), students need to apply computational thinking (CT), which, in short, denotes reasoning processes used in solving problems when solutions are represented in forms that can efficiently be performed by computers [63]. CT clearly involves some degree of algorithmic thinking (AT) that is applied in the work with algorithms. This is because algorithms are used to describe, in a precise manner, which steps (e.g., calculations, visualizations, logical inferences) need to be taken and in what order, to solve the problem under consideration (e.g., [9]).

Recent educational research evidences a growing number of researchers and educators who call for cultivating CT, not only in teaching computer science (informatics) but also in teaching other subjects, such as mathematics and statistics. To illustrate this state, the role of CT in three international projects is summarized below.

2020 Mathematics Subject Classification. Primary 97C70; Secondary 97B99, 97C30, 97K40, 97P99.
Keywords. Algorithmic thinking, computational thinking, data modeling, interactive displays, school mathematics.

- Students' knowledge and skills regarding computer and information literacy have been evaluated worldwide using International Computer and Information Literacy Study carried out by the International Association for the Evaluation of Educational Achievement (https://www.iea.nl/). In a 2018 study cycle, students' CT was assessed for the first time, using tasks that required them to analyze problems, divide these into subproblems, and then find steps that lead to their solutions [18].
- The Organization for Economic Co-operation and Development (https://www.oecd.org/) has evaluated educational achievements in reading, mathematics, and science worldwide using the well-known PISA study (Program for International Student Assessment). In its current 2021 cycle [48], students' CT is assessed for the first time by including it in the steps of mathematical modeling that have already been used in this study (e.g., in the step of employing, one may apply technology to find exact/approximate solutions).
- The growing need for experts in the field of data science, i.e., for the so-called data scientists (e.g., [34]), demands the development of skills required by such a profession, including CT. Within an international project named International Data Science in Schools Project (http://www.idssp.org/), the content of the high school subject on data science has been developed, aiming at the integration of computational and statistical thinking. The development of various resources to support teachers in the realization of this subject is planned [23].

In the remaining text of this contribution, we first consider critical CT/AT issues concerning their definition, state of research, and curricular integration. This consideration is primarily based on a recently published encyclopedia entry [56]. Then we present a way to cultivate CT through data practice to support the position that other learning practices (not only programming, as is often assumed) could be used to develop CT/AT. This presentation is mainly based on two recently published contributions: a chapter in an edited book [31] and an entry in an encyclopedia [33]. The contribution concludes with a summary of the issues examined and implications for research and practice.

2. Critical CT/AT issues

This section comprises three subsections. The first clarifies the notion of CT/AT, the second summarizes the current state of CT/AT-oriented educational research, whereas the third examines the integration of CT/AT in the school mathematics curriculum.

2.1. Definition

As mentioned in Section 1, algorithms are used to describe, in a precise manner, which steps (e.g., calculations, visualizations, logical inferences) need to be taken and in what order to solve the problem under consideration. It is usually assumed that the notion of AT is used to describe reasoning processes applied in work with algorithms, which require their user/developer to comprehend, test, evaluate, correct, improve, or design them. CT clearly involves work with algorithms because to apply computers in problem solving, problem solutions have to be represented in an algorithmic fashion ("do this, do that, in the following order") using forms that are recognized by the computer programs applied.

CT is widely spread and increasingly used in educational research. As CT is based on AT, it may be expected that most researchers agree on what the notion of CT stands for. This is not true, however, because a widely accepted definition of CT is lacking [45].

Various CT definitions have been proposed in the literature. To define CT, researchers have referred to its main facets, practices, concepts, components, and dimensions, and examined them within the specific educational context. This context ranged from specific subject area(s), such as programming or STEM (Science, Technology, Engineering, and Mathematics) education, to a general educational setting such as K-12 subjects [56]. In a high school STEM context, for example, CT may be applied in (and thus cultivated by) various learning practices, including *data practices* (e.g., preparing data and visualizing them), *modeling and simulation practices* (e.g., building and using computational models), and *computational problem-solving practices* (e.g., programming, troubleshooting) [61].

A recent review showed that to clarify the notion of CT, researchers have used and combined entities of different sorts. Most of these entities refer to thinking processes (e.g., abstraction), problem solving methods (e.g., simulation), standard implementation practice (e.g., debugging), and general skills (e.g., technology solution design) [43]. To make progress in CT/AT-related research, researchers may focus on similarities in proposed CT definitions rather than on their differences. CT entities are still common in many of these definitions, such as decomposition (i.e., breaking a problem down into subproblems), abstraction (i.e., making general statements concerning particular examples), and algorithms [55]. These three entities are highly relevant to mathematics learning through programming because during this learning, CT makes use of decomposition, abstraction, pattern recognition, and algorithmic thinking [22].

Regarding CT's main entities (cornerstones), instead of decomposition, abstraction, pattern recognition, and algorithmic thinking, researchers may consider decomposition, abstraction, algorithmization, and automation. There are two reasons for this conceptual shift. Firstly, pattern recognition may be viewed as an instance of abstraction and generalization [53]. Secondly, CT relies on automation of calculations, i.e., using computers that apply certain computational models; a human may formulate a problem solution, but this solution is primarily carried out by a computer not by a human [39]. Although the term CT was coined more than forty years ago [49], it can

```
Known facts about triangle
side(a).
side(b).
angle(alpha).
angle(beta).
opposite(alpha, a).
opposite(beta, b).
greater_side(a, b).

New fact added
greater_angle(alpha, beta).

Discovered rule afterwards
greater_angle(X, Y) :- angle(X), angle(Y), opposite(X, X1), opposite(Y, Y1),
    side(X1), side(Y1), greater_side(X1, Y1).
```

Figure 1. Rule discovery.

be said that CT has been used for centuries to design computational procedures and computing machines to automate them; to formalize a computing procedure, mathematicians have usually described its steps using an algorithm [12], which may also deal with (frequently overlooked) model of computation (to be) applied [11].

If we accept the conceptual shift mentioned above, AT main entities (cornerstones) might then be defined as decomposition, abstraction, and algorithmization, and it is precisely the application of automation that separates AT from CT. This means that AT is not equal to CT but is rather included in it [28, 32]. Interestingly, mathematics educators/researchers may prefer to use AT even when technology is applied, whereas computer science educators/researchers may prefer to use CT even when technology is not used (see [6, 42] for this preference), which may be the result of distinguishing (securing) the position and role of AT/CT in their discipline.

Although the relevance of automation to CT cannot be questioned, its importance to the development of mathematical thinking might be lower than that of other CT cornerstones (decomposition, abstraction, algorithmization), which were also critical learning activities in Pólya's [50] approach to problem solving [13]. Such a state that puts automation in the CT background was found in a recent study involving twenty-five mathematics and computer science experts. They considered CT aspects in mathematics courses and reached a much lower consensus for applying automation than that for using decomposition, abstraction, and algorithmic thinking [36].

It might be that the role of automation is devaluated in general for a few reasons but this position is questionable. By applying abstraction (e.g., through selecting variables), we provide building blocks for automation to be carried out. However, computer programs may not only provide means to support abstraction (e.g., through the work with classes in object-oriented programming), but also they may do abstraction themselves as well. Think about a computer program that enables rule discovery (e.g., [10]). Figure 1 presents facts that may be needed to support the discovery of the well-known rule that says that a greater angle of a triangle is opposite a greater side.

A similar bi-directional relationship holds true for decomposition. By applying decomposition (through identifying substantial or relational sub-problems [52]), we also provide building blocks for automation to be carried out. Regarding the contribution of automation to decomposition, consider computer programs (the so-called expert systems) that instruct their users which sub-problems to solve first and how to use their solutions in order to solve the initial problems (e.g., [24, 25]). A video available at https://www.mi.sanu.ac.rs/~djkadij/FRA20.avi presents the work with such a system concerning problems on multiple proportion (e.g., if three workers repaired 6 windows working 8 hours, how many workers are needed to repair 9 windows working 6 hours?)

2.2. State of research

The notion of CT originated from learning mathematics with technology, i.e., the work with Turtle Geometry through LOGO programming more than forty years ago [49]. Since 2000, due to the elaboration of CT done by a computer scientist Wing [63, 64], this notion has been mostly used by computer science experts, who link it with computer science topics, mostly programming [21]. As a result, during the previous decade, CT has become a critical curricular component in computer science (informatics) education in a number of countries worldwide (e.g., [59]).

Due to limited research on CT in mathematics learning, CT has not had a similar status in mathematics education. Studies on linking CT and learning mathematics in an explicit way are rather rare, and in doing so they mostly refer to areas that are traditionally connected to programming, including numbers and operations, algebra, and geometry. There are, of course, other areas suitable for this linking, such as functions, probability, and statistics. Functions might be explored through modeling, probability through simulations, whereas statistics could better be understood through data analysis [21]. In solving problems, these areas are often combined. Data analysis may, for example, reveal the most probable distribution of the values of a particular variable, and this distribution might be used to build a mathematical model with simulation.

To pursue these explorations successfully, appropriate learning paths need to be followed. Such paths have been proposed outside the mathematics education community, such as an understand-debug-extend path [8], or a use-modify-create path [40], which, when combined, may result in the following path: use problem solutions (to understand or evaluate them) – modify problem solutions (to debug or extend them) – create problem solutions, i.e., develop problem solutions from scratch. Mathematics educators have proposed CT pedagogy for the work with various conceptual or digital objects in the classroom [38]. The proposed pedagogy assumes that this work makes use of four overlapping activities: *unplugging* (not using computers), *tinkering* (dividing existing objects into their components and changing or modifying these components), *making* (constructing new objects), and *remixing* (producing new objects through the appropriation of existing objects or their components). As an example of unplugging, consider sorting mathematical expressions. Tinkering is applied when the content of a spreadsheet is modified, whereas remixing is practiced when a dashboard (a set of interactive reports) is created through combining and modifying existing interactive reports. Although these four activities (unplugging, tinkering, making, remixing) are present in the combined learning path mentioned above, this pedagogy can be applied without using computers, which opens the question "What is actually being developed when computers are not used: CT or (nevertheless) AT only?"

AT is a central activity in mathematics. Although AT is widely practiced in mathematics classes (though mostly implicitly), research on AT in mathematics learning is also limited. There are, however, several studies whose valuable findings may contribute to fostering both AT and mathematics learning. It was found, for example, that procedural knowledge rich in connections could be developed through designing and implementing procedures and algorithms [42]. AT may also be used to develop conceptual knowledge representing a deeper conceptual understanding when a special case of an algorithm in general, or a formula in particular, is considered in detail to ask advanced questions about its result [1]. In other words, AT may contribute to developing and relating procedural and conceptual mathematical knowledge. When AT is supported by technology (i.e., when CT is practiced in our terms), it is important to understand in what ways mathematics learning could be mediated by technology [14], especially in developing and relating these two types of mathematical knowledge (e.g., [2, 30]). To develop AT gradually, the following learning path (derived from the combined path mentioned above) could be applied: consider formulas, procedures, and algorithms given (to understand or evaluate them) – modify formulas, procedures, and algorithms given (to debug or extend them) – create formulas, procedures, and algorithms, i.e., develop them from scratch. Furthermore, as in case with CT, the activities comprising this path are not realized separately but, as a rule, overlap each other.

Although research on CT/AT in mathematics learning is limited at present, it seems to be a growing research area as evidenced, for example, with the inclusion of CT in PISA 2021 [48]. Also, in 2021, research and practice regarding CT/AT were explicitly represented (probably for the first time) at an international congress on mathematical education. In particular, at the 14th International Congress on Mathematical Education" (ICME-14, https://www.icme14.org), there was a topic study group titled "Teaching and learning of programming and algorithms" (TSG-14) and a discussion group titled "Computational and algorithmic thinking, programming and coding in the school mathematics curriculum: Sharing ideas and implications for practice" (DG-1), whose participants emphasized the importance of fostering CT/AT in mathematics education. This might be done through problem solving using a CT/AT lens described in this contribution. Such an approach would result in more focused (and explicit!) instruction on AT and its core components (decomposition, abstraction, algorithmization), possibly supported by particular computer programs. As assumed by a model of mathematical thinking based on the triad abstraction-modeling-problem solving [15], these components denote critical activities applied in mathematics learning.

2.3. Curricular CT/AT integration

This subsection comprises three parts. The first explains the rationale for this curricular CT/AT integration, the second summarizes different models of integrations applied worldwide, while the third examines various educational implications of this integration.

2.3.1. Rationale for integration. More and more workplaces require specialized knowledge based on the use of modern information-communication technology (ICT); tens of millions of specialists with this knowledge are needed today worldwide [4]. Among them data scientists are particularly important, whose competencies (e.g., [3]) are essentially supported by CT/AT. An increasing demand to (better) prepare students for a range of ICT-based jobs (with many future ones unknown at present) clearly provides a good rationale for the inclusion of CT/AT in school mathematics. There is another good rationale for this inclusion. Due to an increasing reliance on computations in scientific inquiry (e.g., [17]), students should learn how to solve problem with technology for the development of their mathematical thinking. To this end, they should act as information-processing agents (e.g., [64]). Although these two rationales clearly represent different perspectives (a societal one *vs* a professional one [7]), they are not separated, obviously influencing each other. Note that an extensive rationale for including CT in school mathematics (at least for senior high school students) was elaborated in a discussion paper developed by four mathematical and computer science academies in France by using the following arguments: (1) CT is becoming increasingly embedded in university courses in mathematics; (2) certain areas such as graphs, combinatorics, and logic could be used to establish creative interfaces between mathematics and computer science; (3) CT can strengthen students' mathematical development [19]. This paper also contains a number of examples that can be used to foster creative interfaces between mathematics and informatics (computer science).

2.3.2. Models of integration. Various models of the integration of CT in the school mathematics curriculum have been applied worldwide. Let us provide some examples. To integrate CT/AT across different school subjects, a cross-curriculum model may be applied like in Finland. If this integration is realized within the curriculum of an information technology (IT) subject, an IT model may be in use like in Australia and England. CT/AT may be integrated in mathematics and other school subjects in several grades gradually, meaning that a gradualist model is being applied, like in Japan where the focus is on programming thinking not on CT/AT. Finally, the integration may be realized within a new school subject, like in France (subject *Algorithmique et Programmation* in the middle grades taught by mathematics and IT teachers) and Australia (subject *Algorithmics* in the senior high school). Clearly, although CT/AT integration has often been realized within one or several existing subjects, it could be done within new subjects as well. Although all these models remain unexamined and are by and large untested, certain *pros* and *cons* can be identified. For example, the cross-curriculum model might be implemented in a shallow way; in the IT model, teachers may focus more on using technology than on mathematical connections; the gradualist model allows time for teacher preparation, but creating interfaces between school subjects with entrenched boundaries would be challenging; a separate subject, especially if taught by mathematics and IT teachers, can provide opportunities for exploring interfaces between mathematics and computer science, but may, at a higher educational level, require rich prior experience with CT/AT [56]. For a thorough evaluation, the curricular integration of CT may be examined in terms of critical curricular components (e.g., goals, content, materials, forms of teaching, student activities, assessment [47]). Note that a detailed integration of CT in the school mathematics curriculum is planned in Australia. The new Australian F-10 curriculum for mathematics (from Foundation to Year 10) calls for the application of CT in problem solving, and gives examples and instructions of doing that from Year 4 to Year 10 [5]. In this document, the phrase *computational thinking* occurs almost forty times (e.g., Year 10: "apply computational thinking to model and solve algebraic problems graphically or numerically"). In July 2021, the status of this document was "waiting for approval."

2.3.3. Educational implications. Due to technological advances, computational mathematics has been increasingly used in research mathematics; there are great number of respectable research publications with the words *computational* and *mathematics* in their titles, whose authors, stated briefly, primarily examine various algorithms carried out by computers. Such a reliance on computations has changed the practice of scientific inquiry in which "together with theory and experimentation, a third pillar of scientific inquiry of complex systems has emerged in the form of a combination of modeling, simulation, optimization, and visualization" [17, p. 2]. Hence, the development of CT/AT in mathematical classes should cultivate such an inquiry by applying different kinds of practice, such as those already mentioned *data practices* (e.g., preparing data and visualizing them), *modeling and simulation practices* (e.g., building and using computational models), and *computational problem-solving practices* (e.g., programming, troubleshooting) [61]. To this end, instruction may relate (integrate) content, technology, and pedagogy through, for example, identifying relevant CT/AT practice(s) for each curriculum strand content descriptor and computer tool available (e.g., [51]).

Although classroom practice should be different from disciplinary practice (increasingly using computation to support experimentation, approximation, conjecture testing, visualization, and other aspects of mathematicians' work), the latter should inform the former and help design it [41]. Hence, students may in general use CT/AT to define (construct) objects, identify their possible properties (of algebraic, geometric, or other nature), and verify these properties. Furthermore, like mathematicians who apply computation to find approximate solutions to intractable problems, students may use CT/AT to approximate solutions of mathematical models that cannot (easily) be solved in the context of school mathematics (e.g., [37]). Regarding the use of algorithms in particular, it may, for example, support students to (1) unpack concepts and procedures, (2) identify the mathematical structure of a given problem and generalize its solution, (3) familiarize themselves with modeling, optimization, operations research, and experimental mathematics, and (4) generate examples of problems for which the given algorithm works or does not work [56].

Some readers may insist on the position that despite the fact that various CT/AT-based practices might be applied to solve a variety of task types, CT/AT should nevertheless be promoted primarily through programming. The following examples may help these readers make this position less strict. In preparing these examples, it was supposed that we apply CT whenever we recognize aspects of computations in problem solving and deals with them in appropriate ways by using tools and techniques from computers science [57]. Regarding this computing support, the examples make us of Wolfram Alpha (https://www.wolframalpha.com/). Of course, the use of computing support in general may generate various learning challenges and appropriate didactic treatments need to be applied to alleviate them (e.g., [20, 26]).

Example 1. To determine the greatest common divisor, one can simply use a built-in command `gcd`, such as `gcd(24,16)` that yields 8. Another way to do this is to apply a four-step-approach: (1) find the set of the first number divisors, (2) find the set of the second number divisors, (3) determine the intersection of these sets, and (4) find the maximum value in the intersection set (with each step supported by an important algorithm). To combine these steps, clearly in an algorithmic fashion, use the following commands: `Max[intersect[divisors(24),divisors(16)]]`.

Example 2. To discover functional dependence that connects two arrays of natural numbers, we may apply a curve fitting approach with perfect fit. The number of diagonals in a triangle, quadrilateral, and pentagon are 0, 2, and 5, respectively. If the number pairs $(3, 0)$, $(4, 2)$, $(5, 5)$ are fitted with a quadratic model, the following dependence is found $0.5x^2 - 1.5x$, and this fit is perfect because $R^2 = 1$. When this dependence is factorized, the result is $0.5(x-3)x$, directing students what key elements to consider: the role of x is clear, but why 0.5 and $x-3$ are included? Relevant commands are `quadratic fit`$\{3,0\}$,$\{4,2\}$,$\{5,5\}$ and `factor`$(0.5\mathtt{x}^2-1.5)$.

As AT is critical to the processes of conjecturing and proving, the development of algorithms may be connected with these processes. There are some areas of discrete mathematics (e.g., combinatorics, graph theory) that are particularly suitable for fostering creative interfaces between mathematics and computer science through exploring relations between algorithm, proof, logic, and programming [44]. In this exploration, different conceptions of algorithm might emerge: an algorithm is implicitly included in the proof of a theorem (if the activity is from a problem to a theorem to a proof, or, in short, problem-theorem-proof); a proof of the correctness of an algorithm is given (problem-algorithm-proof); an algorithm is given as a computer program whose validity is established in some way (problem-program-validation) [16].

Although the exploration sketched in the previous paragraph may only be suitable for senior high school students, an algorithm should be considered not only as a useful tool that can solve certain problems, but also as a separate entity that can be investigated in itself. For example, apart from applying the algorithm for determining the greatest common divisors of two natural numbers when we use this algorithm as a tool, we may examine its applicability to whole or other numbers (or its complexity in terms of the number of operations needed to complete it) when we treat the algorithm as a separate entity (e.g, [16]). Such an approach calls for considering the so-called process-object nature of algorithm, whenever this approach is appropriate and accessible to students. This dual nature also characterizes other mathematical entities, such as relations and functions (e.g., [54]).

3. Cultivating CT through data practice

3.1. Preliminaries

As mentioned in Section 2.2, CT has mostly been cultivated through programming (e.g., [21]), which is hence often assumed as a dominant learning practice that would support CT development. However, to this end, other learning practices might be applied as well (e.g., [61]). Among these are data practices (e.g., data preparation and visualization) that may activate different CT components, such as abstraction, decomposition, and pattern recognition.

The relevance of data practices to developing CT is, for example, recognized by a CT definition that refers to core CT facets, assuming that these facets might be: abstraction (data collection and analysis, pattern recognition, modeling), decomposition, algorithms (algorithm design, parallelism, efficiency, automation), iteration, debugging, and generalization [55]. Bearing in mind this relevance, CT assessment may also include some aspects of data practice. This was, for example, done in a large worldwide assessment named ICILS 2018 (International Computer and Information Literacy Study completed in 2018), which used tasks that called for programming as well as structuring and manipulating datasets [18].

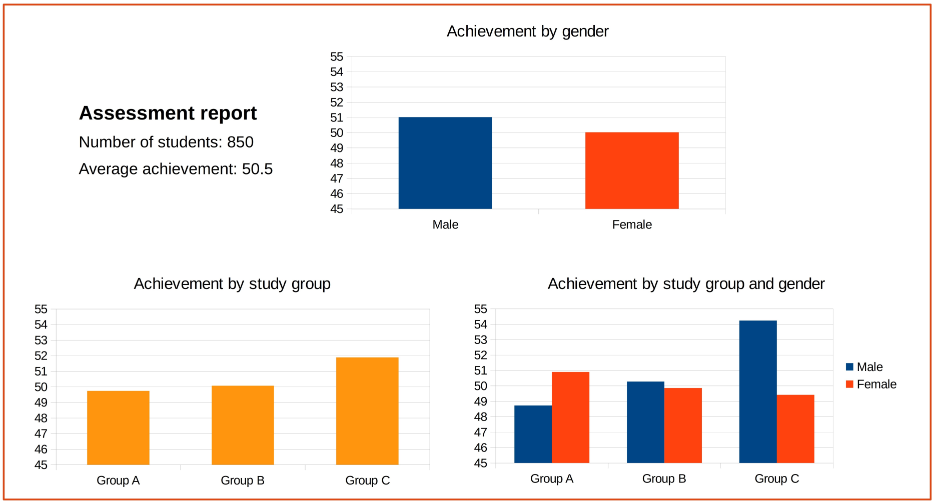
As mentioned in Section 2.3.1, tens of millions of workers with specialized ICT knowledge are needed worldwide today, among whom data scientists are particularly important, applying various (often complex) techniques from mathematics, statistics, and computer science to obtain useful information from (big) datasets. It is reasonable to expect that in their future jobs most students would have to work with data as a foundation for their claims and actions regarding various professional issues, and, to this end, they may primarily use some simple data science techniques. Among these is exploratory data analysis that is applied to summarize the main characteristics of the dataset analyzed by using data visualization methods, primarily charts, aiming at discovering what the data can tell us not at formal data modeling or hypothesis testing [58]. This expectation regarding such use of exploratory data analysis is supported by the increasing application of dashboards (e.g., [62]), which are particularly suitable tools for this kind of analysis. In a specialized computer environment, building charts and dashboards (combining various types of charts and summary measures) can be (relatively effortlessly) done visually using the drag-and-drop approach.

Dashboards are interactive displays that are composed of two or more interactive reports, mostly charts, whose content updates automatically whenever there are changes in data or variables considered [33]. Dashboards are today used in various industries and areas (for a gallery of dashboards, visit https://www.yellowfinbi.com/analytics-best-practice/dashboard-gallery). Among them is learning analytics in education (e.g., [60]), where such interactive displays summarize the values of various learning indicators. Dashboards may also be used in education to support the work with data in various school subjects and university courses. If this work is practiced within a suitable learning cycle (e.g., a mathematical modeling cycle [29]), it would not only support the understanding of this cycle and the realization of its values in capturing the main features of disciplinary thinking (i.e., thinking applied in the particular discipline), but also support the development of important (disciplinary or general) notions, such as variable and functional dependence [33]. In other words, although interactive displays are primarily a means for visualizing data, they can also be a learning tool if used within an appropriate learning cycle [31]. Note that a growing demand for the inclusion of data science in secondary education (e.g., [23]) may, at introductory levels, profit from the work with interactive displays, whose visualizations (although mostly based on simple mathematical models such as frequencies, sums, and means) can support the discovery of useful (interesting) patterns, trends, effects, and interactions in the data examined. To find an interesting interaction in a tourism dataset, the reader may examine the visualizations available at https://www.mi.sanu.ac.rs/~djkadij/Dashboard.htm.

3.2. Data modeling using dashboards

The key activities in data modeling using dashboards (key stages) may be Asking questions, Preparing data, Visualizing data, Answering questions, Validating modeling, and Recommending changes. Apart from Visualizing data, the use of dashboards would support other data modeling steps, especially Answering questions and Validating modeling. In Answering questions, students have to match patterns, trends, effects, and interactions found with the questions posed, whereas in Validating modeling they might improve the modeling applied by using other variables or charts, or even other data or another dashboard. When datasets to model are not given to students, the use of dashboards may also support (though not primarily) the stage of Preparing data, because they may signal some oddities in data (e.g., outliers, missing or inappropriate data) that should be addressed before the stage of Visualizing data is applied. In most cases, datasets to model should be given to data modelers, especially novices, because removing these oddities is a very challenging task, even for data scientists who usually spend most of their time preparing data, i.e., collecting, cleaning, and organizing data [29].

Data modeling using dashboards clearly calls for abstraction (e.g., in using variables), decomposition (e.g., in deciding what charts to include in a dashboard, or what variables to use in a chart and in what role), and pattern recognition (e.g., in recognizing an effect or a trend in data). Apart from decomposition, this modeling would promote other computational strategies, such as top-down and bottom-up approaches [31], recalling that these approaches are relevant to mathematical problem solving proposed Pólya's [50]. A top-down approach is applied when the modeler goes from a dashboard as a whole to its individual reports as parts, whereas a bottom-up approach is used when he/she starts from some individual reports and combine them to create a dashboard; instead of a single approach, their combination is often applied. In addition, building a dashboard may make use of another computational strategy called rapid prototyping, which denotes an iterative process through which the modeler incrementally presents what the dashboard under development will look like in order to get feedback and validation from peers and future users [31]. This strategy is, in general, relevant to mathematical modeling whenever models of increased complexity are developed in an incremental fashion. To consider a way to promote these computational strategies, the reader may consider the development of a dashboard whose content is presented in Figure 2, but it should be kept in mind that only a basic understanding of these strategies may be promoted because the applied dashboard development (as is the case most often) calls for simple system engineering



Assessment report

Number of students: 850

Average achievement: 50.5

Figure 2. Assessment dashboard.

[31]. Note that although the three computational strategies, especially rapid prototyping, have been under-represented in CT-related research, there is a CT facet named iteration [55] under which these strategies might be discussed.

As the previous consideration shows, the presented work with data offers a number of learning opportunities: cultivating a modeling (or a data inquiry) cycle; supporting the development of important disciplinary notions (e.g., variable and functional dependence); promoting a basic understanding of CT strategies, such as decomposition as well as rapid prototyping, and top-down and bottom-up approaches. To be practiced skillfully, this work requires the modeler to demonstrate a range of skills, such as choosing relations to examine, identifying dependent and independent variables (Asking questions), selecting charts and measure to use (Visualizing data), recognizing regularities in charts produced, and connecting regularities to questions asked (Answering questions) [31]. A number of challenges would be faced in the development and use of these skills. Among them are the following: using appropriate sets of variables to answer questions; selecting appropriate charts and measures; considering context properly to interpret findings. There are several possible reasons for these challenges, such as complexity of this data practice when considered as a design task; limited experience in using various charts and measures; and complex interactions of knowledge from different domains [27, 29]. To alleviate these and other challenges, hints and supports (the so-called scaffolds) need to be provided to modelers, which would hopefully enable them to complete successfully, on their own, data modeling using dashboards. These scaffolds may connect key stages using their underlying skills (e.g., variables selection with charts production; charts production with regularities recognition) and link contextual/conceptual and technology-related issues (e.g., between questions to ask and chart types to use, visualizations produced and questions to answer, or modeling features to validate and technology components used) [29, 31].

Data modeling using dashboards should be practiced within a rich computational environment that supports various CT assets, such as ZOHO Analytics (https://www.zoho.com/analytics/). Bearing in mind the learning paths examined in Section 2.2, to support (and empower) this practice, the following learning path may be applied: examine dashboards to understand or evaluate data modeling (DM) completed – modify dashboards to debug or extend DM done – create dashboards to perform DM by yourself. To assess the outcome of data modeling using dashboards, the instructor may examine students' portfolios about dashboards evaluated, improved, or fully developed (done individually or through a cooperative work), focusing on success of pursuing each key DM stage and connecting these stages in terms of major skills underlying them and their links [31].

Although the presentation of data modeling using dashboards is linked to fostering CT in a mathematical context, this modeling may also contribute to fostering CT in other school subjects if embedded in another disciplinary context using an appropriate learning cycle (e.g., in statistics using a data inquiry cycle). Such a data practice is also in accord with a CT pedagogy regarding a range of disciplines that calls for focusing on interactive visualizations or simulations, modeling and troubleshooting of datasets, and searching for patterns in large datasets [46]. Regarding the work with data in general, this focus aligns with an already underlined today's practice of scientific inquiry, whose three pillars are theory, experimentation, and a combination of modeling, simulation, optimization, and visualization [17].

4. Closing remarks

After briefly presenting an emerging educational context regarding the application of CT, this contribution first examined critical issues of CT/AT concerning the notion of CT/AT, the state of CT/AT-oriented educational research, and the integration of CT/AT in the school mathematics curriculum. Although a widely accepted definition of CT is lacking, it was argued that CT cornerstones might be decomposition, abstraction, algorithmization, and automation, where the first three might comprise AT. The examination of the state of CT/AT-oriented educational research showed that research on CT/AT in mathematics learning is limited but growing, being concerned with exploring various areas through different activities to foster this learning, especially developing and relating procedural and conceptual mathematical knowledge. Regarding the integration of CT/AT in the school mathematics curriculum, the rationale for doing that is supported by both societal and professional needs. Various models have been used (within one or several existing school subjects or within new school subjects), but they remain unexamined and are by and large untested. To develop CT/AT, instruction should, whenever appropriate and accessible to students, be based on applying a range of activities designed in accord with disciplinary practice (increasingly empowered by computations), exploring interfaces between mathematics and computer science, and considering the dual nature of algorithm (a tool to apply as well a separate entity to investigate). In doing that, suitable learning paths may be followed, such as use problem solutions (to understand or evaluate them) – modify problem solutions (to debug or extend them) – create problem solutions, i.e., develop problem solutions from scratch.

After the examination of these critical CT/AT issues concerning their notion, examination in educational research, and integration in the school mathematics curriculum, this contribution presented a way to cultivate CT through data practice. This practice, which has been increasingly advocated in educational research, is based on using sets of interactive reports called dashboards. The rationale for using such interactive displays is supported by an expectation that in their future jobs, most students would have to work with data as a foundation for their claims and actions regarding various professional issues, and to this end, they may primarily apply exploratory data analysis with dashboards, because on one hand, this analysis, as an introductory data science technique, could be accessible to most students, and, on the other, dashboards, which have been increasingly applied in various industries and areas, are particularly suitable tools for this kind of analysis. After describing the key stages in data modeling with dashboards, CT components involved in this modeling are discussed (e.g., pattern recognition), especially computational strategies (e.g., top-down approach), which, despite their educational relevance, have been under-represented in CT-related research. Next, various learning issues concerning the proposed data modeling with dashboards were discussed, including learning opportunities, underlying skills required, expected challenges in practicing this modeling and possible reasons for these challenges, scaffolds that would alleviate these challenges, as well as a learning path that may be followed in practicing this modeling. Finally, it was considered whether the advocated data practice is aligned, in a pedagogical way, with today's practice of scientific inquiry. All in all, the presentation showed that data modeling using dashboards may be a promising way to cultivate CT, provided that the discussed learning issues are adequately treated.

The content of this contribution has evidenced that more research is needed on linking CT with mathematics learning. Although it showed how CT could be developed through exploring the area of statistics using exploratory data analysis with dashboards, this is just an initial research step in this research direction. Further research may be (more) concerned with, for example, exploring various areas (e.g., probability) through different activities (e,g., simulation), designing these activities in accord with disciplinary practice (e.g., experimenting and approximation), exploring interfaces between mathematics and computer science (e.g., computational geometry), and considering algorithm in (more) explicit and detailed way (e.g., its dual nature). The outcomes of such a directed research would considerably inform instruction and help designing it.

Bearing in mind that the models of curricular CT integration remain largely unexamined, research is also needed on this integration, and, to this end, research could apply a detailed evaluation, which may, as already suggested, examine critical curricular components, such as goals, content, materials, forms of teaching, student activities, and assessment. As materials considerably influence teaching, learning, and assessment, applying appropriate materials, developed in lines that align with the proposed goals and content, seems to be the most critical component not only for this integration, but also for teacher education and further professional development.

To summarize, as there is a long-standing reliance on algorithms in mathematics, CT/AT should be cultivated in mathematics education, especially today with a growing application of computer tools in almost every areas of our work and life. Although thinking supported by technology has been named differently in the literature – computational, algorithmic, or even programming thinking – and defined in a number of ways, the focus in mathematics education should be on cultivating the aspects of mathematical thinking using tools and techniques from computer science. If this cultivation, supported by various suitable materials describing CT/AT based activities, is realized in appropriate ways in mathematical classes, the integration of CT/AT in the school curriculum would be a success. Undoubtedly, such an integration calls for international cooperation and sharing among educators and researchers at all educational levels. In doing that, special care may be taken about the following issues: how to define thinking with technology in a precise way; how to cultivate this thinking accordingly, focusing on the development of mathematical reasoning; and how to assess its contribution to this development in an adequate way [35].

Acknowledgments. The author dedicates the contribution to his son Aleksandar.

Funding. This contribution resulted from the author's research funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Contract No. 451-03-68/2022-14/200018).

References

[1] S. Abramovich, Mathematical problem posing as a link between algorithmic thinking and conceptual knowledge. *Teach. Math.* **18** (2015), no. 2, 45–60

[2] M. Artigue, The future of teaching and learning mathematics with digital technologies. In *Mathematics Education and Technology – Rethinking the Terrain. The 17th ICMI Study*, edited by C. Hoyles and J. B. Lagrange, pp. 463–476, Springer, New York, 2010

[3] Asia Pacific Economic Cooperation (APEC), *Data science and analytics skills shortage: Equipping the APEC workforce with the competencies demanded by employers*. APEC, Singapore, 2017, https://www.apec.org/Publications/2017/11/Data-Science-and-Analytics-Skills-Shortage

[4] Asia Pacific Economic Cooperation (APEC), *Project DARE (Data Analytics Raising Employment)*. APEC, Singapore, 2018, https://www.apec.org/Press/News-Releases/2018/1109_dare

[5] Australian Curriculum, Assessment and Reporting Authority (ACARA), *Australian curriculum: Mathematics – All elements F–10 consultation curriculum*. NSW, Sydney, 2021, https://www.australiancurriculum.edu.au/media/7044/mathematics_all_elements_f-10.pdf

[6] T. Bell and J. Vahrenhold, CS unplugged—How is it used, and does it work? In *Adventures Between Lower Bounds and Higher Altitudes*, edited by H. J. Böckenhauer, D. Komm, and W. Unger, pp. 497–521, Lect. Notes Comput. Sci. 11011, Springer, Cham, 2018

[7] S. Bocconi, A. Chioccariello, G. Dettori, A. Ferrari, and K. Engelhardt, *Developing computational thinking in compulsory education – Implications for policy and practice*. Joint Research Centre, European Commission, European Union, Luxemburg, 2016, https://publications.jrc.ec.europa.eu/repository/bitstream/JRC104188/jrc104188_computhinkreport.pdf

[8] K. Brennan and M. Resnick, New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 Annual Meeting of the American Educational Research Association (Vancouver, Canada)*, 2012

[9] C. Clapham and J. Nicholson, *The Concise Oxford Dictionary of Mathematics*. 5th edn., Oxford University Press, Oxford, 2014 Zbl 1290.00009 MR 3186178

[10] L. de Raedt and M. Bruynooghe, Interactive concept-learning and constructive induction by analogy. *Mach. Learn.* **8** (1992), no. 2, 107–150 Zbl 0751.68051

[11] P. J. Denning, Remaining trouble spots with computational thinking. *Commun. ACM* **60** (2017), no. 6, 33–39

[12] P. J. Denning and M. Tedre, *Computational Thinking*. MIT Press, Cambridge, MA, 2019

[13] A. A. DiSessa, Computational literacy and "the big picture" concerning computers in mathematics education. *Math. Think. Learn.* **20** (2018), no. 1, 3–31

[14] P. Drijvers, Tools and taxonomies: a response to Hoyles. *Res. Math. Edu.* **20** (2018), no. 3, 229–235

[15] P. Drijvers, H. Kodde-Buitenhuis, and M. Doorman, Assessing mathematical thinking as part of curriculum reform in the Netherlands. *Educ. Stud. Math.* **102** (2019), 435–456

[16] V. Durand-Guerrier, A. Meyer, and S. Modeste, Didactical issues at the interface of mathematics and computer science. In *Proof Technology in Mathematics Research and Teaching*, edited by G. Hanna, D. Reid, and M. de Villiers, pp. 115–138, Math. Educ. Digit. Era 14, Springer, Cham, 2019

[17] European Mathematical Society (EMS), Position paper on the European Commission's contributions to European research. 2011

[18] J. Fraillon, J. Ainley, W. Schulz, D. Duckworth, and T. Friedman, *IEA International Computer and Information Literacy Study 2018 Assessment Framework*. Springer, Cham, 2019

[19] Groupe de travail des sociétés savantes de mathématiques et d'informatique, *Propositions pour le futur programme de mathématiques du lycée*. Sociétéinformatique de France, Grenoble, France, 2016

[20] D. Guin, K. Ruthven, and L. Trouche (eds.), *The Didactical Challenge of Symbolic Calculators: Turning a Computational Device into a Mathematical Instrument*. Springer, New York, 2005

[21] D. Hickmott, E. Prieto-Rodriguez, and K. Holmes, A scoping review of studies on computational thinking in K-12 mathematics classrooms. *Digit. Exp. Math. Educ.* **4** (2018), no. 1, 48–69

[22] C. Hoyles and R. Noss, Revisiting programming to enhance mathematics learning. 2015, paper presented at Math + Coding Symposium, Western University, London, Canada

[23] International Data Science in Schools Project (IDSSP) Curriculum Team, Curriculum frameworks for introductory data science. 2019, http://www.idssp.org/files/IDSSP_Frameworks_1.0.pdf

[24] D. M. Kadijevich, Can mathematics students be successful knowledge engineers? *J. Interact. Learn. Res.* **9** (1998), no. 3–4, 235–248

[25] D. M. Kadijevich, An approach to learning mathematics through knowledge engineering. *J. Comput. Assist. Learn.* **15** (1999), no. 4, 291–301

[26] D. M. Kadijevich, Neglected critical issues of effective CAS utilization. *J. Symb. Comput.* **61–62** (2014), 85–99 Zbl 1284.97033

[27] D. M. Kadijevich, Data modelling with dashbards: Opportunities and challenges. In *Promoting Understanding of Statistics About Society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), Berlin, Germany, July 2016*, edited by J. Engel, ISI/IASE, The Haag, the Netherlands, 2016

[28] D. M. Kadijevich, A cycle of computational thinking. In *Proceedings of the 9th International Conference on e-Learning*, edited by B. Trebinjac and S. Jovanović, pp. 75–77, Metropolitan University, Belgrade, 2018

[29] D. M. Kadijevich, Data modelling using interactive charts. *Teach. Math.* **21** (2018), no. 2, 55–72

[30] D. M. Kadijevich, Relating procedural and conceptual knowledge. *Teach. Math.* **21** (2018), no. 1, 15–28

[31] D. M. Kadijevich, Cultivating computational thinking through data practice. In *Empowering Learners for Life in the Digital Age*, edited by D. Passey, R. Bottino, C. Lewin, and E. Sanchez, pp. 24–33, Springer, Cham, 2019

[32] D. M. Kadijevich, A cycle of computational thinking and its relevance: An empirical study. In *Proceedings of the 10th Conference on e-Learning*, edited by S. Jovanović and B. Trebinjac, pp. 136–138, Metropolitan University, Belgrade, 2019

[33] D. M. Kadijevich, Interactive displays: Use of interactive charts and dashboards in education. In *Encyclopedia of Education and Information Technologies*, edited by A. Tatnall, pp. 968–973, Springer, Cham, 2020

[34] D. M. Kadijevich and M. Stephens, Modern statistical literacy, data science, dashboards, and automated analytics and its applications. *Teach. Math.* **23** (2020), no. 1, 71–80

[35] D. M. Kadijevich, M. Stephens, and A. Rafiepour, Emergence of computational/algorithmic thinking and its impact on the mathematics curriculum. In *Mathematics curriculum reforms around the world*, edited by Y. Shimizu and R. Vithal, Springer, Cham, 2023

[36] M. Kallia, S. P. van Borkulo, P. Drijvers, E. Barendsen, and J. Tolboom, Characterising computational thinking in mathematics education: a literature-informed Delphi study. *Res. Math. Educ.* **23** (2021), no. 2, 159–187

[37] P. S. Kenderov, Powering knowledge versus pouring facts. In *Invited Lectures from the 13th International Congress on Mathematical Education*, edited by G. Kaiser, H. Forgasz, M. Graven, A. Kuzniak, E. Simmt, and B. Xu, pp. 289–306, CME-13 Monographs, Springer, Cham, 2018

[38] D. Kotsopoulos et al., A pedagogical framework for computational thinking. *Digit. Exp. Math. Educ.* **3** (2017), no. 2, 154–171

[39] I. Lee, Reclaiming the roots of CT. *CSTA Voice* **12** (2016), no. 1, 3–5

[40] I. Lee et al., Computational thinking for youth in practice. *ACM Inroads* **2** (1), no. 2011, 33–37

[41] E. Lockwood, A. F. DeJarnette, and M. Thomas, Computing as a mathematical disciplinary practice. *J. Math. Behav.* **54** (2019), Paper No. 100688

[42] E. E. Lockwood, A. DeJarnette, A. Asay, and M. Thomas, Algorithmic thinking: An initial characterization of computational thinking in mathematics. In *Proceedings of the 38th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*, edited by M. B. Wood, E. E. Turner, M. Civil, and J. A. Eli, pp. 1588–1595, The University of Arizona, Tucson, AZ, 2016

[43] M. Lodi, Informatical thinking. *Olymp. Inform.* **14** (2020), 113–132

[44] S. Modeste, Impact of informatics on mathematics and its teaching. In *History and philosophy of computing (HaPoC 2015)*, edited by F. Gadducci and M. Tavosanis, pp. 243–255, IFIP Adv. Inf. Commun. Technol. 487, Springer, Cham, 2016

[45] C. Mouza, H. Yang, Y.-C. Pan, S. Y. Ozden, and L. Pollock, Resetting educational technology coursework for pre-service teachers: A computational thinking approach to the development of technological pedagogical content knowledge (TPACK). *Australas. J. Educ. Technol.* **33** (2017), no. 3, 61–76

[46] National Research Council, *Report of a workshop of pedagogical aspects of computational thinking*. The National Academies Press, Washington, DC, 2011

[47] M. Niss, Mathematical standards and curricula under the influence of digital affordances-different notions, meanings and roles in different parts of the world. In *Digital Curricula in School Mathematics*, edited by M. Bates and Z. Usiskin, pp. 239–250, Information Age Publishing, Charlotte, NC, 2016

[48] Organization for Economic Co-operation and Development (OECD), *PISA 2021 mathematics framework (draft)*. OECD, Paris, 2018, https://www.oecd.org/pisa/pisaproducts/pisa-2021-mathematics-framework-draft.pdf

[49] S. Papert, *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books, New York, 1980

[50] G. Pólya, *How to Solve It*. Princeton University Press, Princeton, NJ, 1945

[51] E. Prieto and K. Holmes, Working mathematically and thinking computationally: Capitalizing on commonalities for integrated teaching. 2021, paper presented at Topic Study Group 14 “Teaching and learning of programming and algorithms” at the 14th International Congress on Mathematical Education, https://www.icme14.org

[52] P. J. Rich, G. Egan, and J. Ellsworth, A framework for decomposition in computational thinking. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE ’19)*, pp. 416–421, ACM, New York, 2019

[53] T. Scantamburlo, *Philosophical aspects in pattern recognition research*. Ph.D. thesis, Department of Informatics, Ca’ Foscari University of Venice, Venice, Italy, 2013

[54] A. Sfard, On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educ. Stud. Math.* **22** (1991), no. 1, 1–36

[55] V. J. Shute, C. Sun, and J. Asbell-Clarke, Demystifying computational thinking. *Educ. Res. Rev.* **22** (2017), 142–158

[56] M. Stephens and D. M. Kadijevich, Computational/algorithmic thinking. In *Encyclopedia of Mathematics Education*, edited by S. Lerman, pp. 117–123, Springer, Cham, 2020

[57] The Royal Society, *Shut down or restart? The way forward for computing in UK schools*. The Royal Society, London, 2012, https://royalsociety.org/~/media/education/computing-in-schools/2012-01-12-computing-in-schools.pdf

[58] J. M. Tukey, *Exploratory Data Analysis*. Addison-Wesley, Reading, PA, 1977
Zbl 0409.62003

[59] M. Webb et al., Computer science in K-12 school curricula of the 2lst century: Why, what and when? *Educ. Inf. Technol. (Dordr.)* **22** (2017), no. 2, 445–468

[60] M. E. Webb et al., Challenges for IT-enabled formative assessment of complex 21st century skills. *Technol. Knowl. Learn.* **23** (2018), no. 3, 441–456

[61] D. Weintrop et al., Defining computational thinking for mathematics and science classroom. *J. Sci. Educ. Technol.* **25** (2016), no. 1, 127–141

[62] S. Wexler, J. Shaffer, and A. Cotgreave, *The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*. 1st edn., Wiley, Hoboken, NJ, 2017

[63] J. M. Wing, Computational thinking. *Commun. ACM* **49** (2006), no. 3, 33–35

[64] J. M. Wing, Research notebook: Computational thinking—What and why? *Link Newsl.* **6** (2011), 1–32

Djordje M. Kadijevich
Institute for Educational Research, Dobrinjska 11/III, Belgrade, Serbia; djkadijevic@ipi.ac.rs

DOI 10.4171/8ECM/11

Graph and hypergraph colouring via nibble methods: A survey

Dong Yeap Kang, Tom Kelly, Daniela Kühn, Abhishek Methuku, and Deryk Osthus

Abstract. This paper provides a survey of methods, results, and open problems on graph and hypergraph colourings, with a particular emphasis on semi-random "nibble" methods. We also give a detailed sketch of some aspects of the recent proof of the Erdős–Faber–Lovász conjecture.

1. Introduction

The theory of graph and hypergraph colouring is fundamental to combinatorics, with numerous applications to other areas of combinatorics and beyond. It has also given rise to the introduction and development of techniques that have had a major impact far beyond the settings for which they were initially developed. In this paper, we survey results, open problems, and methods in the area, with a focus on one such technique called the "Rödl nibble" or the "semi-random method." We also provide a detailed outline of some ideas involved in the authors' recent proof of the Erdős–Faber–Lovász conjecture [96], with the Rödl nibble playing an important role.

1.1. Background

A *hypergraph* $\mathcal{H}$ is a pair $\mathcal{H} = (V, E)$, where V is a set of elements called *vertices* and $E \subseteq 2^V$ is a set of subsets of V called *edges*. For convenience, we often identify a hypergraph $\mathcal{H}$ with its edge set and use $V(\mathcal{H})$ to denote its vertex set. A *proper edge-colouring* of a hypergraph $\mathcal{H}$ is an assignment of colours to the edges of $\mathcal{H}$ such that no two edges of the same colour share a vertex, and a *proper vertex-colouring* (often simply called a proper colouring) of $\mathcal{H}$ is an assignment of colours to the vertices of $\mathcal{H}$ such that no edge contains vertices all of the same colour. The *chromatic index* of $\mathcal{H}$, denoted by $\chi'(\mathcal{H})$, is the minimum number of colours used by a proper edge-colouring of $\mathcal{H}$, and the *chromatic number*, denoted by $\chi(\mathcal{H})$, is the

2020 Mathematics Subject Classification. Primary 05C15; Secondary 05D40.
Keywords. Graph colouring, hypergraph colouring, probabilistic method, nibble, Erdős–Faber–Lovász conjecture.

minimum number of colours used by a proper vertex-colouring of $\mathcal{H}$. A hypergraph $\mathcal{H}$ is k-*uniform* if every edge $e \in \mathcal{H}$ satisfies $|e| = k$, and a *graph* is a 2-uniform hypergraph. A *matching* $M \subseteq \mathcal{H}$ in a hypergraph $\mathcal{H}$ is a set of disjoint edges, and an independent set $I \subseteq V(\mathcal{H})$ in $\mathcal{H}$ is a set of vertices that contains no edge of $\mathcal{H}$ (as a subset). The maximum size of an independent set in $\mathcal{H}$, denoted by $\alpha(\mathcal{H})$, is called the *independence number* of $\mathcal{H}$.

Bounding the chromatic index of a graph or hypergraph is closely related to the problem of finding large matchings (note that the colour classes of a proper edge-colouring form a matching). Matching theory is a classical subject in the study of graphs and is well developed, dating back to the work of König [111], Egerváry [47], and Hall [75] in the 1930s. Tutte's theorem [140] provides a characterization of graphs that contain a perfect matching, and Edmonds' [46] "blossom algorithm" finds a maximum matching in a graph in polynomial time. In contrast, there is no polynomial-time algorithm known to compute the independence number or chromatic number of a graph, or the size of a largest matching in a k-uniform hypergraph for $k \geq 3$. Indeed, these problems were all among Karp's [100] original twenty-one NP-complete problems. It is also NP-complete to compute the chromatic index $\chi'(G)$ of a graph G [76]. However, every graph G trivially satisfies $\chi'(G) \geq \Delta(G)$, where $\Delta(G) := \max_{v \in V(G)} d_G(v)$ and $d_G(v) := |\{e \in E(G) : e \ni v\}|$, and Vizing's theorem [142] implies that $\chi'(G) \leq \Delta(G) + 1$ ($\Delta(G)$ is called the *maximum degree of* G, $d_G(v)$ is the *degree of* v *in* G, and these definitions, as well as the lower bound, extend naturally to hypergraphs). More generally, it is natural to seek similar bounds for hypergraphs.

Consequently, there is a rich literature and active research on proving bounds on the chromatic index and chromatic number of hypergraphs. As we will describe in this survey, the "Rödl nibble" method has played a major role in the growth of this field. Though this survey is mainly concerned with hypergraphs (rather than graphs), several results and colouring problems for hypergraphs arise naturally from the graph case, so we also provide the relevant context on the latter. Similarly, we provide background on the study of matchings and independent sets in graphs and hypergraphs. Some of the earlier developments in the area are described in the surveys of Füredi [63] and Kahn [86,90]. Some aspects are also covered in the book of Molloy and Reed [121] on graph colouring with the probabilistic method. For some recent surveys on perfect matchings in hypergraphs, see [104, 113, 126, 146].

1.2. The Rödl nibble

In its basic form, the Rödl nibble is a probabilistic approach for constructing a combinatorial substructure, such as a matching or independent set, within some host structure (such as a hypergraph) which exhibits some weak quasirandom properties.

The substructure is built bit by bit by iterating a step called a “nibble,” in which elements of the host structure are selected randomly. This approach enabled Rödl [125] to prove the conjecture of Erdős and Hanani [50] on the existence of approximate combinatorial designs (see Theorem 2.1) in 1985.

Preceding Rödl’s [125] work, Ajtai, Komlós, and Szemerédi [5] showed in 1981 that a similar approach produces large independent sets in graphs with bounded average degree and no *triangles* (i.e., three pairwise adjacent vertices). Ultimately, the work of Rödl [125] and of Ajtai, Komlós, and Szemerédi [5] led to numerous important developments in the theory of hypergraph colouring. Frankl and Rödl [58] and Pippenger (unpublished) showed that Rödl’s “nibble” method produces nearly perfect matchings in hypergraphs under much more general conditions than those considered in [125]. In particular, every regular uniform hypergraph with comparatively small codegree has a nearly perfect matching; Pippenger and Spencer [122], coining the term “nibble,” generalized this further in 1989 by showing that D-regular hypergraphs have chromatic index tending to D as $D \to \infty$, as long as the codegree is $o(D)$. (Here a hypergraph $\mathcal{H}$ is *D-regular* if all of its vertices have degree D and *regular* if it is D-regular for some D, and the *codegree* of $\mathcal{H}$ is the maximum of the codegrees of all the pairs of distinct vertices of $\mathcal{H}$, where the codegree of distinct vertices $u, v \in V(\mathcal{H})$ is $|\{e \in \mathcal{H} : e \supseteq \{u, v\}\}|$.) The Pippenger–Spencer theorem was further generalized to list edge-colourings by Kahn [87] in 1996. Meanwhile, the Ajtai–Komlós–Szemerédi theorem [5] was generalized in 1982 by Ajtai, Komlós, Pintz, Spencer, and Szemerédi [3], who showed that the bound also holds for uniform hypergraphs, and also by Johansson [82] in 1996, who proved a bound on the chromatic number of triangle-free graphs of bounded maximum degree. In 2013, Frieze and Mubayi [60] showed that both of these results have a common generalization in the setting of vertex-colouring hypergraphs.

These two threads of research, of edge-colouring and of vertex-colouring with the “nibble” method, have developed somewhat in parallel, sometimes intertwining. They also both converge in the authors’ [96] recent resolution of the Erdős–Faber–Lovász conjecture. Indeed, in [96], we apply generalizations of the Pippenger–Spencer theorem [122] as well as results inspired by Johansson’s theorem [82] on vertex-colouring “locally sparse” graphs.

1.3. Organization of the paper

This paper is organized as follows. In Section 2, we survey results on hypergraph matchings and edge-colouring hypergraphs, and in Section 3, we survey results on independent sets and vertex-colourings of graphs and hypergraphs. In Section 4, we present the history of the Erdős–Faber–Lovász conjecture, and in Section 5 we describe ideas involved in its recent proof [96].

1.4. Basic definitions and notation

We say a vertex $v \in V(\mathcal{H})$ is *covered* by a matching M if there is an edge $e \in M$ such that $e \ni v$, and we say a set $X \subseteq V(\mathcal{H})$ is *covered* by M if every vertex in X is covered by M. A matching M in $\mathcal{H}$ is *perfect* if it covers $V(\mathcal{H})$. The maximum size of a matching in $\mathcal{H}$, denoted by $\nu(\mathcal{H})$, is called the *matching number* of $\mathcal{H}$. Note that in a proper edge-colouring, each colour is assigned to the edges of a matching, and in a proper vertex-colouring, each colour is assigned to the vertices of an independent set.

We usually denote a graph by G, with vertex set $V(G)$ and edge set $E(G)$. The *line graph* of a hypergraph $\mathcal{H}$, denoted by $L(\mathcal{H})$, is the graph $G := L(\mathcal{H})$ where $V(G)$ is the edge set of $\mathcal{H}$, and $e, f \in V(G)$ are adjacent in G if $e \cap f \neq \varnothing$. For an edge $e \in \mathcal{H}$, we write $N_{\mathcal{H}}(e)$ for short instead of $N_{L(\mathcal{H})}(e)$ to denote the neighbourhood of e in the line graph of $\mathcal{H}$. Note that the matchings in $\mathcal{H}$ are in one-to-one correspondence with the independent sets of $L(\mathcal{H})$ and that $\chi'(\mathcal{H}) = \chi(L(\mathcal{H}))$.

The *fractional chromatic number* of a hypergraph $\mathcal{H}$, denoted by $\chi_f(\mathcal{H})$, is the smallest $k \in \mathbb{R}$ for which there exists a probability distribution on the independent sets of $\mathcal{H}$ satisfying $\mathbb{P}[v \in I] \geq 1/k$ for every $v \in V(\mathcal{H})$ if I is drawn according to the distribution, and the *fractional chromatic index* of $\mathcal{H}$ is defined as $\chi'_f(\mathcal{H}) = \chi_f(L(\mathcal{H}))$. The *list chromatic number* of a hypergraph $\mathcal{H}$, denoted by $\chi_\ell(\mathcal{H})$, is the minimum $k \in \mathbb{N}$ such that the following holds: if C is an assignment of "lists of colours" $C(v) \subseteq \mathbb{N}$ for each $v \in V(\mathcal{H})$ satisfying $|C(v)| \geq k$ for all $v \in V(\mathcal{H})$, then $\mathcal{H}$ has a proper vertex-colouring ϕ such that $\phi(v) \in C(v)$ for every $v \in V(\mathcal{H})$. The *list chromatic index* of $\mathcal{H}$ is defined as $\chi'_\ell(\mathcal{H}) = \chi_\ell(L(\mathcal{H}))$. It is well known that every hypergraph $\mathcal{H}$ satisfies $|V(\mathcal{H})|/\alpha(\mathcal{H}) \leq \chi_f(\mathcal{H}) \leq \chi(\mathcal{H}) \leq \chi_\ell(\mathcal{H})$ and $|\mathcal{H}|/\nu(\mathcal{H}) \leq \chi'_f(\mathcal{H}) \leq \chi'(\mathcal{H}) \leq \chi'_\ell(\mathcal{H})$.

Some authors define a hypergraph to be a pair $\mathcal{H} = (V, E)$ where E is a multi-set of subsets of V; in this survey, we refer to such an object as a *multi-hypergraph*, and if every $e \in \mathcal{H}$ has size two, then $\mathcal{H}$ is a *multigraph*.

For $n \in \mathbb{N}$, we write $[n] := \{k \in \mathbb{N} : 1 \leq k \leq n\}$. We write $c = a \pm b$ if $a - b \leq c \leq a + b$. In Sections 4 and 5, we use the "$\ll$" notation in proofs. Whenever we write a hierarchy of constants, they have to be chosen from right to left. More precisely, if we claim that a result holds whenever $0 < a \ll b \leq 1$, then this means that there exists a non-decreasing function $f : (0, 1] \mapsto (0, 1]$ such that the result holds for all $0 < a, b \leq 1$ with $a \leq f(b)$. We will not calculate these functions explicitly. Hierarchies with more constants are defined in a similar way. We use "log" to denote the natural logarithm, which is relevant in Section 3.

Our graph theory notation is standard, but one may refer to [96, Section 3] for a comprehensive list of the notation we use.

2. Matchings and edge-colouring

2.1. Early results

A *partial Steiner system* with parameters (t,k,n) is a k-uniform n-vertex hypergraph such that every set of t vertices is contained in at most one edge, and a *(full) Steiner system* with parameters (t,k,n) is a k-uniform n-vertex hypergraph such that every set of t vertices is contained in precisely one edge. Note that a Steiner system with parameters (t,k,n) has $\binom{n}{t}/\binom{k}{t}$ edges, which implies that $\binom{k}{t}\big|\binom{n}{t}$. The so-called existence conjecture for designs asserts that, apart from finitely many exceptions, this and a few other trivial divisibility conditions are sufficient to ensure the existence of a Steiner system with parameters (t,k,n). In 1963, Erdős and Hanani [50] asked for an approximate version of this conjecture, which was confirmed by Rödl [125] in 1985, initiating the use of the celebrated "nibble" method, as follows.

Theorem 2.1 (Rödl [125]). *For every $k > t \geq 1$ and $\varepsilon > 0$, there exists n_0 such that the following holds. For every $n \geq n_0$, there exists a partial Steiner system with parameters (t,k,n) and at least $(1-\varepsilon)\binom{n}{t}/\binom{k}{t}$ edges.*

The existence conjecture was proved by Keevash [103] in 2014, by combining a generalization of Theorem 2.1 (briefly discussed in Section 2.3), with an "absorption" technique called "randomized algebraic construction." Glock, Kühn, Lo, and Osthus [66] provided a purely combinatorial proof, using "iterative absorption" instead of the algebraic approach of Keevash.

A partial Steiner system $\mathcal{H}$ with parameters (t,k,n) corresponds to a matching $M := \{\binom{e}{t} : e \in \mathcal{H}\}$ in the $\binom{k}{t}$-uniform auxiliary hypergraph with vertex set $\binom{V(\mathcal{H})}{t}$ and edge set $\{\binom{X}{t} : X \in \binom{V(\mathcal{H})}{k}\}$. In particular, a Steiner system with parameters (t,k,n) exists if and only if the hypergraph $\mathcal{H}^*_{t,k,n}$ with vertex set $\binom{[n]}{t}$ and edge set $\{\binom{X}{t} : X \in \binom{[n]}{k}\}$ has a perfect matching, and Theorem 2.1 is equivalent to the statement that $\mathcal{H}^*_{t,k,n}$ contains a matching covering all but a vanishing proportion of its vertices as $n \to \infty$. This result holds much more generally for hypergraphs satisfying mild pseudorandomness conditions involving the degrees and codegrees. Indeed, Frankl and Rödl [58] proved that if $\mathcal{H}$ is an N-vertex, D-regular hypergraph with codegree at most $D/(\log N)^4$, then $\mathcal{H}$ has a matching covering all but $o(N)$ vertices as $N \to \infty$. Since $\mathcal{H}^*_{t,k,n}$ is $\binom{n-t}{k-t}$-regular and has codegree at most $\binom{n-t-1}{k-t-1}$, this result generalizes Theorem 2.1. Pippenger generalized this result further, by relaxing the codegree condition, as follows. (Pippenger's result was not published, but a proof is given, e.g., in [63, Theorem 8.4].)

Theorem 2.2 (Pippenger). *For every $k, \varepsilon > 0$, there exists $\delta > 0$ such that the following holds. If $\mathcal{H}$ is an n-vertex k-uniform D-regular hypergraph with codegree at most δD, then there is a matching in $\mathcal{H}$ covering all but at most εn vertices.*

As mentioned in Section 1.2, Theorem 2.2 is proved with the nibble method, which we now sketch in more detail. Each step of the nibble process produces a matching in a nearly D_i-regular subhypergraph $\mathcal{H}_i \subseteq \mathcal{H}$ (beginning with $\mathcal{H}_1 := \mathcal{H}$ and $D_1 := D$), as follows. First, select a set of edges $X_i \subseteq \mathcal{H}_i$ randomly, where each edge $e \in \mathcal{H}_i$ is included in X_i independently with probability ε'/D_i where $\varepsilon' > 0$ is a small constant depending on k and ε. Then, let $N_i \subseteq X_i$ be the matching consisting of the edges of X_i that are disjoint from the rest. Crucially, each vertex is in an edge of X_i with probability close to $1 - e^{-\varepsilon'}$, and moreover, the codegree condition ensures that two distinct vertices are in an edge of X_i somewhat independently. This fact implies that the hypergraph $\mathcal{H}_{i+1}$ obtained by removing every vertex in an edge of X_i is nearly D_{i+1}-regular, where $D_{i+1} := e^{-\varepsilon'(k-1)}D_i$, which in turn allows the nibble process to continue. Each edge $e \in \mathcal{H}$ is in X_i with probability roughly $(\varepsilon'/D_i)e^{-\varepsilon' k(i-1)} = (\varepsilon'/D)e^{-\varepsilon'(i-1)}$ (indeed, e is in $\mathcal{H}_i$ with probability roughly $e^{-\varepsilon' k(i-1)}$, and conditioning on this, is selected in X_i with probability ε'/D_i). Each edge in X_i is then kept in N_i with probability roughly $e^{-\varepsilon' k}$. Thus, after t steps of the nibble, each edge $e \in \mathcal{H}$ is contained in $M := \bigcup_{i=1}^{t} N_i$ with probability close to $\sum_{i=1}^{t}(\varepsilon'/D)e^{-\varepsilon'(i-1)}e^{-\varepsilon' k} = \alpha/D$, where $\alpha := \varepsilon' e^{-\varepsilon' k}(1 - e^{-\varepsilon' t})/(1 - e^{-\varepsilon'})$. In particular, M is a matching and the expected number of uncovered vertices is essentially at most $(1 - \alpha)n \leq \varepsilon n$ (if ε' and t^{-1} are small enough). Kahn [89] and Kahn and Kayll [93] proved generalizations of Theorem 2.2 where the regularity and codegree conditions are replaced with the existence of a fractional matching satisfying a certain “local sparsity” condition, which can be used to guide the nibble process.

It is natural to wonder if the “random greedy algorithm” (which would select X_i to consist of a single edge chosen uniformly at random from $\mathcal{H}_i$ in the process above) also produces a nearly perfect matching under the conditions of Theorem 2.2. Indeed, this result was obtained independently by Spencer [138] and by Rödl and Thoma [127]. To prove this, Spencer [138] considered a branching process, and Rödl and Thoma [127] showed that the random greedy algorithm produces a matching with a similar distribution as the nibble process. Note that these results immediately yield a (randomized) polynomial-time Monte Carlo algorithm for finding the matching guaranteed by Theorem 2.2. The proof of Theorem 2.2 given in [122] also yields such an algorithm. Rödl and Thoma also asked if there is an NC-algorithm (and in particular, a deterministic, polynomial-time algorithm) for finding such a matching, and Grable [70] answered their question in the affirmative.

In 1989, Pippenger and Spencer [122] generalized Theorem 2.2 to edge-colouring, as follows.

Theorem 2.3 (Pippenger and Spencer [122]). *For every $k, \varepsilon > 0$, there exists $\delta > 0$ such that the following holds. If $\mathcal{H}$ is a k-uniform D-regular hypergraph of codegree at most δD, then $\chi'(\mathcal{H}) \leq (1 + \varepsilon)D$.*

Since every hypergraph $\mathcal{H}$ satisfies $\nu(\mathcal{H}) \geq |\mathcal{H}|/\chi'(\mathcal{H})$ and moreover $|\mathcal{H}| = D|V(\mathcal{H})|/k$ if $\mathcal{H}$ is D-regular and k-uniform, Theorem 2.3 implies Theorem 2.2. In fact, in Pippenger and Spencer's [122] proof of Theorem 2.3, they used the argument described above to select nearly perfect matchings randomly with the nibble process, which ultimately form most of the colour classes. Roughly, they show that after selecting D such matchings—in groups of size $o(D)$, selected iteratively in a "semi-random" way (which could also be considered a nibble process)—the remaining hypergraph has small maximum degree and can thus be properly edge-coloured with at most εD colours in a "greedy" fashion.

Pippenger and Spencer [122] actually proved the slightly stronger version of Theorem 2.3 that applies if every vertex of $\mathcal{H}$ has degree $(1 \pm \delta)D$, rather than precisely D. Kahn [84] observed that Theorem 2.3 holds more generally for k-bounded hypergraphs of maximum degree at most D, by showing that such hypergraphs can be "embedded" in a nearly D-regular k-uniform hypergraph with the same or larger chromatic index (a hypergraph $\mathcal{H}$ is *k-bounded* if every $e \in \mathcal{H}$ satisfies $|e| \leq k$). This sequence of results culminated in Kahn's [87] generalization of the Pippenger–Spencer theorem to list colouring, as follows.

Theorem 2.4 (Kahn [87]). *For every $k, \varepsilon > 0$, there exists $\delta > 0$ such that the following holds. If $\mathcal{H}$ is a k-bounded hypergraph of maximum degree at most D and codegree at most δD, then $\chi'_\ell(\mathcal{H}) \leq (1+\varepsilon)D$.*

The so-called "List Edge Colouring conjecture"—first posed by Vizing in 1975 and asked by many others since (see, e.g., [81])—asserts that every graph G satisfies $\chi'_\ell(G) = \chi'(G)$, and Theorem 2.4 for $k = 2$ confirms this conjecture asymptotically. Kahn's proof of Theorem 2.4 is also based on a nibble argument but is notably different from Pippenger and Spencer's [122] proof of Theorem 2.3. In particular, rather than selecting colour classes one by one, in each step of the nibble, edges are assigned a colour randomly from their lists, so the colour classes are constructed in parallel.

For *linear* hypergraphs (i.e., hypergraphs of codegree one), Molloy and Postle [119, Theorem 10] recently generalized Theorem 2.4 to the setting of "correspondence colouring" (also known as DP-colouring), and Bonamy, Delcourt, Lang, and Postle [21, Theorem 7] generalized this result further by proving a "local version."

Several results also strengthen Theorems 2.2–2.4 by improving the asymptotic error terms. This is the focus of Section 2.2. We conclude this subsection by discussing two open problems from the late 1990s. Both of these are conjectured to hold for multi-hypergraphs. In fact, Theorems 2.2–2.4 also hold for multi-hypergraphs (but the codegree conditions also bound the number of copies of each edge).

Conjecture 2.5 (Kahn [87]). *For every $k, \varepsilon > 0$, there exists K such that the following holds. If $\mathcal{H}$ is a k-bounded multi-hypergraph, then $\chi'_\ell(\mathcal{H}) \leq \max\{(1+\varepsilon)\chi'_f(\mathcal{H}), K\}$.*

Even the weaker version of this conjecture, with the list chromatic index replaced by the chromatic index, is wide open. Only the case $k = 2$ is known. For 2-bounded hypergraphs (i.e., graphs with edge-multiplicity 1), the conjecture follows from Vizing's theorem [142] for the chromatic index and from Theorem 2.4 for the list chromatic index. As shown by Seymour [131] using Edmonds' Matching Polytope theorem [45], every multigraph G satisfies $\chi'_f(G) = \max\{\Delta(G), \Gamma(G)\}$, where

$$\Gamma(G) := \max\left\{\frac{2|E(H)|}{|V(H)| - 1} : H \subseteq G,\ |V(H)| \geq 3 \text{ and is odd}\right\}.$$

Kahn [88] proved that every multigraph G satisfies $\chi'(G) \leq (1 + o(1))\chi_f(G)$ and in [91] extended this result to list colouring, thus confirming Conjecture 2.5 in full for the case $k = 2$. For the ordinary chromatic index, even more is now known in this case. In the 1970s, Goldberg [69] and Seymour [131] independently conjectured that every multigraph G satisfies $\chi'(G) \leq \max\{\Delta(G) + 1, \lceil\Gamma(G)\rceil\}$. Thus, Kahn's result [88] confirmed the Goldberg–Seymour conjecture asymptotically. Recently, a full proof (which does not rely on probabilistic arguments) of the Goldberg–Seymour conjecture was obtained by Chen, Jing, and Zang [31].

The next conjecture was posed by Alon and Kim [9]. A hypergraph $\mathcal{H}$ is called *t-simple* if every two distinct edges of $\mathcal{H}$ have at most t vertices in common; in particular, a hypergraph is 1-simple if and only if it is linear.

Conjecture 2.6 (Alon and Kim [9]). *For every $k \geq t \geq 1$ and $\varepsilon > 0$, there exists D_0 such that the following holds. For every $D \geq D_0$, if $\mathcal{H}$ is a k-uniform, t-simple multi-hypergraph with maximum degree at most D, then*

$$\chi'(\mathcal{H}) \leq (t - 1 + 1/t + \varepsilon)D.$$

The conjecture is true for $k = 2$ by Vizing's theorem [142] for $t = 1$ and by a result of Shannon [133] for $t = 2$. For $k > t = 1$, the conjecture follows from Theorem 2.3 (together with the observation of Kahn in [87] mentioned above). Kahn (see [9]) conjectured that the t-simple condition in Conjecture 2.6 can be relaxed to requiring that the "$(t + 1)$-codegrees" are small (i.e., every set of $t + 1$ vertices is contained in at most δD edges, for some $\delta > 0$), which if true, would generalize Theorem 2.3. The remaining cases are still open. The case $k = t$ (without the "$+\varepsilon$" in the bound) was proved by Füredi, Kahn, and Seymour [64] for the fractional chromatic index.

Alon and Kim [9] showed that Conjecture 2.6 holds for *intersecting* hypergraphs (i.e., hypergraphs with matching number one), and they gave a construction to show that if Conjecture 2.6 is true, then it would be asymptotically tight for every $k \geq t$ for which there exists a projective plane of order $t - 1$. We sketch their construction here. Let D be a large integer divisible by t, let $m := t^2 - t + 1$, and fix a projective plane P of order $t-1$ with m lines $\ell_1, \ell_2, \dots, \ell_m$ on a set of m points. For each of the lines ℓ_i, let $\mathcal{F}_i$ be a collection of D/t sets of size k containing ℓ_i, so that all the mD/t sets $\{A \setminus \ell_i : 1 \le i \le m$ and $A \in \mathcal{F}_i\}$ are pairwise disjoint and disjoint from P. Let $\mathcal{H}$ be the k-uniform hypergraph whose edge set is $\bigcup_i \mathcal{F}_i$. Then clearly $\mathcal{H}$ is intersecting, k-uniform, and t-simple; its maximum degree is at most D; and it has $mD/t = (t-1+1/t)D$ edges. Thus, $\chi'(\mathcal{H}) \ge (t-1+1/t)D$.

2.2. Asymptotic improvements

Let $\mathcal{H}$ be a k-uniform, D-regular hypergraph on n vertices. Recall that Pippenger's theorem (Theorem 2.2) shows that if the codegree of $\mathcal{H}$ is $o(D)$, then there is a matching in $\mathcal{H}$ covering all but at most $o(n)$ vertices. However, his proof does not supply an explicit estimate for the error term $o(n)$. Also recall that Theorems 2.3 and 2.4 imply that if the codegree of $\mathcal{H}$ is $o(D)$, then the chromatic index of $\mathcal{H}$ is $D + o(D)$. Sharpening these error terms is useful for many applications, and considerable progress has been made towards this end with improved analysis and variations of the nibble method, with more powerful concentration inequalities. In this subsection, we will discuss many such results.

Grable [71] proved that if the codegree is at most $D^{1-\delta}$ in Theorem 2.2, then there is a matching covering all but at most $n(D/\log n)^{-\delta/(2k-1+o(1))}$ vertices. In 1997, Alon, Kim, and Spencer [10] improved this bound for linear hypergraphs by showing the following.

Theorem 2.7 (Alon, Kim, and Spencer [10]). *Let $k \ge 3$. Let $\mathcal{H}$ be a k-uniform D-regular n-vertex linear hypergraph. Then $\mathcal{H}$ has a matching containing all but at most $O(nD^{-\frac{1}{k-1}} \log^{c_k} D)$ vertices, where $c_k = 0$ for $k > 3$ and $c_3 = 3/2$.*

Based on computer simulations (see, e.g., [13]), Alon, Kim, and Spencer conjectured that the simple random greedy algorithm outlined in the previous subsection should also produce a matching containing all but at most $O(nD^{-\frac{1}{k-1}} \log^{O(1)} D)$ vertices. The results of Spencer [138] and of Rödl and Thoma [127] mentioned after Theorem 2.2 only show that the random greedy algorithm produces a matching covering all but $o(n)$ vertices. This error term was sharpened by Wormald [145] and Bennett and Bohman [15] but the conjecture is still open.

Kostochka and Rödl [112] extended Theorem 2.7 to hypergraphs with small codegrees C (i.e., satisfying $C \le D^{1-\gamma}$ for some $\gamma > 0$). In 2000, Vu [143] further extended the result of Kostochka and Rödl [112] by removing the assumption $C < D^{1-\gamma}$ on the codegree. More precisely, he showed that every k-uniform D-regular n-vertex hypergraph with codegree at most C contains a matching covering all but at most $O(n(D/C)^{-\frac{1}{k-1}} \log^{c} D)$ vertices for some constant $c > 0$. He also obtained stronger bounds if one makes additional assumptions on the "t-codegrees" for $t > 2$.

Very recently, Kang, Kühn, Methuku, and Osthus [99] improved Theorem 2.7 and the results of Kostochka and Rödl [112] and of Vu [143] for hypergraphs with small codegree. In the case when $\mathcal{H}$ is linear, they showed the following.

Theorem 2.8 (Kang, Kühn, Methuku, and Osthus [99]). *For every $k > 3$, $\mu \in (0, 1)$, and $\eta < \frac{k-3}{(k-1)(k^3-2k^2-k+4)}$, there exists n_0 such that the following holds for all $n \geq n_0$ and $D \geq \exp(\log^{\mu} n)$.*

If $\mathcal{H}$ is a k-uniform D-regular linear hypergraph on n vertices, then $\mathcal{H}$ contains a matching covering all but at most $nD^{-\frac{1}{k-1}-\eta}$ vertices.

Their approach consists of showing that the Rödl nibble process not only constructs a large matching but it also produces many well-distributed "augmenting stars" which can then be used to significantly augment the matching constructed by the Rödl nibble process.

Below we discuss results concerning improvements on the chromatic index of hypergraphs. In 2000, Molloy and Reed [120] sharpened the error term in Theorem 2.4. For linear hypergraphs their result can be stated as follows.

Theorem 2.9 (Molloy and Reed [120]). *If $\mathcal{H}$ is a k-uniform linear hypergraph with maximum degree at most D, then $\chi'_{\ell}(\mathcal{H}) \leq D + O(D^{1-1/k} \log^4 D)$.*

For graphs, this result improves a result of Häggkvist and Janssen [74] and provides the best-known general bound for the List Edge Colouring conjecture. Molloy and Reed [120] actually proved a more general result showing that every k-uniform hypergraph $\mathcal{H}$ with maximum degree at most D and codegree at most C has list chromatic index at most $D + O(D(D/C)^{-1/k}(\log D/C)^4)$, which also gave the best-known bound on the ordinary chromatic index $\chi'(\mathcal{H})$. Very recently, Kang, Kühn, Methuku, and Osthus [99] showed that this bound on the chromatic index can be improved further. For linear hypergraphs their result can be stated as follows.

Theorem 2.10 (Kang, Kühn, Methuku, and Osthus [99]). *For every $k \geq 3$, $\mu \in (0, 1)$, and $\eta < \frac{k-2}{k(k^3+k^2-2k+2)}$, there exists n_0 such that the following holds for all $n \geq n_0$ and $D \geq \exp(\log^{\mu} n)$.*

If $\mathcal{H}$ is a k-uniform linear hypergraph on n vertices with maximum degree at most D, then $\chi'(\mathcal{H}) \leq D + D^{1-1/k-\eta}$.

Theorems 2.8–2.10 are unlikely to be best possible. The best lower bounds we know come from the following construction, in which every matching leaves $\Omega(n/D)$ vertices uncovered. Consider an m-vertex k-uniform D-regular linear hypergraph $\mathcal{H}$ such that $m = O(D)$ and $m - 1$ is divisible by k (e.g., using a Steiner system $S(2, k, m)$), so the union of n/m disjoint copies of $\mathcal{H}$ yields an n-vertex k-uniform D-regular hypergraph with at least $\Omega(n/D)$ vertices uncovered by any matching.

If we know more about the hypergraph $\mathcal{H}$, then the bound given in Theorem 2.8 can be improved further. For example, if $\mathcal{H}$ is a Steiner triple system (i.e., a Steiner system with parameters $(2, 3, n)$), then $\mathcal{H}$ is $(n-1)/2$-regular and linear, and Brouwer [25] conjectured the following in 1981.

Conjecture 2.11 (Brouwer [25]). *Every Steiner triple system with n vertices has a matching of size at least $\frac{n-4}{3}$.*

Recently, combining the nibble method with the robust expansion properties of edge-coloured pseudorandom graphs, Keevash, Pokrovskiy, Sudakov, and Yepremyan [105] showed that every Steiner triple system has a matching covering all but at most $O(\log n/\log\log n)$ vertices.

A related problem is a famous conjecture of Ryser, Brualdi, and Stein [26, 129, 139] which states that every $n \times n$ Latin square has a transversal of order $n-1$ and moreover, if n is odd, then it has a full transversal. The best-known bound for this problem was given in [105] where the authors showed that every $n \times n$ Latin square contains a transversal of order $n - O(\log n/\log\log n)$. The problem of finding large transversals in Latin squares can be rephrased as a problem about finding large matchings in hypergraphs. Indeed, we can construct a 3-uniform hypergraph $\mathcal{H}_{\mathcal{L}}$ on $3n$ vertices from an $n \times n$ Latin square $\mathcal{L}$ as follows. The vertex set of $\mathcal{H}_{\mathcal{L}}$ is $V(\mathcal{H}_{\mathcal{L}}) = R \cup C \cup S$ where R, C, and S are the rows, columns, and symbols of $\mathcal{L}$. For every entry of $\mathcal{L}$ we add an edge to $\mathcal{H}_{\mathcal{L}}$—if the (i, j)-th entry of $\mathcal{L}$ contains a symbol s, then we add the edge $\{i, j, s\}$ to $\mathcal{H}_{\mathcal{L}}$. Clearly, $\mathcal{H}_{\mathcal{L}}$ is n-regular and 3-partite, and a matching in $\mathcal{H}$ corresponds to a transversal in $\mathcal{L}$. Thus, the Ryser–Brualdi–Stein conjecture can be regarded as the "partite-version" of Brouwer's conjecture.

Similarly, it is interesting to determine the maximum chromatic index of an n-vertex Steiner triple system (or an $n \times n$ Latin square). Meszka, Nedela, and Rosa [116] conjectured the following in 2006.

Conjecture 2.12 (Meszka, Nedela, and Rosa [116]). *If $\mathcal{H}$ is a Steiner triple system with $n > 7$ vertices, then $\chi'(\mathcal{H}) \leq (n-1)/2 + 3$ and moreover, if $n \equiv 3 \pmod 6$, then $\chi'(\mathcal{H}) \leq (n-1)/2 + 2$.*

Since an n-vertex Steiner triple system is $(n-1)/2$-regular, it is obvious that $\chi'(\mathcal{H}) \geq (n-1)/2$, and equality holds if and only if $\mathcal{H}$ can be decomposed into perfect matchings. Hence, if $n \equiv 1 \pmod 6$, then $\chi'(\mathcal{H}) \geq (n+1)/2$. In fact, there are constructions of Steiner triple systems with n vertices which show that Conjecture 2.12, if true, is tight [28, 115, 123, 141]. Similarly, for Latin squares the following conjecture was posed independently by Cavenagh and Kuhl [29] in 2015 and Besharati, Goddyn, Mahmoodian, and Mortezaeefar [18] in 2016.

Conjecture 2.13. *Let $\mathcal{L}$ be an $n \times n$ Latin square. If $\mathcal{H}_{\mathcal{L}}$ is the corresponding* 3*-uniform* 3*-partite hypergraph, then $\chi'(\mathcal{H}_{\mathcal{L}}) \le n + 2$ and moreover, if n is odd, then $\chi'(\mathcal{H}_{\mathcal{L}}) \le n + 1$.*

Note that Conjecture 2.12 implies Conjecture 2.11, and Conjecture 2.13 implies the Ryser–Brualdi–Stein conjecture. Theorem 2.8 implies that every n-vertex Steiner triple system has chromatic index at most $n/2 + O(n^{2/3-1/100})$ and every hypergraph corresponding to an $n \times n$ Latin square has chromatic index at most $n + O(n^{2/3-1/100})$; currently these bounds are the best known.

2.3. Pseudorandom hypergraph matchings

Let $\mathcal{H}$ be a k-uniform D-regular hypergraph on n vertices, and let $M \subseteq \mathcal{H}$ be a random matching generated by the nibble process, such that M covers all but at most εn vertices of $\mathcal{H}$ (with high probability), where $\varepsilon \in (0, 1)$. A heuristic argument suggests that each vertex of $\mathcal{H}$ is left uncovered by M roughly independently with probability ε. In many applications (including our proof in [96]), it is useful to find a nearly perfect matching guaranteed by Theorem 2.2 with additional "pseudorandom" properties that are compatible with this heuristic. In this subsection, we discuss some results that provide nearly perfect pseudorandom hypergraph matchings and some of their applications. In particular, we show how a "pseudorandom version" of Pippenger's theorem (Theorem 2.2) is in fact equivalent to the Pippenger–Spencer theorem (Theorem 2.3).

The first pseudorandom hypergraph matching result of this sort was proved by Alon and Yuster [14] in 2005. With a slightly stronger assumption regarding codegrees, Ehard, Glock, and Joos [48] recently proved a stronger and more flexible version. The following is an immediate corollary of [48, Theorem 1.2].

Theorem 2.14 (Ehard, Glock, and Joos [48]). *For every $k \ge 2$ and $\delta \in (0, 1)$, there exists D_0 such that the following holds for all $D \ge D_0$ and $\varepsilon := \delta/(50k^2)$. Suppose that $\mathcal{H}$ is a k-uniform hypergraph and $\mathcal{F}$ is a collection of subsets of $V(\mathcal{H})$ such that $|\mathcal{F}| \le \exp(D^{\varepsilon^2})$ and $\sum_{v \in S} d_{\mathcal{H}}(v) \ge kD^{1+\delta}$ for every $S \in \mathcal{F}$. If $\mathcal{H}$ has maximum degree at most D, codegree at most $D^{1-\delta}$, and $e(\mathcal{H}) \le \exp(D^{\varepsilon^2})$, then there exists a matching M of $\mathcal{H}$ such that every $S \in \mathcal{F}$ satisfies $|S \cap V(M)| = (1 \pm D^{-\varepsilon}) \sum_{v \in S} d_{\mathcal{H}}(v)/D$.*

Note that if $\mathcal{H}$ is D-regular in Theorem 2.14 and $V(\mathcal{H}) \in \mathcal{F}$, then M covers all but at most $nD^{-\varepsilon}$ vertices of $\mathcal{H}$. Moreover, for every $S \in \mathcal{F}$, at most $|S|D^{-\varepsilon}$ vertices are uncovered by M, as we would expect if each vertex was uncovered with probability $D^{-\varepsilon}$. Ehard, Glock, and Joos [48] actually proved a stronger version of Theorem 2.14 involving weight functions on the edges of $\mathcal{H}$ of the form

$\omega : \mathcal{H} \to \mathbb{R}_{\geq 0}$. The "pseudorandomness" heuristic suggests that every edge is in M with probability $1/D$, and thus the expected total weight of edges in M should be $\sum_{e \in \mathcal{H}} \omega(e)/D$.

Hypergraph matching results, particularly ones with pseudorandomness guarantees, are widely applicable in combinatorics and beyond. We give some examples here. Ford, Green, Konyagin, Maynard, and Tao [57] proved a pseudorandom generalization of Theorem 2.2 (stated for *coverings* rather than matchings and which allows for non-uniform hypergraphs). They used it to improve bounds on gaps between prime numbers. As we saw in Section 2.2, in [99, 105], pseudorandom properties of hypergraph matchings can be "bootstrapped" to produce a larger matching. Furthermore, in some applications of the "absorption method," such as [34, 55, 66, 67, 103], a matching in an auxiliary hypergraph is used to construct a nearly spanning structure which is complemented by an "absorbing structure," so that the pseudorandom properties can be exploited for "absorption," which results in a spanning structure. Hypergraph matchings with pseudorandomness properties can also be used to construct approximate decompositions (see, e.g., [65, 66, 103, 107]) or edge-colourings. Indeed, in the proof of the Erdős–Faber–Lovász conjecture, we use Theorem 2.14 to obtain a partial edge-colouring of a linear hypergraph in which each colour class has pseudorandom properties that enable some of the uncovered vertices to be absorbed (see Section 5.1 for more details). As an illustration of this approach, we show how Theorem 2.14 implies a version of the Pippenger–Spencer theorem. First we need the following definition and observation.

Definition 2.15. For every hypergraph $\mathcal{H}$ and $t \in \mathbb{N}$, we define the *t-wise incidence hypergraph* $\mathcal{H}^* := \mathrm{inc}_t(\mathcal{H})$ to be the hypergraph with

- vertex set $\mathcal{H} \cup ([t] \times V(\mathcal{H}))$ and
- edge set $\{\{e\} \cup (\{i\} \times e) : e \in \mathcal{H}, i \in [t]\}$.

That is, for every $e = \{v_1, \dots, v_k\} \in \mathcal{H}$, we include t edges in the t-wise incidence hypergraph $\mathcal{H}^* := \mathrm{inc}_t(\mathcal{H})$, where each such edge is of the form $\{e, (i, v_1), \dots, (i, v_k)\}$ for some $i \in [t]$.

Observation 2.16. Let $\mathcal{H}$ be a hypergraph, and let $\mathcal{H}^* := \mathrm{inc}_t(\mathcal{H})$ be the t-wise incidence hypergraph. The following holds.

(a) If $\mathcal{H}$ is k-uniform, then $\mathcal{H}^*$ is $(k+1)$-uniform.

(b) The codegree of $\mathcal{H}^*$ is at most the codegree of $\mathcal{H}$.

(c) For every $v \in V(\mathcal{H})$ and $i \in [t]$, $d_{\mathcal{H}^*}((i, v)) = d_{\mathcal{H}}(v)$, and for every $e \in \mathcal{H}$, $d_{\mathcal{H}^*}(e) = t$.

(d) A set $M \subseteq \mathcal{H}^*$ is a matching if and only if $M_1, \dots, M_t$, where $M_i := \{e \in \mathcal{H} : \exists f \in M,\ f \supseteq \{i\} \times e\}$ are pairwise edge-disjoint matchings in $\mathcal{H}$. In particular, the chromatic index of $\mathcal{H}$ is at most t if and only if $\mathcal{H}^*$ contains a matching covering $\mathcal{H}$.

Using Observation 2.16, we can show that, under a slightly stronger codegree condition, Theorem 2.14 implies Theorem 2.3 (i.e., that the chromatic index of a D-regular hypergraph of small codegree tends to D), as follows. Let $k \geq 2$, $\delta \in (0, 1)$, and $\varepsilon := \delta/(50(k+1)^2)$, and suppose that $\mathcal{H}$ is a k-uniform, n-vertex, D-regular hypergraph, with codegree at most $D^{1-\delta}$, such that $D \geq \log^{\varepsilon^{-2}} n$. By Observation 2.16 (a)–(c), $\mathcal{H}^* := \mathrm{inc}_D(\mathcal{H})$ is a $(k+1)$-uniform, D-regular hypergraph with codegree at most $D^{1-\delta}$, and $e(\mathcal{H}^*) = D \cdot e(\mathcal{H}) = D^2 n/k \leq \exp(D^{\varepsilon^2})$. Let $\mathcal{F} := \{[D] \times \{v\} : v \in V(\mathcal{H})\}$, and note that $\mathcal{F}$ is a collection of subsets of $V(\mathcal{H}^*)$ such that $|\mathcal{F}| \leq n \leq \exp(D^{\varepsilon^2})$ and every $S \in \mathcal{F}$ satisfies $\sum_{v \in S} d_{\mathcal{H}^*}(v) = D^2$. Thus, if D is sufficiently large, then, by Theorem 2.14, there exists a matching M in $\mathcal{H}^*$ such that $|S \cap V(M)| \geq (1 - D^{-\varepsilon})|S|$ for every $S \in \mathcal{F}$. By Observation 2.16 (d), $M_1, \dots, M_D$, where $M_i := \{e \in \mathcal{H} : \exists f \in M,\ f \supseteq \{i\} \times e\}$ are pairwise edge-disjoint matchings, and moreover, by the construction of $\mathcal{F}$, every $v \in V(\mathcal{H})$ is covered by all but $D^{1-\varepsilon}$ of these matchings. In particular, $\chi'(\mathcal{H}') \leq D$ and $\Delta(\mathcal{H} \setminus \mathcal{H}') \leq D^{1-\varepsilon}$, where $\mathcal{H}' := \bigcup_{i=1}^{D} M_i$. Hence, $\chi'(\mathcal{H} \setminus \mathcal{H}') \leq k(\Delta(\mathcal{H} \setminus \mathcal{H}') - 1) + 1 \leq kD^{1-\varepsilon}$, so $\chi'(\mathcal{H}) \leq \chi'(\mathcal{H}') + \chi'(\mathcal{H} \setminus \mathcal{H}') \leq D + kD^{1-\varepsilon} = D + o(D)$, as desired.

Theorem 2.14 is actually proved via a generalization of Theorem 2.9 (which implies Theorem 2.3). Thus, the above argument is based on "circular logic," but it demonstrates that in the setting of Theorem 2.2, the existence of nearly perfect pseudorandom hypergraph matchings is in some sense equivalent to the existence of a nearly optimal proper edge-colouring (the above comments about the proof of Theorem 2.14 also apply to the result of Alon and Yuster [14] on pseudorandom matchings, which is proved via Theorem 2.3). Moreover, Kahn's [87] proof of Theorem 2.4 (in the case when all lists are the same) more closely resembles the approach described here, wherein a nibble process is used to construct a matching in the incidence hypergraph, than it does Pippenger and Spencer's [122] proof of Theorem 2.3.

Note that one could also prove Theorem 2.14 "more directly" by a more careful analysis of the proof of Theorem 2.2—the reason being essentially that the matchings chosen in each step of the nibble intersect the sets in $\mathcal{F}$ as one would expect a random set would. This intuition is made rigorous in [99], where Theorem 2.10 is derived from [99, Theorem 7.1], a pseudorandom version of Theorem 2.8. Moreover, the approach of finding an edge-colouring via Theorem 2.14 and Observation 2.16 is very versatile and was used, e.g., in [96, 99] (see Section 5.1).

3. Independent sets and vertex-colouring

3.1. Independence number

Prior to Rödl's [125] proof of the Erdős–Hanani conjecture [50], in 1981, Ajtai, Komlós, and Szemerédi [5] employed a similar semi-random approach to show that every triangle-free graph has a large independent set.

Theorem 3.1 (Ajtai, Komlós, and Szemerédi [4, 5]). *There exists an absolute constant $c > 0$ such that the following holds. If G is an n-vertex triangle-free graph of average degree at most d, then*

$$\alpha(G) \geq c\left(\frac{n}{d}\right) \log d.$$

This result has spawned intensive research over the last four decades. Theorem 3.1 and a hypergraph analogue of it due to Komlós, Pintz, Spencer, and Szemerédi [110] have surprising applications to number theory and geometry, respectively. Improving and generalizing Theorem 3.1 is also a problem of major importance within combinatorics, in part due to connections to Ramsey theory and to the study of random graphs and algorithms.

In [5], Ajtai, Komlós, and Szemerédi used Theorem 3.1 to construct an infinite *Sidon sequence* (i.e., a sequence of positive integers in which the pairwise sums are all distinct) with "high density;" in particular, for every n, the sequence contains $\Omega((n \log n)^{1/3})$ integers less than n. Erdős conjectured that for every $\varepsilon > 0$, there exists an infinite Sidon sequence containing $\Omega(n^{1/2-\varepsilon})$ integers less than n, and this problem is still open. The best-known result is due to Rusza [128], who proved the weaker version with $1/2$ replaced with $\sqrt{2} - 1$ in the exponent, and Cilleruelo [33] provided an explicit construction of such a sequence.

A new proof of Theorem 3.1 was given in [4] by Ajtai, Komlós, and Szemerédi (written by Spencer), which uses the Cauchy–Schwarz inequality to build the independent set deterministically, rather than with a random nibble process. Theorem 3.1 is used in [4] to prove the *Ramsey number* bound $R(3, k) = O(k^2/\log k)$. (The Ramsey number $R(\ell, k)$ is the smallest n such that every red-blue edge-colouring of the n-vertex complete graph contains either a red copy of K_ℓ or a blue copy of K_k.) The matching lower bound $R(3, k) = \Omega(k^2/\log k)$ was later established by Kim [109], also using a semi-random approach. Theorem 3.1 was improved by Shearer [134, 135], who showed that the constant c can be replaced with $1 - o(1)$ in Theorem 3.1, as conjectured by Ajtai, Komlós, and Szemerédi [5]. Although Shearer's proof is more similar to the Cauchy–Schwartz approach of [4] than the random nibble approach of [5], his proof implies that the random greedy algorithm produces an independent set with expected size at least $(1 - o(1))(n/d) \log d$ in every n-vertex triangle-free graph of average degree d.

Improving the value of the leading constant in Theorem 3.1, or determining if $1 - o(1)$ is best possible, is an interesting open problem. Bollobás [20] proved that there are n-vertex d-regular triangle-free graphs with $\alpha(G) \leq 2(n/d)\log d$ (by considering random d-regular graphs), so Shearer's bound [134] is within a factor of about at most two of best possible. Shearer's result also implies that $R(3, k) \leq (1 + o(1))k^2/\log k$, which is still the best-known upper bound, and any further improvement to the value of c in Theorem 3.1 would improve this bound on the Ramsey number as well and be a major breakthrough. Fiz Pontiveros, Griffiths, and Morris [56] and independently Bohman and Keevash [19] showed that $R(3, k) \geq (1/4 - o(1))k^2/\log k$, so Shearer's bound on $R(3, k)$ is also within a factor of about at most four of best possible.

Theorem 3.1 holds more generally for k-uniform hypergraphs, where $k \geq 2$, as follows. An *ℓ-cycle* in a k-uniform hypergraph is a set of ℓ edges spanned by at most $\ell(k - 1)$ vertices, which does not contain an ℓ'-cycle for $\ell' < \ell$, and the *girth* of a k-uniform hypergraph is the length of its shortest cycle (or infinity if there is no cycle). In 1982, Komlós, Pintz, Spencer, and Szemerédi [110] proved an analogue of Theorem 3.1 for 3-uniform hypergraphs of girth at least five and used this result to disprove Heilbronn's conjecture on the Heilbronn triangle problem, which asks for the minimum area of a triangle formed by any three points out of a set of n points placed in the unit disk. Heilbronn conjectured that this area is at most $O(n^{-2})$ for any set of n points, but Komlós, Pintz, Spencer, and Szemerédi [110] used their hypergraph analogue of Theorem 3.1 to construct a set of n points in which the minimum area of a triangle with its vertices among those points is at least $\Omega(n^{-2}\log n)$. Ajtai, Komlós, Pintz, Spencer, and Szemerédi [3] later generalized the result of [110] by showing that every k-uniform hypergraph on n vertices with girth at least five and average degree at most d contains an independent set of size at least $\Omega(n(\log d/d)^{1/(k-1)})$. Duke, Lefmann, and Rödl [44] strengthened this result by showing that for $k \geq 3$, this bound holds for hypergraphs of girth at least three (that is, for linear hypergraphs), confirming a conjecture of Spencer [137] in a strong sense. Note that for $k = 2$, this bound matches the one in Theorem 3.1. Notably, the proofs in [110] and [3] use a random nibble approach like in [5]. The proof in [44] proceeds by a reduction to the case of hypergraphs of girth at least 5, whence the result follows from the result in [3].

Ajtai, Erdős, Komlós, and Szemerédi [2] suggested that Theorem 3.1 may still hold for K_r-free graphs for any fixed r (and it may even hold more generally for vertex-colouring—see Conjecture 3.3), and they proved the weaker result that K_r-free graphs on n vertices of average degree at most d have an independent set of size at least $\Omega((n/d)\log\log d)$. Later, a breakthrough of Shearer [136] in 1995 improved this bound to $\Omega((n/d)\log d/\log\log d)$, which, up to the leading constant factor, is still the best known. Alon [6] proved that Theorem 3.1 holds more generally for graphs where the neighbourhood of every vertex has bounded chromatic number. These results of Shearer [136] and of Alon [6] actually bound the average size of an independent set. In this vein, Davies, Jenssen, Perkins, and Roberts [40] recently proved that the average size of an independent set in a triangle-free graph of maximum degree at most Δ is at least $(1-o(1))(n/\Delta)\log\Delta$, which also generalizes Theorem 3.1 and even matches the earlier bound of Shearer [134] for the special case of regular graphs.

3.2. Chromatic number

Nearly all of the results bounding the independence number mentioned in the previous subsection can be generalized to bounds on the chromatic number. In 1995, Kim [108] proved that every graph of girth at least five and maximum degree at most Δ has (list) chromatic number at most $(1+o(1))\Delta/\log\Delta$. Independently, Johansson [82] proved that every triangle-free graph of maximum degree at most Δ has chromatic number at most $O(\Delta/\log\Delta)$, which generalizes Theorem 3.1. In 2019, Molloy [118] simultaneously generalized both Kim's [108] and Johansson's [82] result by improving the leading constant in Johansson's result to match that of Kim, as follows.

Theorem 3.2 (Molloy [118]). *For every $\varepsilon > 0$, there exists Δ_0 such that the following holds for every $\Delta \geq \Delta_0$. If G is a triangle-free graph of maximum degree at most Δ, then*

$$\chi_\ell(G) \leq (1+\varepsilon)\frac{\Delta}{\log\Delta}.$$

Theorem 3.2 also matches Shearer's bound [134] for regular graphs. Improving the leading constant in Theorem 3.2, or determining if it is best possible, is another major open problem. By the same argument as in the previous subsection, the bound in Theorem 3.2 is within a factor of at most two of best possible. In fact, Frieze and Łuczak [61] proved that random Δ-regular graphs have chromatic number $(1/2 \pm o(1))\Delta/\log\Delta$ with high probability, and it is an open problem whether there is a polynomial-time algorithm which almost surely finds a proper vertex-colouring of such a graph with at most $(1-\varepsilon)\Delta/\log\Delta$ colours for some $\varepsilon > 0$ (see [118]). Since random regular graphs of bounded degree have $O(1)$ cycles with high probability, the affirmative would follow if there exists such an algorithm for colouring triangle-free graphs of maximum degree at most Δ (again, see [118]). A related longstanding open problem of Karp [101] is whether there exists a polynomial-time algorithm for finding an independent set of size within a factor two of best possible in a binomial random graph.

The proofs of Kim [108] and Johansson [82] use a nibble approach inspired by Kahn's [87] proof of Theorem 2.4, in which a small random selection of vertices are assigned a colour randomly in each step of the nibble. Johansson [82] never published his proof, but Molloy and Reed [121, Chapters 12 and 13] provided simpler proofs of the results of both Kim [108] and Johansson [82]. Molloy's [118] proof of Theorem 3.2, which uses the "entropy compression" method, is even simpler, and Bernshteyn [17] simplified this proof further by showing that the "Lopsided Local Lemma" can be used instead of "entropy compression." However, Bernshteyn's proof is non-constructive, and Molloy's "entropy compression" argument provides an efficient randomized algorithm for finding a proper colouring using $(1 + o(1))\Delta/\log\Delta$ colours, matching the "algorithmic barrier" for colouring random graphs described above. Molloy's proof has inspired further algorithmic results such as in [1, 41].

All of these proofs rely on a "coupon collector"-type approach. Roughly speaking, this means that a useful heuristic is to consider a random colouring, where each vertex $v \in V(G)$ is assigned a colour uniformly at random from a set of colours C. If G is triangle-free, then $G[N(v)]$ is an independent set for every $v \in V(G)$ and is thus properly coloured. Moreover, the well-known solution to the coupon collector's problem implies that if $d(v) \leq (1 - o(1))|C| \log |C|$, then there is a colour in C not assigned to a neighbour of v, which we could potentially use to "recolour" v. In particular, if G has maximum degree Δ and $|C| \geq (1 + o(1))\Delta/\log\Delta$, then with non-zero probability, for every vertex $v \in V(G)$, less than $|C|$ colours are assigned to a vertex in $N(v)$. This is of course not sufficient to prove Theorem 3.2 but is a useful intuition for the bound.

It is also believed that at the expense of a worse leading constant, Theorem 3.2 holds for K_r-free graphs for every fixed r, as follows.

Conjecture 3.3. *For every $r \in \mathbb{N}$, there exists a constant c_r such that the following holds. If G is a K_r-free graph with maximum degree at most Δ, then $\chi_\ell(G) \leq c_r\Delta/\log\Delta$.*

The resulting bound on the independence number is already a major open problem proposed earlier by Ajtai, Erdős, Komlós, and Szemerédi [2] (as mentioned in Section 3.1) and is still open even for $r = 4$, and the resulting bound on the chromatic number was conjectured by Alon, Krivelevich, and Sudakov [12]. In this direction, Johansson [83] proved that for every fixed r, every K_r-free graph of maximum degree at most Δ has list chromatic number $O(\Delta \log\log\Delta/\log\Delta)$, which generalizes the result of Shearer [136] mentioned at the end of Section 3.1. Johansson also proved that for every fixed r, if G is a graph of maximum degree at most Δ that satisfies $\chi(G[N(v)]) \leq r$ for every $v \in V(G)$, then $\chi_\ell(G) = O(\Delta/\log\Delta)$, generalizing the result of Alon [6] mentioned at the end of Section 3.1. These results of Johansson were also not published, but Molloy [118] gave a new proof of the former, and the latter was proved (using the approach of Bernshteyn [17]) by Bonamy, Kelly, Nelson, and Postle [22]. Alon, Krivelevich, and Sudakov [12] generalized Johansson's result to "locally sparse graphs" by proving the following: if G is a graph of maximum degree at most Δ such that the neighbourhood of any vertex spans at most Δ^2/f edges, then $\chi(G) = O(\Delta/\log\sqrt{f})$ for $f \leq \Delta^2 + 1$, and Vu [144] generalized this result to list colouring. Davies, Kang, Pirot, and Sereni [42] improved this result by showing that it holds with a leading constant of $1 + o(1)$ as $f \to \infty$, thus generalizing Theorem 3.2.

Theorem 3.4 (Davies, Kang, Pirot, and Sereni [42]). *For every $\varepsilon > 0$, there exists Δ_0 such that the following holds for every $\Delta \geq \Delta_0$. If G is a graph of maximum degree at most Δ such that the neighbourhood of any vertex spans at most Δ^2/f edges for $f \leq \Delta^2 + 1$, then*

$$\chi_\ell(G) \leq (1+\varepsilon)\frac{\Delta}{\log\sqrt{f}}.$$

We note that the aforementioned results of Kim [108], Johansson [82, 83], and Vu [144] all use the nibble method. Davies, Kang, Pirot, and Sereni [42] provided a generalization of all of these results (and also Theorem 3.4) by introducing the "local occupancy method." This method reduces these colouring problems to optimization problems involving relevant local properties of the "hard-core model," which is a family of probability distributions over the independent sets of a graph with origins in statistical physics. Their approach builds on the work of Molloy [118] and Bernshteyn [17] and subsequent work in [22, 38], and the approach used to prove the results of [6, 40, 136] bounding the average size of independent sets mentioned in the previous subsection and also of [39] may be viewed as a precursor to these methods. The main result of Davies, Kang, Pirot, and Sereni [42] is proved using the Lopsided Local Lemma as in Bernshteyn's [17] proof of Theorem 3.2. It can also be proved using entropy compression as in the original proof of Theorem 3.2 of Molloy [118], and indeed, Davies, Kang, Pirot, and Sereni [41] used this approach to obtain additional algorithmic coloring results.

All of the results mentioned so far in this subsection provide a bound of $o(\Delta)$ on the chromatic number of graphs of maximum degree Δ under a "local sparsity" condition. Trivially, every graph G satisfies $\chi(G) \leq \Delta(G) + 1$, and Brooks [24] famously showed that equality holds if and only if G is a complete graph or an odd cycle (when G is connected). With a considerably relaxed "local sparsity" condition, we can still bound the chromatic number away from Δ, as in the following result.

Theorem 3.5 (Molloy and Reed [121]). *For every $\zeta > 0$, there exists Δ_0 such that the following holds for every $\Delta \geq \Delta_0$. If G is a graph of maximum degree at most Δ and every $v \in V(G)$ satisfies $|E(G[N(v)])| \leq (1-\zeta)\binom{\Delta}{2}$, then $\chi(G) \leq (1-\zeta/e^6)\Delta$.*

This result was improved by Bruhn and Joos [27] and by Bonamy, Perrett, and Postle [23]. Recently, Hurley, de Joannis de Verclos, and Kang [78] improved it further by proving the bound $\chi(G) \leq (1 - \zeta/2 + \zeta^{3/2}/6 + o(1))\Delta$, which gives the correct dependence on ζ as $\zeta \to 0$. Determining the best possible bound in Theorem 3.5 for larger ζ is an interesting problem; any further improvements would also improve the best-known bound for Reed's ω, Δ, χ conjecture [124] and for the Erdős–Nešetřil conjecture [51]. We also use Theorem 3.5 in our proof of the Erdős–Faber–Lovász conjecture (see Section 5.2), but we do not need the improvements of [23, 27, 78]. The following related problem was posed by Vu [144] in 2002.

Conjecture 3.6 (Vu [144]). *For every $\zeta, \varepsilon > 0$, there exists Δ_0 such that the following holds for every $\Delta \geq \Delta_0$. If G is a graph of maximum degree at most Δ and every two distinct vertices have at most $\zeta\Delta$ common neighbours in G, then $\chi_\ell(G) \leq (\zeta + \varepsilon)\Delta$.*

This conjecture is still open if we replace $\chi_\ell(G)$ by $\chi(G)$, and even the much weaker conjecture, that G satisfies $\alpha(G) \geq (1/\zeta - \varepsilon)(n/\Delta)$, is still open. The results of [78] give nontrivial bounds when ζ is close to one. If true, Conjecture 3.6 with $\zeta = 1/k$ implies Theorem 2.4 for linear hypergraphs, as follows. Let $\mathcal{H}$ be a k-bounded linear hypergraph with maximum degree at most D. It is clear that $\Delta(L(\mathcal{H})) \leq kD$, and every two distinct vertices in $L(\mathcal{H})$ have at most $\max\{k^2, (D-2) + (k-1)^2\} \leq \zeta k(D + k^2)$ common neighbours. Letting $\Delta := k(D + k^2)$, Conjecture 3.6 would imply $\chi'(\mathcal{H}) = \chi(L(\mathcal{H})) \leq (\zeta + \varepsilon)\Delta = D + o(D)$ when k is fixed as $D \to \infty$. Recently, Kelly, Kühn, and Osthus [106] confirmed a special case of Conjecture 3.6 that also recovers this application to Theorem 2.4.

Our final problem on vertex-colouring graphs is the following conjecture of Alon and Krivelevich [11] from 1998 on the list chromatic number of bipartite graphs.

Conjecture 3.7 (Alon and Krivelevich [11]). *There exists K such that the following holds. If G is a bipartite graph of maximum degree at most Δ, then $\chi_\ell(G) \leq K \log \Delta$.*

The best-known bound for this conjecture is provided by Theorem 3.2; however, this bound can also be proved more directly with the "coupon collector" argument described earlier. Alon, Cambie, and Kang [8] used this argument to prove a stronger result for list colouring bipartite graphs when each vertex in one of the parts has a list of available colours of the conjectured size. Alon and Krivelevich [11] also suggested that the stronger bound $\chi_\ell(G) \leq (1 + o(1)) \log_2 \Delta$ may also hold, which would be best possible for complete bipartite graphs. In fact, Saxton and Thomason [130] proved that every graph of minimum degree at least d has list chromatic number at least $(1 - o(1)) \log_2 d$, improving an earlier result of Alon [7].

3.3. Hypergraph colourings

Theorem 3.1 cannot only be generalized to vertex-colouring in the graphic setting but also for hypergraphs. In 2013, Frieze and Mubayi [60] proved the following result, which generalizes both Johansson's theorem [82] and Theorem 3.1.

Theorem 3.8 (Frieze and Mubayi [60]). *For every $k \geq 2$ there exists $c, \Delta_0 > 0$ such that the following holds for every $\Delta \geq \Delta_0$. If $\mathcal{H}$ is a k-uniform hypergraph with* *maximum degree at most* Δ *and girth at least four, then*

$$\chi(\mathcal{H}) \le c\left(\frac{\Delta}{\log \Delta}\right)^{\frac{1}{k-1}}.$$

To prove this result, Frieze and Mubayi [60] analyzed a nibble procedure inspired by the proof of Johansson [82]. Molloy [117] conjectured that for $k = 3$ the result holds for $c = \sqrt{2} + o(1)$, as this value is suggested by the "coupon collector" heuristic described in Section 3.2, and he asked more broadly if it can also be proved with either the "entropy compression" or the "local occupancy" approach. Iliopoulos [79] showed that the bound $\chi_\ell(\mathcal{H}) \le (1 + o(1))(k-1)(\Delta/\log\Delta)^{1/(k-1)}$ holds in Theorem 3.8 if $\mathcal{H}$ has girth at least five.

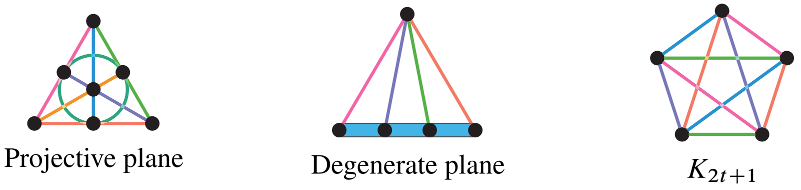
For $k \ge 3$, Frieze and Mubayi [60] "bootstrapped" Theorem 3.8 to show that it actually holds for linear hypergraphs (that is, hypergraphs of girth at least three), by applying it with a distinct set of colours to vertex-disjoint induced subgraphs of girth at least four whose vertices partition $V(\mathcal{H})$. This latter result generalizes the bound of Duke, Lefmann, and Rödl [44] on the independence number mentioned in Section 3.1 to vertex-colouring. Cooper and Mubayi [35] also generalized Theorem 3.8 for $k = 3$ by showing that the girth hypothesis can be replaced with the condition that $\mathcal{H}$ has no *triangle*, where a triangle is a set of three edges e, f, and g such that there exist vertices u, v, and w satisfying $\{u, v\} \subseteq e$, $\{v, w\} \subseteq f$, $\{u, w\} \subseteq g$, and $\{u, v, w\} \cap e \cap f \cap g = \varnothing$. Cooper and Mubayi [36] later showed that both of these results hold under more general "local sparsity" conditions similar to that of Theorem 3.4 for graphs. Frieze and Mubayi [59, 60] conjectured a generalization of Conjecture 3.3 for k-uniform hypergraphs; however, Cooper and Mubayi [37] disproved this conjecture for all $k \ge 3$.

4. The Erdős–Faber–Lovász conjecture

In this section, we introduce and provide background for the Erdős–Faber–Lovász conjecture, which we abbreviate to the EFL conjecture. Earlier developments related to the EFL conjecture are also detailed in the surveys of Kahn [86, 90] and of Kayll [102]. The EFL conjecture states the following (recall that a hypergraph is linear if it has codegree one):

(EFL1) Every n-vertex linear hypergraph has chromatic index at most n.

Erdős often wrote that this was one of his "three favourite combinatorial problems" (see, e.g., [90]). Erdős, Faber, and Lovász famously formulated this conjecture at a tea party in 1972. The simplicity and elegance of the EFL conjecture initially led them to believe it would be easily solved (see, e.g., the discussion in [32, 49]). How-



Projective plane

Degenerate plane

K_{2t+1}

Figure 1. Extremal examples for the Erdős–Faber–Lovász conjecture.

ever, as the difficulty became apparent Erdős offered successively increasing rewards for a proof of the conjecture, which eventually reached $500.

The following three infinite families of hypergraphs are extremal for this conjecture (see Figure 1):

- finite projective planes of order k (known to exist when k is a prime power), which are $(k+1)$-uniform, linear, intersecting hypergraphs on n vertices with n edges where $n := k^2 + k + 1$;
- degenerate planes, also called "Near Pencils," which are linear, intersecting hypergraphs on n vertices with n edges for any $n \in \mathbb{N}$ consisting of one edge of size $n-1$ and $n-1$ edges of size two; and
- complete graphs on n vertices where $n \in \mathbb{N}$ is odd (as well as some "local" modifications of these).

The vastly different structure of these extremal examples contributes to the difficulty of the EFL conjecture. Note in particular that the first two examples have edges of unbounded size as $n \to \infty$, whereas complete graphs are 2-uniform. Let us note that we can (and will) assume without loss of generality that hypergraphs have no edges of size one in the EFL conjecture, since any proper edge-colouring of the edges of size at least two in an n-vertex linear hypergraph $\mathcal{H}$ with at most n colours can be extended to the remaining size-one edges of $\mathcal{H}$ (again with at most n colours). Without this assumption, the hypergraph obtained from an n-vertex star by adding the edge of size one containing the center vertex is also an extremal example.

4.1. Equivalent formulations

Part of the beauty of this conjecture lies in the fact that it can be equivalently stated in several simple, yet seemingly unconnected, ways. The following are all in fact equivalent to the EFL conjecture:

(EFL2) If $\mathcal{H}$ is a linear hypergraph with n edges, each of size at most n, then the vertices of $\mathcal{H}$ can be coloured with at most n colours such that no edge contains two vertices of the same colour.

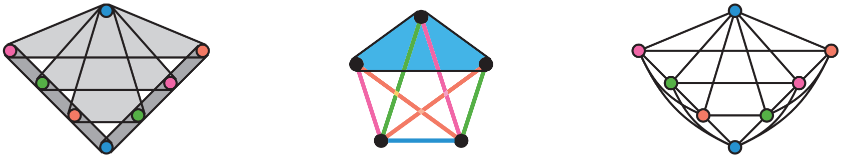


Figure 2. The hypergraph dual (left) of the 5-vertex hypergraph in the center, and the line graph (right).

(EFL3) If $A_1, \ldots, A_n$ are sets of size n such that every pair of them shares at most one element, then the elements of $\bigcup_{i=1}^n A_i$ can be coloured with n colours so that all colours appear in each A_i.

(EFL4) If $G_1, \ldots, G_n$ are complete graphs, each on at most n vertices, such that $|V(G_i) \cap V(G_j)| \leq 1$ for every $1 \leq i \neq j \leq n$, then the chromatic number of $\bigcup_{i=1}^n G_i$ is at most n.

We show that the EFL conjecture is equivalent to (EFL2)–(EFL4) by showing the following implications: (EFL1)⇒(EFL2)⇒(EFL3)⇒(EFL4)⇒(EFL1).

For the first implication, we need to introduce the notion of *hypergraph duality*. The *dual* of a hypergraph $\mathcal{H}$ is the hypergraph $\mathcal{H}^*$ with vertex set $\mathcal{H}$ and edge set $\{\{e \ni v : e \in \mathcal{H}\} : v \in V(\mathcal{H})\}$ (see Figure 2 for an example). Clearly, the dual of $\mathcal{H}^*$ is isomorphic to $\mathcal{H}$ itself. Note that $\mathcal{H}$ is linear if and only if $\mathcal{H}^*$ is linear. Now suppose that $\mathcal{H}$ is a linear hypergraph with n edges, each of size at most n. We may assume without loss of generality that every vertex of $\mathcal{H}$ has degree at least two. Since $\mathcal{H}$ has n edges and is linear, $\mathcal{H}^*$ has n vertices and is also linear, so (EFL1) implies that there is proper edge-colouring of $\mathcal{H}^*$ using at most n colours. By assigning each vertex of $\mathcal{H}$ the colour of the corresponding edge of $\mathcal{H}^*$, we obtain the desired colouring, proving (EFL2).

To show that (EFL2)⇒(EFL3), let $\mathcal{H}$ be the hypergraph with vertex set $\bigcup_{i=1}^n A_i$ and edge set $\{A_i : i \in [n]\}$. Since $A_1, \ldots, A_n$ have size n and every pair of them shares at most one element, $\mathcal{H}$ is linear with n edges, each of size n. By (EFL2), the vertices of $\mathcal{H}$ can be coloured with at most n colours such that no edge contains two vertices of the same colour. Since every edge has size n, every edge contains a vertex of every colour, so this colouring satisfies (EFL3). To prove (EFL3)⇒(EFL4), first note that by possibly adding new vertices to each G_i, we may assume without loss of generality that $|V(G_i)| = n$ for each $i \in [n]$. Letting $A_i = V(G_i)$ for each $i \in [n]$, (EFL3) implies there is a colouring of $\bigcup_{i=1}^n A_i$ with n colours so that all colours appear in each A_i. In particular, if $u, v \in A_i$, then u and v are assigned different colours, so this colouring is also a proper vertex-colouring of $\bigcup_{i=1}^n G_i$, proving (EFL4).

Finally, to prove that the EFL conjecture follows from (EFL4), let $\mathcal{H}$ be a linear hypergraph on n vertices, and for each $v \in V(\mathcal{H})$, let G_v be the complete graph with vertex set $\{e \ni v : e \in \mathcal{H}\}$. Since $\mathcal{H}$ is linear, each G_v for $v \in V(\mathcal{H})$ has at most n vertices, and $|V(G_u) \cap V(G_v)| \leq 1$ for distinct $u, v \in V(\mathcal{H})$. Since $\bigcup_{v \in V(\mathcal{H})} G_v$ is in fact the line graph of $\mathcal{H}$ (see Figure 2), we have $\chi'(\mathcal{H}) = \chi(\bigcup_{v \in V(\mathcal{H})} G_v)$, so (EFL4)⇒(EFL1), as desired.

An interpretation of (EFL4) in terms of a scheduling problem was given by Haddad and Tardif [73].

4.2. Results

Recently, we confirmed the EFL conjecture for all but finitely many hypergraphs.

Theorem 4.1 (Kang, Kelly, Kühn, Methuku, and Osthus [96]). *For every sufficiently large n, every n-vertex linear hypergraph has chromatic index at most n.*

The proof of Theorem 4.1 can be turned into a randomized polynomial-time algorithm. The necessary modifications are discussed in detail in [98]. We also proved the following "stability result," predicted by Kahn [86].

Theorem 4.2 (Kang, Kelly, Kühn, Methuku, and Osthus [96]). *For every $\delta > 0$, there exist $n_0, \sigma > 0$ such that the following holds for every $n \geq n_0$. If $\mathcal{H}$ is an n-vertex linear hypergraph such that*

(i) *$\mathcal{H}$ has maximum degree at most $(1-\delta)n$ and*

(ii) *the number of edges of size $(1 \pm \delta)\sqrt{n}$ in $\mathcal{H}$ is at most $(1-\delta)n$,*

then the chromatic index of $\mathcal{H}$ is at most $(1-\sigma)n$.

The hypothesis (i) in Theorem 4.2 ensures that $\mathcal{H}$ does not too closely resemble the degenerate plane or the complete graph, while (ii) ensures that $\mathcal{H}$ does not too closely resemble a projective plane, since projective planes on n vertices have n edges of size roughly $\sqrt{n}$.

Let us overview previous progress leading up to these results. Predating the EFL conjecture, in 1948 de Bruijn and Erdős [43] showed that every intersecting n-vertex linear hypergraph has at most n edges. Equivalently, the line graph of an n-vertex linear hypergraph contains no clique of size greater than n. Seymour [132] proved that every n-vertex linear hypergraph $\mathcal{H}$ contains a matching of size at least $|\mathcal{H}|/n$, which implies the de Bruijn–Erdős theorem, as an intersecting hypergraph has matching number one. Kahn and Seymour [94] strengthened this result by proving that every n-vertex linear hypergraph has fractional chromatic index at most n. (Recall that every hypergraph $\mathcal{H}$ satisfies $\chi'_f(\mathcal{H}) \leq \chi'(\mathcal{H})$, so all of these results are relaxations of the EFL conjecture.) Chang and Lawler [30] proved that every n-vertex linear hypergraph has chromatic index at most $\lceil 3n/2 - 2 \rceil$.

Interestingly, results from both Section 2 and Section 3 have the following immediate applications to the EFL conjecture, which are illustrative to note.

(4.1) For every $\varepsilon, k > 0$, there exists n_0 such that the following holds for every $n \geq n_0$. If $\mathscr{H}$ is a k-bounded, n-vertex, linear hypergraph, then $\chi'(\mathscr{H}) \leq n + \varepsilon n$ and moreover, if every $e \in \mathscr{H}$ satisfies $|e| \geq 3$, then $\chi'(\mathscr{H}) \leq n$.

(4.2) For every $\varepsilon > 0$, there exist $\delta, n_0 > 0$ such that the following holds for every $n \geq n_0$ and $k := \delta\sqrt{n}$. If $\mathscr{H}$ is a k-uniform, n-vertex, linear hypergraph, then $\chi'(\mathscr{H}) \leq \varepsilon n$.

(4.3) For every $\varepsilon > 0$, there exist $\delta, n_0 > 0$ such that the following holds for every $n \geq n_0$ and $k := (1-\delta)\sqrt{n}$. If $\mathscr{H}$ is a k-uniform, n-vertex, linear hypergraph, then $\chi'(\mathscr{H}) \leq (1-\varepsilon)n$.

To prove (4.1), it suffices to note that since $\mathscr{H}$ is linear, it has maximum degree at most $n/(\min_{e\in\mathscr{H}} |e| - 1)$ (assuming $\mathscr{H}$ has no size-one edges), whence (4.1) follows immediately from Theorem 2.4. To prove (4.2) and (4.3), it suffices to note that since $\mathscr{H}$ is linear, the line graph $L := L(\mathscr{H})$ has maximum degree at most $k(n-k)/(k-1) \leq (1+2/k)n$ and every pair of adjacent vertices in L has at most $(k-1)^2 + n/k$ common neighbours. Hence, (4.2) follows immediately from Theorem 3.4, and (4.3) follows immediately from Theorem 3.5.

In 1992, a breakthrough by Kahn [84] confirmed the EFL conjecture asymptotically, by showing that every n-vertex linear hypergraph has chromatic index at most $n + o(n)$. Note that this result strengthens the first part of (4.1) by showing that the k-boundedness assumption is not necessary. Kahn's argument in [84] relies on a "restricted" list colouring result which strengthens the Pippenger–Spencer theorem (Theorem 2.3) but is still weaker than Theorem 2.4, and thus can be viewed as a "stepping stone" towards Theorem 2.4. Moreover, Kahn's argument from [84], combined with Theorem 2.4 that he proved later in [87], can be adapted to prove that every n-vertex linear hypergraph has *list* chromatic index at most $n + o(n)$, which we explain further in Section 4.4. The second part of (4.1) was strengthened in 2019 by Faber and Harris [54], who proved that for some absolute constant c, the EFL conjecture holds if every edge has size at least three and at most $c\sqrt{n}$. In fact, the main result of [54] also implies (4.2). Their argument relies on Theorem 2.4 and the result of Vu [144] mentioned before Theorem 3.4. That the works [12,144] have applications to the EFL conjecture was first observed by Faber [52], namely to prove a result similar to (4.2).

Nevertheless, none of the results prior to Theorem 4.1 confirmed the conjecture for any nontrivial class of hypergraphs containing one of the extremal families. In particular, the case of k-bounded (or even 3-bounded) hypergraphs was still open (and was highlighted as a challenging problem by Kahn [86]). Similarly, the case of hypergraphs in which all edges have size $\omega(1)$ was also still open. Both of these cases turned out to be significant stepping stones towards the proof of Theorem 4.1, and their proofs contain several of the main ideas. To highlight these ideas, we provide a detailed sketch of the following two results in Section 5.

Theorem 4.3. *There exists $n_0 > 0$ such that the following holds. If $\mathcal{H}$ is an n-vertex linear hypergraph such that every $e \in \mathcal{H}$ satisfies $|e| \in \{2, 3\}$ and $n > n_0$, then $\chi'(\mathcal{H}) \leq n + 1$.*

Theorem 4.4. *For every $\delta > 0$, there exist $n_0, r, \sigma > 0$ such that the following holds. If $\mathcal{H}$ is an n-vertex linear hypergraph where $n > n_0$ and every $e \in \mathcal{H}$ satisfies $|e| > r$, then $\chi'(\mathcal{H}) \leq n$. Moreover, if $\chi'(\mathcal{H}) > (1 - \sigma)n$, then $|\{e \in \mathcal{H} : |e| = (1 \pm \delta)\sqrt{n}\}| \geq (1 - \delta)n$.*

Notice that in Theorem 4.3, the bound on the chromatic index is one larger than in the EFL conjecture. In Section 5.1, we prove Theorem 4.3 and briefly explain the additional ideas required to prove the stronger bound of the EFL conjecture in this case. The proof of Theorem 4.3 can also be adapted with little additional effort to prove the same result for k-bounded hypergraphs, for any fixed k. We focus on the case of 3-bounded hypergraphs as it is slightly cleaner yet complex enough to capture many of the important ideas.

Note also that Theorem 4.4 implies Theorem 4.2 in the case when all edges of $\mathcal{H}$ are sufficiently large. This "stability result" is needed to combine the arguments of Theorems 4.3 and 4.4 to obtain Theorem 4.1. Roughly, we apply (a stronger version of) Theorem 4.4 first to the "large" edges of $\mathcal{H}$, and then we apply the arguments of Theorem 4.3 to find a proper edge-colouring of the "small" edges of $\mathcal{H}$ that is compatible with the colouring of the "large" edges. If Theorem 4.4 only requires $(1 - \sigma)n$ colours, then only minor adaptations to the arguments of Theorem 4.3 are required, which we briefly describe in Section 5.2 after proving Theorem 4.4. If Theorem 4.4 only guarantees a proper edge-colouring of the large edges of $\mathcal{H}$ with n colours, then additional ideas are required, for which we refer the interested reader to [96] (which in particular contains a sketch of the overall argument).

4.3. Open problems

We now discuss some open problems related to the EFL conjecture. First, it would be interesting to characterize when equality holds in Theorem 4.1. As mentioned, finite projective planes, degenerate planes, and complete graphs on an odd number of vertices are extremal examples. In fact, any n-vertex hypergraph with more than $(n - 1)^2/2$ size-two edges has chromatic index at least n when n is odd, and any hypergraph $\mathcal{H}$ obtained from K_n by replacing a complete subgraph with a single edge e is linear and has chromatic index n if $|V(\mathcal{H}) \setminus e|$ is odd (note that the degenerate plane is obtained in this way). These may include all of the extremal examples.

Berge [16] and Füredi [62] independently posed the following beautiful conjecture.

Conjecture 4.5 (Berge [16], Füredi [62]). *If $\mathcal{H}$ is a linear hypergraph with vertex set V, then $\chi'(\mathcal{H}) \leq \max_{v\in V} |\bigcup_{e\ni v} e|$.*

If true, Conjecture 4.5 implies the EFL conjecture, since every linear hypergraph $\mathcal{H}$ satisfies $\max_{v\in V(\mathcal{H})} |\bigcup_{e\ni v} e| \leq n$. Note also that if $G_{\mathcal{H}}$ is the graph obtained from $\mathcal{H}$ by replacing each edge $e \in \mathcal{H}$ with a complete graph on the vertices of e (sometimes called the *shadow* of $\mathcal{H}$), then $|\bigcup_{e\ni v} e| = \Delta(G_{\mathcal{H}}) + 1$. In particular, if $\mathcal{H}$ is 2-uniform, then $G_{\mathcal{H}} = \mathcal{H}$, so Conjecture 4.5, if true, also implies Vizing's theorem. The fractional relaxation of Conjecture 4.5 is still open. The following related conjecture was posed by Füredi, Kahn, and Seymour [64]: if $\mathcal{H}$ is a multi-hypergraph with vertex set V, then $\chi'_f(\mathcal{H}) \leq \max_{v\in V} \sum_{e\ni v}(|e| - 1 + 1/|e|)$. This conjecture may even hold for the chromatic index, which, if true, would generalize Shannon's theorem [133] and imply Conjecture 2.6 in the case $k = t$.

It is also natural to ask whether the EFL conjecture holds more generally for list colouring. Faber [53] conjectured that it does, as follows.

Conjecture 4.6 (The "List" EFL conjecture [53]). *Every n-vertex linear hypergraph has list chromatic index at most n.*

Conjecture 4.6 was recently confirmed by the authors [97] for the special case of hypergraphs of maximum degree at most $n - o(n)$, and their result also implies that in this case projective planes are the only extremal examples. The main result in [97] also solves a conjecture of Erdős on the chromatic index of hypergraphs of small codegree.

A related problem to Conjecture 4.6 is an algebraic strengthening of the EFL conjecture involving the Combinatorial Nullstellensatz, posed by Janzer and Nagy [80].

As mentioned, the arguments of Kahn [84, 87] can be adapted to prove that the List EFL conjecture holds asymptotically. Kahn's proof in [84] also implies that Conjecture 4.5 holds asymptotically. In fact, assuming the sizes of the lists are polylogarithmic in n, it is easy to show that the argument can be adapted to work for list colouring, as follows.

Theorem 4.7. *For every $\varepsilon > 0$, there exists n_0 such that the following holds for every $n, D \geq n_0$. If $\mathcal{H}$ is an n-vertex linear hypergraph such that $|\bigcup_{e\ni v} e| \leq D$ for every $v \in V(\mathcal{H})$, then $\chi'(\mathcal{H}) \leq (1+\varepsilon)D$. Moreover, if $D \geq \log^2 n$, then $\chi'_\ell(\mathcal{H}) \leq (1+\varepsilon)D$.*

For completeness, we prove Theorem 4.7 in Section 4.4. It would be interesting to prove that the bound on the list chromatic index in Theorem 4.7 holds without the assumption $D \geq \log^2 n$.

The next open problem is the following special case of a conjecture of Larman.

Conjecture 4.8 ("Restricted" Larman's conjecture). *If $\mathcal{H}$ is an n-vertex intersecting hypergraph, then there exists a decomposition of $\mathcal{H}$ into $\mathcal{F}_1, \dots, \mathcal{F}_n \subseteq \mathcal{H}$ such that $|F \cap F'| \geq 2$ for every $F, F' \in \mathcal{F}_i$ and $i \in [n]$.*

The full version of Larman's conjecture was a combinatorial relaxation of Borsuk's conjecture from 1933, which states that every set of diameter at most one in $\mathbb{R}^d$ can be partitioned into at most $d + 1$ sets of diameter strictly less than one. However, in 1993, Kahn and Kalai [92] disproved Larman's conjecture (and thus in turn Borsuk's conjecture). Nevertheless, they asked whether the special case of Larman's conjecture presented in Conjecture 4.8 still holds (see also [95]), in part because of its resemblance to the EFL conjecture.

Finally, we note that Alon, Saks, and Seymour (see Kahn [85]) conjectured the following "bipartite version" of (EFL4): if a graph G can be decomposed into k edge-disjoint bipartite graphs, then the chromatic number of G is at most $k + 1$. This conjecture was a generalization of the Graham–Pollak theorem [72] on edge decompositions of complete graphs into bipartite graphs, which has applications to communication complexity. However, it was disproved by Huang and Sudakov [77] in a strong form, i.e., the conjectured bound on the chromatic number is far from being true.

4.4. Asymptotic list colouring version of the Berge–Füredi conjecture

In this subsection, we prove Theorem 4.7. We only prove the bound on the list chromatic index when $D \geq \log^2 n$, as the proof of the general bound on the chromatic index will be evident from the argument we provide here. Our proof closely follows the approach of [84], but with a simple additional trick, and using Theorem 2.4 instead of [84, Theorem 1.3].

Proof of Theorem 4.7. Let

$$1/n_0 \ll 1/r_0 \ll 1/r_1 \ll \gamma \ll \varepsilon \ll 1,$$

let $n \geq n_0$, let $D \geq \log^2 n$, and let $\mathcal{H}$ be an n-vertex linear hypergraph such that $|\bigcup_{e \ni v} e| \leq D$ for every $v \in V(\mathcal{H})$. It suffices to show that if C is an assignment of lists $C(e)$ to every $e \in \mathcal{H}$, such that every $e \in \mathcal{H}$ satisfies $|C(e)| \geq (1 + \varepsilon)D$, then $\mathcal{H}$ has a proper edge-colouring ϕ such that $\phi(e) \in C(e)$ for every $e \in \mathcal{H}$. We assume without loss of generality that $|C(e)| = (1 + \varepsilon)D \pm 1$.

Let $\preceq$ be a linear ordering of the edges of $\mathcal{H}$ satisfying $e \preceq f$ if $|e| > |f|$, and decompose $\mathcal{H}$ into the following spanning subhypergraphs:

- $\mathcal{H}_{\text{sml}} := \{e \in \mathcal{H} : |e| \leq r_1\}$,
- $\mathcal{H}_{\text{med}} := \{e \in \mathcal{H} : r_1 < |e| \leq r_0\}$, and
- $\mathcal{H}_{\text{lrg}} := \{e \in \mathcal{H} : |e| > r_0\}$.

Since $\mathcal{H}$ is linear,

(4.4) every $e \in \mathcal{H}_{\mathrm{lrg}}$ satisfies $|\{f \in N_{\mathcal{H}}(e) : f \preceq e\}| \leq |e|(D - |e|)/(|e| - 1) \leq (1 + \varepsilon/3)D$,

(4.5) $\Delta(\mathcal{H}_{\mathrm{med}}) \leq D/r_1$ (and thus, by Theorem 2.4, $\chi'_\ell(\mathcal{H}_{\mathrm{med}}) \leq 2D/r_1 \leq \gamma D/2$), and

(4.6) every $e \in \mathcal{H}_{\mathrm{sml}}$ satisfies $|N_{\mathcal{H}}(e) \cap \mathcal{H}_{\mathrm{lrg}}| \leq r_1 D/r_0 \leq \varepsilon D/4$.

Now we show there exists a set $R \subseteq \bigcup_{e \in \mathcal{H}} C(e)$ such that

$$\text{every } e \in \mathcal{H} \text{ satisfies } \big|R \cap C(e)\big| = (1 \pm 1/2)\gamma\big|C(e)\big|. \tag{4.7}$$

Include every colour in R randomly and independently with probability γ. By a standard application of the Chernoff bound, every $e \in \mathcal{H}$ satisfies

$$\big|R \cap C(e)\big| = (1 \pm 1/2)\gamma\big|C(e)\big|$$

with probability at least $1 - 2\exp(-\gamma|C(e)|/12) \geq 1 - 2\exp(-\gamma \log^2 n/12)$. Thus by the Union Bound, (4.7) holds with high probability, and hence there indeed exists such a set R.

Fix R satisfying (4.7), and for every $e \in \mathcal{H}$, let $C'(e) := C(e) \setminus R$ and $R(e) := C(e) \cap R$. By (4.7),

(4.8) every $e \in \mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{sml}}$ satisfies $|C'(e)| \geq (1 + \varepsilon/2)D$, and

(4.9) every $e \in \mathcal{H}_{\mathrm{med}}$ satisfies $|R(e)| \geq \gamma D/2$.

Therefore, by (4.4) and (4.8), there exists a proper edge-colouring ϕ_{lrg} of $\mathcal{H}_{\mathrm{lrg}}$ such that $\phi_{\mathrm{lrg}}(e) \in C'(e)$ for every $e \in \mathcal{H}_{\mathrm{lrg}}$, and by (4.5) and (4.9), there exists a proper edge-colouring ϕ_{med} of $\mathcal{H}_{\mathrm{med}}$ such that $\phi_{\mathrm{med}}(e) \in R(e)$ for every $e \in \mathcal{H}_{\mathrm{med}}$. Now for each $e \in \mathcal{H}_{\mathrm{sml}}$, let $C''(e) := C'(e) \setminus \{\phi_{\mathrm{lrg}}(f) : f \in N_{\mathcal{H}}(e) \cap \mathcal{H}_{\mathrm{lrg}}\}$. By (4.6) and (4.8), $|C''(e)| \geq (1 + \varepsilon/4)D$ for every $e \in \mathcal{H}_{\mathrm{sml}}$. Therefore, by Theorem 2.4, there exists a proper edge-colouring ϕ_{sml} of $\mathcal{H}_{\mathrm{sml}}$ such that $\phi_{\mathrm{sml}}(e) \in C''(e)$ for every $e \in \mathcal{H}_{\mathrm{sml}}$. By combining ϕ_{lrg}, ϕ_{med}, and ϕ_{sml}, we obtain the desired colouring. ■

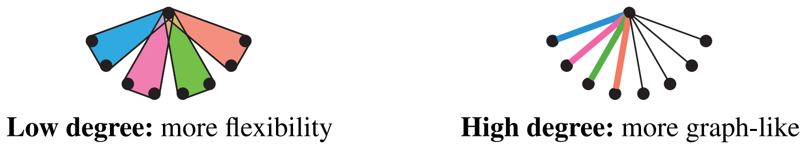
5. Proving the Erdős–Faber–Lovász conjecture

In this section, we give detailed sketches of the proofs of Theorems 4.3 and 4.4, the special cases of the proof of the EFL conjecture in [96] discussed in Section 4.2.

5.1. Using $n + 1$ colours when edge-sizes are bounded

We begin with Theorem 4.3, which we restate for the readers' convenience.

Theorem 4.3. *There exists $n_0 > 0$ such that the following holds. If $\mathcal{H}$ is an n-vertex linear hypergraph such that every $e \in \mathcal{H}$ satisfies $|e| \in \{2, 3\}$ and $n > n_0$, then $\chi'(\mathcal{H}) \leq n + 1$.*



Low degree: more flexibility

High degree: more graph-like

Figure 3. Two partial edge-colourings using 4 colours (when $n = 9$, say). The uncoloured edges form a graph of maximum degree at most 4 and can be coloured with at most 5 colours by Vizing's theorem.

In this subsection, we fix constants satisfying the hierarchy

$$0 < 1/n_0 \ll \xi \ll \kappa \ll \gamma \ll \varepsilon \ll 1, \tag{5.1}$$

we let $n \geq n_0$, and we let $\mathcal{H}$ be an n-vertex linear hypergraph such that every $e \in \mathcal{H}$ satisfies $|e| \in \{2, 3\}$. We assume without loss of generality that every pair of vertices of $\mathcal{H}$ is contained in an edge, since otherwise we can add a size-two edge to $\mathcal{H}$ to obtain an n-vertex linear hypergraph with chromatic index greater than or equal to $\chi'(\mathcal{H})$. Let G be the graph with $V(G) := V(\mathcal{H})$ and $E(G) := \{e \in \mathcal{H} : |e| = 2\}$, and let $U := \{u \in V(\mathcal{H}) : d_G(u) \geq (1-\varepsilon)n\}$. Since every pair of vertices is contained in precisely one edge, we have

$$\begin{aligned} n - 1 &= 2\big(d_{\mathcal{H}}(v) - d_G(v)\big) + d_G(v) \\ &= 2d_{\mathcal{H}}(v) - d_G(v) \quad \text{for every vertex } v \in V(\mathcal{H}). \end{aligned} \tag{5.2}$$

Our strategy to prove Theorem 4.3 is to reduce it to Vizing's theorem. In order to do that, it suffices to partially colour $\mathcal{H}$ with $k < n$ colours (for some suitable $k \sim n/2$) such that every edge of $\mathcal{H} \setminus E(G)$ is coloured and the remaining uncoloured edges of G form a graph of maximum degree at most $n - k$ (see Figure 3 and Lemma 5.9). Roughly speaking, each colour class of this partial colouring will be obtained by first constructing a large matching via the Rödl nibble and then extending it to cover (essentially) all of U. The latter step is of course necessary in order to obtain an (uncoloured) leftover graph of small maximum degree. (It is also sufficient since U consists of precisely those $v \in V(\mathcal{H})$ with $d_{\mathcal{H}}(v) \sim n$.) On the other hand, while this is the reason we need to pay special attention to the vertices in U, the definition of U also means that we have many (graph) edges at our disposal, which allow us to carry out the extension step mentioned above. To make this precise, we introduce the following important definition.

Definition 5.1 (Perfect and nearly perfect coverage). Let $\mathcal{M}$ be a set of edge-disjoint matchings in $\mathcal{H}$, and let $S \subseteq U$.

- We say that $\mathcal{M}$ has *perfect coverage* of U if each $M \in \mathcal{M}$ covers U.

- We say that $\mathcal{M}$ has *nearly perfect coverage of U with defects in S* if
  (i) each $u \in U$ is covered by at least $|\mathcal{M}| - 1$ matchings in $\mathcal{M}$ and
  (ii) each $M \in \mathcal{M}$ covers all but at most one vertex in U, and $U \setminus V(M) \subseteq S$.

More precisely, using $k := \lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil$ colours, we will partially colour $\mathcal{H}$ such that

- every edge of $\mathcal{H} \setminus E(G)$ is coloured,
- at least $d_G(v)/2 - 2\xi n$ edges of G containing v are coloured for every $v \in V(G)$, and
- the colour classes have nearly perfect coverage of U (with defects in U).

As we will show, these conditions ensure that the partial colouring can be extended via Vizing's theorem to all of $\mathcal{H}$ using at most $n + 1$ total colours.

The first step of the proof is to randomly construct a "reservoir" consisting of edges of G (which will be used for the extension step), as in the following lemma.

Lemma 5.2 (Reservoir lemma). *There exists $R \subseteq E(G)$ satisfying the following:*

(R1) *(Typicality) every $v \in V(\mathcal{H})$ satisfies*

$$\begin{aligned} \left|N_R(v) \cap U\right| &= \left|N_G(v) \cap U\right|/2 \pm \xi n, \\ \left|N_R(v) \setminus U\right| &= \left|N_G(v) \setminus U\right|/2 \pm \xi n; \end{aligned}$$

(R2) *(Upper regularity) for every pair of disjoint sets $S, T \subseteq V(\mathcal{H})$ with $|S|, |T| \geq \xi n$, we have*

$$\left|E_G(S,T) \cap R\right| \leq (1/2 + \xi)|S||T|. \qquad \blacksquare$$

This lemma can be proved with a straightforward application of the Chernoff Bound and the Union Bound, by considering the set R to be chosen randomly, where each edge of G is included independently and with probability $1/2$, so we omit the details. For the remainder of the subsection, we fix R satisfying Lemma 5.2.

By (R1), every vertex $v \in V(\mathcal{H})$ satisfies $d_{\mathcal{H}\setminus R}(v) = d_{\mathcal{H}}(v) - d_G(v)/2 \pm 2\xi n$. Hence, by (5.2),

$$\text{every vertex } v \in V(\mathcal{H}) \text{ satisfies } d_{\mathcal{H}\setminus R}(v) = \frac{n-1}{2} \pm 2\xi n. \tag{5.3}$$

Note that by (5.3), the Pippenger–Spencer theorem already implies $\chi'(\mathcal{H} \setminus R) \leq (1/2 + \gamma^{1/3})n$, but we need to prove the stronger result that there is a set of pairwise edge-disjoint matchings $\mathcal{M} = \{M_1, \ldots, M_k\}$ such that $M_1 \cup \cdots \cup M_k \supseteq \mathcal{H} \setminus R$ and $\mathcal{M}$ has nearly perfect coverage of U (with defects in U).

5.1.1. Absorption. To obtain these matchings with nearly perfect coverage of U, we combine the nibble method with an absorption strategy. We first find matchings in $\mathcal{H} \setminus R$ covering almost all of U using Theorem 2.14, and then for each such matching, we find a vertex-disjoint matching in R covering (all but at most one of) the remaining vertices of U. We extend the first matching by adding the matching in R, and in this way we "absorb" the uncovered vertices of U. It will be convenient to work with the following definitions, which will apply to the matchings produced by Theorem 2.14.

Definition 5.3 (Pseudorandom matchings). For a family $\mathcal{F}$ of subsets of $V(\mathcal{H})$, a matching M in $\mathcal{H}$ is (γ, κ)*-pseudorandom* with respect to $\mathcal{F}$ if every $S \in \mathcal{F}$ satisfies $|S \setminus V(M)| = \gamma|S| \pm \kappa n$.

Definition 5.4 (Absorbable matchings). Let $R' \subseteq R$, let $S \subseteq U$, and let M be a matching in $\mathcal{H} \setminus R$. We say (M, R', S) is *absorbable* if

(AB1) $|S| \geq \min\{|U|, n/5\}$,

(AB2) $\Delta(R \setminus R') \leq \gamma n$, and

(AB3) either $|V(M)| \leq \sqrt{\gamma} n$, or M is (γ, κ)-pseudorandom with respect to $\mathcal{F}(R') \cup \{U, S\}$, where

$$\mathcal{F}(R') := \left\{N_{R'}(u) \cap U : u \in U\right\} \cup \left\{N_{R'}(u) \setminus U : u \in U\right\}.$$

If the former holds in (AB3), we say that (M, R', S) is absorbable by the *smallness* of M, and if the latter holds, we say that (M, R', S) is absorbable by the *pseudorandomness* of M.

In the proof of Theorem 4.3, we apply our absorption argument successively to each matching constructed by the nibble. Hence, in each step we will consider absorbable tuples (M, R', S) where M was obtained via a nibble process, R' consists of reservoir edges not used in previous absorption steps, and S consists of vertices of U that are not the "defect" from any of the previous absorption steps. Now we can state our main absorption lemma, but first we note the following proposition, which is used in its proof.

Proposition 5.5. *Let $0 < 1/m_0 \ll \alpha \ll 1$, and let $m \geq m_0$ be even. If H is an m-vertex graph such that*

(i) *every $v \in V(H)$ satisfies $d_H(v) \geq 3m/8$ and*

(ii) *every pair of disjoint sets $S, T \subseteq V(H)$ with $|S|, |T| \geq \alpha m$ satisfies $e_H(S, T) \leq (1/2 + \alpha)|S||T|$,*

then H has a perfect matching.

To prove Proposition 5.5, one can consider a random equitable partition of $V(H)$ and apply Hall's theorem.

Lemma 5.6 (Absorption lemma). *Let $R' \subseteq R$, let $S \subseteq U$, and let $\mathcal{N} := \{N_1, \ldots, N_k\}$ be a set of pairwise edge-disjoint matchings in $\mathcal{H} \setminus R$. If either*

(i) *$k \leq \kappa n$ and, for every $i \in [k]$, (N_i, R', S) is absorbable by the pseudorandomness of N_i or*

(ii) *$k \leq \lceil \gamma^{1/3} n \rceil$ and, for every $i \in [k]$, (N_i, R', S) is absorbable by the smallness of N_i,*

then there is a set of pairwise edge-disjoint matchings $\mathcal{M} := \{M_1, \ldots, M_k\}$ in $\mathcal{H}$ such that

- *$M_i \supseteq N_i$ and $M_i \setminus N_i \subseteq R'$ for all $i \in [k]$, and*
- *$\mathcal{M}$ has nearly perfect coverage of U with defects in S, and moreover, if $|U| < 3n/4$, then $\mathcal{M}$ has perfect coverage of U.*

Proof. Let $\mathcal{F} := \mathcal{F}(R') \cup \{U, S\}$, and for each $i \in [k]$, let $U_i := U \setminus V(N_i)$. In both cases, the proof proceeds roughly as follows. If $|U| < n/100$, then one-by-one for each $i \in [k]$, we can greedily find a matching N_i^{abs} of edges in R', with precisely one end in U_i and the other end not in $V(N_i)$, edge-disjoint from those previously chosen, that covers U_i. Letting $M_i := N_i \cup N_i^{\text{abs}}$ for each $i \in [k]$, $\{M_1, \ldots, M_k\}$ has perfect coverage of U, as desired. If $|U| \geq n/100$, then one-by-one for each $i \in [k]$, using Proposition 5.5, we can find a matching N_i^{abs} of edges in R', with both ends in U_i, edge-disjoint from those previously chosen, that contain all but at most one vertex of U_i. Moreover, we ensure that the vertices in each U_i not covered by N_i^{abs} are distinct, and if $|U| < 3n/4$, we can also augment each N_i^{abs} with an edge of R' that has an end in $V(\mathcal{H}) \setminus (U \cup V(N_i) \cup V(N_i^{\text{abs}}))$ to cover U. Hence, $\{M_1, \ldots, M_k\}$ has nearly perfect coverage of U with defects in S and perfect coverage of U if $|U| < 3n/4$, where $M_i := N_i \cup N_i^{\text{abs}}$ for each $i \in [k]$, as desired. We only provide a formal proof of the case when (i) holds and $|U| \geq n/100$, as this case is the most challenging.

For each $i \in [k]$, let G_i be the graph with $V(G_i) := U_i$ and $E(G_i) := \{e \in R' : e \subseteq U_i\}$. Since N_i is (γ, κ)-pseudorandom with respect to $\mathcal{F} \ni U$, we have

$$|U_i| = \gamma |U| \pm \kappa n \text{ and, in particular, } |U_i| \geq \gamma n/200. \tag{5.4}$$

We claim that for each $i \in [k]$ there exists $u_i \in U_i$ and a matching N_i^{abs} in G_i such that the following holds. The vertices $u_1, \ldots, u_k$ are distinct, the matchings $N_1^{\text{abs}}, \ldots, N_k^{\text{abs}}$ are pairwise edge-disjoint, and N_i^{abs} covers every vertex of $U_i \setminus \{u_i\}$ for each $i \in [k]$. Moreover, if $|U| < 3n/4$, then N_i^{abs} covers every vertex of U_i for each $i \in [k]$, and otherwise $u_i \in S$.

To that end, we choose distinct $u_i \in U_i$ for each $i \in [k]$, as follows.

- If $|U| \leq 3n/4$, then by (R1) and (AB2), every $u \in U_i$ satisfies $|N_{R'}(u) \setminus U| \geq (1/4 - \varepsilon)n/2 - \xi n - \gamma n \geq n/10$. By (AB3), since N_i is (γ, κ)-pseudorandom with respect to $\mathcal{F} \supseteq \mathcal{F}(R')$ for each $i \in [k]$, this inequality implies that every $u \in U_i$ satisfies $|N_{G_i}(u) \setminus U| \geq \gamma n/20$. Since $k \leq \kappa n$ and $\kappa \ll \gamma$, by (5.4), we can choose $u_i \in U_i$ one-by-one such that there is a matching $\{u_i v_i : i \in [k]\}$ where $v_i \in N_{G_i}(u_i) \setminus U$ for each $i \in [k]$.

- Otherwise, (AB1) implies $|S| \geq \gamma n$, and since N_i is (γ, κ)-pseudorandom with respect to $\mathcal{F} \ni S$, by (AB3), we have $|S \setminus V(N_i)| \geq \gamma|S| - \kappa n \geq \gamma^2 n/2 > \kappa n$ for each $i \in [k]$. So we can choose $u_i \in U_i \cap S = S \setminus V(N_i)$ one-by-one such that they are distinct, as required.

Now let $U_i' := U_i \setminus \{u_i\}$ if $|U_i|$ is odd. Otherwise, let $U_i' := U_i$. By the choice of the vertices $u_1, \ldots, u_k$, it suffices to find pairwise edge-disjoint perfect matchings N'^{abs}_i in $G_i[U_i']$ for each $i \in [k]$. Indeed if $|U| \leq 3n/4$ and $|U_i|$ is odd, then $N^{\text{abs}}_i := N'^{\text{abs}}_i \cup \{u_i v_i\}$ satisfies the claim, and otherwise $N^{\text{abs}}_i := N'^{\text{abs}}_i$ satisfies the claim.

We find these matchings one-by-one using Proposition 5.5. To this end, we assume that for some $\ell \leq k$, we have found such matchings N'^{abs}_i for $i \in [\ell - 1]$, and we show that there exists such a matching N'^{abs}_ℓ, which proves the claim. Let $G_\ell' := G_\ell[U_\ell'] \setminus \bigcup_{i \in [\ell-1]} N'^{\text{abs}}_i$. Since $|U| \geq n/100$, by (R1) and (AB2), every $u \in U$ satisfies

$$\big|N_{R'}(u) \cap U\big| \geq \big|N_R(u) \cap U\big| - \gamma n \geq \big(|U| - \varepsilon n\big)/2 - 2\gamma n \geq 49|U|/100. \quad (5.5)$$

Note that N_ℓ is (γ, κ)-pseudorandom with respect to $\mathcal{F} \supseteq \mathcal{F}(R') \cup \{U\}$ by (AB3). Together with (5.5), this implies that every $u \in U_\ell'$ satisfies $d_{G_\ell[U_\ell']}(u) \geq \gamma|N_{R'}(u) \cap U| - \kappa n - 1 \geq 48\gamma|U|/100$. Since $\ell \leq k \leq \kappa n$, we have

$$d_{G_\ell'}(u) \geq d_{G_\ell[U_i']}(u) - \kappa n \geq 47\gamma|U|/100. \quad (5.6)$$

By (5.4), we also have

$$|U_\ell'| \pm 1 = |U_\ell| = \gamma|U| \pm \kappa n \text{ and, in particular, } |U_\ell'| \leq 5\gamma|U|/4. \quad (5.7)$$

Combining (5.6) and (5.7), we have $d_{G_\ell'}(u) \geq 3|U_\ell'|/8$ for every $u \in U_\ell'$. So by (R2) and (5.7), we can apply Proposition 5.5 to G_ℓ', with $200\xi/\gamma$ as α, to obtain a perfect matching N'^{abs}_ℓ in G_ℓ', as desired.

Therefore we have pairwise edge-disjoint matchings N^{abs}_i in G_i, as claimed, which by construction are edge-disjoint from $N_1, \ldots, N_k$. For each $i \in [k]$, let $M_i := N_i \cup N^{\text{abs}}_i$ and let $\mathcal{M} = \{M_1, \ldots, M_k\}$. Now $M_i \supseteq N_i$ and $M_i \setminus N_i \subseteq R$ for each $i \in [k]$, and $\mathcal{M}$ has nearly perfect coverage of U with defects in S, as desired. Moreover, if $|U| < 3n/4$, then $\mathcal{M}$ has perfect coverage of U, as desired. ■

5.1.2. Finding absorbable matchings. Lemma 5.6 allows us to apply absorption for up to κn "pseudorandom" matchings at a time. We construct these collections of matchings in the following lemma using Theorem 2.14 and the strategy described in Section 2.3.

Lemma 5.7 (Nibble lemma). *Let $D \in [n^{1/2}, n]$, and let $\mathcal{H}' \subseteq \mathcal{H}$ be a spanning subhypergraph such that for every $w \in V(\mathcal{H}')$ we have $d_{\mathcal{H}'}(w) = (1 \pm \sqrt{\xi})D$. If $\mathcal{F}_V$ and $\mathcal{F}_E$ are families of subsets in $V(\mathcal{H}')$ and $E(\mathcal{H}')$, respectively, such that $|\mathcal{F}_V|, |\mathcal{F}_E| \leq n^{\log n}$, then there exist pairwise edge-disjoint matchings $N_1, \dots, N_D$ in $\mathcal{H}'$ such that*

(N1) *N_i is (γ, κ)-pseudorandom with respect to $\mathcal{F}_V$ for every $i \in [D]$, and*

(N2) *$|F \setminus \bigcup_{i=1}^{D} N_i| \leq \gamma|F| + \kappa \max(|F|, D)$ for each $F \in \mathcal{F}_E$.*

Proof (sketch). First we embed $\mathcal{H}'$ into a 3-uniform linear hypergraph $\mathcal{H}''$ with $O(n^4)$ vertices, in which every vertex has degree $(1 \pm \sqrt{\xi})D$. We then let $\mathcal{H}^* := \mathrm{inc}_D(\mathcal{H}'')$ be the D-wise incidence hypergraph of $\mathcal{H}''$ (recall Definition 2.15). By Observation 2.16 (a)–(c), $\mathcal{H}^*$ is 4-uniform, linear, and every vertex of $\mathcal{H}^*$ has degree $(1 \pm \sqrt{\xi})D$. Thus, we can apply Theorem 2.14 to $\mathcal{H}^*$ with $\delta = 1/4$ and an appropriately chosen $\mathcal{F}$, determined by $\mathcal{F}_V$ and $\mathcal{F}_E$, to obtain a matching M in $\mathcal{H}^*$ such that every $S \in \mathcal{F}$ satisfies $|S \setminus V(M)| \leq \kappa \max\{|S|, D\}/2$. Next, we "sparsify" M, by randomly and independently removing each edge with probability γ, to obtain a new matching N. For each $i \in [D]$, we let $N_i := \{e \in \mathcal{H} : \exists f \in N,\ f \supseteq \{i\} \times e\}$. By Observation 2.16 (d), the matchings $N_1, \dots, N_D$ are pairwise edge-disjoint, and the pseudorandomness property of M guaranteed by Theorem 2.14 ensures that (N1) and (N2) are satisfied. ■

We will use Lemmas 5.7 and 5.6 to construct $\lceil n/2 \rceil$ pairwise edge-disjoint matchings with nearly perfect coverage of U such that the remaining edges of $\mathcal{H} \setminus R$ comprise a subhypergraph of small maximum degree. We apply the following lemma to this subhypergraph to decompose it into matchings which are absorbable by "smallness."

Lemma 5.8 (Leftover colouring lemma). *If $\mathcal{H}' \subseteq \mathcal{H} \setminus R$ is a spanning subhypergraph such that $\Delta(\mathcal{H}') \leq \gamma n$, then there exist pairwise edge-disjoint matchings $N_1, \dots, N_k$ where $k \leq \lceil \gamma^{1/3} n \rceil$ such that*

(L1) *$|V(N_i)| \leq \sqrt{\gamma} n$ for every $i \in [k]$, and*

(L2) *$\mathcal{H}' = \bigcup_{i=1}^{k} N_i$.*

Proof. Let $D := \lceil \gamma n \rceil$. For every $e \in \mathcal{H}$, since $|e| \leq 3$, we have $\sum_{v \in e} d_{\mathcal{H}'}(v) \leq 3D$. Thus, $\chi'(\mathcal{H}) \leq 3D + 1$, so there exist pairwise edge-disjoint matchings $M_1, \dots, M_{3D+1}$ such that $\bigcup_{i=1}^{3D+1} M_i = \mathcal{H}'$. Let $\ell := \lceil \gamma^{-1/2} \rceil + 1$. For each $i \in [3D + 1]$, there exist pairwise edge-disjoint matchings $N_{i,1}, \dots, N_{i,\ell}$ such that $\bigcup_{j=1}^{\ell} N_{i,j} = M_i$ and $|V(N_{i,j})| \leq \sqrt{\gamma} n$ for each $j \in [\ell]$. By reindexing, $\bigcup_{i=1}^{3D+1} \{N_{i,1}, \dots, N_{i,\ell}\}$ is the desired set of matchings, since $(3D + 1)\ell \leq \gamma^{1/3} n$. ■

5.1.3. Proof of Theorem 4.3. By combining Lemmas 5.6, 5.7, and 5.8, we prove the following lemma, which effectively reduces Theorem 4.3 to Vizing's theorem.

Lemma 5.9 (Main colouring lemma). *There exists $\mathcal{H}' \subseteq \mathcal{H}$ and a proper edge-colouring of $\mathcal{H}'$ using $\lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil$ colours such that*

- *$\mathcal{H}' \supseteq \mathcal{H} \setminus R$ and*
- *the colour classes have nearly perfect coverage of U.*

Moreover, $\mathcal{H} \setminus \mathcal{H}'$ is a graph and satisfies $\Delta(\mathcal{H} \setminus \mathcal{H}') \leq n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$.

Proof. The proof proceeds in two steps.

Step 1. Using Lemmas 5.7 and 5.6, find a set $\mathcal{M}$ of $\lceil n/2 \rceil$ pairwise edge-disjoint matchings $M_1, \ldots, M_{\lceil n/2 \rceil}$ such that the following holds:

(M1) $\mathcal{M}$ has nearly perfect coverage of U, and moreover, if $|U| < 3n/4$, then $\mathcal{M}$ has perfect coverage of U,

(M2) $\Delta(\mathcal{H} \setminus (R \cup \bigcup_{i=1}^{\lceil n/2 \rceil} M_i)) \leq \gamma n$, and

(M3) $\Delta(R \cap \bigcup_{i=1}^{\lceil n/2 \rceil} M_i) \leq \gamma n$.

First we partition $\mathcal{H} \setminus R$ into $K := \lceil 1/\kappa \rceil$ pairwise edge-disjoint hypergraphs $\mathcal{H}_1, \ldots, \mathcal{H}_K$ such that $\bigcup_{i=1}^{K} \mathcal{H}_i = \mathcal{H} \setminus R$, and every vertex has degree $(1/2 \pm 3\xi)n/K$ in $\mathcal{H}_i$ for each $i \in [K]$. (To show that the desired partition exists, consider a partition chosen uniformly at random which, by (5.3), will satisfy the vertex degree condition with high probability.) We iteratively apply alternating applications of Lemmas 5.7 and 5.6 to each $\mathcal{H}_i$.

Now, for each $i \in [K]$, we choose n_i to be either $\lfloor \lceil n/2 \rceil / K \rfloor$ or $\lceil \lceil n/2 \rceil / K \rceil$ such that $\sum_{j=1}^{K} n_j = \lceil n/2 \rceil$, and we partition the set $[\lceil n/2 \rceil]$ into K disjoint parts $I_1, \ldots, I_K$ such that $|I_i| = n_i$. Note that $n_i \leq \kappa n$ and every vertex in $\mathcal{H}_i$ has degree $(1 \pm 7\xi)n_i$ for every $i \in [K]$.

For $j \in [K] \cup \{0\}$, let us define the following inductive properties, where $\mathcal{M}_k := \{M_c : c \in I_k\}$ is a set of matchings in $\mathcal{H}$ for each $k \in [j]$.

(M1$_j$) For every $k \in [j]$, $M_c \subseteq \mathcal{H}_k \cup R$ for every $c \in I_k$ and, moreover, the matchings in $\bigcup_{k=1}^{j} \mathcal{M}_k$ are pairwise edge-disjoint.

(M2$_j$) For every $w \in V(\mathcal{H})$,

$$\left| E_R(w) \cap \bigcup_{k \in [j]} \bigcup_{M \in \mathcal{M}_k} M \right| \leq (\gamma + 3\kappa) \sum_{k \in [j]} n_k,$$

$$\left| E_{\bigcup_{k=1}^{j} \mathcal{H}_k}(w) \setminus \bigcup_{k \in [j]} \bigcup_{M \in \mathcal{M}_k} M \right| \leq (\gamma + 3\kappa) \sum_{k \in [j]} n_k.$$

(M3$_j$) If $|U| < 3n/4$, then $\bigcup_{k=1}^{j} \mathcal{M}_k$ has perfect coverage of U. Otherwise, $\bigcup_{k=1}^{j} \mathcal{M}_k$ has nearly perfect coverage of U with defects in U.

Using induction on j, we will show that there exist sets of matchings $\mathcal{M}_1, \dots, \mathcal{M}_K$ satisfying (M1$_j$)–(M3$_j$) for $j = K$. Note that (M1$_j$)–(M3$_j$) trivially hold for $j = 0$. Let $i \in [K]$, and suppose that $\mathcal{M}_1, \dots, \mathcal{M}_{i-1}$ satisfy (M1$_j$)–(M3$_j$) for $j = i - 1$. Our goal is to find a collection $\mathcal{M}_i$ of n_i pairwise edge-disjoint matchings in $\mathcal{H}$ satisfying (M1$_j$)–(M3$_j$) for $j = i$.

Let $R_i := R \setminus \bigcup_{k=1}^{i-1} \bigcup_{M \in \mathcal{M}_k} M$, let $S_i := U \setminus \bigcup_{k=1}^{i-1} \bigcup_{M \in \mathcal{M}_k} (U \setminus V(M))$, and let $\mathcal{W} := \mathcal{F}(R_i) \cup \{U, S_i\}$, where $\mathcal{F}(R_i) := \{N_{R_i}(u) \cap U : u \in U\} \cup \{N_{R_i}(u) \setminus U : u \in U\}$.

Now we apply Lemma 5.7 with $\mathcal{H}_i$, $\mathcal{W}$, $\{E_{\mathcal{H}_i}(w) : w \in V(\mathcal{H})\}$, and n_i playing the roles of $\mathcal{H}'$, $\mathcal{F}_V$, $\mathcal{F}_E$, and D, respectively to obtain a set $\mathcal{N}_i := \{N_c : c \in I_i\}$ of n_i pairwise edge-disjoint matchings in $\mathcal{H}_i$ such that the following hold.

(N'1) For every $c \in I_i$, N_c is (γ, κ)-pseudorandom with respect to $\mathcal{W}$.

(N'2) For every $w \in V(\mathcal{H})$, $d_{\mathcal{H}_i \setminus \bigcup_{c \in I_i} N_c}(w) \leq (\gamma + 2\kappa)n_i$.

Now we show that for every $c \in I_i$, (N_c, R_i, S_i) is absorbable by pseudorandomness of N_c, as follows.

- By (M3$_j$) for $j = i - 1$, if $|U| < 3n/4$, then $S_i = U$, and otherwise $|S_i| \geq |U| - \sum_{k=1}^{i-1} n_k \geq n/4 - 1$, so (AB1) holds.
- By (M2$_j$) for $j = i - 1$, since $(\gamma + 3\kappa)\lceil n/2 \rceil \leq \gamma n$, (AB2) holds.
- By (N'1), N_c is (γ, κ)-pseudorandom with respect to $\mathcal{W}$ so (AB3) holds, as required.

Therefore we can apply Lemma 5.6 to obtain a set $\mathcal{M}_i := \{M_c : c \in I_i\}$ of n_i pairwise edge-disjoint matchings in $\mathcal{H}$ such that the following hold.

- For every $c \in I_i$, $M_c \supseteq N_c$, and $M_c \setminus N_c \subseteq R_i$, and consequently (M1$_j$) holds for $j = i$.
- By (N'2), for every $w \in V(\mathcal{H})$, $|E_{\mathcal{H}_i}(w) \setminus \bigcup_{c \in I_i} M_c| \leq (\gamma + 2\kappa)n_i$. Moreover, again by (N'2), all but at most $n_i - (d_{\mathcal{H}_i}(w) - d_{\mathcal{H}_i \setminus \bigcup_{c \in I_i} N_c}(w)) \leq (\gamma + 3\kappa)n_i$ of the matchings in $\mathcal{N}_i$ cover w, so $|E_{R_i}(w) \cap \bigcup_{c \in I_i} M_c| \leq (\gamma + 3\kappa)n_i$. Hence, (M2$_j$) holds for $j = i$.
- If $|U| < 3n/4$, then $\mathcal{M}_i$ has perfect coverage of U, and, otherwise, $\mathcal{M}_i$ has nearly perfect coverage of U with defects in $S_i \subseteq U$. Hence, (M3$_j$) holds for $j = i$.

Therefore, by induction, there exist sets of matchings $\mathcal{M}_1, \dots, \mathcal{M}_K$ such that for every $i \in [K]$, $\mathcal{M}_i$ satisfies (M1$_j$)–(M3$_j$) for $j = i$, as claimed. Now $\mathcal{M} := \bigcup_{i=1}^{K} \mathcal{M}_i$ satisfies (M1)–(M3). Indeed, by (M1$_j$) and (M3$_j$) for $j = K$, $\mathcal{M}$ satisfies (M1), and by (M2$_j$) for $j = K$, $\mathcal{M}$ satisfies (M2) and (M3), since $(\gamma + 3\kappa)\lceil n/2 \rceil \leq \gamma n$.

Step 2. Using Lemmas 5.8 and 5.6, find a set $\mathcal{M}'$ of $\lceil \gamma^{1/3} n \rceil$ pairwise edge-disjoint matchings $M'_1, \ldots, M'_{\lceil \gamma^{1/3} n \rceil}$ such that the following holds:

(M'1) $\bigcup_{M \in \mathcal{M}} M \cap \bigcup_{M' \in \mathcal{M}'} M' = \varnothing$,

(M'2) $\mathcal{M} \cup \mathcal{M}'$ has nearly perfect coverage of U, and

(M'3) $\mathcal{H} \setminus R \subseteq \bigcup_{M \in \mathcal{M}} M \cup \bigcup_{M' \in \mathcal{M}'} M'$.

By (M2) and Lemma 5.8 applied with $\mathcal{H} \setminus (R \cup \bigcup_{M \in \mathcal{M}} M)$ playing the role of $\mathcal{H}'$, there exists a set $\mathcal{N}' := \{N'_1, \ldots, N'_{\lceil \gamma^{1/3} n \rceil}\}$ of pairwise edge-disjoint matchings such that

(L'1) $|V(N'_i)| \leq \sqrt{\gamma} n$ for every $i \in [\lceil \gamma^{1/3} n \rceil]$ and

(L'2) $\mathcal{H} \setminus (R \cup \bigcup_{M \in \mathcal{M}} M) = \bigcup_{i=1}^{\lceil \gamma^{1/3} n \rceil} N'_i$.

Now we show that for every $i \in \lceil \gamma^{1/3} n \rceil$, (N'_i, R', S') is absorbable by smallness of N'_i, where $R' := R \setminus \bigcup_{M \in \mathcal{M}} M$ and $S' := U \setminus \bigcup_{M \in \mathcal{M}} (U \setminus V(M))$. Indeed,

- by (M1), if $|U| < 3n/4$, then $S' = U$, and otherwise $|S'| \geq |U| - \lceil n/2 \rceil \geq n/4 - 1$, so (AB1) holds,
- by (M3), $\Delta(R \setminus R') = \Delta(R \cap \bigcup_{M \in \mathcal{M}} M) \leq \gamma n$, so (AB2) holds, and
- by (L'1), (AB3) holds.

Therefore we can apply Lemma 5.6 to obtain a set $\mathcal{M}' := \{M'_1, \ldots, M'_{\lceil \gamma^{1/3} n \rceil}\}$ of pairwise edge-disjoint matchings in $\mathcal{H}$ such that the following hold:

- for every $i \in [\lceil \gamma^{1/3} n \rceil]$, $M'_i \supseteq N'_i$ and $M'_i \setminus N'_i \subseteq R'$, and
- $\mathcal{M}'$ has nearly perfect coverage of U with defects in S'

Therefore, by the choice of R', $\mathcal{M}'$ satisfies (M'1), by the choice of S', $\mathcal{M}'$ satisfies (M'2), and by (L'2), $\mathcal{M}'$ satisfies (M'3), as desired.

Now let $\mathcal{H}' := \bigcup_{M \in \mathcal{M}} M \cup \bigcup_{M' \in \mathcal{M}'} M'$, assign colour c to each edge in M_c for every $c \in [\lceil n/2 \rceil]$, and assign colour $c = \lceil n/2 \rceil + i$ to each edge in M'_i for every $i \in [\lceil \gamma^{1/3} n \rceil]$. By (M'1), we have a proper edge-colouring of $\mathcal{H}'$ using at most $\lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil$ colours, as required. By (M'3), $\mathcal{H}' \supseteq \mathcal{H} \setminus R$, as desired, and by (M'2), the colour classes $\mathcal{M} \cup \mathcal{M}'$ of $\mathcal{H}'$ have nearly perfect coverage of U, as desired. Since $\mathcal{H}' \supseteq \mathcal{H} \setminus R$, it follows that $\mathcal{H} \setminus \mathcal{H}' \subseteq R$ is a graph. Since $\mathcal{M} \cup \mathcal{M}'$ has nearly perfect coverage of U, every vertex $w \in U$ satisfies $d_{\mathcal{H} \setminus \mathcal{H}'}(w) \leq (n-1) - (\lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil - 1) = n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$, and by (R1), every vertex $w \in V(\mathcal{H}) \setminus U$ satisfies $d_{\mathcal{H} \setminus \mathcal{H}'}(w) \leq (1 - \varepsilon) n/2 + 2\xi n \leq n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$. Hence, $\Delta(\mathcal{H} \setminus \mathcal{H}') \leq n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$, as desired. ■

Now we can immediately deduce Theorem 4.3.

Proof of Theorem 4.3. By Lemma 5.9, there exists $\mathcal{H}' \subseteq \mathcal{H}$ such that $\chi'(\mathcal{H}') \leq \lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil$ and $\mathcal{H} \setminus \mathcal{H}'$ is a graph with $\Delta(\mathcal{H} \setminus \mathcal{H}') \leq n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$.

By Vizing's theorem, $\chi'(\mathcal{H} \setminus \mathcal{H}') \leq \Delta(\mathcal{H} \setminus \mathcal{H}') + 1 \leq n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil + 1$. Therefore

$$\chi'(\mathcal{H}) \leq \chi'(\mathcal{H}') + \chi'(\mathcal{H} \setminus \mathcal{H}') \leq n + 1,$$

as desired. ■

We conclude this subsection by briefly discussing how Theorem 4.3 can be improved to show that $\chi'(\mathcal{H}) \leq n$. First, we note that the same argument combined with Vizing's theorem proves $\chi'(\mathcal{H}) \leq n$ if at least one of the following holds:

(a) the colour classes of $\mathcal{H}'$ in Lemma 5.9 have perfect coverage of U or

(b) every $v \in U$ which is a "defect vertex" of some colour class of $\mathcal{H}'$ in Lemma 5.9 satisfies $d_G(v) < n - 1$.

Indeed, in either case, $\Delta(\mathcal{H} \setminus \mathcal{H}') \leq n - k - 1$ for $k = \lceil n/2 \rceil + \lceil \gamma^{1/3} n \rceil$, and since $\chi'(\mathcal{H}') \leq k$, we have $\chi'(\mathcal{H}) \leq \chi'(\mathcal{H}') + \chi'(\mathcal{H} \setminus \mathcal{H}') \leq k + (n - k - 1 + 1) = n$, as desired.

Recall that Lemma 5.6 actually guarantees (a) if $|U| < 3n/4$. In fact, this argument even works as long as $|U| \leq (1 - 10\varepsilon)n$. Moreover, the proof of Lemma 5.9 also guarantees (b) if $|\{v \in U : d_G(v) < n - 1\}| \geq (1/2 + 2\gamma^{1/3})n$. In particular, with only minor modifications to the proof of Lemma 5.9, we can prove either (a) or (b) unless U consists of nearly all of the vertices of $\mathcal{H}$ and nearly half of the vertices of $\mathcal{H}$ have degree $n - 1$. Note that this means that $\mathcal{H}$ resembles the complete graph, one of the extremal examples for the EFL conjecture. In this case, additional ideas are needed to prove that $\chi'(\mathcal{H} \setminus \mathcal{H}') = \Delta(\mathcal{H} \setminus \mathcal{H}')$, which then ensures that $\chi'(\mathcal{H}) \leq \chi'(\mathcal{H}') + \chi'(\mathcal{H} \setminus \mathcal{H}') \leq k + (n - k) = n$, as desired. To obtain this improved bound on $\chi'(\mathcal{H} \setminus \mathcal{H}')$, we modify the above approach to ensure that $\mathcal{H} \setminus \mathcal{H}'$ is quasirandom and then apply an edge-colouring result of Glock, Kühn, and Osthus [68]. (This edge-colouring result in turn is deduced from the theorem of Kühn and Osthus [114] that dense even-regular robustly expanding graphs have a Hamilton decomposition. This deduction is based on the fact that a Δ-regular graph of even order with a Hamilton decomposition has chromatic index Δ.)

5.2. Proving the EFL conjecture when all edges are large

This subsection is devoted to the proof of Theorem 4.4, which we restate here.

Theorem 4.4. *For every $\delta > 0$, there exist $n_0, r, \sigma > 0$ such that the following holds. If $\mathcal{H}$ is an n-vertex linear hypergraph where $n > n_0$ and every $e \in \mathcal{H}$ satisfies $|e| > r$, then $\chi'(\mathcal{H}) \leq n$. Moreover, if $\chi'(\mathcal{H}) > (1 - \sigma)n$, then $|\{e \in \mathcal{H} : |e| = (1 \pm \delta)\sqrt{n}\}| \geq (1 - \delta)n$.*

If $\preceq$ is a linear ordering of the edges of a hypergraph $\mathcal{H}$, for each $e \in \mathcal{H}$, we define $N^{\preceq}_{\mathcal{H}}(e) := \{f \in N_{\mathcal{H}}(e) : f \preceq e\}$ and $d^{\preceq}_{\mathcal{H}}(e) := |N^{\preceq}_{\mathcal{H}}(e)|$. We omit the subscript $\mathcal{H}$ when it is clear from the context. For each $e \in \mathcal{H}$, we also let $\mathcal{H}^{\preceq e} := \{f \in \mathcal{H} : f \preceq e\}$.

For every n-vertex hypergraph $\mathcal{H}$ and $W \subseteq \mathcal{H}$, the *normalized volume* of W is $\mathrm{vol}_{\mathcal{H}}(W) := \sum_{e\in W}\binom{|e|}{2}/\binom{n}{2}$. If $\mathcal{H}$ is linear, then $\mathrm{vol}_{\mathcal{H}}(W) \in [0,1]$ for every $W \subseteq \mathcal{H}$.

As in (4.4), if $\mathcal{H}$ is an n-vertex linear hypergraph such that $|e| > r$ for every $e \in \mathcal{H}$, then every $e \in \mathcal{H}$ satisfies $d^{\preceq}_{\mathcal{H}}(e) \leq (1 + 2/r)n$ if $\preceq$ is *size-monotone-decreasing* (i.e., satisfying $e \preceq f$ if $|e| > |f|$). In the following key lemma, we showed that we can either obtain an improved bound on $d^{\preceq}_{\mathcal{H}}(e)$ (by modifying the ordering if necessary) or find a highly structured set $W \subseteq \mathcal{H}$. In particular, the edges of W have similar size, and W has large volume. The fact that the edges have similar size will allow us to colour W efficiently via Theorem 3.5, unless W closely resembles a projective plane.

Lemma 5.10 (Reordering lemma [96]). *Let $0 < 1/r \ll \tau, 1/K$ where $\tau < 1$, $K \geq 1$, and $1 - \tau - 7\tau^{1/4}/K > 0$. If $\mathcal{H}$ is an n-vertex linear hypergraph where every $e \in \mathcal{H}$ satisfies $|e| \geq r$, then there exists a linear ordering $\preceq$ of the edges of $\mathcal{H}$ such that at least one of the following holds.*

(5.10:a) *Every $e \in \mathcal{H}$ satisfies $d^{\preceq}(e) \leq (1 - \tau)n$.*

(5.10:b) *There is a set $W \subseteq \mathcal{H}$ such that*

(W1) $\max_{e\in W} |e| \leq (1 + 3\tau^{1/4}K^4) \min_{e\in W} |e|$ *and*

(W2) $\mathrm{vol}_{\mathcal{H}}(W) \geq \frac{(1-\tau-7\tau^{1/4}/K)^2}{1+3\tau^{1/4}K^4}$.

Moreover, if e^* is the last edge of W, then

(O1) *for all $f \in \mathcal{H}$ such that $e^* \preceq f$ and $f \neq e^*$, we have $d^{\preceq}(f) \leq (1 - \tau)n$ and*

(O2) *for all $e, f \in \mathcal{H}$ such that $f \preceq e \preceq e^*$, we have $|f| \geq |e|$.*

We do not provide a proof of Lemma 5.10, but we briefly sketch the idea. Beginning with a size-monotone-decreasing ordering $\preceq$, we “reorder” $\preceq$ as follows. Let e^* be the last edge of $\mathcal{H}$ that does not satisfy (5.10:a). If there exists $f \in N^{\preceq}(e^*)$ such that $|N_{\mathcal{H}}(f) \cap \mathcal{H}^{\preceq e^*}| \leq (1 - \tau)n - 1$, then let $\preceq'$ be the ordering obtained from $\preceq$ by moving f to be the successor of e^*. If $\preceq$ satisfies (O1) and (O2), then $\preceq'$ does as well, and, moreover, $|\mathcal{H}^{\preceq' e^*}| < |\mathcal{H}^{\preceq e^*}|$. Thus, by iterating this argument, we may assume that there is no such $f \in N^{\preceq}(e^*)$. Moreover, since we started with a size-monotone-decreasing ordering, we may also assume that e^* satisfies (O1) and (O2). Now a double-counting argument shows that $W := \{f \in \mathcal{H}^{\preceq e^*} : |f| \leq (1 + 3\tau^{1/4}K)|e^*|\}$ satisfies (W2).

By applying Lemma 5.10 twice, we obtain the following.

Lemma 5.11. *Let $0 < 1/r \ll \sigma \ll 1$. Let $\mathcal{H}$ be an n-vertex linear hypergraph such that every $e \in \mathcal{H}$ satisfies $|e| > r$. If $\chi'(\mathcal{H}) > (1 - \sigma)n$, then there exists a partition of $\mathcal{H}$ into three spanning subhypergraphs, $\mathcal{H}_1$, W, and $\mathcal{H}_2$ such that*

(P1) $\max_{e\in W} |e| \leq (1 + 4\sigma^{1/4}) \min_{e\in W} |e|$, (P2) $\mathrm{vol}_{\mathcal{H}}(W) \geq 1 - 4\sigma^{1/5}$, *and*

(P3) $|e| \geq \max_{f \in W} |f|$ *for all* $e \in \mathcal{H}_2$,

and a linear ordering $\preceq$ *of the edges of* $\mathcal{H}$ *such that*

(FD1) *every* $e \in \mathcal{H}_1$ *satisfies* $d^{\preceq}_{\mathcal{H}}(e) \leq (1-2\sigma)n$ *and* $f \preceq e$ *for every* $f \in \mathcal{H}_2 \cup W$, *and*

(FD2) $d^{\preceq}_{\mathcal{H}}(e) \leq n/2000$ *for all* $e \in \mathcal{H}_2$.

Proof. We apply Lemma 5.10 twice and combine the resulting orderings to obtain the desired ordering $\preceq$ of $\mathcal{H}$. First, we apply Lemma 5.10 to $\mathcal{H}$ with 2σ and 1 playing the roles of τ and K, respectively, to obtain an ordering $\preceq_1$. If $\preceq_1$ satisfies (5.10:a), then $\chi'(\mathcal{H}) < (1-\sigma)n$, and so we assume that (5.10:b) holds. Let W be the set W obtained from (5.10:b), let e^* be the last edge of W in $\preceq_1$, and let $\mathcal{H}_1 := \mathcal{H} \setminus \mathcal{H}^{\preceq_1 e^*}$. Let f^* be the edge of W which comes first in $\preceq_1$, and let $\mathcal{H}_2 := \mathcal{H} \setminus \{e \in \mathcal{H} : f^* \preceq_1 e\}$. By the choices of τ and K, and since $\sigma \ll 1$, we have $\max_{e \in W} |e| \leq (1 + 4\sigma^{1/4})|e^*|$ and $\mathrm{vol}_{\mathcal{H}}(W) \geq (1 - \sigma^{1/5})^3 \geq 1 - 4\sigma^{1/5}$, and so W satisfies (P1) and (P2), as desired. Also by (O2) of (5.10:b), we may assume without loss of generality that every $e \in \mathcal{H}$ satisfying $f^* \preceq_1 e \preceq_1 e^*$ is in W, and so $\mathcal{H}$ is partitioned into $\mathcal{H}_1$, W, and $\mathcal{H}_2$, as required, and $\mathcal{H}_2$ satisfies (P3), as desired.

Now we reapply Lemma 5.10 to $\mathcal{H}_2$ and show that the resulting ordering satisfies (FD1) and (FD2), as follows. Apply Lemma 5.10 with $\mathcal{H}_2$, $1 - 1/2000$, and 2000^2 playing the roles of $\mathcal{H}$, τ, and K, respectively, to obtain an ordering $\preceq_2$. Since $W \cap \mathcal{H}_2 = \varnothing$, we have $\mathrm{vol}_{\mathcal{H}}(W) + \mathrm{vol}_{\mathcal{H}}(\mathcal{H}_2) \leq 1$. Thus, $\preceq_2$ satisfies (5.10:a), because (5.10:b) would imply that there is a set $W' \subseteq \mathcal{H}_2$ disjoint from W with $\mathrm{vol}_{\mathcal{H}}(W') > 4\sigma^{1/5}$, contradicting (P2). Combine $\preceq_1$ and $\preceq_2$ to obtain an ordering $\preceq$ of $\mathcal{H}$ where

- if $f \in \mathcal{H}_1 \cup W$, then $e \preceq f$ for every $e \in \mathcal{H}^{\preceq_1 f}$, and
- if $f \in \mathcal{H}_2$, then $e \preceq f$ for every $e \in \mathcal{H}_2^{\preceq_2 f}$.

Since $\mathcal{H}_1$ and $\preceq_1$ satisfy (O1) of (5.10:b) with $\tau = 2\sigma$, (FD1) holds, and since $\mathcal{H}_2$ and $\preceq_2$ satisfy (5.10:a) with $\tau = 1 - 1/2000$, (FD2) holds, as desired. ■

To prove Theorem 4.4, we apply Lemma 5.11 and consider two cases depending on the size of the edges in W. In either case, by (FD1), it suffices to show that $\chi'(\mathcal{H}_2 \cup W) \leq n$. When the edges in W have size close to or larger than $\sqrt{n}$, we apply the following lemma to colour $\mathcal{H}_2 \cup W$. As its proof covers the case when $\mathcal{H}$ is close to a projective plane, the argument is quite delicate and we refer the reader to [96, Lemma 5.1] for a proof.

Lemma 5.12. *Let* $0 < 1/n_0 \ll \delta \ll 1$, *and let* $n \geq n_0$. *If* $\mathcal{H}$ *is an* n*-vertex linear hypergraph where every* $e \in \mathcal{H}$ *satisfies* $|e| \geq (1-\delta)\sqrt{n}$, *then* $\chi'(\mathcal{H}) \leq n$. ■

When the edges in W have size bounded away from $\sqrt{n}$, we apply the following lemma [96, Corollary 6.5], which we prove using Theorem 3.5, to colour W.

Lemma 5.13. *Let* $0 < 1/n_0, 1/r \ll \alpha \ll \zeta < 1$, *let* $n \geq n_0$, *and suppose that* $r \leq (1-\zeta)\sqrt{n}$. *If* $\mathcal{H}$ *is an* n*-vertex linear hypergraph such that every* $e \in \mathcal{H}$ *satisfies* $|e| \in [r, (1+\alpha)r]$, *then* $\chi'(\mathcal{H}) \leq (1-\zeta/500)n$.

Proof. Let $\Delta := (1+\alpha)r(n-r)/(r-1)$, and let $L := L(\mathcal{H})$. For every edge $e \in \mathcal{H}$, there are at most $(1+\alpha)r(n-r)$ pairs of vertices $\{u, v\}$ of $\mathcal{H}$ where $u \notin e$ and $v \in e$. Thus, since $\mathcal{H}$ is linear and every edge has size at least r, we have $\Delta(L) \leq \Delta$. Similarly, if $e, f \in \mathcal{H}$ share a vertex, then

$$\left|N_L(e) \cap N_L(f)\right| \leq n/(r-1) + (1+\alpha)^2 r^2 \leq (1-5\zeta/6)n.$$

Thus, every $v \in V(L)$ satisfies $e(L[N(v)]) \leq \Delta(1-5\zeta/6)n/2 \leq (1-5\zeta/6)\binom{\Delta}{2}$. Therefore, by Theorem 3.5, $\chi'(\mathcal{H}) = \chi(L) \leq (1-5\zeta/(6e^6))\Delta \leq (1-\zeta/500)n$, as desired. ■

Now we can combine Lemmas 5.11–5.13 to prove Theorem 4.4.

Proof of Theorem 4.4. We may assume without loss of generality that $\delta \ll 1$, and we let $0 < 1/n_0 \ll 1/r \ll \sigma \ll \delta$. We assume that $\chi'(\mathcal{H}) > (1-\sigma)n$ or else there is nothing to prove.

Apply Lemma 5.11 to obtain a partition of $\mathcal{H}$ into $\mathcal{H}_1$, W, and $\mathcal{H}_2$ satisfying (P1)–(P3) and an ordering $\preceq$ of the edges of $\mathcal{H}$ satisfying (FD1) and (FD2), and let $r' := \min_{e \in W} |e|$. We assume that

$$r' \leq \sqrt{n/(1-4\sigma)}, \tag{5.8}$$

as otherwise the fact that $\mathrm{vol}_{\mathcal{H}}(\mathcal{H}_2 \cup W) \leq 1$ and (P3) together would imply $e(\mathcal{H}_2 \cup W) \leq (1-2\sigma)n$. Together with (FD1), this fact would imply that every $e \in \mathcal{H}$ satisfies $d^{\preceq}_{\mathcal{H}}(e) \leq (1-2\sigma)n$, in which case $\chi'(\mathcal{H}) \leq (1-\sigma)n$, a contradiction.

We now consider two cases: $r' < (1-\delta)\sqrt{n}$ and $r' \geq (1-\delta)\sqrt{n}$. In the former case, we derive a contradiction by showing $\chi'(\mathcal{H}) \leq (1-\sigma)n$, and in the latter case, we prove that $\chi'(\mathcal{H}) \leq n$ and $|\{e \in \mathcal{H} : |e| = (1 \pm \delta)\sqrt{n}\}| \geq (1-\delta)n$.

Case 1. $r' < (1-\delta)\sqrt{n}$.

Let $\zeta := 1 - r'/\sqrt{n}$. Since $r' < (1-\delta)\sqrt{n}$, we have $\zeta > \delta$. By (P1) and Lemma 5.13 with r' and $4\sigma^{1/4}$ playing the roles of r and α, respectively, we have $\chi'(W) \leq (1-\zeta/500)n$.

Now we claim that $\chi'(\mathcal{H}_2) \leq \zeta n/1000$. To that end, let $k := e(\mathcal{H}_2)$. If $k \leq \zeta n/1000$, then we can simply assign each edge of $\mathcal{H}_2$ a distinct colour and the claim holds; so we assume that $k > \zeta n/1000$. Since $\zeta > \delta$, we have $k > \zeta n/1000 > 2\delta^2 n$. By

(P3), every edge of $\mathcal{H}_2$ has size at least r', so we have $\mathrm{vol}_{\mathcal{H}}(\mathcal{H}_2) \geq k(r'-1)^2/n^2$. On the other hand, by (P2), and since $\mathcal{H}_2 \cap W = \varnothing$, we have $\mathrm{vol}_{\mathcal{H}}(\mathcal{H}_2) \leq 4\sigma^{1/5} \leq \delta^3$. Thus, $2\delta^2 n < k \leq \delta^3 n^2/(r'-1)^2$, so $r' < \delta^{1/4}\sqrt{n}$. Therefore $\zeta > 1000/1001$. Now by (FD2), we can properly colour $\mathcal{H}_2$ greedily in the ordering provided by $\preceq$ using at most $n/2000 + 1 \leq \zeta n/1000$ colours, as claimed.

Since $\chi'(\mathcal{H}_2) \leq \zeta n/1000$ and $\chi'(W) \leq (1-\zeta/500)n$, there is a proper edge-colouring of $\mathcal{H}_2 \cup W$ using at most $(1-\zeta/1000)n \leq (1-\sigma)n$ colours, and by (FD1), we can extend such a colouring to $\mathcal{H}_1$ greedily without using any additional colours, contradicting that $\chi'(\mathcal{H}) > (1-\sigma)n$.

Case 2. $r' \geq (1-\delta)\sqrt{n}$.

By (P3) and Lemma 5.12, there is a proper edge-colouring of $\mathcal{H}_2 \cup W$ using at most n colours, and as before, by (FD1), we can extend such a colouring to $\mathcal{H}_1$ greedily without using any additional colours. Hence, $\chi'(\mathcal{H}) \leq n$, as desired.

Since $r' \geq (1-\delta)\sqrt{n}$, by (P1) and (5.8), the edges in W have size $(1 \pm \delta)\sqrt{n}$. In fact, the edges in W have size at most $(1+\delta^2)\sqrt{n}$, so by (P2), since $\mathrm{vol}_{\mathcal{H}}(W) \geq 1-\delta^2$, we have $e(W) \geq \mathrm{vol}_{\mathcal{H}}(W)(n-1)/(1+\delta^2)^2 \geq (1-\delta)n$, as desired. ■

We conclude by briefly discussing how to combine the arguments of Theorems 4.3 and 4.4 to obtain Theorem 4.1. First, we merge the hierarchy used in the proof of Theorem 4.4 with (5.1) and also introduce constants r_1, r_0, β, and ρ into the hierarchy, letting

$$1/n_0 \ll 1/r_0 \ll \xi \ll 1/r_1 \ll \beta \ll \kappa \ll \gamma \ll \varepsilon \ll \rho \ll \sigma \ll \delta \ll 1.$$

As in the proof of Theorem 4.7, we decompose $\mathcal{H}$ into three spanning subhypergraphs $\mathcal{H}_{\mathrm{sml}} := \{e \in \mathcal{H} : |e| \leq r_1\}$, $\mathcal{H}_{\mathrm{med}} := \{e \in \mathcal{H} : r_1 < |e| \leq r_0\}$, and $\mathcal{H}_{\mathrm{lrg}} := \{e \in \mathcal{H} : |e| > r_0\}$. We apply a stronger version of Theorem 4.4 to $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$ in which

(a) every colour class either covers at most βn vertices or consists of a single edge, and

(b) at most γn colours are used to colour $\mathcal{H}_{\mathrm{med}}$.

This strengthening of Theorem 4.4 enables us to modify the proof of Lemma 5.9 to find a colouring of some $\mathcal{H}'$ satisfying $\mathcal{H}_{\mathrm{sml}} \setminus R \subseteq \mathcal{H}' \subseteq \mathcal{H}_{\mathrm{sml}}$ compatible with the colouring of $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$. As in the proof of Theorem 4.3, we can ensure that $\mathcal{H} \setminus \mathcal{H}'$ is a graph of maximum degree at most $n - \lceil n/2 \rceil - \lceil \gamma^{1/3} n \rceil$; however, Vizing's theorem does not guarantee a colouring of $\mathcal{H} \setminus \mathcal{H}'$ that avoids conflicts with $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$. To that end, we modify Lemma 5.9 further to colour $\mathcal{H}'$ with $k := \lceil (1-\rho)n \rceil$ colours. If $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$ can be coloured with at most $(1-\sigma)n$ colours, then we can colour $\mathcal{H} \setminus \mathcal{H}'$ with $n-k$ colours that are not used on $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$ (either using Vizing's theorem or the more involved argument discussed at the end of Section 5.1). If $\mathcal{H}_{\mathrm{lrg}} \cup \mathcal{H}_{\mathrm{med}}$ requires more than $(1-\sigma)n$ colours, then we need a different approach, using the fact that in this case we know that $|\{e \in \mathscr{H} : |e| = (1 \pm \delta)\sqrt{n}\}| \geq (1-\delta)n$; i.e., that $\mathscr{H}$ is close to a projective plane. See [96] for the full proof.

Acknowledgments. We are grateful to Jeff Kahn for helpful comments on a preliminary version of the survey and to Dhruv Mubayi for informing us that the conjecture of Frieze and Mubayi [60] listed in a previous version of the survey was disproved in [37]. We also thank an anonymous reviewer for their helpful comments.

Funding. This project has received partial funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 786198, D. Kühn and D. Osthus). The research leading to these results was also partially supported by the EPSRC, grant nos. EP/N019504/1 (D. Kang, T. Kelly, and D. Kühn) and EP/S00100X/1 (A. Methuku and D. Osthus).

References

[1] D. Achlioptas, F. Iliopoulos, and A. Sinclair, Beyond the Lovász local lemma: point to set correlations and their algorithmic applications. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*, pp. 725–744, IEEE Comput. Soc. Press, Los Alamitos, CA, 2019 MR 4228196

[2] M. Ajtai, P. Erdős, J. Komlós, and E. Szemerédi, On Turán's theorem for sparse graphs. *Combinatorica* **1** (1981), no. 4, 313–317 Zbl 0491.05038 MR 647980

[3] M. Ajtai, J. Komlós, J. Pintz, J. Spencer, and E. Szemerédi, Extremal uncrowded hypergraphs. *J. Combin. Theory Ser. A* **32** (1982), no. 3, 321–335 Zbl 0485.05049 MR 657047

[4] M. Ajtai, J. Komlós, and E. Szemerédi, A note on Ramsey numbers. *J. Combin. Theory Ser. A* **29** (1980), no. 3, 354–360 Zbl 0455.05045 MR 600598

[5] M. Ajtai, J. Komlós, and E. Szemerédi, A dense infinite Sidon sequence. *European J. Combin.* **2** (1981), no. 1, 1–11 Zbl 0474.10038 MR 611925

[6] N. Alon, Independence numbers of locally sparse graphs and a Ramsey type problem. *Random Structures Algorithms* **9** (1996), no. 3, 271–278 Zbl 0876.05049 MR 1606837

[7] N. Alon, Degrees and choice numbers. *Random Structures Algorithms* **16** (2000), no. 4, 364–368 Zbl 0958.05049 MR 1761581

[8] N. Alon, S. Cambie, and R. J. Kang, Asymmetric list sizes in Bipartite graphs. *Ann. Comb.* **25** (2021), no. 4, 913–933 Zbl 07449490 MR 4346745

[9] N. Alon and J. H. Kim, On the degree, size, and chromatic index of a uniform hypergraph. *J. Combin. Theory Ser. A* **77** (1997), no. 1, 165–170 Zbl 0868.05037 MR 1426745

[10] N. Alon, J.-H. Kim, and J. Spencer, Nearly perfect matchings in regular simple hypergraphs. *Israel J. Math.* **100** (1997), 171–187 Zbl 0882.05107 MR 1469109

[11] N. Alon and M. Krivelevich, The choice number of random bipartite graphs. *Ann. Comb.* **2** (1998), no. 4, 291–297 Zbl 0927.05028 MR 1774970

[12] N. Alon, M. Krivelevich, and B. Sudakov, Coloring graphs with sparse neighborhoods. *J. Combin. Theory Ser. B* **77** (1999), no. 1, 73–82 Zbl 1026.05043 MR 1710532

[13] N. Alon and J. H. Spencer, *The Probabilistic Method*. 4th edn., Wiley Ser. Discrete Math. Optim., John Wiley & Sons, Hoboken, NJ, 2016 Zbl 1333.05001 MR 3524748

[14] N. Alon and R. Yuster, On a hypergraph matching problem. *Graphs Combin.* **21** (2005), no. 4, 377–384 Zbl 1090.05051 MR 2209008

[15] P. Bennett and T. Bohman, A natural barrier in random greedy hypergraph matching. *Combin. Probab. Comput.* **28** (2019), no. 6, 816–825 Zbl 1436.05079 MR 4015657

[16] C. Berge, On the chromatic index of a linear hypergraph and the Chvátal conjecture. In *Combinatorial Mathematics: Proceedings of the Third International Conference (New York, 1985)*, pp. 40–44, Ann. New York Acad. Sci. 555, New York Acad. Sci., New York, 1989 Zbl 0726.05055 MR 1018607

[17] A. Bernshteyn, The Johansson–Molloy theorem for DP-coloring. *Random Structures Algorithms* **54** (2019), no. 4, 653–664 Zbl 1417.05059 MR 3957361

[18] N. Besharati, L. Goddyn, E. S. Mahmoodian, and M. Mortezaeefar, On the chromatic number of Latin square graphs. *Discrete Math.* **339** (2016), no. 11, 2613–2619 Zbl 1339.05036 MR 3518411

[19] T. Bohman and P. Keevash, Dynamic concentration of the triangle-free process. *Random Structures Algorithms* **58** (2021), no. 2, 221–293 MR 4201797

[20] B. Bollobás, The independence ratio of regular graphs. *Proc. Amer. Math. Soc.* **83** (1981), no. 2, 433–436 Zbl 0474.05057 MR 624948

[21] M. Bonamy, M. Delcourt, R. Lang, and L. Postle, Edge-colouring graphs with local list sizes. 2020, arXiv:2007.14944

[22] M. Bonamy, T. Kelly, P. Nelson, and L. Postle, Bounding χ by a fraction of Δ for graphs without large cliques. J. Combin. Theory Ser. B, to appear

[23] M. Bonamy, T. Perrett, and L. Postle, Colouring graphs with sparse neighbourhoods: Bounds and applications. 2018, arXiv:1810.06704

[24] R. L. Brooks, On colouring the nodes of a network. *Proc. Cambridge Philos. Soc.* **37** (1941), 194–197 Zbl 0027.26403 MR 12236

[25] A. E. Brouwer, On the size of a maximum transversal in a Steiner triple system. *Canadian J. Math.* **33** (1981), no. 5, 1202–1204 Zbl 0481.05016 MR 638375

[26] R. A. Brualdi and H. J. Ryser, *Combinatorial Matrix Theory*. Encyclopedia Math. Appl. 39, Cambridge University Press, Cambridge, 1991 Zbl 0746.05002 MR 1130611

[27] H. Bruhn and F. Joos, A stronger bound for the strong chromatic index. *Combin. Probab. Comput.* **27** (2018), no. 1, 21–43 Zbl 1378.05047 MR 3734328

[28] D. Bryant, C. J. Colbourn, D. Horsley, and I. M. Wanless, Steiner triple systems with high chromatic index. *SIAM J. Discrete Math.* **31** (2017), no. 4, 2603–2611 Zbl 1375.05034 MR 3723319

[29] N. J. Cavenagh and J. Kuhl, On the chromatic index of Latin squares. *Contrib. Discrete Math.* **10** (2015), no. 2, 22–30 Zbl 1341.05017 MR 3499074

[30] W. I. Chang and E. L. Lawler, Edge coloring of hypergraphs and a conjecture of Erdős, Faber, Lovász. *Combinatorica* **8** (1988), no. 3, 293–295 Zbl 0661.05026 MR 963120

[31] G. Chen, G. Jing, and W. Zang, Proof of the Goldberg–Seymour conjecture on edge-colorings of multigraphs. 2019, arXiv:1901.10316

[32] F. Chung and R. Graham, *Erdős on Graphs: His Legacy of Unsolved Problems*. A K Peters, Wellesley, MA, 1998 Zbl 0890.05049 MR 1601954

[33] J. Cilleruelo, Infinite Sidon sequences. *Adv. Math.* **255** (2014), 474–486 Zbl 1302.11006 MR 3167490

[34] P. Condon, A. Espuny Díaz, A. Girão, D. Kühn, and D. Osthus, Hamiltonicity of random subgraphs of the hypercube. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 889–898, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2021 MR 4262489

[35] J. Cooper and D. Mubayi, List coloring triangle-free hypergraphs. *Random Structures Algorithms* **47** (2015), no. 3, 487–519 Zbl 1325.05075 MR 3385743

[36] J. Cooper and D. Mubayi, Coloring sparse hypergraphs. *SIAM J. Discrete Math.* **30** (2016), no. 2, 1165–1180 Zbl 1338.05078 MR 3507547

[37] J. Cooper and D. Mubayi, Sparse hypergraphs with low independence number. *Combinatorica* **37** (2017), no. 1, 31–40 Zbl 1399.05168 MR 3638331

[38] E. Davies, R. de Joannis de Verclos, R. J. Kang, and F. Pirot, Coloring triangle-free graphs with local list sizes. *Random Structures Algorithms* **57** (2020), no. 3, 730–744 Zbl 1464.05148 MR 4144082

[39] E. Davies, M. Jenssen, W. Perkins, and B. Roberts, Independent sets, matchings, and occupancy fractions. *J. Lond. Math. Soc. (2)* **96** (2017), no. 1, 47–66 Zbl 1370.05160 MR 3687939

[40] E. Davies, M. Jenssen, W. Perkins, and B. Roberts, On the average size of independent sets in triangle-free graphs. *Proc. Amer. Math. Soc.* **146** (2018), no. 1, 111–124 Zbl 1375.05194 MR 3723125

[41] E. Davies, R. J. Kang, F. Pirot, and J.-S. Sereni, An algorithmic framework for colouring locally sparse graphs. 2020, arXiv:2004.07151

[42] E. Davies, R. J. Kang, F. Pirot, and J.-S. Sereni, Graph structure via local occupancy. 2020, arXiv:2003.14361

[43] N. G. de Bruijn and P. Erdös, On a combinatorial problem. *Indagationes Math.* **10** (1948), 421–423 Zbl 0032.24405 MR 28289

[44] R. A. Duke, H. Lefmann, and V. Rödl, On uncrowded hypergraphs. *Random Structures Algorithms* **6** (1995), no. 2-3, 209–212 Zbl 0822.05051 MR 1370956

[45] J. Edmonds, Maximum matching and a polyhedron with 0, 1-vertices. *J. Res. Nat. Bur. Standards Sect. B* **69B** (1965), 125–130 Zbl 0141.21802 MR 183532

[46] J. Edmonds, Paths, trees, and flowers. *Canadian J. Math.* **17** (1965), 449–467 Zbl 0132.20903 MR 177907

[47] J. Egerváry, Matrixok kombinatorius tulajdonságairól. *Matematikai és Fizikai Lapok* **38** (1931), 16–28

[48] S. Ehard, S. Glock, and F. Joos, Pseudorandom hypergraph matchings. *Combin. Probab. Comput.* **29** (2020), no. 6, 868–885 Zbl 1466.05147 MR 4173135

[49] P. Erdős, On the combinatorial problems which I would most like to see solved. *Combinatorica* **1** (1981), no. 1, 25–42 Zbl 0486.05001 MR 602413

[50] P. Erdős and H. Hanani, On a limit theorem in combinatorial analysis. *Publ. Math. Debrecen* **10** (1963), 10–13 Zbl 0122.24802 MR 166116

[51] P. Erdős, A. Sárközy, and V. T. Sós, On a conjecture of Roth and some related problems. I. In *Irregularities of Partitions (Fertőd, 1986)*, pp. 47–59, Algorithms Combin. Study Res. Texts 8, Springer, Berlin, 1989 Zbl 0689.10061 MR 999930

[52] V. Faber, The Erdős–Faber–Lovász conjecture—the uniform regular case. *J. Comb.* **1** (2010), no. 2, 113–120 Zbl 1225.05190 MR 2732509

[53] V. Faber, Linear hypergraph list edge coloring – generalizations of the EFL conjecture to list coloring. 2017, arXiv:1701.03774

[54] V. Faber and D. G. Harris, Edge-coloring linear hypergraphs with medium sized edges. *Random Structures Algorithms* **55** (2019), no. 1, 153–159 Zbl 1423.05066 MR 3974196

[55] A. Ferber, V. Jain, and B. Sudakov, Number of 1-factorizations of regular high-degree graphs. *Combinatorica* **40** (2020), no. 3, 315–344 Zbl 1474.05320 MR 4121149

[56] G. Fiz Pontiveros, S. Griffiths, and R. Morris, The triangle-free process and the Ramsey number $R(3,k)$. *Mem. Amer. Math. Soc.* **263** (2020), no. 1274, v+125 Zbl 1439.05001 MR 4073152

[57] K. Ford, B. Green, S. Konyagin, J. Maynard, and T. Tao, Long gaps between primes. *J. Amer. Math. Soc.* **31** (2018), no. 1, 65–105 Zbl 1392.11071 MR 3718451

[58] P. Frankl and V. Rödl, Near perfect coverings in graphs and hypergraphs. *European J. Combin.* **6** (1985), no. 4, 317–326 Zbl 0624.05055 MR 829351

[59] A. Frieze and D. Mubayi, On the chromatic number of simple triangle-free triple systems. *Electron. J. Combin.* **15** (2008), no. 1, Research Paper 121 Zbl 1165.05324 MR 2443136

[60] A. Frieze and D. Mubayi, Coloring simple hypergraphs. *J. Combin. Theory Ser. B* **103** (2013), no. 6, 767–794 MR 3127593

[61] A. M. Frieze and T. Łuczak, On the independence and chromatic numbers of random regular graphs. *J. Combin. Theory Ser. B* **54** (1992), no. 1, 123–132 Zbl 0771.05088 MR 1142268

[62] Z. Füredi, The chromatic index of simple hypergraphs. *Graphs Combin.* **2** (1986), no. 1, 89–92 Zbl 0589.05036 MR 1554349

[63] Z. Füredi, Matchings and covers in hypergraphs. *Graphs Combin.* **4** (1988), no. 2, 115–206 Zbl 0820.05051 MR 943753

[64] Z. Füredi, J. Kahn, and P. D. Seymour, On the fractional matching polytope of a hypergraph. *Combinatorica* **13** (1993), no. 2, 167–180 Zbl 0779.05030 MR 1237040

[65] S. Glock, F. Joos, J. Kim, D. Kühn, and D. Osthus, Resolution of the Oberwolfach problem. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 8, 2511–2547 Zbl 1473.05241 MR 4269420

[66] S. Glock, D. Kühn, A. Lo, and D. Osthus, The existence of designs via iterative absorption: Hypergraph F-designs for arbitrary F. *Mem. Amer. Math. Soc.*, to appear

[67] S. Glock, D. Kühn, R. Montgomery, and D. Osthus, Decompositions into isomorphic rainbow spanning trees. *J. Combin. Theory Ser. B* **146** (2021), 439–484 Zbl 1457.05087 MR 4177962

[68] S. Glock, D. Kühn, and D. Osthus, Optimal path and cycle decompositions of dense quasirandom graphs. *J. Combin. Theory Ser. B* **118** (2016), 88–108 Zbl 1332.05078 MR 3471846

[69] M. K. Goldberg, On multigraphs of almost maximal chromatic class. *Diskret. Analiz* **23** (1973), 3–7

[70] D. A. Grable, Nearly-perfect hypergraph packing is in NC. *Inform. Process. Lett.* **60** (1996), no. 6, 295–299 Zbl 1336.68127 MR 1432991

[71] D. A. Grable, More-than-nearly-perfect packings and partial designs. *Combinatorica* **19** (1999), no. 2, 221–239 Zbl 0929.05065 MR 1723040

[72] R. L. Graham and H. O. Pollak, On embedding graphs in squashed cubes. In *Graph Theory and Applications (Proc. Conf., Western Michigan Univ., Kalamazoo, Mich., 1972; Dedicated to the Memory of J. W. T. Youngs)*, pp. 99–110, Lecture Notes in Math. 303, Springer, Berlin, 1972 Zbl 0251.05123 MR 0332576

[73] L. Haddad and C. Tardif, A clone-theoretic formulation of the Erdős–Faber–Lovász conjecture. *Discuss. Math. Graph Theory* **24** (2004), no. 3, 545–549 Zbl 1065.05040 MR 2120637

[74] R. Häggkvist and J. Janssen, New bounds on the list-chromatic index of the complete graph and other simple graphs. *Combin. Probab. Comput.* **6** (1997), no. 3, 295–313 Zbl 0880.05035 MR 1464567

[75] P. Hall, On representatives of subsets. *J. Lond. Math. Soc.* **10** (1935), 26–30 Zbl 0010.34503

[76] I. Holyer, The NP-completeness of edge-coloring. *SIAM J. Comput.* **10** (1981), no. 4, 718–720 Zbl 0473.68034 MR 635430

[77] H. Huang and B. Sudakov, A counterexample to the Alon–Saks–Seymour conjecture and related problems. *Combinatorica* **32** (2012), no. 2, 205–219 Zbl 1299.05123 MR 2927639

[78] E. Hurley, R. de Joannis de Verclos, and R. J. Kang, An improved procedure for colouring graphs of bounded local density. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 135–148, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2021 MR 4262443

[79] F. Iliopoulos, Improved bounds for coloring locally sparse hypergraphs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, p. Art. No. 39, LIPIcs. Leibniz Int. Proc. Inform. 207, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2021 MR 4366594

[80] O. Janzer and Z. L. Nagy, Coloring linear hypergraphs: the Erdős–Faber–Lovász conjecture and the Combinatorial Nullstellensatz. *Des. Codes Cryptogr.* **90** (2022), no. 9, 1991–2001 Zbl 1496.05052 MR 4473947

[81] T. R. Jensen and B. Toft, *Graph Coloring Problems*. Wiley-Interscience Ser. Discrete Math. Optim., John Wiley & Sons, New York, 1995 Zbl 0855.05054 MR 1304254

[82] A. Johansson, Asymptotic choice number for triangle free graphs. Dimacs technical report, The Center for Discrete Mathematics and Theoretical Computer Science, Piscataway, NJ, 1996

[83] A. Johansson, The choice number of sparse graphs. 1996, unpublished manuscript

[84] J. Kahn, Coloring nearly-disjoint hypergraphs with $n + o(n)$ colors. *J. Combin. Theory Ser. A* **59** (1992), no. 1, 31–39 Zbl 0774.05073 MR 1141320

[85] J. Kahn, Recent results on some not-so-recent hypergraph matching and covering problems. In *Extremal Problems for Finite Sets (Visegrád, 1991)*, pp. 305–353, Bolyai Soc. Math. Stud. 3, János Bolyai Math. Soc., Budapest, 1994 Zbl 0820.05050 MR 1319170

[86] J. Kahn, Asymptotics of hypergraph matching, covering and coloring problems. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pp. 1353–1362, Birkhäuser, Basel, 1995 Zbl 0842.05067 MR 1404037

[87] J. Kahn, Asymptotically good list-colorings. *J. Combin. Theory Ser. A* **73** (1996), no. 1, 1–59 Zbl 0851.05081 MR 1367606

[88] J. Kahn, Asymptotics of the chromatic index for multigraphs. *J. Combin. Theory Ser. B* **68** (1996), no. 2, 233–254 Zbl 0861.05026 MR 1417799

[89] J. Kahn, A linear programming perspective on the Frankl–Rödl–Pippenger theorem. *Random Structures Algorithms* **8** (1996), no. 2, 149–157 Zbl 0842.05066 MR 1607104

[90] J. Kahn, On some hypergraph problems of Paul Erdős and the asymptotics of matchings, covers and colorings. In *The Mathematics of Paul Erdős, I*, pp. 345–371, Algorithms Combin. 13, Springer, Berlin, 1997 Zbl 0865.05056 MR 1425195

[91] J. Kahn, Asymptotics of the list-chromatic index for multigraphs. *Random Structures Algorithms* **17** (2000), no. 2, 117–156 Zbl 0956.05038 MR 1774747

[92] J. Kahn and G. Kalai, A counterexample to Borsuk's conjecture. *Bull. Amer. Math. Soc. (N.S.)* **29** (1993), no. 1, 60–62 Zbl 0786.52002 MR 1193538

[93] J. Kahn and P. M. Kayll, Fractional v. integral covers in hypergraphs of bounded edge size. *J. Combin. Theory Ser. A* **78** (1997), no. 2, 199–235 Zbl 0884.05067 MR 1445415

[94] J. Kahn and P. D. Seymour, A fractional version of the Erdős–Faber–Lovász conjecture. *Combinatorica* **12** (1992), no. 2, 155–160 Zbl 0774.05072 MR 1179253

[95] G. Kalai, Some old and new problems in combinatorial geometry I: around Borsuk's problem. In *Surveys in Combinatorics 2015*, pp. 147–174, London Math. Soc. Lecture Note Ser. 424, Cambridge Univ. Press, Cambridge, 2015 Zbl 1361.51008 MR 3497269

[96] D. Y. Kang, T. Kelly, D. Kühn, A. Methuku, and D. Osthus, A proof of the Erdős–Faber–Lovász conjecture. 2021, arXiv:2101.04698

[97] D. Y. Kang, T. Kelly, D. Kühn, A. Methuku, and D. Osthus, Solution to a problem of Erdős on the chromatic index of hypergraphs with bounded codegree. 2021, arXiv:2110.06181

[98] D. Y. Kang, T. Kelly, D. Kühn, A. Methuku, and D. Osthus, A proof of the Erdős–Faber–Lovász conjecture: Algorithmic aspects. In *IEEE 62th Annual Symposium on Foundations of Computer Science (FOCS)*, 2022

[99] D. Y. Kang, D. Kühn, A. Methuku, and D. Osthus, New bounds on the size of nearly perfect matchings in almost regular hypergraphs. 2020, arXiv:2010.04183

[100] R. M. Karp, Reducibility among combinatorial problems. In *Complexity of Computer Computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972)*, pp. 85–103, Plenum, New York, 1972 Zbl 1467.68065 MR 0378476

[101] R. M. Karp, The probabilistic analysis of some combinatorial search algorithms. In *Algorithms and Complexity (Proc. Sympos., Carnegie-Mellon Univ., Pittsburgh, Pa., 1976)*, pp. 1–19, Academic Press, New York, 1976 Zbl 0368.68035 MR 0445898

[102] P. M. Kayll, Two chromatic conjectures: one for vertices and one for edges. In *Graph Theory—Favorite Conjectures and Open Problems. 1*, pp. 171–194, Probl. Books in Math., Springer, Cham, 2016 Zbl 1352.05071 MR 3617190

[103] P. Keevash, The existence of designs. 2014, arXiv:1401.3665

[104] P. Keevash, Hypergraph matchings and designs. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited Lectures*, pp. 3113–3135, World Sci. Publ., Hackensack, NJ, 2018 Zbl 1451.05160 MR 3966525

[105] P. Keevash, A. Pokrovskiy, B. Sudakov, and L. Yepremyan, New bounds for Ryser's conjecture and related problems. *Trans. Amer. Math. Soc. Ser. B* **9** (2022), 288–321 Zbl 07512761 MR 4408405

[106] T. Kelly, D. Kühn, and D. Osthus, A special case of Vu's conjecture: Coloring nearly disjoint graphs of bounded maximum degree. 2021, arXiv:2109.11438

[107] J. Kim, D. Kühn, A. Kupavskii, and D. Osthus, Rainbow structures in locally bounded colorings of graphs. *Random Structures Algorithms* **56** (2020), no. 4, 1171–1204 Zbl 1450.05070 MR 4101356

[108] J. H. Kim, On Brooks' theorem for sparse graphs. *Combin. Probab. Comput.* **4** (1995), no. 2, 97–132 Zbl 0833.05030 MR 1342856

[109] J. H. Kim, The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$. *Random Structures Algorithms* **7** (1995), no. 3, 173–207 Zbl 0832.05084 MR 1369063

[110] J. Komlós, J. Pintz, and E. Szemerédi, A lower bound for Heilbronn's problem. *J. London Math. Soc. (2)* **25** (1982), no. 1, 13–24 Zbl 0483.52008 MR 645860

[111] D. König, Gráfok és mátrixok. *Matematikai és Fizikai Lapok* **38** (1931), 116–119

[112] A. V. Kostochka and V. Rödl, Partial Steiner systems and matchings in hypergraphs. *Random Structures Algorithms* **13** (1998), no. 3-4, 335–347 Zbl 0959.05079 MR 1662789

[113] D. Kühn and D. Osthus, Embedding large subgraphs into dense graphs. In *Surveys in Combinatorics 2009*, pp. 137–167, London Math. Soc. Lecture Note Ser. 365, Cambridge Univ. Press, Cambridge, 2009 Zbl 1182.05098 MR 2588541

[114] D. Kühn and D. Osthus, Hamilton decompositions of regular expanders: a proof of Kelly's conjecture for large tournaments. *Adv. Math.* **237** (2013), 62–146 Zbl 1261.05053 MR 3028574

[115] M. Meszka, The chromatic index of projective triple systems. *J. Combin. Des.* **21** (2013), no. 11, 531–540 Zbl 1292.05061 MR 3103360

[116] M. Meszka, R. Nedela, and A. Rosa, Circulants and the chromatic index of Steiner triple systems. *Math. Slovaca* **56** (2006), no. 4, 371–378 Zbl 1141.05023 MR 2267758

[117] M. Molloy, Graph colouring with the probabilistic method. 2019, talk at CanaDAM

[118] M. Molloy, The list chromatic number of graphs with small clique number. *J. Combin. Theory Ser. B* **134** (2019), 264–284 Zbl 1402.05076 MR 3906639

[119] M. Molloy and L. Postle, Asymptotically good edge correspondence colourings. *J. Graph Theory* **100** (2022), no. 3, 559–577 MR 4433315

[120] M. Molloy and B. Reed, Near-optimal list colorings. *Random Structures Algorithms* **17** (2000), no. 3-4, 376–402 Zbl 0971.05047 MR 1801140

[121] M. Molloy and B. Reed, *Graph Colouring and the Probabilistic Method*. Algorithms Combin. 23, Springer, Berlin, 2002 Zbl 0987.05002 MR 1869439

[122] N. Pippenger and J. Spencer, Asymptotic behavior of the chromatic index for hypergraphs. *J. Combin. Theory Ser. A* **51** (1989), no. 1, 24–42 Zbl 0729.05038 MR 993646

[123] D. K. Ray-Chaudhuri and R. M. Wilson, Solution of Kirkman's schoolgirl problem. In *Combinatorics (Proc. Sympos. Pure Math., Vol. XIX, Univ. California, Los Angeles, Calif., 1968)*, pp. 187–203, Amer. Math. Soc., Providence, RI, 1971 Zbl 0248.05009 MR 0314644

[124] B. Reed, ω, Δ, and χ. *J. Graph Theory* **27** (1998), no. 4, 177–212 Zbl 0980.05026 MR 1610746

[125] V. Rödl, On a packing and covering problem. *European J. Combin.* **6** (1985), no. 1, 69–78 Zbl 0565.05016 MR 793489

[126] V. Rödl and A. Ruciński, Dirac-type questions for hypergraphs—a survey (or more problems for Endre to solve). In *An Irregular Mind*, pp. 561–590, Bolyai Soc. Math. Stud. 21, János Bolyai Math. Soc., Budapest, 2010 Zbl 1221.05255 MR 2815614

[127] V. Rödl and L. Thoma, Asymptotic packing and the random greedy algorithm. *Random Structures Algorithms* **8** (1996), no. 3, 161–177 Zbl 0856.05074 MR 1605397

[128] I. Z. Ruzsa, An infinite Sidon sequence. *J. Number Theory* **68** (1998), no. 1, 63–71 Zbl 0927.11005 MR 1492889

[129] H. J. Ryser, Neuere Probleme der Kombinatorik. In *Vorträge über Kombinatorik*, pp. 69–91, Oberwolfach, 24–29 July (1967)

[130] D. Saxton and A. Thomason, Hypergraph containers. *Invent. Math.* **201** (2015), no. 3, 925–992 Zbl 1320.05085 MR 3385638

[131] P. D. Seymour, Some unsolved problems on one-factorizations of graphs. In *Graph Theory and Related Topics*, edited by J. A. Bondy and U. S. R. Murty, pp. 367–368, Academic Press, New York, 1979

[132] P. D. Seymour, Packing nearly disjoint sets. *Combinatorica* **2** (1982), no. 1, 91–97 Zbl 0494.05015 MR 671149

[133] C. E. Shannon, A theorem on coloring the lines of a network. *J. Math. Physics* **28** (1949), 148–151 Zbl 0032.43203 MR 0030203

[134] J. B. Shearer, A note on the independence number of triangle-free graphs. *Discrete Math.* **46** (1983), no. 1, 83–87 Zbl 0516.05053 MR 708165

[135] J. B. Shearer, A note on the independence number of triangle-free graphs. II. *J. Combin. Theory Ser. B* **53** (1991), no. 2, 300–307 Zbl 0753.05074 MR 1129557

[136] J. B. Shearer, On the independence number of sparse graphs. *Random Structures Algorithms* **7** (1995), no. 3, 269–271 Zbl 0834.05030 MR 1369066

[137] J. Spencer, Uncrowded graphs. In *Mathematics of Ramsey Theory*, pp. 253–262, Algorithms Combin. 5, Springer, Berlin, 1990 Zbl 0738.05052 MR 1083606

[138] J. Spencer, Asymptotic packing via a branching process. *Random Structures Algorithms* **7** (1995), no. 2, 167–172 Zbl 0847.60068 MR 1369062

[139] S. K. Stein, Transversals of Latin squares and their generalizations. *Pacific J. Math.* **59** (1975), no. 2, 567–575 Zbl 0302.05015 MR 387083

[140] W. T. Tutte, The factorization of linear graphs. *J. London Math. Soc.* **22** (1947), 107–111 Zbl 0029.23301 MR 23048

[141] S. A. Vanstone, D. R. Stinson, P. J. Schellenberg, A. Rosa, R. Rees, C. J. Colbourn, M. W. Carter, and J. E. Carter, Hanani triple systems. *Israel J. Math.* **83** (1993), no. 3, 305–319 Zbl 0783.05023 MR 1239064

[142] V. G. Vizing, The chromatic class of a multigraph. *Cybernetics* **1** (1965), 32–41

[143] V. H. Vu, New bounds on nearly perfect matchings in hypergraphs: higher codegrees do help. *Random Structures Algorithms* **17** (2000), no. 1, 29–63 Zbl 0953.05056 MR 1768848

[144] V. H. Vu, A general upper bound on the list chromatic number of locally sparse graphs. *Combin. Probab. Comput.* **11** (2002), no. 1, 103–111 Zbl 0991.05041 MR 1888186

[145] N. C. Wormald, The differential equation method for random graph processes and greedy algorithms. In *Lectures on Approximation and Randomized Algorithms*, pp. 73–155, Polish Scientific Publishers, Warsaw, 1999 Zbl 0943.05073

[146] Y. Zhao, Recent advances on Dirac-type problems for hypergraphs. In *Recent Trends in Combinatorics*, pp. 145–165, IMA Vol. Math. Appl. 159, Springer, Cham, 2016
Zbl 1354.05100 MR 3526407

Dong Yeap Kang
School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK;
d.y.kang.1@bham.ac.uk

Tom Kelly
School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK;
t.j.kelly@bham.ac.uk

Daniela Kühn
School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK;
d.kuhn@bham.ac.uk

Abhishek Methuku
School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK;
abhishekmethuku@gmail.com

Deryk Osthus
School of Mathematics, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK;
d.osthus@bham.ac.uk

DOI 10.4171/8ECM/10

From matrix pivots to graphs in surfaces: Exploring combinatorics through partial duals

Iain Moffatt

Abstract. To what extent is a graph determined by the trees in it? What changes if we ask this question not for graphs in the abstract, but graphs that are embedded on surfaces? By considering these questions we will see how a collection of seemingly disjoint topics in mathematics are brought together through the idea of a partial dual.

1. Introduction

Consider two graphs $\mathbb{G}$ and $\mathbb{H}$ each of which is drawn on a plane so that its edges do not intersect (or consider two spherical polyhedra if you prefer). Then $\mathbb{G}$ and $\mathbb{H}$ are *geometric duals* if the vertices in one correspond to the faces in the other, and the edges between vertices in one correspond to the edges between faces in the other. (See Figure 2 for an example.)

Now consider two graphs G and H (not drawn on the plane this time). Each contains a set of *spanning trees*, these are the maximal acyclic subgraphs contained in them. Then G and H are *algebraic duals* if their sets of spanning trees correspond through complementation (i.e., the edge set of a spanning tree of one is the complement of the edge set of a spanning tree of the other).

It is a classical result of H. Whitney that a graph has an algebraic dual if and only if it can be drawn on the plane without its edges crossing, in which case the algebraic dual is exactly a geometric dual. This sets up a fundamental relationship between planarity, duality, and spanning trees.

But what happens if the graphs cannot be drawn on the plane in this way? It is this situation we examine here. We shall see that it is inexorably linked to graphs drawn on surfaces, duals and partial duals, matroids and delta-matroids, principal pivot transforms of matrices, and pivot-minors of simple graphs.

2020 Mathematics Subject Classification. Primary 05C10; Secondary 05B35.
Keywords. Delta-matroid, dual, embedded graph, matroid, partial dual, pivot, quasi-tree, ribbon graph, tree.

This exposition is aimed at a general mathematical reader. A familiarity with elementary graph theory and with orientable surfaces is assumed. We note that graphs here may have *multiple edges* (edges that have the same ends) and *loops* (an edge with both ends being the same vertex). For simplicity we shall only consider orientable surfaces, but (almost) everything here can be extended to non-orientable surfaces.

2. Graphs and their spanning trees

We start with a classical question with a well-known answer. Recall that a graph is a *tree* if it is connected and contains no cycles. A *spanning tree* of a graph G is a subgraph that is a tree and that contains every vertex of G. (For example, the bold edges in the left and right images in Figure 2 define spanning trees.) Only connected graphs have spanning trees, and to simplify terminology here we shall generally restrict ourselves to connected graphs. This restriction does not result in any real loss of generality. This is since most of our results extend trivially and obviously to non-connected graphs by considering the *maximal spanning forests* of a graph, which are the subgraphs that restrict to a spanning tree in each connected component.

Our initial interest is in the question:

Is a graph determined by its spanning trees?

There are a few ways to interpret this question resulting in different answers. Here we are interested in what happens if the only information you have about any given spanning tree is the edges that are in it. But since loops will never appear in a spanning tree, we will also need to know if there are any loops. So our precise question is: *If you know the edge set of each spanning tree of a connected graph as well as any loops in the graph, do you then know the graph?* It is not hard to see that the answer is no. For example, consider the two non-isomorphic trees on three edges. But this “no” is really a “more or less, yes.”

Consider the moves of *vertex identification*, *vertex cleaving*, and *Whitney twisting* illustrated in Figure 1. Vertex identification is a move that identifies two vertices that lie in different connected components of a graph, and vertex cleaving is the inverse operation. For Whitney twisting, suppose u_1 and v_1 are vertices in a graph G_1, and u_2 and v_2 are vertices in a graph G_2. Construct a graph G by identifying u_1 and u_2, and v_1 and v_2. Construct also a graph G' by identifying u_1 and v_2, and v_1 and u_2. Then we say G and G' are related by Whitney twists. Two graphs are said to be 2-*isomorphic* if one can be obtained from the other through isomorphism, vertex identification, vertex cleaving, and Whitney twisting.

Whitney's 2-*isomorphism theorem* [57] provides an answer to our question. It states that if you know the edge set of each spanning tree of a graph as well as any

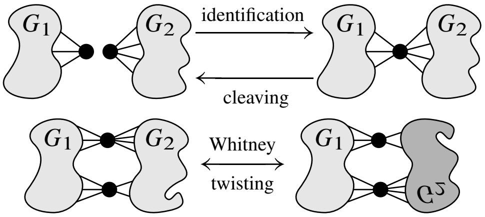


Figure 1. The moves for 2-isomorphism: vertex identification, vertex cleaving, and Whitney twisting.

loops in the graph, then you know the graph up to 2-isomorphism. Conversely, the collections of edge sets of spanning trees and loops in two 2-isomorphic graphs are equal. (We shall give a cleaner statement of Whitney's 2-isomorphism theorem below.)

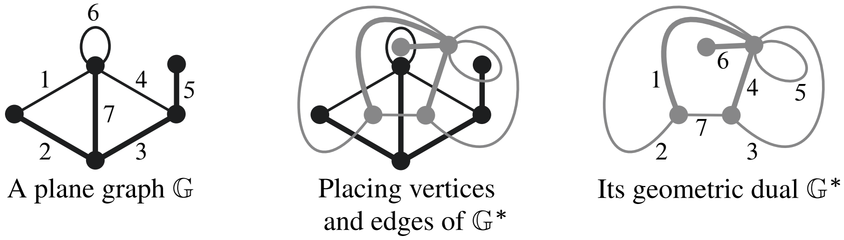
Thus the spanning tree structure determines the graph up to some simple moves. In particular, it completely determines 3-connected graphs (ones in which there are three internally disjoint paths between each pair of vertices) up to isomorphism as the moves cannot be applied to such graphs. It turns out that many graph properties and results do not distinguish between 2-isomorphic graphs, and so can be understood in terms of spanning tree structure. In fact, considering the spanning tree structure of a graph, rather than the graph itself, turns out to be an extremely fruitful thing to do.

The spanning trees in a connected graph G have many nice standard properties. For example, every non-loop edge of G is in some spanning tree; all spanning trees have the same number of edges; and if G has n vertices, a spanning subgraph is a spanning tree if and only if it is connected and has exactly $n-1$ edges. Spanning trees also satisfy an *exchange property*: if T and T' are spanning trees and e is an edge in T but not T', then there is always some edge f in T' but not T such that removing e from T then adding f results in another spanning tree. (A reader may spot that this exchange property also applies to the bases of a vector space.) These properties on the collection of spanning trees lead us to *matroids*.

Definition 2.1. Let E be a finite set, and let $\mathcal{B}$ be a non-empty collection of subsets of E. Then the pair $M := (E, \mathcal{B})$ is called a *matroid* if for distinct $A, B \in \mathcal{B}$ and for all $a \in A \setminus B$ there exists $b \in B \setminus A$ such that $(A \setminus a) \cup b \in \mathcal{B}$.

By the properties of trees mentioned above, if G is a connected graph with edge set E and $\mathcal{B}$ is the set consisting of all edge sets of its spanning trees, then $C(G) := (E, \mathcal{B})$ is a matroid. It is called the *cycle matroid* of G.

Example 2.2. The graph on the left of Figure 2 has cycle matroid $(E, \mathcal{B})$ with $E = \{1,2,3,4,5,6,7\}$ and $\mathcal{B} = \{\{1,2,3,5\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\},\{1,3,5,7\},\{1,4,5,7\},\{2,3,5,7\},\{2,4,5,7\}\}$.



A plane graph $\mathbb{G}$

Placing vertices and edges of $\mathbb{G}^*$

Its geometric dual $\mathbb{G}^*$

Figure 2. Forming the geometric dual $\mathbb{G}^*$ of a plane graph $\mathbb{G}$.

Our initial question of whether the spanning trees determine the graph then becomes a matroid theoretic one: *If you have a cycle matroid, can you determine the graph it came from?* We can rephrase our previous answer as follows. (For the statement, matroid isomorphism is defined in the obvious way.)

Theorem 2.3 (Whitney's 2-isomorphism theorem). *Let G and H be connected graphs. Then $C(G)$ and $C(H)$ are isomorphic matroids if and only if G and H are* 2*-isomorphic.*

Whitney's 2-isomorphism theorem nails down the connection between cycle matroids and graphs. Cycle matroids give rise to a class of matroids, but almost all matroids are *not* cycle matroids [41]. Nevertheless, cycle matroids are important in matroid theory and graph theory. On one hand, insights from matroid theory can lead to new results about graphs. On the other hand, graph theory can serve as an excellent guide for studying matroids. A good introduction to the mutually enriching relationship between graph theory and matroid theory can be found in [42].

Bibliographic remarks. The topics discussed in this section are classical. An excellent resource for this material is Chapter 5 of J. Oxley's book [43]. Whitney's 2-isomorphism theorem dates from the 1930s and is due to H. Whitney, [57] (see also [50, 54]), and Theorem 2.3 is a modern formulation in terms of matroids.

Our motivational question was whether a graph is determined by its spanning trees or its cycle matroid. We restrict discussion here to characterising graphs that have the same cycle matroid, ignoring the algorithmic question about constructing the graphs from the cycle matroid. Discussion of the latter problem can be found in [53] (for what will follow, the equivalent problem for quasi-trees can be answered through the circle graph recognition methods of [30, 34, 46]).

H. Whitney introduced matroids in the 1930s (see [58]) to capture ideas of dependence common to linear algebra and graph theory. There are many ways to define matroids and Definition 2.1 provides their definition in terms of "bases." The cycle matroid $C(G)$ can also be defined through the cycles in a graph (using a "circuit definition" of a matroid), hence the name. Matroid theory is a major topic of study in combinatorics. Our encounter with matroids here is extremely brief and we refer the reader to the books [43,55] for more on them.

A spectacular illustration of the mutually enriching relationship between graph theory and matroid theory can be found in J. Geelen, B. Gerards, and G. Whittle's recent and, at the time of writing, unpublished result that, for any finite field, the class of matroids that are representable over that field is well-quasi-ordered by the minor relation. Their results generalise N. Robertson and P. Seymour's graph minors project where it is shown that graphs are well-quasi-ordered by the minor relation [45]. In [31] Geelen, Gerards, and Whittle wrote "it would be inconceivable to prove a structure theorem for matroids without the Graph Minors Structure Theorem as a guide."

3. The appearance of topology

We want to make contact with topological graph theory, which is the study of graphs embedded in surfaces. We shall do this by considering duals. Suppose $M = (E, \mathcal{B})$ is a matroid. Define a collection of sets $\mathcal{B}^*$ by taking the complement of each member of $\mathcal{B}$, so $\mathcal{B}^* := \{E \setminus B : B \in \mathcal{B}\}$. It is not hard to check that the pair $(E, \mathcal{B}^*)$ also forms a matroid. This is called the *dual* of M and is denoted by M^*.

Example 3.1. The dual of the matroid in Example 2.2 has $\mathcal{B}^* = \{\{4,6,7\}, \{3,6,7\}, \{2,6,7\}, \{1,6,7\}, \{2,4,6\}, \{2,3,6\}, \{1,4,6\}, \{1,3,6\}\}$.

If G is a graph and $C(G)$ its cycle matroid, then the dual matroid $C(G)^*$ is always a matroid. However, it is not always the cycle matroid of a graph. If $C(G) = (E, \mathcal{B})$, the graph G is connected, and $C(G)^* = (E, \mathcal{B}^*)$, then $\mathcal{B}$ consists of the edge sets of all the spanning trees of G. For $C(G)^*$ to be the cycle matroid of a graph we require the existence of some graph H on the edge set E such that the sets in $\mathcal{B}^*$ define exactly the spanning trees of H. That is, we require H to have the property that T is a spanning tree of G if and only if $E \setminus T$ is a spanning tree of H. Such a graph H, if it exists, is called an *algebraic dual* (or *abstract dual* or *combinatorial dual*) of G. If it does exist, it may or may not be unique.

The existence of algebraic duals is tied to the topological properties of a graph. A connected *plane graph* consists of a connected graph drawn, or *embedded*, in the sphere (or, equivalently, the plane) in such a way that vertices are distinct points and edges only intersect at their ends. (So each vertex is a point on the sphere, each edge is a simple curve between these points, and these curves do not intersect except when their ends share a vertex.) Plane graphs are *equivalent* if there is a homeomorphism of the sphere taking one graph drawing to the other (i.e., inducing a graph isomorphism). A plane graph divides the sphere into regions called*faces*. For example, with the page representing a portion of the sphere, the left-hand image of Figure 2 shows a plane graph with four faces. A connected graph is said to be *planar* if it can be drawn in the sphere in the above way. (So a plane graph *is* drawn on the sphere, and a planar graph *can* be drawn on the sphere.) Inequivalent plane graphs can be drawings of the same planar graph. These definitions are extended to non-connected graphs by drawing each graph component in its own copy of the sphere.

Plane graphs have another type of dual. If $\mathbb{G}$ is a plane graph, then its *geometric dual*, denoted $\mathbb{G}^*$, is the plane graph obtained from $\mathbb{G}$ by placing one vertex in each of its faces, and embedding an edge of $\mathbb{G}^*$ between two of these vertices whenever the faces of $\mathbb{G}$ they lie in meet at an edge. Edges of $\mathbb{G}^*$ are embedded so that they cross only the corresponding edge of $\mathbb{G}$. An example is given in Figure 2.

For a plane graph $\mathbb{G} = (V, E)$, Euler's formula gives that $|V| - |E| + |F| = 2$, where $|F|$ is the number of faces. Thus if A is the edge set of a spanning tree in $\mathbb{G}$, then $|V| - |A| = 1$ and so $|F| - |E \setminus A| = 1$ giving that $E \setminus A$ is the edge set of a spanning tree of $\mathbb{G}^*$. As $(\mathbb{G}^*)^* = \mathbb{G}$, it follows that geometric duals of plane graphs are algebraic duals, and so for plane graphs $C(\mathbb{G})^* = C(\mathbb{G}^*)$.

The converse is also true: if G and H are algebraic duals, then the correspondence between their spanning tree structures guarantees that there are plane graphs $\mathbb{G}$ and $\mathbb{H}$ that are embeddings (i.e., drawings) of G and H that are geometric duals, $\mathbb{H} = \mathbb{G}^*$. Collecting all this together gives the following result of Whitney [56].

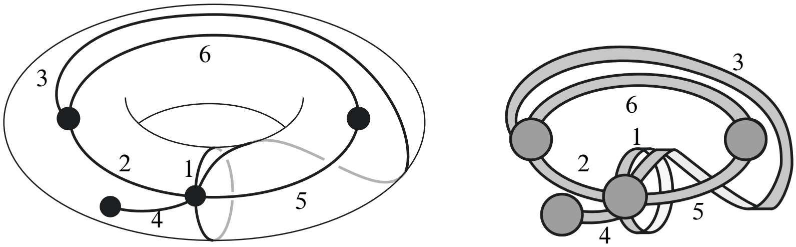
Theorem 3.2. *Let G be a connected graph with cycle matroid $C(G)$. Then the dual matroid $C(G)^*$ is the cycle matroid of a graph if and only if G is planar. Moreover, if G is planar, then*

$$C(\mathbb{G})^* = C(\mathbb{G}^*),$$

where $\mathbb{G}$ is any plane embedding of G, and $\mathbb{G}^*$ its geometric dual.

In this theorem we see how the spanning tree (or cycle matroid) structure of a graph captures its topological properties. However, Theorem 3.2 illustrates that many of these properties are tied to planarity. What if you do not want to restrict yourself to plane or planar graphs? Let us examine what changes when you consider graphs on surfaces other than the plane.

As noted above, for expositional simplicity we shall only consider orientable surfaces. However (almost) everything here extends to non-orientable surfaces (with varying degrees of difficulty) and details of how to do this can be found in the references. We will often omit the work "orientable", although we shall add it when it is crucial. We recall that the classification of surfaces states that every closed orientable surface is homeomorphic to a sphere with handles (or n-torus). Every orientable surface with boundary is homeomorphic to a sphere with handles with the interiors of some discs removed from it. In both cases the number of handles is its *genus*.



(a) A graph embedded in the torus

(b) A ribbon graph

Figure 3. Realisations of the same embedded graph.

An *embedded graph* $\mathbb{G}$ is a graph drawn on a closed surface Σ in such a way that edges only intersect at their ends, and the drawing divides Σ into regions that are homeomorphic to discs. (As in the plane case, each vertex is a point on the surface, each edge is a simple curve between these points, and these curves do not intersect except when their ends share a vertex.) The regions of Σ determined by the graph drawing are called *faces* of $\mathbb{G}$. Thus a plane graph is a graph embedded in the sphere. We note that if $\mathbb{G}$ has more than one component, then each component of the graph lies in its own surface. Figure 3a shows a graph embedded in a torus. It has two faces.

The *geometric dual* $\mathbb{G}^*$ of an embedded graph $\mathbb{G}$ is formed just as in the plane case by placing vertices in the faces and drawing edges between these vertices when the faces meet at an edge. Note that $\mathbb{G}$ and $\mathbb{G}^*$ are embedded in the same surface.

Suppose $\mathbb{G}$ is a connected embedded graph and $\mathbb{G}^*$ its geometric dual. Since the edge sets of $\mathbb{G}$ and $\mathbb{G}^*$ correspond, we may assume that a graph and its geometric dual have the same edge set E. The operation $* : A \mapsto E \setminus A$ sends edge sets of $\mathbb{G}$ to edge sets of $\mathbb{G}^*$, or equivalently the set of spanning subgraphs of $\mathbb{G}$ to the set of spanning subgraphs of $\mathbb{G}^*$.[1] (As an example, the bold edges in Figure 2 indicate a pair of spanning trees identified under this map.) Theorem 3.2 and the characterisation of planar graphs in terms of algebraic duals depend upon the fact that if $\mathbb{G}$ (and so $\mathbb{G}^*$) is a plane graph, then $*$ sends spanning trees to spanning trees, and this happens if and only if $\mathbb{G}$ is a plane graph.

[1] At this point we are glossing over the issue of exactly how a subgraph of $\mathbb{G}$ should be considered as an embedded graph. The difficulty is that restricting the drawing of $\mathbb{G}$ to the edges and vertices in the subgraph may result in faces that are not discs, in which case the surface will need to be altered, by removing any redundant handles, to obtain an embedded graph. This issue will be resolved in the next section by switching to the language of ribbon graphs. For the present discussion it is safe, although not quite correct, to think of restricting the drawing of $\mathbb{G}$ to the edges and vertices in the subgraph.

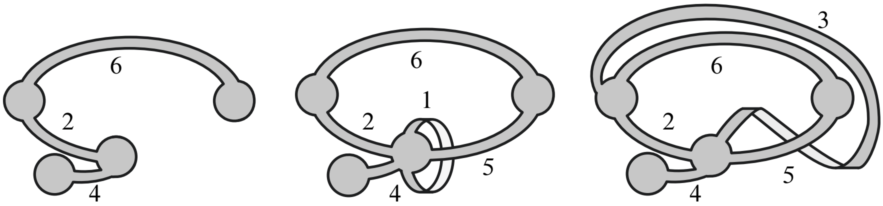


Figure 4. Neighbourhoods of the subgraphs on $\{2, 4, 6\}$, $\{1, 2, 4, 5, 6\}$, and $\{2, 3, 4, 5, 6\}$.

Suppose that $\mathbb{G}$ is embedded in an arbitrary closed surface Σ and A is the edge set of one of its spanning trees $\mathbb{T}$. Let $*(\mathbb{T})$ be the spanning subgraph of $\mathbb{G}^*$ on the edge set $*(A)$. Then it is easy to see (e.g., by drawing a picture; as an example consider the bold edges in the middle image of Figure 2) that Σ can be written as the union of a neighbourhood of $\mathbb{T}$ and a neighbourhood of $*(\mathbb{T})$. Since $\mathbb{T}$ is a spanning tree, its neighbourhood is a disc. Thus the neighbourhood of $*(\mathbb{T})$ consists of a once-punctured copy of Σ. In particular, it is a subgraph whose neighbourhood has exactly one boundary component. This is the property that is important to us.

A spanning subgraph of an embedded graph $\mathbb{G}$ is said to be a *spanning quasi-tree* if its neighbourhood has exactly one boundary component. Notice that every spanning tree is a spanning quasi-tree, although in general an embedded graph will have many other spanning quasi-trees. The *genus* of a quasi-tree is the genus of its neighbourhood considered as a surface with boundary. (We shall reformulate these definitions in the next section.) If $\mathbb{G}$ is in a surface Σ of genus n, then it will have spanning quasi-trees of genus $0, 1, 2, \ldots, n$, and the spanning trees are just those of genus zero. The map $*$ then sends a tree to a quasi-tree of maximal genus n. More generally, $*$ will send a spanning quasi-tree of genus g to a spanning quasi-tree of genus $n - g$.

Example 3.3. For the embedded graph shown in Figure 3a, each of the sets $\{2, 4, 6\}$, $\{1, 2, 4, 5, 6\}$, and $\{2, 3, 4, 5, 6\}$ induces a spanning quasi-tree. The neighbourhoods are shown in Figure 4. The set $\{2, 4, 6\}$ defines a spanning quasi-tree of genus zero, and the other two sets induce spanning quasi-trees of genus one.

We started with the question of whether the spanning trees in a graph determine the graph itself. Whitney's theorem provided a complete answer to this question, and Theorem 3.2 tied together duality, spanning tree structure, and planarity. If instead we want to work with non-plane embedded graphs, rather than looking at spanning trees, we should consider quasi-trees. Thus we are led to ask:

Is an embedded graph determined by its spanning quasi-trees?

Just as in the spanning trees case we formalise this by asking: *If you know the edge set of each spanning quasi-tree of an embedded graph, as well as any edges that appear in no quasi-trees, then do you then know the embedded graph?*

Again the immediate answer is no. For example, if $\mathbb{G}$ is a plane graph, then its set of spanning quasi-trees is exactly its set of spanning trees, and we already know that these do not necessarily determine a plane embedding. In this plane case, however, Whitney's 2-isomorphism theorem will provide a way to characterise all plane graphs that have the same set of spanning quasi-trees. What if $\mathbb{G}$ is not embedded in the plane? In this case Whitney's 2-isomorphism theorem does not help.

Bibliographic remarks. Dual matroids date back to H. Whitney's foundational work on matroids [58]. The construction of a geometric dual is classical and can be seen in J. Kepler's work on dual polyhedra (see p. 181 of his *Harmonices mundi* dating from 1619). Algebraic duals, as well as their connection with planarity and geometric duals, are due to H. Whitney [56]. Theorem 3.2 provides a modern statement of his results.

Embedded graphs are standard objects in graph theory. They have several alternative names and formulations including *combinatorial maps*, *rotation systems*, *ribbon graphs*, *graph encoded maps*, and so on. Excellent introductions to embedded graphs and topological graph theory are the works of J. Gross and T. Tucker [35], and B. Mohar and C. Thomassen [40].

4. Partial duals

Duality tied spanning tree structure to planarity. For non-plane embedded graphs and quasi-trees we consider a generalisation of geometric duality called *partial duality*. For our discussion of partial duals it is convenient to describe embedded graphs as *ribbon graphs*.

A ribbon graph is a structure that arises by taking a regular neighbourhood of a graph embedded in a surface, but without throwing away the vertex-edge structure of the graph; see Figure 3. We can think of them informally as "graphs whose vertices consist of discs, and whose edges consist of ribbons," as in Figure 3b. They can be defined formally as follows.

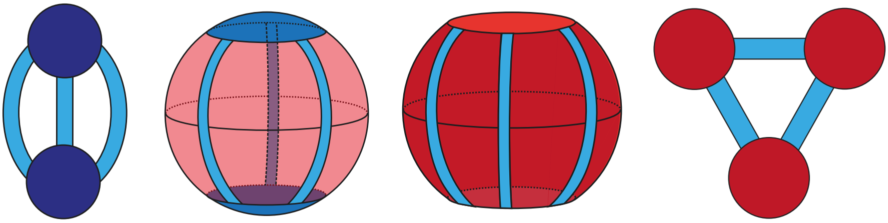
Definition 4.1. A *ribbon graph* $\mathbb{G} = (V, E)$ is a surface with boundary represented as the union of two sets of discs, a set V of *vertices*, and a set E of *edges* such that (1) the vertices and edges intersect in disjoint line segments; (2) each such line segment lies on the boundary of precisely one vertex and precisely one edge; (3) every edge contains exactly two such line segments.

Ribbon graphs are equivalent to embedded graphs. Above we described how a ribbon graph can be obtained from an embedded graph. In the other direction, given a ribbon graph, the classification of surfaces with boundary ensures there is a unique way (up to homeomorphism) to embed it in a closed surface by "filling in the holes." This gives an embedding of the ribbon graph in a closed surface from which it is clear how to obtain the embedded graph. Two ribbon graphs are *equivalent* if there is a homeomorphism from one to the other that sends vertices to vertices and edges to edges. Thus ribbon graphs are equivalent precisely when their corresponding embedded graphs are. Thus any result about ribbon graphs is a result about embedded graphs, and vice versa.

Graph theory terminology is extended to ribbon graphs in the obvious way. A *ribbon subgraph* $\mathbb{H}$ of $\mathbb{G}$ is a ribbon graph obtained from $\mathbb{G}$ by removing some of its vertices and edges. It is *spanning* if it has the same vertices as $\mathbb{G}$. The spanning ribbon subgraph obtained from $\mathbb{G}$ by deleting an edge e is denoted by $\mathbb{G}\backslash e$. Ribbon graphs have topological parameters in addition to their graph theoretic ones. Here we defined ribbon graphs to be *orientable* meaning that they are orientable when considered as a surface with boundary. (Recall that for expositional simplicity we restricted ourselves to orientable surfaces, and therefore to orientable ribbon graphs.) In general, ribbon graphs may be *non-orientable* as well, and at times we will comment on this case. The *genus* of a ribbon graph is its genus as a surface. A connected ribbon graph is *plane* if it has genus 0 (i.e., if it corresponds to a graph on a sphere). We are often interested in the *boundary components* of a ribbon graph, which are just the components of its boundary when it is considered as a surface with boundary. A ribbon graph that has exactly one vertex is called a *bouquet*. These form an important class of ribbon graphs.

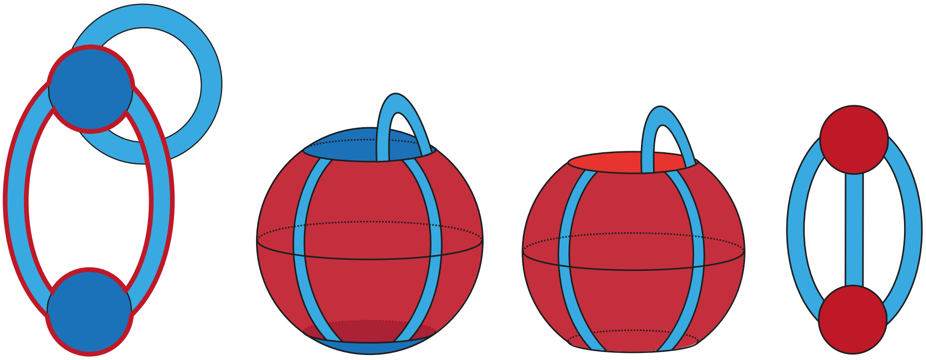
Geometric duality has a very neat description in terms of ribbon graphs. If $\mathbb{G} = (V, E)$ is a ribbon graph, then its *geometric dual* $\mathbb{G}^*$ is the ribbon graph formed by taking one disc for each boundary component of $\mathbb{G}$ (these will form the vertices of the dual); for each boundary component of $\mathbb{G}$ (which is topologically a circle), identify it with the boundary of one of these discs (resulting in a surface without boundary); finally, in the resulting surface, delete the interiors of the vertex discs in V. This results in the ribbon graph $\mathbb{G}^*$. The discs that were added during the construction form the vertices of $\mathbb{G}^*$, and the edges of $\mathbb{G}$ form the edges of $\mathbb{G}^*$ but the parts of their boundary that are and are not attached to vertices are switched. This construction is illustrated in Figure 5.

It is not too hard to see our two constructions for geometric duals agree. The construction of $\mathbb{G}^*$ in terms of embedded graphs is a global construction in the sense that it applies to the whole of $\mathbb{G}$ at the same time. However, once you have switched to the language of ribbon graphs, the construction is easily adapted to give a local construction, where local here means that you can form the geometric dual $\mathbb{G}^*$ at



(a) $\mathbb{G}$ **(b)** Sewing in discs **(c)** Removing original vertices to get $\mathbb{G}^*$ **(d)** Redrawing $\mathbb{G}^*$

Figure 5. Forming the geometric dual of a ribbon graph.



(a) $\mathbb{G} = (V, E)$ with the boundary of (V, A) highlighted **(b)** Adding discs to the boundary of (V, A) **(c)** Deleting vertices in V gives $\mathbb{G}^A$ **(d)** Redrawing $\mathbb{G}^A$

Figure 6. Forming a partial dual $\mathbb{G}^A$ where A consists of the two non-loop edges of $\mathbb{G}$.

individual edges. Then, with this local construction in hand, we can form the dual at just some of edges while leaving the remaining edges alone. This observation leads to the surprising idea of *partial duals*.

Partial duals arise by modifying the description of geometric duality for ribbon graphs so that the dual is formed with respect to only a subset of edges. Let $\mathbb{G} = (V, E)$ be a ribbon graph and $A \subseteq E$. The *partial dual* of $\mathbb{G}$ with respect to A, denoted by $\mathbb{G}^A$, is the ribbon graph formed as follows. Consider the spanning ribbon subgraph (V, A) as a subset of $\mathbb{G}$. The boundary of (V, A) defines a set of closed curves on $\mathbb{G}$. For each of these closed curves, take a disc (which will form a vertex of $\mathbb{G}^A$) and identify the curve and the boundary of this disc. Finally, delete the interior of each vertex disc in V. The resulting ribbon graph is $\mathbb{G}^A$. This construction is illustrated in Figure 6.

The following properties of partial duals follow directly from the definition: $\mathbb{G}^* = \mathbb{G}^{E(\mathbb{G})}$; $\mathbb{G}^{\emptyset} = \mathbb{G}$; $(\mathbb{G}^A)^B = \mathbb{G}^{(A \cup B)\backslash(A \cap B)}$ and so partial duals can be formed one edge at a time; partial duality acts disjointly on the connected components of a ribbon graph; and $\mathbb{G}^A$ is orientable if and only if $\mathbb{G}$ is. Another useful fact for us is that if $\mathbb{H}$ is a spanning ribbon subgraph of $\mathbb{G}$ with exactly one boundary component (for example, if $\mathbb{H}$ is a spanning tree) and A is the edge set of $\mathbb{H}$, then $\mathbb{G}^A$ is a bouquet (i.e., has exactly one vertex). This is because the vertices of $\mathbb{G}^A$ correspond to the boundary components of $\mathbb{H}$ (just as the vertices of $\mathbb{G}^*$ correspond to the boundary components of $\mathbb{G}$).

Bibliographic remarks. As with embedded graphs, ribbon graphs are standard objects in graph theory. They arise in several settings and under different names including *fat graphs*, *dessins d'enfants*, and *reduced band decompositions*. However, it should be remembered that they are just one of the many descriptions of embedded graphs. J. Ellis-Monaghan and I. Moffatt's book [29] offers an introduction to ribbon graphs and partial duals. Although we described partial duals in terms of ribbon graphs here, they can, of course, be described in other the models for embedded graphs. In particular, their local nature is prominent when they are defined in the languages of arrow presentations [18], graph encoded maps [28], or permutation models [20].

Partial duality was introduced by S. Chmutov in [18] in order to reconcile the various results in [19, 20, 26] which constructed the Jones polynomial of a knot or link as an evaluation of the Bollobás–Riordan polynomial of a ribbon graph. The Bollobás–Riordan polynomial of [4, 5] is a graph polynomial that offers an analogue of the Tutte polynomial [52] for embedded graphs. The connections between ribbon graphs and knot theory extend Thistlethwaite's well-known connection [48] between the Tutte polynomial of a plane graph and the Jones polynomial of an alternating link; a connection that was integral to his proof of the Tait conjectures. Chmutov used the term "generalized duality" in his original paper. Its adopted name 'partial duality' was suggested to the author of the present article by D. Archdeacon and has been used in all subsequent papers. Partial duality has since entered topological graph theory as a topic of study in its own right and is proving to be a fundamental operation on embedded graphs.

5. Ribbon graphs and their spanning quasi-trees

In the language of ribbon graphs, a *quasi-tree* is a ribbon graph that has exactly one boundary component. A ribbon subgraph $\mathbb{H}$ is a *spanning quasi-tree* of $\mathbb{G}$ if it is a quasi-tree that contains all of the vertices of $\mathbb{G}$. A ribbon graph of genus g has a spanning quasi-tree of genus $0, 1, \dots, g$, and its spanning trees are exactly its spanning quasi-trees of genus zero.

Recall from Section 2 that the set of spanning trees in a graph satisfies an exchange property: if T and T' are spanning trees and e is an edge in T but not T', then there is always some edge f in T' but not T such that removing e from T then adding f results in another spanning tree. This exchange property does not hold for spanning quasi-trees in general.

However, spanning quasi-trees satisfy a more general *symmetric exchange property*. If $\mathbb{H}$ and $\mathbb{H}'$ are spanning quasi-trees and e is an edge that is in exactly one of $\mathbb{H}$ or $\mathbb{H}'$, then there is always an edge f that is in exactly one of $\mathbb{H}'$ or $\mathbb{H}$ such that adding or removing each of e and f from $\mathbb{H}$ results in a spanning quasi-tree. Proving that this symmetric exchange property holds does require a little work. A proof can be found in [22] or implicitly in [12], or see [38, Figure 16] for a pictorial explanation. We shall return to this symmetric exchange property in the next section.

In Section 2 we used matroids to capture the spanning tree structure of a graph. A minor modification of the definition of a cycle matroid gives a way to similarly record the spanning quasi-trees in a ribbon graph.

Definition 5.1. Let $\mathbb{G} = (V, E)$ be a connected ribbon graph, and let

$$\mathcal{F} := \{F \subseteq E : F \text{ is the edge set of a spanning quasi-tree of } \mathbb{G}\}.$$

We call $D(\mathbb{G}) := (E, \mathcal{F})$ the *delta-matroid of* $\mathbb{G}$.

Example 5.2. Let $\mathbb{G}$ be the ribbon graph of Figure 3b. Then $D(\mathbb{G}) = (E, \mathcal{F})$ where $E = \{1, 2, \ldots, 6\}$ and $\mathcal{F} = \{\{2,4,5\}, \{2,4,6\}, \{3,4,5\}, \{3,4,6\}, \{4,5,6\}, \{1,2,3,4,5\}, \{1,2,3,4,6\}, \{1,2,4,5,6\}, \{2,3,4,5,6\}\}$.

Euler's formula gives that if $\mathbb{H}$ is an orientable quasi-tree with v vertices and e edges, then $(1 - v + e)/2$ gives the genus of $\mathbb{H}$ (or half its genus if $\mathbb{H}$ is non-orientable). As the spanning quasi-trees of $\mathbb{G}$ have the same number of vertices, this relates the sizes of the sets in $\mathcal{F}$ to the topology of the spanning quasi-trees. In particular, it follows that every set in $\mathcal{F}$ has the same parity (i.e., is of odd or even size) if and only if $\mathbb{G}$ is orientable, that the genus of $\mathbb{G}$ is one half of the differences in sizes between the largest and smallest sets in $\mathcal{F}$, and that for $\mathbb{G}$ connected, $D(\mathbb{G}) = C(\mathbb{G})$ if and only if $\mathbb{G}$ is plane.

Rephrased in terms of ribbon graphs, the map $*$ from Section 3 sends a spanning ribbon subgraph (V, A) of $\mathbb{G} = (V, E)$ to the spanning ribbon subgraph $(V^*, E \setminus A)$ of $\mathbb{G}^*$. Moreover, this map sends a spanning quasi-tree of genus g to a spanning quasi-tree of genus $n - g$ where n here is the genus of $\mathbb{G}$. Thus if $D(\mathbb{G}) = (E, \mathcal{F})$ and we define $\mathcal{F}^* := \{E \setminus F : F \in \mathcal{F}\}$, then for *any* ribbon graph $\mathbb{G}$ we have that $D(\mathbb{G}^*) = (E, \mathcal{F}^*)$. The main insights for quasi-tree structure, however, come from partial duals rather than geometric duals.

Partial duality preserves the quasi-tree structure of a ribbon graph. Let $\mathbb{G} = (V, E)$ be a ribbon graph and $B \subseteq E$. We shall relate the quasi-trees of $\mathbb{G}$ to those of its partial dual $\mathbb{G}^B$. For this recall that the *symmetric difference* $X \bigtriangleup Y$ of sets X and Y is $(X \cup Y) \setminus (X \cap Y)$. Then $A \subseteq E$ is the edge set of a quasi-tree of $\mathbb{G}$ if and only if $A \bigtriangleup B$ is the edge set of a quasi-tree of $\mathbb{G}^B$. It is not hard to see why this is the case—essentially it follows from the observation that the boundary components of $\mathbb{G}^{\{e\}}$ and $\mathbb{G}\backslash e$ correspond. In terms of the delta-matroids, this means that if $D(\mathbb{G}) = (E, \mathcal{F})$ and we set $\mathcal{F}^B := \{B \bigtriangleup F : F \in \mathcal{F}\}$, then $D(\mathbb{G}^B) = (E, \mathcal{F}^B)$.

The significance of this result is that if we wish to study the spanning quasi-trees of $\mathbb{G}$, we may equivalently study the spanning quasi-trees of any of its partial duals $\mathbb{G}^B$. The partial duals of a ribbon graph can have quite different properties from each other and from the original ribbon graph. This means that we have some ability to choose the ribbon graphs to work with without losing any generality, which is something we did not have much scope to do when working with geometric duals alone. A specific instance of this principle, and one that we shall make much use of here, is that every ribbon graph has a partial dual that is a bouquet (i.e., a one-vertex ribbon graph). Thus we only ever need to consider the spanning quasi-tree structure of bouquets. But to make use of this, we need a better understanding of $D(\mathbb{G})$.

Bibliographic remarks. The definition and approach to the delta-matroids of ribbon graphs that we follow here is due to C. Chun, I. Moffatt, S. Noble, and R. Rueckriemen [22, 23]. However, these delta-matroids are equivalent to A. Bouchet's delta-matroids of maps from [12]. There Bouchet associated a delta-matroid with the 4-regular medial graph of an embedded graph. The delta-matroid arises from its Eulerian circuits, and the Eulerian circuits correspond to the quasi-trees of the embedded graph. That $D(\mathbb{G})$ determines genus and orientability can be deduced from [12] through the correspondence ([22] gives the form stated here). The behaviour of $D(\mathbb{G})$ under partial duals is from [22].

6. Delta-matroids and quasi-tree structure

Recall from Section 3 that the dual of a matroid $M = (E, \mathcal{B})$ is $M^* = (E, \mathcal{B}^*)$ where $\mathcal{B}^* = \{E \setminus B : B \in \mathcal{B}\}$. We can write $\mathcal{B}^*$ as $\{E \bigtriangleup B : B \in \mathcal{B}\}$, and, in light of the above, it becomes obvious that we can form a *partial dual* of a matroid by replacing E with a subset X of E. So we can define a partial dual of $M = (E, \mathcal{B})$ as $M^X := (E, \mathcal{B}^X)$, where, as above, $\mathcal{B}^X := \{X \bigtriangleup B : B \in \mathcal{B}\}$.

For example, if $M = (\{1, 2\}, \{\{1\}, \{2\}\})$ and $X = \{1\}$, then a partial dual is $M^X = (\{1, 2\}, \{\{\emptyset\}, \{1, 2\}\})$. The difficulty, as can be seen in this example, is that M^X may no longer be a matroid. Instead it is an example of a more general structure called a *delta-matroid*.

Definition 6.1. A *delta-matroid* D consists of a pair $(E, \mathcal{F})$ where E is a finite set and $\mathcal{F}$ a non-empty collection of subsets of E. Furthermore, $\mathcal{F}$ is required to satisfy the *symmetric exchange axiom* which states that

$$(\forall X, Y \in \mathcal{F})\ (\forall u \in X \triangle Y)\ (\exists\, v \in X \triangle Y)\ \big(X \triangle \{u, v\} \in \mathcal{F}\big).$$

Since the collection of spanning quasi-trees of a ribbon graph $\mathbb{G}$ satisfies the symmetric exchange property, it follows that $D(\mathbb{G})$, as introduced in Definition 5.1, is a delta-matroid (and so the name we gave $D(\mathbb{G})$ is an honest one). Not every delta-matroid arises in this way, as the following example shows. In fact, almost all delta-matroids do not come from ribbon graphs, however those that do play an important role.

Example 6.2. Let $E = \{1,2,3,4\}$, $\mathcal{F} = \{\emptyset, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}\}$, and $\mathcal{F}' = \{\emptyset, \{1,2\}, \{1,4\}, \{2,3\}, \{3,4\}, \{1,2,3,4\}\}$. Then $(E, \mathcal{F})$ and $(E, \mathcal{F}')$ are both delta-matroids but neither is the delta-matroid of a ribbon graph. This can be verified by calculating the delta-matroids of the bouquets on four edges.

Matroids are also examples of delta-matroids: M is matroid if and only if it is a delta-matroid in which every member of $\mathcal{F}$ has the same size. Most delta-matroids are not matroids though.

While the class of matroids is not closed under partial duals, the class of delta-matroids is. Let $D = (E, \mathcal{F})$ be a delta-matroid and $B \subseteq E$. The *partial dual* (or *twist*) D^B of D is defined as the pair $(E, \mathcal{F}^B)$ where $\mathcal{F}^B := \{F \triangle B : F \in \mathcal{F}\}$. The *dual* D^* of D is D^E.

Example 6.3. If D is the delta-matroid from Example 5.2, then $D^{\{3,4\}} = (E, \mathcal{F}^{\{3,4\}})$ where $E = \{1, \dots, 6\}$ and $\mathcal{F}^{\{3,4\}} = \{\{2,3,5\}, \{2,3,6\}, \{5\}, \{6\}, \{3,5,6\}, \{1,2,5\}, \{1,2,6\}, \{1,2,3,5,6\}, \{2,5,6\}\}$.

Matroid duality captures the way that the spanning trees of a *plane* graph $\mathbb{G}$ are transformed into the spanning trees of its geometric dual $\mathbb{G}^*$, giving the identity $C(\mathbb{G}^*) = C(\mathbb{G})^*$ for plane graphs. Delta-matroid duality captures that the spanning quasi-trees of *any* ribbon graph $\mathbb{G}$ are transformed into the spanning quasi-trees of any partial dual $\mathbb{G}^B$. Indeed the following results follow from our previous discussion.

Theorem 6.4. *Let $\mathbb{G}$ be a connected ribbon graph. Then*

(1) $C(\mathbb{G}^*) = C(\mathbb{G})^*$ *if and only if $\mathbb{G}$ is a* plane *ribbon graph,*

(2) $D(\mathbb{G}^*) = D(\mathbb{G})^*$ *for* any *ribbon graph $\mathbb{G}$, and*

(3) $D(\mathbb{G}^B) = D(\mathbb{G})^B$ *for* any *ribbon graph $\mathbb{G}$ and* any *subset of its edges B.*

Just as with ribbon graphs, we can use partial duality to transform a delta-matroid into one with desirable properties. A delta-matroid $D = (E, \mathcal{F})$ is said to be *normal* if $\emptyset \in \mathcal{F}$. Every delta-matroid has a normal partial dual: if $D = (E, \mathcal{F})$ and F is any element of $\mathcal{F}$, then D^F is normal. On the other hand, some properties are preserved by partial duals. For example, a delta-matroid $D = (E, \mathcal{F})$ is said to be *even* if every set in $\mathcal{F}$ has the same parity (i.e., they are all of odd size or all of even size). If a delta-matroid is even, then so is each of its partial duals.

By making use of the properties of spanning quasi-trees we observe that for a connected ribbon graph $\mathbb{G}$, the delta-matroid $D(\mathbb{G})$ is even if and only if $\mathbb{G}$ is orientable, and that $D(\mathbb{G})$ is normal if and only if $\mathbb{G}$ is a bouquet. As we are restricting to orientable ribbon graphs here, we shall focus on even delta-matroids.

Bibliographic remarks. Delta-matroids were introduced in the mid-1980s, independently, by A. Bouchet in [6]; R. Chandrasekaran and S. Kabadi, under the name of *pseudo-matroids*, in [17]; and A. Dress and T. Havel, under the name of *metroids*, in [27]. Delta-matroids are related to many different matroidal-objects, including É. Tardos' g-matroids [47], J. Kung's Pfaffian structures [37], L. Qi's ditroids [44], A. Bouchet's symmetric matroids [6], L. Traldi's transition matroids [49], Bouchet's isotropic systems [7], jump systems [15], and Bouchet's multimatroids [13]. This list is indicative, not exhaustive.

The discipline has adopted Bouchet's terminology and notation (most of the early development of the topic is due to Bouchet and his collaborators) and it is that we follow here except in the following instance. What we have called the "partial dual" and denoted by D^B is usually called a "twist" and denoted by $D * A$, but here we prefer to keep close to the ribbon graph terminology.

Bouchet, in [6], showed that the partial dual of a delta-matroid is indeed a delta-matroid. That $D(\mathbb{G}^*) = D(\mathbb{G})^*$ is implicit in [12] (it was translated into this form in [22]), and that $D(\mathbb{G}^B) = D(\mathbb{G})^B$ is from [22].

Additional background on delta-matroids can be found in the survey [38] or in the source papers.

7. Matrices and representability

We are interested in the spanning quasi-trees of a connected orientable ribbon graph $\mathbb{G}$. Since $D(\mathbb{G}^B) = D(\mathbb{G})^B$, partial duality preserves the spanning quasi-tree structure and so, without loss of generality, we may assume that $\mathbb{G}$ is a bouquet. Then the ribbon subgraph of $\mathbb{G}$ induced by any two of its edges forms either a genus one or a genus zero ribbon graph. We say that two edges of $\mathbb{G}$ are *interlaced* if the ribbon subgraph $\mathbb{G}$ they induce has genus one.

There is a method from algebraic topology (e.g., see [3, Theorem 3] and its subsequent exercises) for determining via a matrix if an orientable bouquet is a quasi-tree. Let $\mathbb{G} = (V, E)$ be an orientable bouquet. Number the edges of $\mathbb{G}$ by travelling around the boundary of the vertex from an arbitrary starting point in either direction and assigning the numbers $1, 2, \dots, |E|$ in the order that you first encounter one of their ends. Now construct an $|E| \times |E|$-matrix $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$ by setting the (i, j)-entry to be $\mathrm{sgn}(i - j)$ if the edges labelled i and j are interlaced, and 0 otherwise. (Here sgn is the signum function.) Then $\det(\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}) = 1$ if $\mathbb{G}$ is a quasi-tree and is 0 otherwise.

This construction can be simplified by working over the field of two elements, GF(2). In this case, as we are forgetting the signs, we can construct an $|E| \times |E|$-matrix $\mathbf{IM}_{\mathbb{G}}$ whose rows and columns are indexed by the edges of $\mathbb{G}$ by setting the (e, f)-entry to be 1 if edges e and f are interlaced, and to be 0 otherwise. Again $\det(\mathbf{IM}_{\mathbb{G}}) = 1$ if $\mathbb{G}$ is a quasi-tree and is 0 otherwise, where here we compute the determinant over GF(2).

The matrices $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$ and $\mathbf{IM}_{\mathbb{G}}$ in fact determine the whole spanning quasi-tree structure of $\mathbb{G}$ (although not $\mathbb{G}$ itself). This is since we can test if a ribbon subgraph $\mathbb{H}$ of $\mathbb{G}$ is a quasi-tree by computing the determinant of the principal submatrix given by the edges of $\mathbb{H}$ (delete any rows and columns of $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$ or $\mathbf{IM}_{\mathbb{G}}$ that correspond to edges not in $\mathbb{H}$).

Thus the delta-matroid $D(\mathbb{G})$ can be recovered from the matrices $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$ or $\mathbf{IM}_{\mathbb{G}}$ by computing determinants of their principal submatrices over $\mathbb{R}$ or GF(2), respectively. These matrices provide what is known as a *representation* of the delta-matroid $D(\mathbb{G})$.

Before continuing let us highlight one issue with this approach to studying spanning quasi-trees via matrices. As the matrices are only defined on bouquets, if we are interested in a ribbon graph $\mathbb{G}$ that has more than one vertex, then we can obtain a matrix by choosing a one-vertex partial dual of $\mathbb{G}$ and computing a matrix from that. However, different choices of partial dual will result in different matrices, so we will need to understand how the matrices change under this choice.

A matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A}^t = \mathbf{A}$ and is *skew-symmetric* if $\mathbf{A}^t = -\mathbf{A}$ and the diagonal entries are zero. (The condition on the diagonal is there for fields of characteristic 2.) Suppose that $\mathbf{A}$ is a symmetric or skew-symmetric matrix over a field $\Bbbk$, and that a set E labels its rows and columns (in the same order). For $X \subseteq E$, let $\mathbf{A}[X]$ denote the principal submatrix of $\mathbf{A}$ given by the rows and columns indexed by X. Define a collection $\mathcal{F}$ of subsets of E by

$$X \in \mathcal{F} \Leftrightarrow \mathbf{A}[X] \text{ is non-singular,}$$

where $\mathbf{A}[\emptyset]$ is considered to be non-singular. Then the pair $D(\mathbf{A}) := (E, \mathcal{F})$ forms a delta-matroid. (This result is due to A. Bouchet [11].)

Since the principal submatrices of $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$ or $\mathbf{IM}_{\mathbb{G}}$ are non-singular precisely when the corresponding edge sets of $\mathbb{G}$ define a quasi-tree, it follows that when $\mathbb{G}$ is an orientable bouquet,

$$D(\mathbb{G}) = D(\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}) = D(\mathbf{IM}_{\mathbb{G}}),$$

where we work over $\mathbb{R}$ for the middle expression and GF(2) for the one on the right.

Since $\mathbf{A}[\emptyset]$ is non-singular, such a delta-matroid $D(\mathbf{A})$ is necessarily normal. We say a normal delta-matroid is *representable* if it can be obtained as the delta-matroid of a matrix. Every delta-matroid is a partial dual of a normal delta-matroid, so we can extend representability to non-normal delta-matroids by saying that a delta-matroid is *representable* if one of its partial duals is the delta-matroid of a matrix.

Definition 7.1. Let $D = (E, \mathcal{F})$ be a delta-matroid. We say that D is *representable* over $\Bbbk$ if there exists some $X \subseteq E$ and a symmetric or skew-symmetric matrix $\mathbf{A}$ over a field $\Bbbk$ such that

$$D^X = D(\mathbf{A}).$$

A delta-matroid is *binary* if it is representable over GF(2), and is *regular* if it is representable over $\mathbb{R}$. Delta-matroids of orientable ribbon graphs are binary since

$$D(\mathbb{G})^X = D(\mathbb{G}^X) = D(\mathbf{IM}_{\mathbb{G}^X}),$$

where X is the edge set of any spanning quasi-tree of $\mathbb{G}$. Similarly, the matrix $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}^X}$ shows that they are regular. (We note that orientability matters here as delta-matroids of non-orientable ribbon graphs are not regular, although they are binary.)

The definition of representability for delta-matroids requires a choice of a set X to make D^X normal. In general, there are many such sets to choose from, and therefore a delta-matroid D will have many representing matrices. How do the different representing matrices of a delta-matroid relate? That is, if $D(\mathbf{A}) = D(\mathbf{B})^X$ what can you say about the matrices $\mathbf{A}$ and $\mathbf{B}$?

The relevant matrix operation predates delta-matroids and can be found in the work of A. Tucker [51] that appeared in 1960. Let $\mathbf{A}$ be a square matrix over a field $\Bbbk$, whose rows and columns are labelled (in the same order) by a set E. Let $X \subseteq E$. Without loss of generality (reordering if necessary), suppose that X labels the first $|X|$ rows and columns of the matrix. Then $\mathbf{A}$ has a block form

$$\begin{array}{c|c|c} & X & E \setminus X \\ X & \alpha & \beta \\ \hline E \setminus X & \gamma & \delta \end{array}.$$

Suppose that $\mathbf{A}[X]$ is non-singular. Then the *principal pivot transform* of $\mathbf{A}$ with respect to X is the matrix $\mathbf{A} * X$ with block form

$$\begin{array}{c|c|c} & X & E \setminus X \\ X & \alpha^{-1} & \alpha^{-1}\beta \\ \hline E \setminus X & -\gamma\alpha^{-1} & \delta - \gamma\alpha^{-1}\beta \end{array}.$$

A. Bouchet, in [11], proved that principal pivot transformations correspond to partial duals of delta-matroids.

Theorem 7.2. *Let* $\mathbf{A}$ *be a symmetric or skew-symmetric matrix over a field* $\Bbbk$*, whose rows and columns are labelled (in the same order) by a set* E*. Let* $X \subseteq E$ *be such that* $\mathbf{A}[X]$ *is non-singular. Then* $\mathbf{A} * X$ *is a symmetric or skew-symmetric matrix (of the same type as* $\mathbf{A}$*), and*

$$D(\mathbf{A} * X) = D(\mathbf{A})^X. \tag{7.1}$$

Thus if $\mathbf{A}$ is a representing matrix for a delta-matroid D, then $\mathbf{B}$ is also a representing matrix for D if and only if $\mathbf{B}$ is a principal pivot transform of $\mathbf{B}$. Thus we have our answer to the problem in this section: all of the representing matrices for an orientable ribbon graph $\mathbb{G}$ are principal pivot transformations of one another.

Bibliographic remarks. That $D(\mathbf{A})$ is a delta-matroid, that $D(\mathbf{A} * X) = D(\mathbf{A})^X$, and the definition of representability is due to A. Bouchet and from [11]. The representations for $D(\mathbb{G})$ can also be deduced from this reference (see also [9] for $\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}$), although changes in language are needed (the interpretation in ribbon graph language is from [22]). However, a different route to showing that $D(\mathbb{G}) = D(\mathbf{IM}^{\mathcal{O}}_{\mathbb{G}}) = D(\mathbf{IM}_{\mathbb{G}})$ was taken in this section. Here we deduced the result from a theorem on weight systems of Vassiliev invariants due to D. Bar-Natan and S. Garoufalidis [3]. This knot theory work seems to be entirely independent of Bouchet's work.

8. The reappearance of graphs

So far we have seen that the spanning quasi-tree structure of an orientable ribbon graph $\mathbb{G}$ is described by its delta-matroid $D(\mathbb{G})$, and also by a binary representing matrix $\mathbf{IM}_{\mathbb{H}}$, where $\mathbb{H}$ is any one-vertex partial dual of $\mathbb{G}$. The matrix $\mathbf{IM}_{\mathbb{H}}$ is a skew-symmetric 0-1 matrix. (Recall that skew-symmetric matrices here must have zeros on their diagonal.) Thus we can consider it as the adjacency matrix of a simple graph G. (A graph is *simple* if it does not have multiple edges or loops.) In this section we consider the properties of these simple graphs and what they tell us about ribbon graphs.

The *adjacency matrix*, $\mathbf{AM}_G$, of a simple graph G is the matrix over GF(2) whose rows and columns correspond to the vertices of G; and whose (u, v)-entry is 1 if there is an edge uv in G and is 0 otherwise.

Adjacency matrices are skew-symmetric, and every skew-symmetric matrix over GF(2) is an adjacency matrix of some simple graph. This results in a 1-1 correspondence between skew-symmetric binary matrices and simple graphs. Every skew-symmetric binary matrix $\mathbf{A}$ gives rise to a normal even binary delta-matroid $D(\mathbf{A})$. (The delta-matroid must be even since odd order skew-symmetric matrices are always

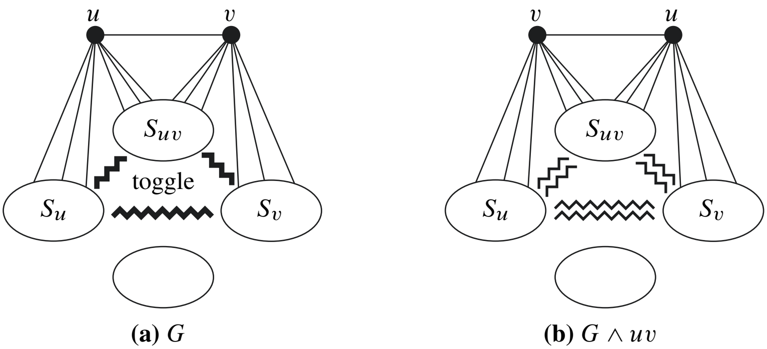


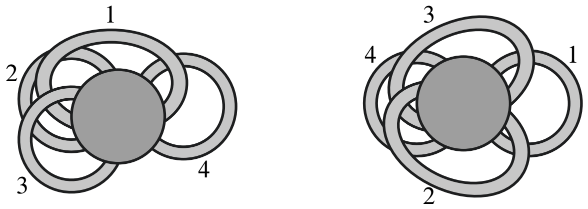
Figure 7. Pivoting (edges between the three sets, S_u, S_v, and $S_{u,v}$, are "toggled," and the names of u and v are switched).

singular.) On the other hand, a normal even binary delta-matroid D determines a unique skew-symmetric matrix $\mathbf{A}$ such that $D = D(\mathbf{A})$. (If $D = (E, \mathcal{F})$ is binary, then it must come from a binary matrix, and the sets of size two in $\mathcal{F}$ determine which entries of this matrix are zero and which are one.) This means that there is a 1-1 correspondence between simple graphs and normal even binary delta-matroids.

However, we want to work with all even binary delta-matroids not just normal ones. Obtaining a representing matrix for an arbitrary binary even delta-matroid D requires choosing a normal partial dual of it. Different choices will result in different matrices, however, from the results of Section 7, we know that these matrices will be related through principal pivot transforms. How are the simple graphs corresponding to these two matrices related? Once again we can find the relevant operation in the literature in a move introduced by A. Bouchet in [10, 14] and rediscovered by R. Arratia, B. Bollobás, and G. Sorkin in [1, 2].

Definition 8.1. Let G be a simple graph and uv an edge. Partition the vertices other than u and v into four classes: (1) vertices adjacent to u but not v, (2) vertices adjacent to v but not u, (3) vertices adjacent to both u and v, (4) vertices adjacent to neither u nor v. The *pivot* of the edge uv is the graph, $G \wedge uv$, constructed from G as follows. For any vertex pair x, y where x is in one of the classes (1)–(3), and y is in a different class (1)–(3), "toggle" the pair xy in the edge set (so if xy was an edge, make it a non-edge; and if xy was a non-edge, make it an edge). Finally, switch the names of the vertices u and v; see Figure 7.

Suppose G is a simple graph with adjacency matrix $\mathbf{AM}_G$, and uv is an edge of G. Then the principal submatrix $\mathbf{AM}_G[\{u, v\}]$ defined by the edge has zeros on the diagonal and ones elsewhere and is hence non-singular. This means we can form the



(a) $\mathbb{G}$ **(b)** $\mathbb{H}$

Figure 8. Two bouquets.

principal pivot transform $\mathbf{AM}_G * \{u, v\}$ of $\mathbf{AM}_G$. This changes the matrix in a very nice way and it is not too hard an exercise (remembering we are working over GF(2)) to track this change through to the corresponding simple graphs: the graphs will be pivots of one another. Passing to delta-matroids, for an edge uv of G we have that

$$D(\mathbf{AM}_G)^{\{u,v\}} = D\big(\mathbf{AM}_G * \{u, v\}\big) = D(\mathbf{AM}_{G\wedge uv}).$$

Thus we can identify even binary delta-matroids up to partial duals with simple graphs up to pivoting:

$$\left\{\begin{matrix}\text{even binary delta-matroids}\\ \text{up to partial duals}\end{matrix}\right\} \xleftrightarrow{\text{1-1}} \left\{\begin{matrix}\text{simple graphs}\\ \text{up to edge pivots}\end{matrix}\right\}.$$

As edge pivoting is of interest in graph theory in its own right, this identification opens up a new body of graph theory for studying delta-matroids, and vice versa. However, there is a catch when we want to use simple graphs and edge pivots to study ribbon graphs and their spanning quasi-trees. Although the delta-matroid $D(\mathbb{G})$ of an orientable ribbon graph is even and binary, not all even and binary delta-matroids arise from ribbon graphs. This means that the delta-matroids of ribbon graphs correspond with a proper subclass of simple graphs. We turn our attention to this class in the next section.

Example 8.2. As an illustration of the discussion from Section 6 onwards, consider the bouquets $\mathbb{G}$ and $\mathbb{H}$ of Figure 8. Both are on the edge set $E = \{1, 2, 3, 4\}$. Their binary representing matrices are

$$\mathbf{IM}_{\mathbb{G}} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4\\ \hline 1 & 0 & 1 & 1 & 1\\ 2 & 1 & 0 & 1 & 0\\ 3 & 1 & 1 & 0 & 0\\ 4 & 1 & 0 & 0 & 0 \end{array} \quad \text{and} \quad \mathbf{IM}_{\mathbb{H}} = \begin{array}{c|cccc} & 1 & 2 & 3 & 4\\ \hline 1 & 0 & 1 & 1 & 0\\ 2 & 1 & 0 & 1 & 1\\ 3 & 1 & 1 & 0 & 1\\ 4 & 0 & 1 & 1 & 0 \end{array}.$$

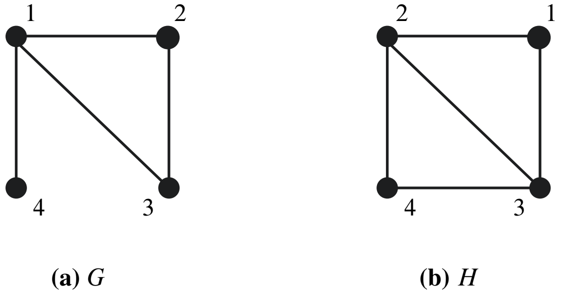


Figure 9. Two simple graphs.

Now let G and H be the simple graphs in Figure 9. It is readily checked that $\mathbf{AM}_G = \mathbf{IM}_{\mathbb{G}}$ and $\mathbf{AM}_H = \mathbf{IM}_{\mathbb{H}}$.

By direct computation from the bouquets and matrices we see that $D(\mathbb{G}) = D(\mathbf{IM}_{\mathbb{G}}) = (E, \mathcal{F}_{\mathbb{G}})$ and $D(\mathbb{H}) = D(\mathbf{IM}_{\mathbb{H}}) = (E, \mathcal{F}_{\mathbb{H}})$ where $\mathcal{F}_{\mathbb{G}} = \{\emptyset, \{1,2\}, \{1,3\}, \{2,3\}, \{1,4\}, \{1,2,3,4\}\}$ and $\mathcal{F}_{\mathbb{H}} = \{\emptyset, \{1,2\}, \{1,3\}, \{2,3\}, \{2,4\}, \{3,4\}\}$.

The bouquets $\mathbb{G}$ and $\mathbb{H}$ are partial duals with $\mathbb{H} = \mathbb{G}^{\{1,2\}}$. In addition, the matrices $\mathbf{IM}_{\mathbb{G}}$ and $\mathbf{IM}_{\mathbb{H}}$ can be verified as principal pivot transforms with $\mathbf{IM}_{\mathbb{H}} = \mathbf{IM}_{\mathbb{G}} * \{1,2\}$, and G and H are pivots with $H = G \wedge 12$. Thus we can see that

$$D\big(\mathbb{G}^{\{1,2\}}\big) = D(\mathbb{G})^{\{1,2\}} = D\big(\mathbf{IM}_{\mathbb{G}} * \{1,2\}\big) = D(\mathbf{AM}_{G\wedge 12}),$$

and we can work with spanning quasi-trees in any of the settings.

Bibliographic remarks. Pivoting is a graph operation related to A. Kotzig's transformations on Eulerian circuits [36]. It was introduced by A. Bouchet in the context of isotropic systems [10] and multimatroids [14], and rediscovered by R. Arratia, B. Bollobás, and G. Sorkin when they introduced the interlace polynomial in [1,2].

Further information on binary delta-matroids can be found in [16]. In particular, this reference contains the result that a normal binary delta-matroid $(D, \mathcal{F})$ is completely determined by the members of $\mathcal{F}$ of size at most two.

The identification of even binary delta-matroids considered up to partial duals with simple graphs considered up to edge pivots can be extended to all binary delta-matroids. They can be identified with looped simple graphs considered up to *elementary pivots* which are pivots on edges not adjacent to loops, and a *local complementation* move (toggle the edges and non-edges, and loops and non-loops in the neighbourhood of a looped vertex). This identification was first written down by J. Geelen in [33] (see also [32]) although he has said that the graph-theoretical point of view was used by both A. Bouchet and W. Cunningham in their discussions with him at the time of writing that paper.

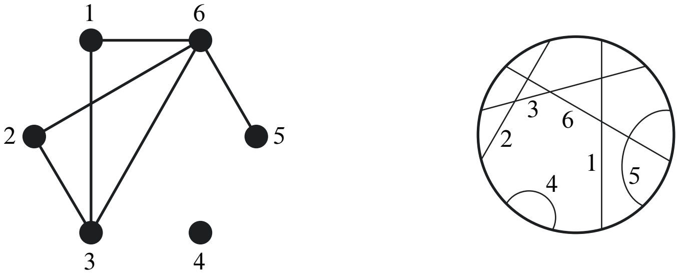


Figure 10. A circle graph and a corresponding chord diagram.

9. Bringing it all together

A *chord diagram* consists of a circle in the plane and a number line segments, called *chords*, whose end-points lie on the circle. The end-points of chords should all be distinct. The *intersection graph* of a chord diagram is the graph $G = (V, E)$ where V is the set of chords, and where $uv \in E$ if and only if the chords u and v intersect. A graph is a *circle graph* if it is the intersection graph of a chord diagram. Figure 10 shows a circle graph and a corresponding chord diagram.

Now suppose that $\mathbb{G}$ is an orientable bouquet. We may regard $\mathbb{G}$ as a chord diagram with the vertex boundary forming the circle and chords defined by where the edges touch this circle. Let $I_{\mathbb{G}}$ denote the corresponding intersection graph. There is an edge ef of $I_{\mathbb{G}}$ whenever the edges e and f are interlaced in $\mathbb{G}$. In terms of the delta-matroid $D(\mathbb{G}) = (E, \mathcal{F})$ this means that there is an edge ef of $I_{\mathbb{G}}$ whenever $\{e, f\}$ is in $\mathcal{F}$. Thus, since $D(\mathbb{G})$ is binary, we can obtain a binary representing matrix $\mathbf{A}$ for $D(\mathbb{G})$ by setting the (e, f)-entry to be 1 if ef is an edge in $I_{\mathbb{G}}$ and 0 otherwise, so $\mathbf{A}$ is the adjacency matrix of $I_{\mathbb{G}}$. Thus the intersection graph $I_{\mathbb{G}}$ of $\mathbb{G}$ is exactly the simple graph corresponding to the delta-matroid $D(\mathbb{G})$. (As an example, it can be checked that $G = I_{\mathbb{G}}$ and $H = I_{\mathbb{H}}$ in Example 8.2.)

We can then conclude that circle graphs are exactly the simple graphs that represent the delta-matroids of orientable ribbon graphs:

$$\left\{\begin{matrix}\text{Delta-matroids of orientable ribbon graphs}\\ \text{up to partial duals}\end{matrix}\right\} \xleftrightarrow{\text{1-1}} \left\{\begin{matrix}\text{circle graphs}\\ \text{up to edge pivots}\end{matrix}\right\}.$$

Circle graphs are well studied in graph theory and their appearance in the present setting provides access to a large body of work that we can apply to ribbon graphs. Let us take advantage of this to characterise the delta-matroids that arise from ribbon graphs.

A *minor* of a graph is any graph that can be obtained from it by edge deletion (remove an edge), vertex deletion (remove a vertex and the edges it meets), and edge contraction (delete the edge then identify its ends). An excluded minor characterisa-

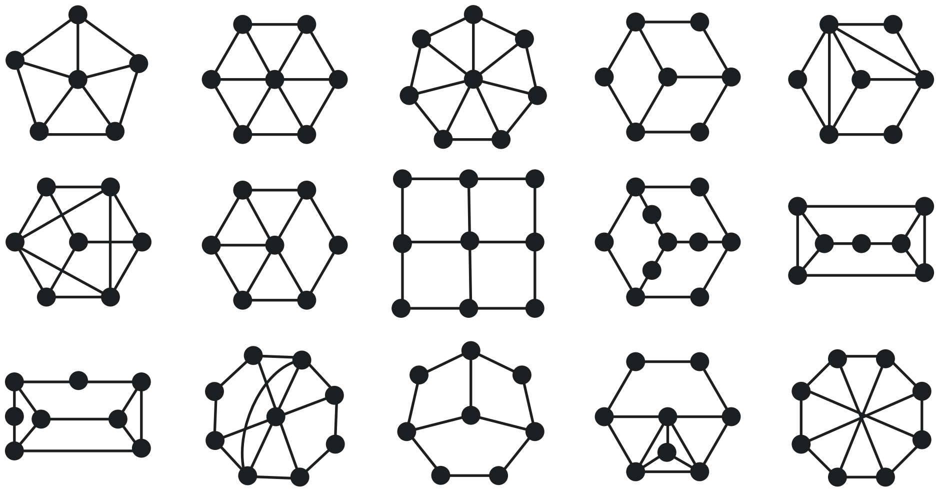


Figure 11. Excluded pivot-minors for circle graphs.

tion of a class of graphs is a result that states that a graph belongs to the class if and only if it has no minor in a given finite list. Possibly the best known example of an excluded minor characterisation is Wagner's theorem which states that a graph is planar if and only if it has no minor isomorphic to K_5 (the graph of five vertices and one edge between each pair of vertices) or $K_{3,3}$ (the graph with two sets of three vertices and an edge between all pairs of vertices in different sets). (The name Kuratowski's theorem, which uses a different type of minor, is often associated with this result.) The spectacular Robertson–Seymour theorem gives that every minor-closed class of graphs has an excluded minor characterisation [45].

Circle graphs, however, are not closed under the usual graph minor operations, and so it does not make sense to ask for an excluded minor characterisation of them with the usual type of graph minor. However, the set of circle graphs is closed under edge pivots and vertex deletions which leads to a different type of graph minor.

A *pivot-minor* of a graph is any graph that can be obtained from it by edge pivots and vertex deletions. Circle graphs have an excluded pivot-minor characterisation. J. Geelen and S. Oum [32] proved that a graph is a circle graph if and only if it has no pivot-minor isomorphic to any of the graphs shown in Figure 11.

We can use the correspondence between delta-matroids and simple graphs to derive an excluded minor characterisation for the class of delta-matroids that arise from ribbon graphs. For this we need delta-matroid versions of the vertex minor operations. We know from Section 8 that the delta-matroid version of an edge pivot is a partial dual. Vertex deletion corresponds to the standard idea of deletion for delta-matroids.

Let $D = (E, \mathcal{F})$ be a delta-matroid, and let $e \in E$. Then D *delete* e, denoted by $D \backslash e$, is defined as $D \backslash e := (E \backslash e, \mathcal{F}')$, where $\mathcal{F}' = \{F : F \in \mathcal{F} \text{ and } e \notin F\}$ when e is not in every member of $\mathcal{F}$; and $\mathcal{F}' = \{F \backslash e : F \in \mathcal{F} \text{ and } e \in F\}$ e is in every member of $\mathcal{F}$. Although we do not use the fact here, it is worth noting that $D(\mathbb{G} \backslash e) = D(\mathbb{G}) \backslash e$. A delta-matroid D' is said to be a *minor* of a delta-matroid D if it can be obtained from D through the operations of deletion and partial duality.

By translating the excluded pivot-minor characterisation of circle graphs we obtain the following characterisation of the even delta-matroids that arise from ribbon graphs.

Theorem 9.1. *Let D be an even delta-matroid. Then $D = D(\mathbb{G})$ for some ribbon graph $\mathbb{G}$ if and only if it has no minor isomorphic to $D(\mathbf{AM}_G)$ where G is one of the graphs shown in Figure 11, or to one of the delta-matroids given in Example 6.2.*

The excluded minors from Example 6.2 are included to ensure that an even delta-matroid is binary and hence comes from a simple graph.

Finally we come to the question from which our journey into delta-matroids began: *Do the spanning quasi-trees of an embedded graph determine it?* In terms of delta-matroids we are asking:

If $D(\mathbb{G}) = D(\mathbb{H})$, then how are the ribbon graphs $\mathbb{G}$ and $\mathbb{H}$ related?

So we are looking for a version of Whitney's theorem that applies to ribbon graphs and their delta-matroids.

Again we can make use of the circle graph literature. There has been extensive work on recovering chord diagrams from circle graphs, and on determining which chord diagrams correspond to the same circle graph. Appearing implicitly in [8, 24, 30], and explicitly in [21], is an operation on chord diagrams called *mutation* that relates all chord diagrams that have the same intersection graph. This operation cuts out a certain substructure in a chord diagram, rotates it then glues it back in (we omit a definition of the move as we do not use its details here). The result uses Cunningham's theory of graph decompositions from [25] to decompose an intersection graph into "prime" graphs that have unique intersection graphs. Mutation then corresponds to the choices that are made when reassembling a corresponding chord diagram from these prime graphs.

In the present setting, if two ribbon graphs $\mathbb{G}$ and $\mathbb{H}$ have equal delta-matroids, then there must be some set of edges X such that the partial duals $\mathbb{G}^X$ and $\mathbb{H}^X$ are both bouquets with the same delta-matroid. The delta-matroids $D(\mathbb{G}^X)$ and $D(\mathbb{H}^X)$ therefore correspond to the same simple graph. As this simple graph can be considered as the intersection graphs of $\mathbb{G}^X$ and $\mathbb{H}^X$, it follows that $\mathbb{G}^X$ and $\mathbb{H}^X$ must be related by mutation (technically, a version of mutation for bouquets). Then by analysing how mutation changes under partial duality, we can pull back the operations

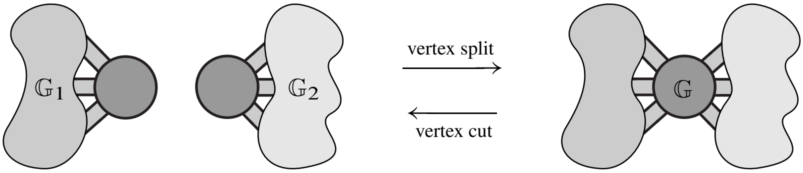


Figure 12. Vertex joins and vertex splits.

to the original ribbon graphs $\mathbb{G}$ and $\mathbb{H}$. This approach results in a characterisation of ribbon graphs that have the same delta-matroid. We describe the relevant moves then the characterisation. The first move is the analogue of the vertex identification and vertex cleaving that are used in Whitney's theorem and illustrated in Figure 1.

Suppose that $\mathbb{G}_1$ and $\mathbb{G}_2$ are ribbon graphs. For $i = 1, 2$, suppose that α_i is an arc that lies on the boundary of $\mathbb{G}_i$ and entirely on a vertex boundary. If a ribbon graph $\mathbb{G}$ can be obtained from $\mathbb{G}_1$ and $\mathbb{G}_2$ by identifying the arc α_1 with α_2 (where the identification merges the vertices), then we say that $\mathbb{G}$ is obtained from $\mathbb{G}_1$ and $\mathbb{G}_2$ by a *vertex join*, and that $\mathbb{G}_1$ and $\mathbb{G}_2$ are obtained from $\mathbb{G}$ by a *vertex split*. The operations are illustrated in Figure 12 and are standard operations in ribbon graph theory. It is important to observe that the definition of a vertex join does not allow for any "interlacing" of the edges of G_1 and G_2.

The next operation we need is called *mutation*. It is illustrated in Figure 13. The figure shows a local change in a ribbon graph (so the ribbon graphs are identical outside of the region shown) and the two parts of vertices that are shown in it may come from the same vertex. To define the move, let $\mathbb{G}_1$ and $\mathbb{G}_2$ be ribbon graphs. For $i = 1, 2$, let α_i and β_i be two disjoint directed arcs that lie on the boundary of $\mathbb{G}_i$ and lie entirely on boundaries of (one or two) vertices. Furthermore, suppose that $\mathbb{G}$ is a ribbon graph that is obtained by identifying the arcs α_1 with α_2, and β_1 with β_2, where both identifications are consistent with the direction of the arcs. (The identification merges the vertices.) Suppose further that $\mathbb{H}$ is a ribbon graph obtained by either (1) identifying α_1 with α_2, and β_1 with β_2, where the identifications are inconsistent with the direction of the arcs, (2) identifying α_1 with β_2, and β_1 with α_2, where the identifications are consistent with the direction of the arcs or (3) identifying α_1 with β_2, and β_1 with α_2, where the identifications are inconsistent with the direction of the arcs. Then we say that $\mathbb{G}$ and $\mathbb{H}$ are related by *mutation*.

With these definitions in hand, we can complete our tour with an answer (due to I. Moffatt and J. Oh [39]) to our original question as to what extent the spanning quasi-trees determine the ribbon graph.

Theorem 9.2. *Let $\mathbb{G}$ and $\mathbb{H}$ be connected orientable ribbon graphs, and let $D(\mathbb{G})$ and $D(\mathbb{H})$ be their delta-matroids. Then $D(\mathbb{G}) = D(\mathbb{H})$ if and only if $\mathbb{G}$ can be obtained from $\mathbb{H}$ by ribbon graph isomorphism, vertex joins, vertex splits or mutation.*

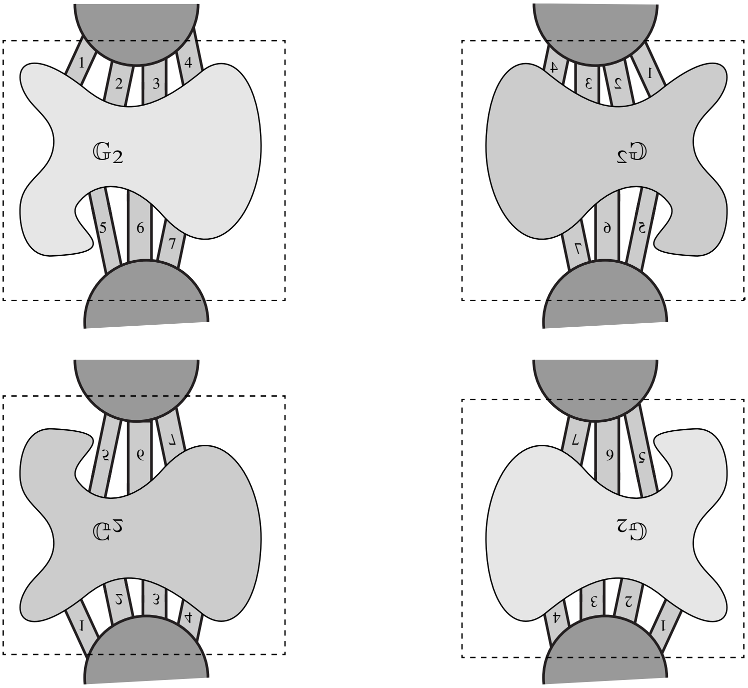


Figure 13. Mutation for ribbon graphs.

As an example of Theorem 9.2, the two non-equivalent ribbon graphs in Figure 14 can be obtained from each other by isomorphism, vertex joins, vertex splits, and mutation. Therefore their delta-matroids are isomorphic.

Bibliographic remarks. The excluded minor characterisation for the delta-matroids of orientable ribbon graphs stated in Theorem 9.1 is implicit in J. Geelen and S. Oum's paper [32]. There it was stated for even Eulerian delta-matroids which, from [22], are equivalent to the delta-matroids of ribbon graphs. The ribbon graph formulation given here is from [22]. The characterisation extends to non-orientable ribbon graphs. Again this was given in for Eulerian delta-matroids in [32] and translated to the ribbon graph setting in [22]. There are 171 excluded minors in this case.

The excluded minor characterisation of binary delta-matroids alluded to after the statement of Theorem 9.1 is due to A. Bouchet and A. Duchamp [16]. There are five excluded minors for binary delta-matroids, and the two appearing in Example 6.2 are the even ones.

Theorem 9.2 is due to I. Moffatt and J. Oh, and from [39]. It is given there more generally for non-orientable and non-connected ribbon graphs. Extending to the non-connected case is straightforward, but additional work is required for the non-orientable case.

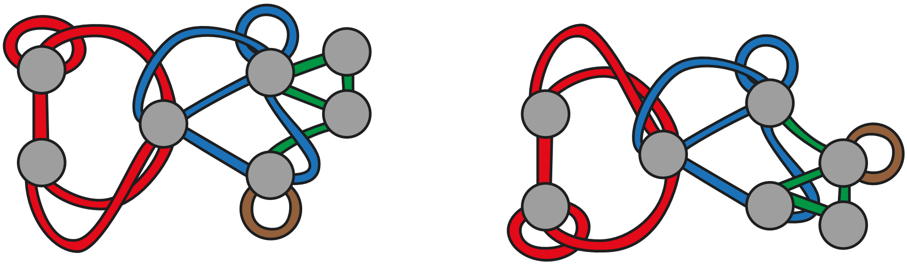


Figure 14. Two ribbon graphs with the same delta-matroid.

10. Now we can get started . . .

We set out with the classical question of whether the spanning trees in a graph determine the graph itself. This led to a topological version of it, if the spanning quasi-trees in a ribbon graph determine it. In answering this question we were guided by the idea of partial duality which appeared in different forms and settings. This took us to ribbon graphs, matroids and delta-matroids, matrices, as well as simple and circle graphs. Moreover, we saw that delta-matroids provided the central unifying framework for all of these ideas. It is this common framework that we should really take away from our journey.

As mentioned earlier, there is a well-known and successful symbiotic relationship between graph theory and matroid theory, with each area informing the other. As reported in [42], W. Tutte famously observed that, “If a theorem about graphs can be expressed in terms of edges and circuits alone it probably exemplifies a more general theorem about matroids.” An analogous correspondence between embedded graphs and delta-matroids was proposed in [22, 23]. This view of delta-matroid is proving to be successful. It has led implicitly and explicitly to advances in, especially, the topics of graph polynomials, and the structural theory of both delta-matroids and ribbon graphs. But we really are only at the beginning of this journey. Many fundamental questions remain unanswered and directions remain unexplored, but our knowledge is rapidly advancing.

Acknowledgements. This exposition draws upon work with many collaborators, to whom I am greatly indebted. These include Carolyn Chun, Rhiannon Hall, Criel Merino, Eunice Mphako-Banda, Jaeseong Oh, Ralf Rueckriemen, and especially Steven Noble.

References

[1] R. Arratia, B. Bollobás, and G. B. Sorkin, The interlace polynomial: a new graph polynomial. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 2000)*, pp. 237–245, ACM, New York, 2000 Zbl 0955.05066 MR 1754863

[2] R. Arratia, B. Bollobás, and G. B. Sorkin, The interlace polynomial of a graph. *J. Combin. Theory Ser. B* **92** (2004), no. 2, 199–233 Zbl 1060.05062 MR 2099142

[3] D. Bar-Natan and S. Garoufalidis, On the Melvin–Morton–Rozansky conjecture. *Invent. Math.* **125** (1996), no. 1, 103–133 Zbl 0855.57004 MR 1389962

[4] B. Bollobás and O. Riordan, A polynomial invariant of graphs on orientable surfaces. *Proc. London Math. Soc. (3)* **83** (2001), no. 3, 513–531 Zbl 1015.05024 MR 1851080

[5] B. Bollobás and O. Riordan, A polynomial of graphs on surfaces. *Math. Ann.* **323** (2002), no. 1, 81–96 Zbl 1004.05021 MR 1906909

[6] A. Bouchet, Greedy algorithm and symmetric matroids. *Math. Programming* **38** (1987), no. 2, 147–159 Zbl 0633.90089 MR 904585

[7] A. Bouchet, Isotropic systems. *European J. Combin.* **8** (1987), no. 3, 231–244 Zbl 0642.05015 MR 919874

[8] A. Bouchet, Reducing prime graphs and recognizing circle graphs. *Combinatorica* **7** (1987), no. 3, 243–254 Zbl 0666.05037 MR 918395

[9] A. Bouchet, Unimodularity and circle graphs. *Discrete Math.* **66** (1987), no. 1–2, 203–208 Zbl 0647.05039 MR 900943

[10] A. Bouchet, Graphic presentations of isotropic systems. *J. Combin. Theory Ser. B* **45** (1988), no. 1, 58–76 Zbl 0662.05014 MR 953895

[11] A. Bouchet, Representability of $\triangle$-matroids. In *Combinatorics (Eger, 1987)*, pp. 167–182, Colloq. Math. Soc. János Bolyai 52, North-Holland, Amsterdam, 1988 Zbl 0708.05013 MR 1221555

[12] A. Bouchet, Maps and $\triangle$-matroids. *Discrete Math.* **78** (1989), no. 1–2, 59–71 Zbl 0719.05019 MR 1020647

[13] A. Bouchet, Multimatroids. I. Coverings by independent sets. *SIAM J. Discrete Math.* **10** (1997), no. 4, 626–646 Zbl 0886.05042 MR 1477659

[14] A. Bouchet, Multimatroids. III. Tightness and fundamental graphs. *European J. Combin.* **22** (2001), no. 5, 657–677 Zbl 0982.05034 MR 1845490

[15] A. Bouchet and W. H. Cunningham, Delta-matroids, jump systems, and bisubmodular polyhedra. *SIAM J. Discrete Math.* **8** (1995), no. 1, 17–32 Zbl 0821.05010 MR 1315956

[16] A. Bouchet and A. Duchamp, Representability of $\triangle$-matroids over GF(2). *Linear Algebra Appl.* **146** (1991), 67–78 Zbl 0738.05027 MR 1083464

[17] R. Chandrasekaran and S. N. Kabadi, Pseudomatroids. *Discrete Math.* **71** (1988), no. 3, 205–217 Zbl 0656.05023 MR 959006

[18] S. Chmutov, Generalized duality for graphs on surfaces and the signed Bollobás–Riordan polynomial. *J. Combin. Theory Ser. B* **99** (2009), no. 3, 617–638 Zbl 1172.05015 MR 2507944

[19] S. Chmutov and I. Pak, The Kauffman bracket of virtual links and the Bollobás–Riordan polynomial. *Mosc. Math. J.* **7** (2007), no. 3, 409–418, 573 Zbl 1155.57004 MR 2343139

[20] S. Chmutov and J. Voltz, Thistlethwaite's theorem for virtual links. *J. Knot Theory Ramifications* **17** (2008), no. 10, 1189–1198 Zbl 1163.57001 MR 2460170

[21] S. V. Chmutov and S. K. Lando, Mutant knots and intersection graphs. *Algebr. Geom. Topol.* **7** (2007), 1579–1598 Zbl 1158.57013 MR 2366172

[22] C. Chun, I. Moffatt, S. D. Noble, and R. Rueckriemen, Matroids, delta-matroids and embedded graphs. *J. Combin. Theory Ser. A* **167** (2019), 7–59 Zbl 1417.05103 MR 3938888

[23] C. Chun, I. Moffatt, S. D. Noble, and R. Rueckriemen, On the interplay between embedded graphs and delta-matroids. *Proc. Lond. Math. Soc. (3)* **118** (2019), no. 3, 675–700 Zbl 1410.05029 MR 3932785

[24] B. Courcelle, Circle graphs and monadic second-order logic. *J. Appl. Log.* **6** (2008), no. 3, 416–442 Zbl 1149.03011 MR 2437322

[25] W. H. Cunningham, Decomposition of directed graphs. *SIAM J. Algebraic Discrete Methods* **3** (1982), no. 2, 214–228 Zbl 0497.05031 MR 655562

[26] O. T. Dasbach, D. Futer, E. Kalfagianni, X.-S. Lin, and N. W. Stoltzfus, The Jones polynomial and graphs on surfaces. *J. Combin. Theory Ser. B* **98** (2008), no. 2, 384–399 Zbl 1135.05015 MR 2389605

[27] A. Dress and T. F. Havel, Some combinatorial properties of discriminants in metric vector spaces. *Adv. in Math.* **62** (1986), no. 3, 285–312 Zbl 0609.05029 MR 866162

[28] M. N. Ellingham and X. Zha, Partial duality and closed 2-cell embeddings. *J. Comb.* **8** (2017), no. 2, 227–254 Zbl 1367.05049 MR 3610736

[29] J. A. Ellis-Monaghan and I. Moffatt, *Graphs on Surfaces. Dualities, Polynomials, and Knots*. SpringerBriefs Math., Springer, New York, 2013 Zbl 1283.57001 MR 3086663

[30] C. P. Gabor, K. J. Supowit, and W. L. Hsu, Recognizing circle graphs in polynomial time. *J. Assoc. Comput. Mach.* **36** (1989), no. 3, 435–473 Zbl 0825.68417 MR 1072233

[31] J. Geelen, B. Gerards, and G. Whittle, Structure in minor-closed classes of matroids. In *Surveys in Combinatorics 2013*, pp. 327–362, London Math. Soc. Lecture Note Ser. 409, Cambridge Univ. Press, Cambridge, 2013 Zbl 1318.05015 MR 3184116

[32] J. Geelen and S.-i. Oum, Circle graph obstructions under pivoting. *J. Graph Theory* **61** (2009), no. 1, 1–11 Zbl 1207.05189 MR 2514095

[33] J. F. Geelen, A generalization of Tutte's characterization of totally unimodular matrices. *J. Combin. Theory Ser. B* **70** (1997), no. 1, 101–117 Zbl 0885.05041 MR 1441261

[34] E. Gioan, C. Paul, M. Tedder, and D. Corneil, Practical and efficient circle graph recognition. *Algorithmica* **69** (2014), no. 4, 759–788 Zbl 1303.05190 MR 3209752

[35] J. L. Gross and T. W. Tucker, *Topological Graph Theory*. Dover Publications, Mineola, NY, 2001 Zbl 0991.05001 MR 1855951

[36] A. Kotzig, Eulerian lines in finite 4-valent graphs and their transformations. In *Theory of Graphs (Proc. Colloq., Tihany, 1966)*, pp. 219–230, Academic Press, New York, 1968 Zbl 0159.54201 MR 0248043

[37] J. P. S. Kung, Bimatroids and invariants. *Adv. in Math.* **30** (1978), no. 3, 238–249 Zbl 0398.05021 MR 520234

[38] I. Moffatt, Delta-matroids for graph theorists. In *Surveys in Combinatorics 2019*, pp. 167–220, London Math. Soc. Lecture Note Ser. 456, Cambridge Univ. Press, Cambridge, 2019 Zbl 07307106 MR 3967296

[39] I. Moffatt and J. Oh, A 2-isomorphism theorem for delta-matroids. *Adv. in Appl. Math.* **126** (2021), Paper No. 102133 Zbl 1462.05053 MR 4224067

[40] B. Mohar and C. Thomassen, *Graphs on Surfaces*. Johns Hopkins Stud. Math. Sci., Johns Hopkins University Press, Baltimore, MD, 2001 Zbl 0979.05002 MR 1844449

[41] P. Nelson, Almost all matroids are nonrepresentable. *Bull. Lond. Math. Soc.* **50** (2018), no. 2, 245–248 Zbl 1384.05065 MR 3830117

[42] J. Oxley, On the interplay between graphs and matroids. In *Surveys in Combinatorics, 2001 (Sussex)*, pp. 199–239, London Math. Soc. Lecture Note Ser. 288, Cambridge Univ. Press, Cambridge, 2001 Zbl 0979.05030 MR 1850709

[43] J. G. Oxley, *Matroid Theory*. Oxford Sci. Publ., Oxford University Press, New York, 1992 Zbl 0784.05002 MR 1207587

[44] L. Q. Qi, Directed submodularity, ditroids and directed submodular flows. *Math. Program., Ser. B* **42** (1988), no. 3, 579–599 Zbl 0665.90075 MR 980724

[45] N. Robertson and P. D. Seymour, Graph minors. XX. Wagner's conjecture. *J. Combin. Theory Ser. B* **92** (2004), no. 2, 325–357 Zbl 1061.05088 MR 2099147

[46] J. Spinrad, Recognition of circle graphs. *J. Algorithms* **16** (1994), no. 2, 264–282 Zbl 0797.68130 MR 1258239

[47] E. Tardos, Generalized matroids and supermodular colourings. In *Matroid Theory (Szeged, 1982)*, pp. 359–382, Colloq. Math. Soc. János Bolyai 40, North-Holland, Amsterdam, 1985 Zbl 0602.05020 MR 843383

[48] M. B. Thistlethwaite, A spanning tree expansion of the Jones polynomial. *Topology* **26** (1987), no. 3, 297–309 Zbl 0622.57003 MR 899051

[49] L. Traldi, The transition matroid of a 4-regular graph: an introduction. *European J. Combin.* **50** (2015), 180–207 Zbl 1319.05034 MR 3361421

[50] K. Truemper, On Whitney's 2-isomorphism theorem for graphs. *J. Graph Theory* **4** (1980), no. 1, 43–49 Zbl 0397.05043 MR 558452

[51] A. W. Tucker, A combinatorial equivalence of matrices. In *Proc. Sympos. Appl. Math., Vol. 10*, pp. 129–140, American Mathematical Society, Providence, RI, 1960 Zbl 0096.00701 MR 0114760

[52] W. T. Tutte, A ring in graph theory. *Proc. Cambridge Philos. Soc.* **43** (1947), 26–40 Zbl 0031.41803 MR 18406

[53] W. T. Tutte, An algorithm for determining whether a given binary matroid is graphic. *Proc. Amer. Math. Soc.* **11** (1960), 905–917 Zbl 0097.38905 MR 117173

[54] D. K. Wagner, On theorems of Whitney and Tutte. *Discrete Math.* **57** (1985), no. 1–2, 147–154 Zbl 0584.05020 MR 816055

[55] D. J. A. Welsh, *Matroid Theory*. L. M. S. Monographs 8, Academic Press, London, 1976 Zbl 0343.05002 MR 0427112

[56] H. Whitney, Non-separable and planar graphs. *Trans. Amer. Math. Soc.* **34** (1932), no. 2, 339–362 Zbl 0004.13103 MR 1501641

[57] H. Whitney, 2-Isomorphic graphs. *Amer. J. Math.* **55** (1933), no. 1–4, 245–254 Zbl 59.1235.01 MR 1506961

[58] H. Whitney, On the abstract properties of linear dependence. *Amer. J. Math.* **57** (1935), no. 3, 509–533 Zbl 0012.00404 MR 1507091

Iain Moffatt
Department of Mathematics, Royal Holloway, University of London, UK;
iain.moffatt@rhul.ac.uk

DOI 10.4171/8ECM/24

Convex bodies all whose sections (projections) are equal

Luis Montejano

Abstract. This work deals with the following question: if all hyperplane sections through the origin (orthogonal projections) of a convex body are "equal", is the convex body "equal" to the ball? where the notion of "equal" changes throughout the paper. Topology, Lie groups, Fourier analysis, and convex geometry interrelates in the solution and understanding of these problems.

1. Introduction

The purpose of this paper is to answer the following question:

If all hyperplane sections through the origin of a convex body are "equal", is the convex body "equal" to the ball?

The meaning of the notion "equal" will change in the course of this paper.

Similarly, we are interested in the following problem:

If all orthogonal projections of a convex body onto hyperplanes are "equal", is the convex body "equal" to the ball?

We believe that topology and convex geometry are deeply and beautifully interrelated in the solution and understanding of these problems.

A good reference for these problems and related problems is the book "Geometric Tomography" by Richard Gardner [11]. In particular, see Problems 3.3 and 7.4.

During this paper, unless otherwise stated, B is always an $(n+1)$-dimensional convex body with the origin as an interior point and $n \geq 2$.

2. Sections with the same area

The first meaning of "equal" is same "area".

2020 Mathematics Subject Classification. Primary 46C15; Secondary 52A20, 22E99, 55Q40.
Keywords. Section, projection, group of affine symmetries, ellipsoid.

If all hyperplane sections through the origin of a convex body B have equal n-dimensional volume, does B have the $(n+1)$-dimensional volume of the corresponding ball?

The answer to this question is by far a resounding no. There exist counterexamples. However, if we add the symmetry hypothesis to the question, the answer becomes yes. More precisely, the following theorem holds.

Theorem 2.1. *If all hyperplane sections through the origin of a centrally symmetric convex body B have equal n-dimensional volume, then the convex body B is a ball centered at the origin.*

Proof. The proof of this theorem uses analysis. We give here a sketch of the proof using harmonic integration. See Falconer's paper [10] or Schneider's book [26].

First of all, let $f : \mathbb{S}^n \to \mathbb{R}$ be a continuous function such that

$$\int_{\langle x,y\rangle=0} f(x)\,dx = 0, \quad \text{for every } y \in \mathbb{S}^n,$$

where integration refers to the usual measure in the $(n-1)$-sphere. Then the classical theorem of Funk-Hecke on spherical harmonics (see [12]) implies that f is an odd function; that is, $-f(x) = f(-x)$, for almost every $x \in \mathbb{S}^n$.

Let now B_1 and B_2 be two $(n+1)$-dimensional convex bodies that are symmetric with center at the origin and assume that the corresponding parallel n-dimensional areas of their sections through the origin are equal. We shall show that $B_1 = B_2$. For this purpose, let $f_1, f_2 \colon \mathbb{S}^n \to \mathbb{R}$ be the radial functions of B_1 and B_2. Note that because B_1 and B_2 are centrally symmetric, f_1 and f_2 are even functions. Moreover, by hypothesis

$$\frac{1}{n}\int_{\langle x,y\rangle=0} f_1(x)^n\,dx = \frac{1}{n}\int_{\langle x,y\rangle=0} f_2(x)^n\,dx,$$

for every $y \in \mathbb{S}^{n-1}$.

By our first argument, $f_1^n - f_2^n$ is an odd function, but since f_1 and f_2 are even functions, $f_1^n = f_2^n$. Moreover, since $f_1 \geq 0$ and $f_2 \geq 0$, we obtain that $f_1 = f_2$ and hence that $B_1 = B_2$.

Suppose now that B is a centrally symmetric convex body with the property that all its hyperplane sections through the origin have equal n-dimensional volume, and let $G \in \mathrm{SO}_{n+1}$ be a linear isometry. Then, by the above $B = GB$, for every $G \in \mathrm{SO}_n$, and consequently B is a ball centered at the origin. ■

3. Congruent and similar sections

The second meaning of "equal" is congruence.

Theorem 3.1 (Schneider's theorem). *If all hyperplane sections through the origin of a convex body B are congruent, then the convex body B is an $(n+1)$-ball centered at the origin.*

This time, the hypothesis of symmetry is not necessary. The theorem was proved by Süss [28] for $n = 2$. In 1970, using topological ideas, Mani [16] proved it for $n =$ even and, in 1979, Burton [7] proved it for $n = 3$. Finally, Rolf Schneider [25] in 1980, using analysis, proved it in general. In 1990, using the topological ideas of Hadwiger and Gromov, Montejano [20] proved the following result which, together with the false center theorem, allows an alternative proof of Schneider's theorem to be given.

Theorem 3.2. *If all hyperplane sections through the origin of a convex body B are affinely equivalent, then every hyperplane section of B through the origin is centrally symmetric.*

The proof of Theorem 3.2 uses topological ideas. Indeed, it uses the notion of field of convex bodies introduced by Hadwiger and developed by Mani in [16].

3.1. Fields of convex bodies

Let $\mathbf{K}^n$ be the space of all compact convex sets in $\mathbb{R}^n$ with the Hausdorff metric topology.

A *field of convex bodies tangent to* $\mathbb{S}^n$ is a continuous function

$$\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1},$$

such that $\kappa(u) \subset u + u^{\perp} \subset \mathbb{R}^{n+1}$, for every $u \in \mathbb{S}^n$, where $u^{\perp}$ denotes the subspace of $\mathbb{R}^{n+1}$ orthogonal to u.

If, in addition, $\kappa(u)$ is congruent (affinely equivalent) to the convex body $K \subset \mathbb{R}^n$, for every $u \in \mathbb{S}^n$, then we obtain a *field of convex bodies tangent to* $\mathbb{S}^n$ and congruent (affinely equivalent) to K. If, in addition, $\kappa(u) - u = \kappa(-u) + u$, then we have a *complete turning* of K in $\mathbb{R}^{n+1}$.

If all hyperplane sections through the origin of a convex body B are congruent (affinely equivalent), then there is a field of convex bodies tangent to $\mathbb{S}^n$ and congruent (affinely equivalent) to $\mathbb{R}^n \cap B$:

$$\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1},$$

defined as follows:

$$\kappa(u) = u + (u^{\perp} \cap B), \quad \text{for every } u \in \mathbb{S}^n.$$

Obviously, this field is a complete turning because $\kappa(u) - u = \kappa(-u) + u$, for every $u \in \mathbb{S}^n$.

Note that given a field of convex bodies $\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1}$ tangent to $\mathbb{S}^n$ and congruent to K, we may always assume without loss of generality that, for every $u \in \mathbb{S}^n$, the circumcenter of $\kappa(u)$ is the point $u \in (u + u^{\perp})$.

The link between Hadwiger's notion of field of convex bodies and the topology of Lie groups traces back to the work of Steenrod [27] and Gromov [13]. Every vector bundle $\xi : E \to \mathbb{S}^n$ with base the sphere $\mathbb{S}^n$, fiber $\mathbb{R}^m$, and structure group $\mathrm{GL}(m, \mathbb{R})$, can be obtained from $\mathbb{B}^n_+ \times \mathbb{R}^m$ disjoint union $\mathbb{B}^n_- \times \mathbb{R}^m$ by gluing the first copy $\mathbb{S}^{n-1} \times \mathbb{R}^m \subset \mathbb{B}^n_+ \times \mathbb{R}^m$ with the second copy $\mathbb{S}^{n-1} \times \mathbb{R}^m \subset \mathbb{B}^n_- \times \mathbb{R}^m$ via a fiber preserving homeomorphism

$$S^{n-1} \times \mathbb{R}^m \to S^{n-1} \times \mathbb{R}^m$$

that glue every fiber $\{x\} \times \mathbb{R}^m$ with the fiber $\{x\} \times \mathbb{R}^m$ using an element $g_x \in \mathrm{GL}(m, \mathbb{R})$, where $\mathbb{B}^n_+$ and $\mathbb{B}^n_-$ are respectively, the north and south closed hemisphere of $\mathbb{S}^n$. The map $g : \mathbb{S}^{n-1} \to \mathrm{GL}(m, \mathbb{R})$, given by $g(x) = g_x$, is called the *characteristic map* of the vector bundle ξ. It is not difficult to see that two vector bundles are equivalent (as fiber bundles) if and only if their corresponding characteristic maps are homotopic.

The existence of a field of convex bodies tangent to $\mathbb{S}^n$ and congruent to K implies that the tangent bundle $\mathrm{T}\mathbb{S}^n$ can be obtained gluing the copies $\mathbb{B}^n_+ \times \mathbb{R}^m$ and $\mathbb{B}^n_- \times \mathbb{R}^m$ using only isometries that fix K. In other words, the following holds:

There exists a field of convex bodies tangent to $\mathbb{S}^n$ and congruent to K if and only if the characteristic map

$$\begin{tikzcd}
\mathbb{S}^{n-1} \arrow[rr, "\chi_n"] \arrow[dr, "f"'] & & \mathrm{SO}_n \\
& G_K \arrow[ur, "i"'] &
\end{tikzcd}$$

factorizes through

$$G_K = \{g \in \mathrm{SO}_n \mid g(K) = K\}.$$

If this is so, then we say that the structure group of $\mathrm{T}\mathbb{S}^n$ *reduces to G_K*.

The main idea in the proof of Theorem 3.2 is that a complete turning of K is only possible if K has a center of symmetry (indeed, if $n = 3, 7$, the fact that the tangent bundle of $\mathbb{S}^3$ and $\mathbb{S}^7$ is parallelizable implies that a complete turning of K is possible if and only if K has a center of symmetry).

Since vector bundles over contractible spaces are trivial, we are going to take advantage of the existence of the field of convex bodies $\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1}$, tangent to the sphere $\mathbb{S}^n$ and congruent to K, to construct a continuous map

$$\Phi : \mathbb{B}^n_+ \to \mathrm{SO}_n,$$

such that $\Phi(x)(K) = \kappa(x)$, for every $x \in \mathbb{B}^n_+$.

Suppose that K is not symmetric. We may assume without generality that there is a point x_0 in the boundary of K such that $-x_0 \notin K$ and hence for every $g \in \mathrm{SO}_n$, $g(x_0) \neq -x_0$.

Note that

$$\{\Phi(x)(x_0)\}$$

is a field of vectors tangent to $\mathbb{B}^n_+$. Furthermore, for every $u \in \mathbb{S}^{n-1}$, we have that $\Phi(u)(x_0) \neq -\Phi(-u)(x_0)$. We are going to add a small annulus to B^n_+ at the boundary to obtain a larger n-dimensional ball $\widetilde{B}^n$ and we are going to take advantage of this annulus to define on it a tangent vector field that coincides with the one we have in B^n and with an additional property. The idea is that for every point $u \in \mathbb{S}^{n-1}$, we will use the annulus to rotate from the vector $\Phi(u)(x_0)$ towards the vector $\Phi(-u)(x_0)$. Since $\Phi(u)(x_0) \neq -\Phi(-u)(x_0)$, we can do this unambiguously in such a way that at the end on the border of $\widetilde{B}^n$, the tangent vector at the point $u \in \partial\widetilde{B}^n$ coincides with the tangent vector at the point $-u \in \partial\widetilde{B}^n$. Using this procedure, we obtain a complete turning of a nonzero vector field in the sphere $\mathbb{S}^n$, which is a contradiction to the well-known result that there is not a section to the canonical vector bundle of n-subspaces in $\mathbb{R}^{n+1}$; see [27].

Suppose that K_1, K_2 are convex bodies who have as ellipsoid of minimal volume containing them the unit ball. It is easy to see that if K_1 and K_2 are affinely equivalent, then they are actually congruent. Suppose now that $K \subset \mathbb{R}^n$ is a convex body with the unit ball as the ellipsoid of minimal volume containing it, and let $\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1}$ be a field of convex bodies tangent to $\mathbb{S}^n$ and affinely equivalent to the convex body $K \subset \mathbb{R}^n$, then there is a field of convex bodies tangent to $\mathbb{S}^n$ congruent to K. For every $x \in \mathbb{S}^n$, let $E_x \subset x + x^\perp$ be the ellipsoid of minimal volume containing $\kappa(x)$ and let h_x be the affine map that translates and dilates the principal axes of E_x to obtain the unit ball. It is easy to observe that the affine map h_x varies continuously with x. Hence $\kappa' : \mathbb{S}^n \to \mathbf{K}^{n+1}$, given by $\kappa'(x) = h_x(\kappa(x))$, is a field of convex bodies tangent to $\mathbb{S}^n$ congruent to K. By all the above, *if $\kappa : \mathbb{S}^n \to \mathbf{K}^{n+1}$ is a field of convex bodies tangent to $\mathbb{S}^n$ and affinely equivalent, then, for every $x \in \mathbb{S}^n$, $\kappa(x)$ is symmetric.*

3.2. The proof of Schneider's theorem and similar sections

Summarizing, Theorem 3.2 is true because a complete turning of K is only possible if K has a center of symmetry. This result, in combination with Larman's beautiful false center theorem [14], gives rise to a topological proof of Schneider's theorem.

Theorem 3.3 (Larman's false center theorem). *If all hyperplane sections through the origin of a convex body B have a center of symmetry, then either B is an ellipsoid or B is symmetric with respect to the origin.*

The proof of Schneider's theorem (Theorem 3.1) goes as follows. If all hyperplane sections through the origin of B are congruent, by Theorem 3.2, then every hyperplane section through the origin is centrally symmetric. By Larman's false center theorem (Theorem 3.3), either B is symmetric with center the origin and Theorem 2.1 implies that B is a ball centered at the origin, or B is an ellipsoid in which case it is easy to see directly that B is again a ball centered at the origin.

The third meaning of equal is similarity. If B is an $(n+1)$-ball with the origin as an interior point but not necessarily centered at the origin, then all hyperplane sections of B through the origin are n-balls and hence all are similar. Our next theorem states that this is always the case.

Theorem 3.4 (Montejano). *If all hyperplane sections through the origin of a convex body B are similar, then the convex body B is an n-ball not necessarily centered at the origin.*

A sketch of the proof is the following. Since similarities are affine equivalences, by Theorem 3.2, all hyperplane sections of B through the origin have a center of symmetry. By Larman's false center theorem (Theorem 3.3), either the origin is the center of symmetry of B or B is an ellipsoid. Using a topological argument, it is possible to prove that, in the first case, all hyperplane sections of B through the origin are not only similar but actually congruent and hence, by Schneider's theorem (Theorem 3.1), B is a ball or, in the second case, if B is an ellipsoid, it is easy to directly verify that our hypothesis implies that B is actually a ball.

4. Affinely equivalent sections and the Banach conjecture

The fourth meaning of equal is affine equivalence.

Conjecture 4.1. *If all hyperplane sections through the origin of a convex body B are affinely equivalent, then the convex body B is an ellipsoid.*

It turns out that Conjecture 4.1 is equivalent to the Banach conjecture over the reals.

4.1. The Banach conjecture

In 1932, in his book [3], Stephan Banach asked the following question:

Let V be a Banach space, real or complex, finite or infinite dimensional, all of whose n-dimensional subspaces, for some fixed integer n, $2 \leq n < \dim(V)$, are isometric to each other. Is it true that V is a Hilbert space?

This conjecture was proved first for $n = 2$ and real V in 1935 by Auerbach, Mazur, and Ulam [2] and in 1959 for all $n \geq 2$ and infinite dimensional real V by A. Dvoretzky [9]. In 1967, M. Gromov [13] proved the conjecture for even n and all V, real or complex, for odd n and real V with $\dim(V) \geq n + 2$, and for odd n and complex V with $\dim(V) \geq 2n$. V. Milman [18] extended Dvoretzky's theorem to the complex case, in particular, reproving Banach's conjecture for infinite dimensional complex space V. Recently, in 2021, Bor, Hernández-Lamoneda, Jiménez-Desantiago, and Montejano [4] proved the Banach conjecture if V is real and $n \equiv 1 \bmod 4$, with the possible exception of $n = 133$, and a little later, Bracho and Montejano [6] proved the Banach conjecture if V is complex and $n \equiv 1 \bmod 4$. A thorough account of the history of this conjecture is found in the notes on Section 9 in [17]. We also recommend [24].

Our next goal is to prove that the Banach conjecture over the reals is equivalent to Conjecture 4.1. First note that Banach's conjecture is a codimension one problem: since every Banach space, all of whose subspaces of a fixed dimension $n \geq 2$ are Hilbert spaces, is itself a Hilbert space, which easily follows from the elementary characterization of a norm coming from an inner product via the "parallelogram law", an affirmative answer for n in codimension one implies immediately an affirmative answer for n in all codimensions.

Note next that two Banach spaces V_1 and V_2 are isometric if there is a linear isomorphism $f : V_1 \to V_2$ that preserves the norm. That is, two Banach spaces V_1 and V_2 are isometric if their unit balls are linearly equivalent. To conclude, note that a finite dimensional Banach space V is a Hilbert space if and only if V is isometric to the Euclidean space, that is, if and only if its unit ball is an ellipsoid.

Finally, in the solution of Conjecture 4.1, we may always assume that not only B but all hyperplane sections of B through the origin have as a center of symmetry the origin. This is so because by Theorem 3.2 every section of B has a center of symmetry and therefore by Larman's false center theorem (Theorem 3.3) either B is an ellipsoid or the origin is the center of B.

4.2. Topology of Lie groups

From now on, until the end of this section, suppose that B is a convex body with the property that all its hyperplane sections through the origin are affinely equivalent. Our first interest is to answer the following question:

What can we say about the sections of B?

For example, due to Theorem 3.2, we know that all these sections have a center of symmetry, but do these sections share some other property?

Choose a convex set $K \subset \mathbb{R}^n$ affinely equivalent to all hyperplane sections of B through the origin with the additional property that the ellipsoid of minimal volume containing K is the unit n-ball. Define G as the group of symmetries of K, that is, G is the subgroup of linear isomorphism in $\mathrm{GL}(n, \mathbb{R})$ keeping fixed K and with positive determinant. Note that every element of G fixes also the unit n-ball that therefore $G \subset \mathrm{SO}_n$. As we shall see, G is a compact Lie group relevant in the solution of our previous question.

As in the sketch of the proof of Theorem 3.2, in Section 3, there is a field of convex bodies tangent to $\mathbb{S}^n$ and affinely equivalent to K. This implies that the structure group of the tangent bundle of the sphere $\mathbb{S}^n$ can be reduced to G or, in other words, that the characteristic map of $T\mathbb{S}^n$

$$\chi_n : \mathbb{S}^{n-1} \to \mathrm{SO}_n$$

can be factorized through G. See Steenrod's book [27] or Mani's paper [16].

If n is even and G is not transitive, the structure group of the tangent bundle of the sphere $\mathbb{S}^n$ cannot be reduced to G. This is so because if there is a map

$$f : \mathbb{S}^{n-1} \to \mathrm{SO}_n$$

homotopic to χ_n, such that $f(\mathbb{S}^{n-1}) \subset G$ and $e : \mathrm{SO}_n \to \mathbb{S}^n$ is the evaluation map (at any point), then ef is homotopic to $e\chi_n$. The non-transitivity of G implies that there are $x, y \in \mathbb{S}^n$ such that $g(x) \neq y$, for every $g \in \mathrm{SO}_n$. If $e : \mathrm{SO}_n \to \mathbb{S}^n$ is the evaluation at x, then the map e is not surjective and therefore ef is null homotopic. Thus, $e\chi_n$ is null homotopic, which is a contradiction in even dimensions, where we can easily calculate the even degree of $e\chi_n$. Consequently, if n is even, a field of convex bodies tangent to $\mathbb{S}^n$ affinely equivalent to K implies that G is transitive and consequently that K is an n-ball. In contrast, for $n = 3$, there is a field of convex bodies tangent to $\mathbb{S}^n$ and congruent to K, for every convex body $K \subset \mathbb{R}^n$, because $\mathbb{S}^3$ is parallelizable.

Summarizing, if n is even, the answer to our question: what can we say about the sections of B? is that all these sections are affinely equivalent to a ball and hence all of them are ellipsoids. This immediately implies that B is an ellipsoid, solving conjecture 1 when n is even and the Banach conjecture when n is even and V is a Banach space over the reals.

The case $n =$ odd is more complicated. First note that if $n = 3, 7$, this topological technique does not give us information about the sections of B, because $\mathbb{S}^3$ and $\mathbb{S}^7$ are parallelizable. We shall prove next that if $n \equiv 1 \mod 4$, with the possible exception of $n = 133$, a field of convex bodies tangent to $\mathbb{S}^n$ affinely equivalent to K implies that K is an affine body of revolution.

Suppose that the characteristic map of the sphere χ_n factorizes through the maximal connected subgroup $G \subset \mathrm{SO}_n$, that is,

$$\mathbb{S}^{n-1} \to G \hookrightarrow \mathrm{SO}_n\ .$$

We have two cases:

(1) G is an irreducible representation, that is, the action of G does not fix any proper subspace, and

(2) the action of G fixes a proper subspace Γ^k; $1 \leq k \leq n-1$.

In the first case, mathematicians have extensively studied irreducible representations, in particular, those for which the structural group of the space tangent to the sphere can be reduced to them. In particular, Leonard [15] proved that if $G \subset \mathrm{SO}_n$ is a maximal connected irreducible representation and the characteristic map of the sphere χ_n factorizes through G, then G is a simple group.

If this is so, we have several options:

- G is a classical group; SO_k, SU_k, Sp_k,
- G is a spin group; Spin_k,
- G is one of the exceptional Lie groups, G_2, F_4, E_6, E_7 or E_8.

Furthermore, in 2006, Cadek and Crabb proved that under the same hypothesis for G, if $n \geq 8$, then G is not isomorphic to SO_k, SU_m, Sp_m, with $k \geq 4$, $m \geq 2$. If $n \equiv 1 \bmod 4$, this rules out the classical groups, with the exception of $n = 5$. We leave this exceptional case for the next section. Furthermore, it can be proved that every irreducible representation of Spin_k, which does not factor through SO_m, is even dimensional. In our case, it is clear that G does not factor through SO_m, so if n is odd, we can rule out the possibility of a spin group for G.

Suppose now that $n \equiv 1 \bmod 4$. If this is the case, $\dim(G)$ is not too small with respect to n and hence G is not an exceptional Lie group, with the possible exception of the Lie group $E_7 \subset O_{133}$. This is so because it can be proved that in this case, $\dim(G) \geq 2n-3$ (see [8, Proposition 3.1]). Hence to rule out the exceptional groups, one can simply check (e.g., in Wikipedia) the following table in which we list the smallest irreducible representation for them, and the smallest irreducible representation congruent to 1 mod 4 is highlighted in red, verifying that in all the cases, with the exception of E_7, $\dim(G) \leq 2n-4$.

Group	G_2	F_4	E_6	E_7	E_8
dim G	14	52	78	133	248
Irreps	7	26	27	56	248
	14	52	78	$\boxed{133}$	3875
	27	273	351	912	⋮
	64	⋮	2925	⋮	1763125
	77	⋮	⋮	⋮	⋮

All the above implies that if G is irreducible and $n \equiv 1 \bmod 4$, then G is E_7 or is conjugate to O_n. Consequently, in this last case, K must be a ball, all the sections must be ellipsoids, and B must be an ellipsoid, as we wished.

The second case is when the action of G fixes a proper subspace Γ^k; $1 \leq k \leq n-1$.

If $n = 4k + 1$, the tangent space of the sphere $T\mathbb{S}^n$ splits:

$$T\mathbb{S}^n = e^1 \oplus \eta^{4k},$$

where e^1 is a vector bundle of dimension 1 and η^{4k} is unsplittable.

From here, we deduce that Γ^k is either 1 or $(n-1)$-dimensional, and G is a subset of a conjugate copy of SO_{n-1}. Furthermore, using an argument very similar to the argument used in the proof that $G = \mathrm{SO}_n$, when n is even (or see the case $n = 5$), it is possible to prove that G is actually a conjugate copy of SO_{n-1}. This gives rise to the case in which K is a body of revolution.

Summarizing, *suppose that B is an $(n+1)$-dimensional convex body with the property that all its hyperplane sections through the origin are affinely equivalent, $n \equiv 1 \bmod 4$, $n \neq 5, 133$. Then, every hyperplane section of B through the origin is an affine body of revolution.*

4.3. The case $n = 5$

This case is an exceptional case in our proof of the Banach conjecture but it is also interesting enough to illustrate the true complexity of the conjecture. This section will be dedicated to its complete proof.

Let B, K, and G be defined as in the previous section but this time B is a centrally symmetric convex body in $\mathbb{R}^6$, and $G = \{g \in \mathrm{SO}_5 \mid g(K) = K\}$ is a compact Lie subgroup of SO_5. Furthermore, we know that the characteristic map of the tangent space of $\mathbb{S}^5$

$$\begin{tikzcd}
\mathbb{S}^4 \arrow[rr, "\chi_5"] \arrow[dr, "f"'] & & \mathrm{SO}_5 \\
& G \arrow[ur, "i"'] &
\end{tikzcd}$$

factorizes through G.

Suppose first that G leaves invariant a proper subspace of $\mathbb{R}^5$. We shall prove that in this case K is a body of revolution.

By hypothesis, there is a k-dimensional subspace Λ invariant under G. This immediately implies that there is a continuous field of k-planes in $\mathbb{S}^5$. By [27, Theorem 27.18], we know that $\mathbb{S}^5$ admits a continuous field of k-planes if and only if $k = 1$ or $k = 4$. So, assume without loss of generality that $k = 1$, and therefore that Λ is a line invariant under G. Suppose without loss of generality that Λ is the line through the origin orthogonal to $\mathbb{R}^4$, in such a way that $G \subset \mathrm{SO}_4$. We will prove that G acts transitively on $\mathbb{R}^4$, thus proving that K is a body of revolution.

Given any 5-dimensional plane through the origin in $\mathbb{R}^6$, it is easy to prove that there is a unique complex plane through the origin contained in it. It is for this reason that there is a field of complex planes tangent to $\mathbb{S}^5$. This implies that the structural group of $\mathrm{T}\mathbb{S}^5$ can be reduced to SU_2. Thus, we may assume that $\chi_5 : \mathbb{S}^4 \to \mathrm{SU}_2$ is the characteristic map of $\mathrm{T}\mathbb{S}^5$. If $e : \mathrm{SO}_4 \to \mathbb{S}^3$ is the evaluation, hence, $e\chi_5 : \mathbb{S}^4 \to \mathbb{S}^3$ is not null homotopic. To see this, note that SU_2 is homeomorphic to $\mathbb{S}^3$ and the evaluation $e : \mathrm{SU}_2 \to \mathbb{S}^3$ is a homeomorphism. Therefore, if $e\chi_5 : \mathbb{S}^4 \to \mathbb{S}^3$ is homotopically trivial, then the same holds for $\chi_5 : \mathbb{S}^4 \to \mathrm{SU}_2$, but this implies that the characteristic map of $\mathrm{T}\mathbb{S}^5$ is homotopic to a constant, and therefore that $\mathrm{T}\mathbb{S}^5$ is parallelizable which is a contradiction.

We know that the structural group of $\mathrm{T}\mathbb{S}^5$ can be reduced to G. Therefore, the characteristic map $\chi_5 : \mathbb{S}^4 \to \mathrm{SU}_2$ is homotopic on SO_4 to a map $f : \mathbb{S}^4 \to G$. This implies that $e\chi_5, ef : \mathbb{S}^4 \to \mathbb{S}^3$ are homotopic. If G does not act transitively on $\mathbb{R}^4$, hence ef is null homotopic, but this is a contradiction to the fact that $e\chi_5$ is not null homotopic. Consequently, G acts transitively on $\mathbb{R}^4$ and K is a body of revolution, as we wished.

Suppose now that $G \subset \mathrm{SO}_5$ does not leave invariant a proper subspace of $\mathbb{R}^6$. That is, we must study the *irreducible representations on* $\mathbb{R}5$.

Consider S the collection of 3×3 real symmetric matrices with zero trace. Then, S is a real vector space of dimension 5 with the following natural interior product: given $A, B \in S$,

$$A \odot B = \mathrm{tr}(AB).$$

The group $G = \mathrm{SO}_3$ defines the following representation: $g(A) = gAg^{-1} = gAg^t$, for every $g \in G$ and $A \in S$.

Clearly, G acts linearly on S and furthermore,

$$g(A) \odot g(B) = \mathrm{tr}(gAg^{-1}gBg^{-1}) = \mathrm{tr}(gABg^{-1}) = A \odot B.$$

It is well known that this is a faithful, irreducible, representation. That is, we may think G is a subgroup of SO_5 with the property that G does not leave invariant any proper subspace. Moreover, it is well known that any other irreducible representation on $\mathbb{R}^5$ factors through G.

The following lemma finally proves that *if B is a 6-dimensional convex body with the property that all its hyperplane sections through the origin are affinely equivalent, then every hyperplane section of B through the origin is an affine body of revolution.*

Lemma 4.2. *Let $\Omega \subset \mathrm{SO}_5$ be a subgroup isomorphic to SO_3, Then, the structural group of $T\mathbb{S}^5$ cannot be reduced to Ω.*

Proof. Suppose that there is $f : \mathbb{S}^4 \to \Omega$ such that $i_\Omega f : \mathbb{S}^4 \to \mathrm{SO}_5$ is homotopic to the characteristic map $\chi_5 : \mathbb{S}^4 \to \mathrm{SO}_5$ of $T\mathbb{S}^5$, where $i_\Omega : \Omega \to \mathrm{SO}_5$ is the inclusion. Let $\pi : \mathbb{S}^3 \to \Omega$ be the double covering map and let $g : \mathbb{S}^3 \to \Omega$ be such that $\pi g = f$.

Let $u : \mathrm{SU}_2 \to \mathrm{SO}_5$ be the inclusion. Hence $\pi_3(\mathrm{SO}_5) = \mathbb{Z}$ (every compact, simple Lie group has $\pi_3 = \mathbb{Z}$) and $u_* : \pi_3(\mathrm{SU}_2) \to \pi_3(\mathrm{SO}_5)$ is an isomorphism. On the other hand, at the level of homology, $H_3(\mathrm{SO}_5, \mathbb{Z}_2)$ is a directed sum of $\mathbb{Z}_2$'s and $u_* : H_3(\mathrm{SU}_2, \mathbb{Z}_2) \to H_3(\mathrm{SO}_5, \mathbb{Z}_2)$ is not zero. Let us consider $[i_\Omega \pi] \in \pi_3(\mathrm{SO}_5) = \mathbb{Z}$. Suppose that $[i_\Omega \pi] = m \in \mathbb{Z}$ and let $\zeta : \mathbb{S}^3 \to \mathrm{SU}_2$ such that the induced homomorphism in homotopy is $\zeta_*(1) = m \in \pi_3(\mathrm{SU}_2) = \mathbb{Z}$. Consequently, $u\zeta : \mathbb{S}^3 \to \mathrm{SO}_5$ is homotopic to $i_\Omega \pi : \mathbb{S}^3 \to \mathrm{SO}_5$. In 3-dimensional homology, $(i_G \pi)_*(1) = 0$ which implies that $(u\zeta)_*(1) = 0$ and therefore, since $u_* : H_3(\mathrm{SU}_2, \mathbb{Z}_2) \to H_3(\mathrm{SO}_5, \mathbb{Z}_2)$ is not zero, that m is even.

Since m is even, the map $\zeta g : \mathbb{S}^4 \to \mathrm{SU}_2$ is null homotopic, because $\zeta_* : \pi_4(\mathbb{S}^3) \to \pi_4(\mathrm{SU}_2)$ is zero. This is a contradiction to the fact that $\mathbb{S}^5$ is not parallelizable. ■

The intuitive claim that

$$u_* : H_3(\mathrm{SU}_2, \mathbb{Z}_2) \to H_3(\mathrm{SO}_5, \mathbb{Z}_2)$$

is not zero, used in the above proof, is not so easy to prove. Indeed, to justify it, it is necessary to use the Dynkin index.

4.4. Affine bodies of revolution

A convex body $K \subset \mathbb{R}^n$ is a *body of revolution* if it admits an *axis of revolution*; i.e., a 1-dimensional line L such that each section of K by an affine hyperplane Δ orthogonal to L is an $(n-1)$-dimensional Euclidean ball in Δ, centered at $\Delta \cap L$ (possibly empty or just a point). If L is an axis of revolution of K, then $L^\perp$ is the associated *hyperplane of revolution*. Clearly, a ball is a body of revolution and any line through its center serves as an axis of revolution.

An axis of revolution of a plane convex figure is an axis of symmetry (or reflexion). Of course, a convex figure may have two different axes of symmetry without being a disk. In dimension $n \geq 3$, the situation is different.

Theorem 4.3. *A convex body of revolution $K \subset \mathbb{R}^n$, $n \geq 3$, with two different axes of revolution must be a ball.*

Proof. Consider $G_K = \{g \in \mathrm{SO}_n \mid g(K) = K\}$ the collection of orientation preserving isometries that fix K and suppose that $L \neq L'$ are two different axes of revolution of K. Without loss of generality, we may assume that L is the 1-dimensional subspace orthogonal to $\mathbb{R}^{n-1}$. Clearly, the collection of orientation preserving isometries of $\mathbb{R}^n$ that fix L also fix K and is equal to $\mathrm{SO}_{n-1} \subset \mathrm{SO}_n$. On the other hand, the group of orientation preserving isometries of $\mathbb{R}^n$ that fixes L' fixes also K and is equal to

$\mathrm{SO}'(n-1)$, a conjugate subgroup of SO_{n-1} in SO_n. Thus, our hypotheses imply that

$$\mathrm{SO}_{n-1} \subsetneq G_K \subset \mathrm{SO}_n,$$

but it is well known that SO_{n-1} is a *maximal connected* subgroup of SO_n (see [23, Lemma 4]). Therefore, $G_K = \mathrm{SO}_n$ and K must be a ball. ■

An *affine body of revolution* is a convex body affinely equivalent to a body of revolution. The images, under an affine equivalence, of an axis of revolution and its associated hyperplane of revolution of the body of revolution are an axis of revolution and associated hyperplane of revolution of the affine body of revolution (not necessarily perpendicular anymore). Clearly, an ellipsoid centered at the origin is an affine body of revolution and any hyperplane through the origin serves as a hyperplane of revolution.

As in the Euclidean case, a non-elliptical body of revolution admits a unique axis of revolution and a unique hyperplane of revolution.

Corollary 4.4. *An affine convex body of revolution $K \subset \mathbb{R}^n$, $n \geq 3$, with two different hyperplanes of revolution must be an ellipsoid.*

Proof. Let E be the ellipsoid of minimal volume containing K. By translation and dilatation of the principal axes of this ellipsoid, we obtain an affine isomorphism $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(E)$ is the unit ball of $\mathbb{R}^n$. Then since every affine isomorphism that fixes $f(K)$ also fixes $f(E)$, we have that $f(K)$ contains two different axes of revolution. By Lemma 4.3, $f(K)$ is a ball and consequently K is an ellipsoid. ■

4.4.1. Sections of affine bodies of revolution. It is not difficult to see that every section of a body of revolution is a body of revolution, that is why sections of affine bodies of revolution are affine bodies of revolution. Is the converse true? As far as I know, nobody knows the answer.

Conjecture 4.5. *Suppose that B is an $(n+1)$-dimensional convex body all whose hyperplane sections through the origin are affine bodies of revolution, $n \geq 3$. Then B is an affine body of revolution.*

We shall give a partial answer to this conjecture which will turn out to be sufficiently good for our purposes. Under the same hypothesis, we shall prove that at least one section of B through the origin is an ellipsoid. If this is so, and if, in addition, B satisfies the hypothesis that every two or its hyperplane section through the origin are affinely equivalent, then every section of B through the origin is an ellipsoid and consequently B is an ellipsoid. The proof of the existence of at least one elliptical section is a very interesting proof that combines ideas of convex geometry and algebraic topology. Before exposing it here, we require three intuitive lemmas, which we will state without proof. We ask the reader to include their own proofs.

Lemma 4.6. *Every hyperplane section $\Gamma \cap K$ of an affine body of revolution $K \subset \mathbb{R}^n$, $n \geq 3$, is an affine body of revolution. Furthermore, if H is the hyperplane of revolution of K, then either Γ is parallel to H or $\Gamma \cap H$ is a hyperplane of revolution of $\Gamma \cap K$.*

Lemma 4.7. *Let $K \subset \mathbb{R}^n$, $n \geq 3$, be an affine body of revolution with axis of revolution the line L and let Γ be a hyperplane containing L. Suppose that $\Gamma \cap K$ is an ellipsoid. Then K is an ellipsoid.*

Lemma 4.8. *Let $B \subset \mathbb{R}^{n+1}$ be a centrally symmetric convex body, all of whose hyperplane sections through the origin are non-elliptical affine symmetric bodies of revolution. For each $x \in \mathbb{S}^n$, let L_x be the (unique) axis of revolution of $x^\perp \cap B$, where $x^\perp$ denotes the subspace orthogonal to x. Then $x \mapsto L_x$ is a continuous function $\mathbb{S}^n \to \mathbb{R}P^n$. Consequently,*

$$\{x + L_x\}_{x \in \mathbb{S}^n}$$

is a field of lines tangent to $\mathbb{S}^n$.

Since every field of tangent lines gives rise to a trivial 1-dimensional fiber bundle over $\mathbb{S}^n$, then there is

$$\psi : \mathbb{S}^n \to \mathbb{S}^n,$$

such that for every $x \in \mathbb{S}^n$

$$L_x \cap \mathbb{S}^n = \big\{ -\psi(x), \psi(x) \big\}.$$

Note that, for every $x \in \mathbb{S}^n$, $\psi(x)$ is orthogonal to x and hence $\psi(x) \neq -x$. This implies that $\psi : \mathbb{S}^n \to \mathbb{S}^n$ is homotopic to the identity map and therefore that ψ is surjective.

From now on, let $B \subset \mathbb{R}^{n+1}$ be a centrally symmetric convex body, all of whose hyperplane sections through the origin are non-elliptical affine symmetric bodies of revolution, and remember that for every $u \in \mathbb{S}^n$, we denote by $u^\perp$ the n-dimensional subspace of $\mathbb{R}^{n+1}$ orthogonal to u. Furthermore, by Lemma 4.4, denote by L_u the unique affine axis of revolution of $u^\perp \cap B$, and by H_u the corresponding $(n-1)$-dimensional hyperplane of revolution of $u^\perp \cap B$. Note that the line L_u contains the origin. The fact that $u^\perp \cap B$ is symmetric implies that the origin is the center of the ellipsoid $H_u \cap B$ and therefore that the origin lies in L_u.

Lemma 4.9. *Let $B \subset \mathbb{R}^{n+1}$ be a symmetric convex body with center at origin, $n \geq 4$, and suppose that every hyperplane section of B through the origin is a non-elliptical affine convex body of revolution. Suppose that $L_v \subset u^\perp$ for some $u, v \in \mathbb{S}^n$. Then*

$$H_v \cap \Gamma_u = H_u \cap \Gamma_v = H_v \cap H_u$$

are highlighted in red.

Proof. Consider $u^\perp \cap v^\perp$, the $(n-1)$-dimensional subspace of $v^\perp$. By hypothesis, $v^\perp \cap B$ is a non-elliptical affine body of revolution with affine axis of revolution L_v. Therefore, since $L_v \subset u^\perp \cap v^\perp$, we have that $u^\perp \cap v^\perp \cap B$ is an affine body of revolution with affine axis of revolution L_v. Furthermore, by Lemma 4.7, $u^\perp \cap v^\perp \cap B$ is not an ellipsoid. Moreover, the principal affine subspace of revolution of $u^\perp \cap v^\perp \cap B$ is $H_v \cap u^\perp$.

On the other hand, $u^\perp \cap v^\perp$ is an $(n-1)$-dimensional subspace of $u^\perp$. Note that $u^\perp \cap v^\perp \neq H_u$, otherwise $u^\perp \cap v^\perp \cap B = H_u \cap B$ would be an ellipsoid, contradicting our previous assumption. Since $u^\perp \cap B$ is a non-elliptical affine body of revolution and $u^\perp \cap v^\perp \neq H_u$, then, by Lemma 4.6, $u^\perp \cap v^\perp \cap B$ is an affine body of revolution with principal affine subspace of revolution $H_u \cap v^\perp$. Consequently, by Lemma 4.4, we have that $H_v \cap u^\perp = H_u \cap v^\perp$. ■

Our main result regarding affine bodies of revolution is the following theorem.

Theorem 4.10. *Let $B \subset \mathbb{R}^{n+1}$ be a symmetric convex body with center at origin, $n \geq 4$, and suppose that every hyperplane section of B through the origin is an affine body of revolution. Then there is a hyperplane section through the origin of B which is an ellipsoid.*

Proof. Suppose not, suppose that B is a symmetric convex body with center at the origin and with the property that every hyperplane section of B through the origin is a non-elliptical affine convex body of revolution.

Let us fix a point $x_0 \in H_{u_0} \cap \mathbb{S}^n$. Since $\psi : \mathbb{S}^n \to \mathbb{S}^n$ is suprayective, let $v_0 \in \mathbb{S}^n$ such that $\psi(v_0) = x_0$. This implies that

$$L_{v_0} \subset H_{u_0}.$$

This is a contradiction to Lemma 4.9 because clearly $L_{v_0} \subset {u_0}^\perp$, hence,

$$L_{v_0} \subset H_{u_0} \cap {v_0}^\perp = H_{v_0} \cap {u_0}^\perp \subset H_{v_0},$$

which is impossible. ■

Theorem 4.10 is also true when $n = 2$. Indeed, in [21] Montejano proved that if B is a 3-dimensional convex body which contains the origin as interior point and every section through the origin is a figure that has a line of reflection (symmetry), then there is a section through the origin that is a disk. The proof also uses topology but it is intrinsically different to the proof of Theorem 4.10. The case $n = 3$ remains open.

With this, we have finished exposing the solution to the Banach conjecture over the reals given by Gromov [13], when $n =$ even and by Bor–Hernández Lamoneda–Jiménez-Desantiago–Montejano [4] when $n \equiv 1 \bmod 4$, $n \neq 133$. We summarize the results below in the following theorem.

Theorem 4.11 (Main theorem). *If all hyperplane sections through the origin of an* $(n+1)$*-dimensional convex body* B *are affinely equivalent,* $n \equiv 0, 1, 2 \bmod 4$*,* $n \neq 133$*, then the convex body* B *is an ellipsoid.*

4.5. The Banach conjecture, when n is odd and $\dim V \geq n+2$

As we have mentioned before, the cases of the Banach conjecture that have yet to be solved are those in which $n \equiv 3 \bmod 4$. That is, the first unsolved case from the Banach conjecture is the following.

Conjecture 4.12. *If all hyperplane sections through the origin of a* 4*-dimensional convex body* B *are affinely equivalent, then the convex body* B *is an ellipsoid.*

Indeed, Gromov in his original paper [13], using topology but a complete different sort of ideas, proved the Banach conjecture over the reals, when $n \equiv 3 \bmod 4$ and $\dim V > n+1$ and the Banach conjecture over the complex numbers, when $n \equiv 3 \bmod 4$ and $\dim V > 2n-1$.

The purpose of this section is to introduce these deep ideas. Let us prove the Banach conjecture over the reals, when $n > 1$ is odd and $\dim V \geq n+2$.

Theorem 4.13 (Gromov). *Let* B *be an* $(n+2)$*-dimensional convex body with the origin as interior point and suppose that all* n*-sections through the origin are linearly equivalent, for* $n > 1$ *odd. Then the convex body* B *is an ellipsoid.*

Denote by $V_{n,k}$ the space of all orthonormal k-frames $(e_1, \dots, e_k)$, where $e_i \in \mathbb{R}^n$, $n \geq k$. For our purpose, consider the space of 4-frames (e_1, e_2, e_3, e_4) in $\mathbb{R}^{n+2}$ and also the two fiber bundles

$$p_1 : V_{n+2,4} \to V_{n+2,2}, \quad p_2 : V_{n+2,4} \to V_{n+2,2},$$

where

$$p_1(e_1, e_2, e_3, e_4) = (e_1, e_2), \quad p_2(e_1, e_2, e_3, e_4) = (e_3, e_4).$$

The fiber in both cases is *the Stiefel Manifold* $V_{n,2}$. For more about Stiefel fiber bundles, see the book [19].

Consider now a nonempty closed subset $V \subset V_{n+2,2}$ and denote

$$\widetilde{V} = p_1^{-1}(V) = \big\{(e_1, e_2, e_3, e_4) \in V_{n+2,4} \mid (e_1, e_2) \in V\big\}.$$

The following lemma is Proposition 3 of Gromov's paper [13].

Lemma 4.14. *If* n *is odd and the restriction* $p_2| : \widetilde{V} \to V_{n+2,2}$ *is a fiber bundle, then* $V = V_{n+2,2}$.

We give only a brief sketch of the main ideas of the proof. We must consider an arbitrary fiber $V_{n,2}$ of p_2 and prove that the intersection $V' = \widetilde{V} \cap V_{n,2}$ coincides with $V_{n,2}$. Note that the dimension of $V_{n,2}$ is equal to $2n-3$. In fact, if $V' \neq V_{n,2}$, then $H^{2n-3}(V;\mathbb{Q}) = 0$ and for $p+q = 2n-3$, the second term $E_2^{p,q}$ in the spectral sequence of the fiber bundle $p_2| : \widetilde{V} \to V_{n+2,2}$ is trivial, which implies that $H^{2n-3}(\widetilde{V};\mathbb{Q})$ is trivial, contradicting an old result of Borel in [5, p. 192] that claims that for $n =$ odd, the homomorphism induced by the inclusion

$$H^{2n-3}(V_{n+4,2};\mathbb{Q}) \to H^{2n-3}(V_{n,2};\mathbb{Q})$$

is non-zero.

Proof of Theorem 4.13. By hypothesis, there is a convex body $K \subset \mathbb{R}^n$ with the property that the ellipsoid of minimal volume containing K is the unit ball of $\mathbb{R}^n$ and such that every n-dimensional section of B through the origin is linearly equivalent to K. Let us denote, as usual, by G_K the Lie group of all linear isomorphisms of $\mathbb{R}^n$ that keep K fixed. Of course, $G_K \subset O_n$.

Let us fix a 2-dimensional plane Δ in $\mathbb{R}^{n+2}$ through the origin and define $V \subset V_{n+2,2}$ as the set of 2-frames (e_1, e_2) in $\mathbb{R}^{n+2}$ such that if $\langle e_1, e_2\rangle$ is the subspace spanned by e_1 and e_2, then the section $\langle e_1, e_2\rangle \cap B$ is linearly equivalent to the section $\Delta \cap B$. Furthermore, let $V' \subset V_{n,2}$ be the set of 2-frames (e_1, e_2) in $\mathbb{R}^n$ such that $\langle e_1, e_2\rangle \cap K$ is linearly equivalent to $\Delta \cap B$. Finally, let

$$\widetilde{V} = p_1^{-1}(V) = \big\{(e_1, e_2, e_3, e_4) \in V_{n+2,4} \mid (e_1, e_2) \in V\big\}.$$

We shall first prove that the restriction $p_2| : \widetilde{V} \to V_{n+2,2}$ is a locally trivial bundle with fiber V'. For that purpose, consider U an open contractible subset of $V_{n+2,2}$. Then, using the contractibility of U and the existence of a field of convex bodies lineally equivalent to K, contained in the fibers of the canonical vector bundle of n-subspaces in $\mathbb{R}^{n+2}$, it is possible to construct a continuous map $\Lambda : U \to \mathrm{GL}(n, n+2)$ satisfying the following properties:

(1) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4} : \mathbb{R}^n \to \mathbb{R}^{n+2}$ is a linear embedding,

(2) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4}(\mathbb{R}^n)$ is orthogonal to both e_3 and e_4,

(3) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4}(K) = \Lambda_{e_3,e_4}(\mathbb{R}^n) \cap B$.

Given a pair of linearly independent vectors (w_1, w_2), denote by $(\mathrm{GS}^1(w_1, w_2), \mathrm{GS}^2(w_1, w_2))$ the 2-frame obtained from (w_1, w_2) by the Gram–Schmidt procedure in such a way that $\langle w_1, w_2\rangle = \langle \mathrm{GS}^1(w_1, w_2), \mathrm{GS}^2(w_1, w_2)\rangle$.

Define the fiber preserving map

$$\Phi : U \times V' \to V_{n+2,4},$$

given by

$$\begin{aligned}\Phi&\big((e_3, e_4), (e_1, e_2)\big)\\ &= \big(\mathrm{GS}^1\left(\Lambda_{e_3,e_4}(e_1), \Lambda_{e_3,e_4}(e_2)\right), \mathrm{GS}^2\left(\Lambda_{e_3,e_4}(e_1), \Lambda_{e_3,e_4}(e_2)\right), e_3, e_4\big).\end{aligned}$$

First of all, by (2),

$$\left(\mathrm{GS}^1\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right),\mathrm{GS}^2\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right),e_3,e_4\right)\in V_{n+2,4}.$$

Moreover, by (1),

$$\left(\mathrm{GS}^1\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right),\mathrm{GS}^2\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right)\right)\in V$$

and therefore

$$\left(\mathrm{GS}^1\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right),\mathrm{GS}^2\left(\Lambda_{e_3,e_4}(e_1),\Lambda_{e_3,e_4}(e_2)\right),e_3,e_4\right)\in p_2|^{-1}(U).$$

Hence, we obtain a fiber preserving homeomorphism:

$$\begin{tikzcd}
U \times V' \arrow[r, hook, "\Phi"] \arrow[d, "\mathrm{proj}"'] & p_2|^{-1}(U) \arrow[d, "p_2|"] \\
U \arrow[r, "\mathrm{id}"'] & U
\end{tikzcd}$$

thus proving that $p_2| : \widetilde{V} \to V_{n+2,2}$ is a locally trivial bundle with fiber V'. Furthermore, p_2 is a fiber bundle with structure group G_K. If this is so, by Lemma 4.14, $V = V_{n+2,2}$. This implies that for every two planes through the origin, the corresponding sections of B are linearly equivalent and hence that B is an ellipsoid. ■

4.6. The complex Banach conjecture

The fifth meaning of equal is complex affinely equivalence.

Let V be a finite dimensional Banach space over the complex numbers all of whose hyperplane subspaces are isometric to each other. Is it true that V is a Hilbert space?

Our next purpose is to prove that the above problem is equivalent to the following geometric problem. We need first some definitions.

Let $\mathbb{S}^1$ be the space of all unit complex numbers $\mathbb{C}$. Let A be a subset of complex space $\mathbb{C}^n$. We say that A is *complex symmetric* if and only if there is a translated copy A' of A such that $\xi A' = A'$, for every $\xi \in \mathbb{S}^1$. In this case, if $A' = A - x_0$, we say that x_0 is the center of complex symmetry of A. If $-A$ is a translated copy of A, then we just say that A is *symmetric*. It will be useful to consider the empty set as a complex symmetric set. Note that a *compact convex set $A \subset \mathbb{C}^n$ is complex symmetric with center at x_0 if and only if for every complex line L through x_0, the section $L \cap A$ is a disk centered at x_0*. Of course, any complex k-plane or a ball in a finite dimensional Banach space over the complex numbers is complex symmetric. A complex ellipsoid is the image of a ball under a complex affine transformation. Thus, balls of finite dimensional Hilbert spaces are complex ellipsoids. Of course, complex ellipsoids are complex symmetric sets. With this definition in mind, we may state the following problem equivalent to the complex Banach conjecture:

If all complex hyperplane sections through the origin of a convex body $B \subset \mathbb{C}^{n+1}$ with the origin as center of complex symmetry are complex linearly equivalent, is the convex body B a complex ellipsoid?

As was already mentioned, this problem has a positive answer when $n =$ even (Gromov [13]) and when $n \equiv 1 \bmod 4$ (Bracho and Montejano [6]). The purpose of this section is to give a brief summary of the ideas and techniques used in the proof.

This time, unlike the real case in which we use the principal bundle

$$\mathrm{SO}_n \hookrightarrow \mathrm{SO}_{n+1} \to \mathbb{S}^n,$$

we will use the corresponding principal bundle $\mathrm{SU}_n \hookrightarrow \mathrm{SU}_{n+1} \to \mathbb{S}^{2n+1}$. Here SU_n is the group of complex isometries of determinant 1 in $\mathbb{C}^n$ and we say that the structure group of the principal bundle $\mathrm{SU}_n \hookrightarrow \mathrm{SU}_{n+1} \to \mathbb{S}^{2n+1}$ can be reduced to $G \subset \mathrm{SU}_n$ if the characteristic map $\chi_n : \mathbb{S}^{2n} \to \mathrm{SU}_n$ of the complex bundle factorizes through G, that is, there is a map $f : \mathbb{S}^{2n} \to G$ such that the following diagram commutes up to homotopy, where $i : G \to \mathrm{SU}_n$ is the inclusion

$$\begin{tikzcd}
\mathbb{S}^{2n} \arrow[rr, "\chi_n"] \arrow[dr, "f"'] & & \mathrm{SU}_n \\
& G. \arrow[ur, "i"'] &
\end{tikzcd}$$

Denote by $\mathrm{GL}'_n(\mathbb{C})$ the group of complex linear isomorphisms of $\mathbb{C}^n$ with determinant a positive real number. Note that if K_1 and K_2 are complex symmetric convex bodies in $\mathbb{C}^n$ which are complex linearly equivalent, then there is $g \in \mathrm{GL}'_n(\mathbb{C})$ such that $g(K_1) = K_2$.

Given a complex symmetric convex body $K \subset \mathbb{C}^n$, let

$$G_K := \left\{ g \in \mathrm{GL}'_n(\mathbb{C}) \mid g(K) = K \right\}$$

be the *group of complex linear isomorphisms of K with positive real determinant*. By Lemma 1 of Gromov [13], there exists a complex ellipsoid of minimal volume containing K centered at the origin. Suppose now that this minimal ellipsoid is the $(2n-1)$-dimensional unit ball, then every $g \in G_K$ is actually an element of SU_n, because it fixes the unit ball, so in this case, $G_K := \{g \in \mathrm{SU}_n \mid g(K) = K\}$.

The link between our geometric problem and the topology is via the following lemma.

Lemma 4.15. *Let $B \subset \mathbb{C}^{n+1}$, $n \geq 2$, be a complex symmetric convex body with center at the origin all of whose complex hyperplane sections through the origin are complex linearly equivalent. Then there exists a complex symmetric convex body $K \subset \mathbb{C}^n$ with center at the origin and with the property that every complex hyperplane section of B is complex linearly equivalent to K and such that the structure group of the principal fiber bundle $\mathrm{SU}_n \hookrightarrow \mathrm{SU}_{n+1} \to \mathbb{S}^{2n+1}$ can be reduced to $G_K \subset \mathrm{SU}_n$.*

Our main interest naturally lies in studying the structure groups of the principal bundle ξ_n: $\mathrm{SU}_n \hookrightarrow \mathrm{SU}_{n+1} \to \mathbb{S}^{2n+1}$. In particular, if $n \equiv 0 \bmod 2$, ξ_n cannot be reduced to a proper subgroup of SU_{n-1} (see Leonard [15, Theorem 1B]). Therefore, under the hypothesis of Lemma 4.15, G_K must be SU_n, and hence K must be a ball. This implies that every section of B is a complex ellipsoid. Of course, every section of a complex symmetric body $B \subset \mathbb{C}^{n+1}$ is a complex ellipsoid only if B is a complex ellipsoid; see [6, Lemma 3.3]. This proves the complex Banach conjecture, when n is even.

For the case $n \equiv 1 \bmod 4$, the proof requires first studying the case in which $G_K \subset \mathrm{SU}_n$ is irreducible. If so, the topology of compact Lie groups over the complex numbers is simpler than over the real numbers and then it is possible to prove, in a similar way to the real case, that $G_K = \mathrm{SU}_n$. If this is the case, then every section of B is an ellipsoid and consequently B is also an ellipsoid. If $G_K \subset \mathrm{SU}_n$ is not irreducible but G_K is a proper subgroup of SU_n, then we can prove that $G_K = \mathrm{SU}_{n-1}$. To understand the convex geometry of the consequences of this result, we need the following definition:

A *complex body of revolution* is a complex symmetric convex body $K \subset \mathbb{C}^n$ for which there exists a 1-dimensional complex subspace L of $\mathbb{C}^n$, called its *axis of revolution*, such that for every affine complex hyperplane H orthogonal to L, we have that $H \cap K$ is either empty, a single point, or a $(2n-2)$-dimensional ball centered at $H \cap L$. Of course, K is a convex body of revolution if and only if $G_K = \mathrm{SU}_{n-1}$.

With this in mind, it is very clear that what we have obtained is the following theorem.

Theorem 4.16. *Let $B \subset \mathbb{C}^{n+1}$, $n \equiv 1 \bmod 4$, $n \geq 5$, be a complex symmetric convex body with center at the origin all of whose complex hyperplane sections through the origin are complex linearly equivalent. Then, there exists a complex body of revolution $K \subset \mathbb{C}^n$ with center at the origin and with the property that every complex hyperplane section of B through the origin is $\mathbb{C}$-linearly equivalent to K.*

To conclude, we need to know what are the geometric consequences of all the complex hyperplane sections of a convex body being complex affine bodies of revolution.

Theorem 4.17. *A complex symmetric convex body $B \subset \mathbb{C}^{n+1}$ with center at the origin, $n \geq 4$, all of whose complex hyperplane sections through the origin are complex affine bodies of revolution, has at least one complex hyperplane section through the origin which is a complex ellipsoid.*

The proof of Theorem 4.17 is similar to the proof of Theorem 4.10 except this time the proofs are just technically more complicated. This concludes an sketch of the proof of the complex Banach conjecture when $n \equiv 0, 1, 2 \bmod 4$, because by Theorems 4.16 and 4.17, every hyperplane section of B through the origin is a complex ellipsoid and therefore, by [6, Theorem 3.3] we obtain that B is a complex ellipsoid as we wished.

The following theorem follows immediately from Theorems 4.16 and 4.17. It proves the Banach conjecture over the complex numbers for $n \equiv 0, 1, 2 \bmod 4$, and $\dim V > n$.

Theorem 4.18 (Bracho–Montejano [6]). *If all complex hyperplane sections through the origin of a complex symmetric convex body $B \subset \mathbb{C}^{n+1}$ are linearly equivalent, $n \equiv 0, 1, 2 \bmod 4$, then the convex body B is a complex ellipsoid.*

5. Convex bodies all whose orthogonal projections are equal

The purpose of this section is to answer the following question:

If all orthogonal projections of a convex body onto hyperplanes are "equal", is the convex body "equal" to the ball?

5.1. Equal area, congruence, and affine equivalence

The first meaning of "equal" is same "area". In 1937, A. D. Aleksandrov [1] proved that if all orthogonal projections of a symmetric convex body have the same area, then not only does the body have the same volume of the corresponding ball but it is actually a ball.

Theorem 5.1 (Aleksandrov's projection theorem [1]). *If all orthogonal projections onto hyperplanes of a symmetric convex body $B \subset \mathbb{R}^{n+1}$ have equal n-dimensional volume, then the convex body B is a ball.*

Without the hypothesis of symmetry, Theorem 5.1 is false. However, a symmetric convex body all whose orthogonal projections have the same area not only has the volume of the corresponding ball but also it is actually a ball. For every $v \in \mathbb{S}^n$, denote by $B|v$ the orthogonal projection of B onto $v^{\perp}$ and let $\nu(B|v)$ be the n-dimensional volume of $B|v$. The proof of Theorem 5.1 follows immediately from the following Aleksandrov result (see [17, Theorem 2.11.1]). Given two convex bodies $B^1, B^2 \subset \mathbb{R}^{n+1}$ symmetric with respect to the origin and such that $\nu(B^1|v) = \nu(B^2|v)$, for every $v \in \mathbb{S}^n$, then B^1 is a translated copy of B^2. The proof of this result is analytic and a little more complicated than the proof of Theorem 2.1.

Using harmonic integration, it can be proved that a centrally symmetric convex body all whose $(n-1)$-dimensional perimeter areas are equal must be a ball. The proof is similar to the proof of Theorem 2.1 but using the support functions instead of the radial functions (see [10, Theorem 4]). Of course, without the symmetry hypothesis, the result is false as it can be observed with 3-dimensional convex bodies of constant width 1, in which the perimeter of all their orthogonal projections is π.

The next meaning of "equal" is congruence. That is, assume that all orthogonal projections onto hyperplanes of the convex body $B \subset \mathbb{R}^{n+1}$ are congruent.

The collection of orthogonal projections of $B \subset \mathbb{R}^{n+1}$,

$$\{B|v\}_{v\in\mathbb{S}^n}$$

give rise, not only to a field of convex bodies congruent to $B|e_1$ and tangent to $\mathbb{S}^n$, but also mainly to a complete turning of $B|e_1$, where $e_1 = \{1, 0, \dots,\} \in \mathbb{R}^{n+1}$. We know that a complete turning is only possible for symmetric convex bodies (see Section 3). So, $B|v$ is symmetric for every $v \in \mathbb{S}^n$ and, consequently, it is not very difficult to prove that B is symmetric, but in this last case Aleksandrov's theorem (Theorem 5.1) implies that B is also a ball. That is, we have the following theorem.

Theorem 5.2. *If all orthogonal projections onto hyperplanes of a convex body $B \subset \mathbb{R}^{n+1}$ are congruent, then the convex body B is a ball.*

Suppose now all orthogonal projections onto hyperplanes of the convex body $B \subset \mathbb{R}^{n+1}$ are affinely equivalent to a convex body K and suppose without loss of generality that the ellipsoid of minimal volume containing K is the unit ball. Denote $G_K := \{g \in \mathrm{GL}_n(\mathbb{R}) \mid g(K) = K \text{ and } \det(g) \text{ is positive}\} \subset \mathrm{SO}_n$. As in the case of the hyperplane sections, we have that the existence of the collection of projections $\{B|v\}_{v\in\mathbb{S}^n}$ gives rise directly to the following lemma which is the link between the topology and the geometric problem. Note that from the arguments given in the preceding paragraph and Theorem 3.2, we may assume without loss of generality that B and K are symmetric with center at the origin.

Lemma 5.3. *Let $B \subset \mathbb{R}^{n+1}$, $n \geq 2$, be a symmetric convex body all of whose orthogonal projections onto hyperplanes are linearly equivalent. Then there exists a symmetric convex body $K \subset \mathbb{R}^n$, with the property that every orthogonal projection of B onto a hyperplane is linearly equivalent to K and such that the structure group of the principal fiber bundle $\mathrm{SO}_n \hookrightarrow \mathrm{SO}_{n+1} \to \mathbb{S}^n$ can be reduced to $G_K \subset \mathrm{SO}_n$.*

Once we have this technical lemma, we are in a position to know, using the topological arguments from Section 4.2, how the projections of B are. That is, we have the following theorem.

Theorem 5.4. *Let $B \subset \mathbb{R}^{n+1}$, $n \equiv 0, 1, 2 \bmod 4$, $n \geq 2$, $n \neq 133$, be a convex body all of whose orthogonal projections onto hyperplanes are affinely equivalent. Then, there exists a body of revolution $K \subset \mathbb{R}^n$, with the property that every orthogonal projection of B is affinely equivalent to K.*

To conclude, we need to know the geometric consequences of all orthogonal projections of a convex body being affine bodies of revolution. Every orthogonal projection of a body of revolution is a body of revolution, this is why projections of affine bodies of revolution are affine bodies of revolution. Is the converse true? As far as I know, nobody knows the answer. The following geometric question is of great interest. *Suppose that B is an $(n+1)$-dimensional convex body all whose orthogonal projections are affine bodies of revolution, $n \geq 3$. Is B an affine body of revolution?*

We shall give a partial answer to this question which will turn out to be sufficiently good for our purposes. Under the same hypothesis of the above question, we shall prove that at least one orthogonal projection of B is an ellipsoid. If this is so, and if, in addition, B satisfies the hypothesis that every two of its orthogonal projections are affinely equivalent, then every orthogonal projection of B is an ellipsoid and consequently B is an ellipsoid. The proof of the next theorem is very similar to the proof of Theorem 4.10, with the different adjustments that are always necessary when trying to adapt a proof for sections to one for projections.

Theorem 5.5. *Let $B \subset \mathbb{R}^{n+1}$ be a symmetric convex body, $n \geq 4$, and suppose that every orthogonal projection onto hyperplanes of K is an affine body of revolution. Then there is an orthogonal projection of B which is an ellipsoid.*

This result, together with Theorem 5.4, immediately implies the following characterization of the ellipsoid first proved by Montejano in [22].

Theorem 5.6. *Let $B \subset \mathbb{R}^{n+1}$, $n \equiv 0, 1, 2 \bmod 4$, $n \geq 2$, $n \neq 133$, be a convex body all of whose orthogonal projections onto hyperplanes are affinely equivalent. Then B is an ellipsoid.*

5.2. The codimension 2 case for orthogonal projections

In this section, we will adapt Gromov's ideas from Section 4.5 to the context of orthogonal projections.

We need first a technical lemma.

Lemma 5.7. *Given a linear embedding $h : \mathbb{R}^n \to \mathbb{R}^m$, $2 < n < m$, there is a continuous map $h^* : V_{n,2} \to V_{m,2}$ such that, for every $u \in V_{n,2}$, (i) $\langle h^*(u)\rangle \subset h(\mathbb{R}^n)$ and (ii) $h(\langle u\rangle^{\perp})$ is orthogonal to $\langle h^*(u)\rangle$, where $\langle u\rangle$ denotes the plane generated by u.*

Furthermore, h^* varies continuously with h, while h varies in the space of linear embeddings from R^n to R^m.

Proof. Let $H \subset h(\mathbb{R}^n)$ be the plane such that H is orthogonal to $h(\langle u\rangle^\perp)$ and let $\pi : h(\mathbb{R}^n) \to H$ be the orthogonal projection. Then, given $u = (u_1, u_2) \in V_{n,2}$, let

$$h^*(u_1, u_2) = \left(\mathrm{GS}^1\left(\pi(u_1), \pi(u_2)\right), \mathrm{GS}^2\left(\pi(u_1), \pi(u_2)\right)\right) \in V_{m,2},$$

where given a pair of linearly independent vector (w_1, w_2), denote by $(\mathrm{GS}^1(w_1, w_2), \mathrm{GS}^2(w_1, w_2))$ the 2-frame obtained from (w_1, w_2) by the Gram–Schmidt procedure in such a way that $\langle w_1, w_2\rangle = \langle \mathrm{GS}^1(w_1, w_2), \mathrm{GS}^2(w_1, w_2)\rangle$. ∎

Here is the analogue of Theorem 4.13 for orthogonal projections:

Theorem 5.8 (Montejano). *Let B be an $(n+2)$-dimensional convex body and suppose that all orthogonal projections onto n-planes are linearly equivalent, for $n > 1$ odd. Then the convex body B is an ellipsoid.*

Proof. There is a convex body $K \subset \mathbb{R}^n$ with the property that the minimal ellipsoid containing K is the unit ball of $\mathbb{R}^n$ and such that all orthogonal projections of B onto an n-dimensional subspace are linearly equivalent to K. Let us fix a 2-dimensional plane $\Delta \subset \mathbb{R}^{n+2}$ through the origin and define $V \subset V_{n+2,2}$ to be the set of 2-frames (e_1, e_2) in $\mathbb{R}^{n+2}$ such that the orthogonal projection of B onto $\langle e_1, e_2\rangle$ is linearly equivalent to the orthogonal projection of B onto Δ. Furthermore, let $V' \subset V_{n,2}$ be the set of 2-frames (e_1, e_2) in $\mathbb{R}^n$ such that the orthogonal projection of K onto $\langle e_1, e_2\rangle$ is linearly equivalent to the orthogonal projection of B onto Δ. Finally, let $\widetilde{V} = p_1^{-1}(V) = \{(e_1, e_2, e_3, e_4) \in V_{n+2,4} \mid (e_1, e_2) \in V\}$.

We shall first prove that the restriction $p_2| : \widetilde{V} \to V_{n+2,2}$ is a locally trivial bundle with fiber V'. For that purpose, consider U an open contractible subset of $V_{n+2,2}$. Then, using the contractibility of U and the existence of a field of convex bodies, lineally equivalent to K, contained in the fibers of the canonical vector bundle of n-subspaces in $\mathbb{R}^{n+2}$, it is possible to construct a continuous map $\Lambda : U \to \mathrm{GL}(n, n+2)$ satisfying the following properties:

(1) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4} : \mathbb{R}^n \to \mathbb{R}^{n+2}$ is a linear embedding,

(2) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4}(\mathbb{R}^n)$ is orthogonal to both e_3 and e_4,

(3) for every $(e_3, e_4) \in U$, $\Lambda_{e_3,e_4}(K)$ is the orthogonal projection of B onto $\Lambda_{e_3,e_4}(\mathbb{R}^n)$.

Define the fiber preserving map

$$\Phi : U \times V' \to V_{n+2,4}$$

given by $\Phi((e_3, e_4), (e_1, e_2)) = (h^*(e_1, e_2), e_3, e_4)$.

First of all, by (2), $(h^*(e_1, e_2), e_3, e_4) \in V_{n+2,4}$. Moreover, by (1) and Lemma 5.7, $(h^*(e_1, e_2)) \in V$ and therefore $(h^*(e_1, e_2), e_3, e_4) \in p_2|^{-1}(U)$. Hence, we obtain a fiber preserving homeomorphism

$$\begin{tikzcd}
U \times V' \arrow[r, hook, "\Phi"] \arrow[d, "\mathrm{proj}"'] & p_2|^{-1}(U) \arrow[d, "p_2"] \\
U \arrow[r, "\mathrm{id}"'] & U.
\end{tikzcd}$$

Thus proving that $p_2| : \widetilde{V} \to V_{n+2,2}$ is a locally trivial bundle with fiber V'. If this is so, by Lemma 4.14, $V = V_{n+2,2}$. This implies that every two orthogonal projections onto 2-dimensional planes are linearly equivalent and hence, by Theorem 5.6, for $n = 2$, that K is an ellipsoid. ■

Acknowledgments. The author thanks Richard Gardner, Rolf Schneider, and the referee for their useful comments.

Funding. The author acknowledges support from CONACyT under project 166306 and from PAPIIT-UNAM under project IN112614.

References

[1] A. D. Aleksandrov, Zur Theorie der gemischten Volumina von konvexen Körpern, II: Neue Ungleichungen zwischen den gemischten Volumina und ihre Anwendungen. *Mat. Sbornik N. S.* **2** (1937), 1205–1238

[2] H. Auerbach, S. Mazur, and S. Ulam, Sur une propriété caractéristique de l'ellipsoïde. *Monatsh. Math. Phys.* **42** (1935), no. 1, 45–48 Zbl 0011.22208 MR 1550413

[3] S. Banach, *Théorie des opérations linéaires*. Monografie Matematyczne 1, PWN – Panstwowe Wydawnictwo Naukowe, Warszawa, 1932; English translation in *Theory of Linear Operations*, Vol. 38, Elsevier, 1987 Zbl 0005.20901

[4] G. Bor, L. Hernández Lamoneda, V. Jiménez-Desantiago, and L. Montejano, On the isometric conjecture of Banach. *Geom. Topol.* **25** (2021), no. 5, 2621–2642 Zbl 07396004 MR 4310896

[5] A. Borel, Sur la cohomologie des espaces fibrés principaux et des espaces homogènes de groupes de Lie compacts. *Ann. of Math. (2)* **57** (1953), 115–207 Zbl 0052.40001 MR 51508

[6] J. Bracho and L. Montejano, On the complex Banach conjecture. *J. Convex Anal.* **28** (2021), no. 4, 1211–1222 Zbl 07470575 MR 4374354

[7] G. R. Burton, Congruent sections of a convex body. *Pacific J. Math.* **81** (1979), no. 2, 303–316 Zbl 0373.52004 MR 547601

[8] M. Čadek and M. Crabb, G-structures on spheres. *Proc. London Math. Soc. (3)* **93** (2006), no. 3, 791–816 Zbl 1110.55008 MR 2266967

[9] A. Dvoretzky, A theorem on convex bodies and applications to Banach spaces. *Proc. Nat. Acad. Sci. U.S.A.* **45** (1959), 223–226; a detailed proof appeared in Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960), 123–160 Zbl 0088.31802 MR 105652

[10] K. J. Falconer, Applications of a result on spherical integration to the theory of convex sets. *Amer. Math. Monthly* **90** (1983), no. 10, 690–693 Zbl 0529.52001 MR 723941

[11] R. J. Gardner, *Geometric Tomography*. 2nd edn., Encyclopedia Math. Appl. 58, Cambridge University Press, New York, 2006 Zbl 1102.52002 MR 2251886

[12] H. Groemer, Fourier series and spherical harmonics in convexity. In *Handbook of Convex Geometry, Vol. A, B*, pp. 1259–1295, North-Holland, Amsterdam, 1993 Zbl 0799.52001 MR 1243009

[13] M. L. Gromov, On a geometric hypothesis of Banach. *Izv. Akad. Nauk SSSR Ser. Mat.* **31** (1967), 1105–1114 Zbl 0162.44402 MR 0217566

[14] D. G. Larman, A note on the false centre problem. *Mathematika* **21** (1974), 216–227 Zbl 0298.52005 MR 362048

[15] P. Leonard, G-structures on spheres. *Trans. Amer. Math. Soc.* **157** (1971), 311–327 Zbl 0217.49201 MR 275468

[16] P. Mani, Fields of planar bodies tangent to spheres. *Monatsh. Math.* **74** (1970), 145–149 Zbl 0189.52901 MR 259753

[17] H. Martini, L. Montejano, and D. Oliveros, *Bodies of Constant Width. An Introduction to Convex Geometry with Applications*. Birkhäuser/Springer, Cham, 2019 Zbl 1468.52001 MR 3930585

[18] V. D. Milman, A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.* **5** (1971), no. 4, 28–37 MR 0293374

[19] J. W. Milnor and J. D. Stasheff, *Characteristic Classes*. Ann. of Math. Stud. 76, Princeton University Press, Princeton, NJ, 1974 Zbl 0298.57008 MR 0440554

[20] L. Montejano, Convex bodies with homothetic sections. *Bull. London Math. Soc.* **23** (1991), no. 4, 381–386 Zbl 0746.52009 MR 1125866

[21] L. Montejano, Two applications of topology to convex geometry. *Tr. Mat. Inst. Steklova* **247** (2004), no. Geom. Topol. i Teor. Mnozh., 182–185 Zbl 1104.52001 MR 2168169

[22] L. Montejano, Convex bodies with affinely equivalent projections and affine bodies of revolution. *J. Convex Anal.* **28** (2021), no. 3, 871–877 Zbl 07470554 MR 4374323

[23] D. Montgomery and H. Samelson, Transformation groups of spheres. *Ann. of Math. (2)* **44** (1943), 454–470 Zbl 0063.04077 MR 8817

[24] A. Pełczyński, On some problems of Banach. *Russ. Math. Surv.* **28** (1973), no. 6, 67–75 Zbl 0288.46016

[25] R. Schneider, Convex bodies with congruent sections. *Bull. London Math. Soc.* **12** (1980), no. 1, 52–54 Zbl 0401.52001 MR 565484

[26] R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*. 2nd expanded edn., Encyclopedia Math. Appl. 151, Cambridge University Press, Cambridge, 2014 Zbl 1287.52001 MR 3155183

[27] N. Steenrod, *The Topology of Fibre Bundles*. Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1999 Zbl 0942.55002 MR 1688579

[28] W. Süss, Kennzeichnende Eigenschaften der Kugel als Folgerung eines Brouwerschen Fixpunktsatzes. *Comment. Math. Helv.* **20** (1947), 61–64 Zbl 0029.32001 MR 21339

Luis Montejano
Instituto de Matemáticas, Universidad Nacional Autónoma de México (UNAM) at Querétaro,
76230 Juriquilla, Querétaro, Mexico; luis@im.unam.mx

DOI 10.4171/8ECM/26

On a class of nonlocal problems with fractional gradient constraint

Assis Azevedo, José-Francisco Rodrigues, and Lisa Santos

Abstract. We consider a Hilbertian and a charges approach to fractional gradient constraint problems of the type $|D^\sigma u| \leq g$, involving the distributional fractional Riesz gradient D^σ, $0 < \sigma < 1$, extending previous results on the existence of solutions and Lagrange multipliers of these nonlocal problems.

We also prove their convergence as $\sigma \nearrow 1$ towards their local counterparts with the gradient constraint $|Du| \leq g$.

1. Introduction

Recently, the distributional partial derivatives of the Riesz potentials of order $1-\sigma$, $0 < \sigma < 1$,

$$(D^\sigma u)_j = \frac{\partial}{\partial x_j}(I_{1-\sigma} u) = D_j(I_{1-\sigma} u), \quad j = 1, \ldots, N,$$

where I_α, $0 < \alpha < 1$, is given by

$$I_\alpha u(x) = (I_\alpha * u)(x) = \gamma_{N,\alpha} \int_{\mathbb{R}^N} \frac{u(y)}{|x-y|^{d-\alpha}}\, dy, \quad \text{with } \gamma_{N,\alpha} = \frac{\Gamma\big(\frac{N-\alpha}{2}\big)}{\pi^{\frac{N}{2}} 2^\alpha \Gamma\big(\frac{\alpha}{2}\big)},$$

are shown to be a useful tool for a fractional vector calculus with the σ-gradient D^σ and σ-divergence $D^\sigma\cdot$ (see [5, 6, 12–14]). It leads to a new class of fractional partial differential equations and new problems in the calculus of variations [4]. As a consequence of the approximation of the identity by the Riesz kernel as $\alpha \to 0$ (see [7]), the σ-gradient converges to the classical gradient D as $\sigma \nearrow 1$, for instance, for smooth functions $u \in \mathcal{C}_0^\infty(\mathbb{R}^N)$ (see also [4,6]). Among the nice properties of D^σ, in [12] it was shown, for $u \in \mathcal{C}_0^\infty(\mathbb{R}^N)$, that

$$D^\sigma u \equiv D(I_{1-\sigma} * u) = I_{1-\sigma} * Du, \tag{1.1}$$

2020 Mathematics Subject Classification. Primary 35R11; Secondary 35J62, 49J40, 35J86, 26A33.
Keywords. Fractional gradient, nonlocal variational inequalities, gradient constraint, nonlocal Lagrange multiplier, elliptic quasilinear equations.

$$(-\Delta)^\sigma u = -D^\sigma \cdot (D^\sigma u), \tag{1.2}$$

where $(-\Delta)^\sigma$ is the classical fractional Laplacian in $\mathbb{R}^N$.

Here we are interested in complementing and extending some results of [10] on elliptic fractional equations of second σ-order, subjected to a σ-gradient constraint

$$|D^\sigma u| \leq g \quad \text{in } \mathbb{R}^N \tag{1.3}$$

and having the distributional form

$$-D^\sigma \cdot (AD^\sigma u + \Lambda^\sigma) = f_\# - D^\sigma \cdot \boldsymbol{f}. \tag{1.4}$$

We consider the homogeneous Dirichlet problem in a bounded open domain $\Omega \subset \mathbb{R}^N$, with Lipschitz boundary, so that the solution u is to be found in the fractional Sobolev space $H^\sigma_0(\Omega)$, $0 < \sigma < 1$, and may be extended by zero, belonging to $H^\sigma(\mathbb{R}^N)$. The Lipschitz boundary is sufficient for the $H^\sigma_0(\Omega)$-extension property, which is required in Section 4. Although in Sections 2 and 3 it is not strictly necessary, we prefer to keep this assumption in order to avoid delicate issues, in particular, with the definition of the classical space $H^\sigma_0(\Omega)$, which is the natural space to treat the Dirichlet boundary condition.

In (1.4), A is a coercive matrix with bounded variable coefficients (see (2.1), (2.2)) and $f_\#$ and $\boldsymbol{f}$ are given functions making the right-hand side an element f' of a suitable dual space.

The vector field Λ^σ is associated with the constraint (1.3) and may have two possible expressions. As we show in Section 2, with a Hilbertian approach, for $g \in L^2(\mathbb{R}^N)$, $g \geq 0$, and $f' \in H^{-\sigma}(\Omega) = (H^\sigma_0(\Omega))'$, $\Lambda^\sigma = D^\sigma \gamma$ for a unique $\gamma \in H^\sigma_0(\Omega)$ and it defines an element of the subdifferential of $\mathbb{K}^\sigma_g$, the convex subset of $H^\sigma_0(\Omega)$ of functions satisfying (1.3). The solution u is then the unique solution to the variational inequality (2.9) in $\mathbb{K}^\sigma_g$ for the operator $-D^\sigma \cdot (AD^\sigma \cdot) - f'$.

In the second case, with a strictly positive $g \in L^\infty(\mathbb{R}^N)$ and $f_\# \in L^1(\Omega)$, $\boldsymbol{f} \in \boldsymbol{L}^1(\mathbb{R}^N) = L^1(\mathbb{R}^N)^N$, in Section 3, by approximating the unique solution u with a suitable quasilinear penalised Dirichlet problem, we show the existence of at least a generalised nonnegative Lagrange multiplier $\lambda^\sigma \in L^\infty(\mathbb{R}^N)'$, such that $\Lambda^\sigma = \lambda^\sigma D^\sigma u$ and $\lambda^\sigma(|D^\sigma u| - g) = 0$ in the sense of charges, i.e., as an element of $L^\infty(\mathbb{R}^N)'$.

We recall (see [15, Example 5, Section 9, Chapter IV]), for instance, that a charge or an element $\chi \in L^\infty(\mathcal{O})'$, in an open set $\mathcal{O} \subset \mathbb{R}^N$, can be represented by a finitely additive measure χ^*, with bounded total variation, which is also absolutely continuous with respect to the Lebesgue measure and may be given by a Radon integral

$$\langle \chi, \varphi \rangle = \int_{\mathcal{O}} \varphi \, d\chi^*, \quad \forall \varphi \in L^\infty(\mathcal{O}). \tag{1.5}$$

As a consequence, it is easy to show the Hölder inequality for nonnegative charges $\chi \in L^\infty(\mathcal{O})'$ and arbitrary functions $\varphi, \psi \in L^\infty(\mathcal{O})$:

$$\big|\langle \chi, \varphi\psi\rangle\big| \leq \langle \chi, |\varphi|^p\rangle^{\frac{1}{p}} \langle \chi, |\psi|^{p'}\rangle^{\frac{1}{p'}}, \quad p > 1,\ p' = \frac{p}{p-1}. \tag{1.6}$$

It was proved in [12] that, similarly to the classical case $\sigma = 1$, the Sobolev, Trudinger, and Morrey inequalities also hold for the fractional D^σ; in particular, there exists a constant $C = C(N, p, \sigma) > 0$, such that, for $1 < p < \infty$, $\sigma \in (0, 1)$,

$$\|u\|_{L^q(\mathbb{R}^N)} \leq C\|D^\sigma u\|_{L^p(\mathbb{R}^N)}, \quad u \in \mathcal{C}_c^\infty(\mathbb{R}^N), \tag{1.7}$$

where $q = \frac{Np}{N-\sigma p}$ if $\sigma < \frac{N}{p}$, $q < \infty$ if $\sigma = \frac{N}{p}$, and $q = \infty$ if $\sigma > \frac{N}{p}$. In addition, when $\sigma > \frac{N}{p}$, we may take in the left-hand side of (1.7) the norm of the Hölder continuous functions $\mathcal{C}_c^\beta(\mathbb{R}^N)$, $0 < \beta = \sigma - \frac{N}{p} < 1$. As a consequence, we consider $H_0^\sigma(\Omega)$ with the equivalent Hilbertian norm $\|D^\sigma u\|_{L^2(\mathbb{R}^N)}$ (see [12]), which is also a consequence of the fractional Poincaré inequality (see [4]).

We observe that our results of Sections 2 and 3 also hold in the limit local case $\sigma = 1$, i.e., in $H_0^1(\Omega)$. We then show in Section 4, where we need to work with generalised sequences or nets, that the charges approach to the constrained problem yields the convergence, as $\sigma \nearrow 1$, of the solution u^σ and the generalised Lagrange multiplier λ^σ to the respective solution $(u, \lambda) \in W_0^{1,\infty}(\Omega) \times L^\infty(\Omega)'$ to the classical problem for D. We remark that, in this case, our results are new for data in L^1 and the general elliptic operator $-D \cdot (AD)$, extending [3], where the charges approach was introduced for $-\Delta$ with $f_\# \in L^2(\Omega)$ and $\boldsymbol{f} = 0$. For a recent survey on gradient type constrained problems, see [11].

2. The Hilbertian approach with σ-gradient constraint in L^2

Let the not necessarily symmetric measurable matrix $A = A(x) : \mathbb{R}^N \to \mathbb{R}^{N\times N}$ satisfy the coercive assumption, for some given $a_*, a^* > 0$,

$$A(x)\boldsymbol{\xi} \cdot \boldsymbol{\xi} \geq a_*|\boldsymbol{\xi}|^2, \quad \text{a.e. } x \in \mathbb{R}^N,\ \forall \boldsymbol{\xi} \in \mathbb{R}^N, \tag{2.1}$$

and the boundedness conditions

$$A(x)\boldsymbol{\xi} \cdot \boldsymbol{\eta} \leq a^*|\boldsymbol{\xi}|\,|\boldsymbol{\eta}|, \quad \text{a.e. } x \in \mathbb{R}^N,\ \forall \boldsymbol{\xi},\ \boldsymbol{\eta} \in \mathbb{R}^N. \tag{2.2}$$

Consider

$$f_\# \in L^{2^\#}(\Omega) \quad \text{and} \quad \boldsymbol{f} = (f_1, \ldots, f_N) \in \boldsymbol{L}^2(\mathbb{R}^N), \tag{2.3}$$

where, by the Sobolev embeddings (1.7), $2^\# = \frac{2N}{N+2\sigma}$ if $0 < \sigma < \frac{N}{2}$, or $2^\# = q$ for any $q > 1$ when $\sigma = \frac{1}{2}$ and $2^\# = 1$ when $\frac{1}{2} < \sigma < 1$, so that

$$\langle f', v\rangle_\sigma = \int_\Omega f_\# v + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma v, \tag{2.4}$$

for arbitrary $v \in H^\sigma_0(\Omega)$, defines the linear form $f' \in H^{-\sigma}(\Omega) = H^\sigma_0(\Omega)'$, $0 < \sigma < 1$. We have

$$\exists! \phi \in H^\sigma_0(\Omega) : \int_{\mathbb{R}^N} D^\sigma \phi \cdot D^\sigma v = \langle f', v\rangle_\sigma, \quad \forall v \in H^\sigma_0(\Omega). \tag{2.5}$$

The validity of (2.5) is a consequence of the Fréchet–Riesz representation theorem and the choice of the left-hand side of this equality as the inner product in $H^\sigma_0(\Omega)$, as stated in Section 1. It follows that $\boldsymbol{F} = D^\sigma \phi \in \boldsymbol{L}^2(\Omega)$ belongs to the image of $H^\sigma_0(\Omega)$ by D^σ:

$$\Psi_\sigma = \big\{\boldsymbol{G} \in \boldsymbol{L}^2(\mathbb{R}^N) : \boldsymbol{G} = D^\sigma v,\ v \in H^\sigma_0(\Omega)\big\} = D^\sigma\big(H^\sigma_0(\Omega)\big), \tag{2.6}$$

which is a strict Hilbert subspace of $\boldsymbol{L}^2(\mathbb{R}^N)$, for the inner product

$$(\boldsymbol{F}, \boldsymbol{G})_{\Psi_\sigma} = \int_{\mathbb{R}^N} D^\sigma \phi \cdot D^\sigma v,$$

and Ψ_σ is isomorphic to $H^{-\sigma}(\Omega)$, by the Riesz theorem (2.5). Actually, this remark extends the well-known case $\sigma = 1$, when D^1 is the classical gradient D.

Consider the nonempty closed convex set

$$\mathbb{K}^\sigma_g = \big\{v \in H^\sigma_0(\Omega) : |D^\sigma v| \leq g \text{ a.e. in } \mathbb{R}^N\big\}, \tag{2.7}$$

where the σ-gradient threshold g is such that

$$g \in L^2(\mathbb{R}^N), \quad g(x) \geq 0 \text{ a.e. } x \in \mathbb{R}^N. \tag{2.8}$$

Under the assumptions (2.1) and (2.2), A defines a continuous bounded coercive bilinear form over $H^\sigma_0(\Omega)$ and, as an immediate consequence of the Stampacchia theorem (see [9, p. 95], for instance), we have the existence, uniqueness, and continuous dependence of the solution u, with respect to the linear form (2.4), of the variational inequality

$$\begin{aligned} u \in \mathbb{K}^\sigma_g : \int_{\mathbb{R}^N} & A D^\sigma u \cdot D^\sigma (v-u) \\ & \geq \int_\Omega f_\#(v-u) + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma(v-u), \quad \forall v \in \mathbb{K}^\sigma_g. \end{aligned} \tag{2.9}$$

In particular, if C_* denotes the Sobolev constant, with $L^{2^*}(\Omega) = L^{2^\#}(\Omega)'$,

$$\|v\|_{L^{2^*}(\Omega)} \le C_* \|D^\sigma v\|_{\boldsymbol{L}^2(\mathbb{R}^N)}, \quad v \in H^\sigma_0(\Omega),\ 0 < \sigma \le 1,$$

and $\hat{u}$ is the solution corresponding to the data $\hat{f}_\#$, $\hat{\boldsymbol{f}}$, we have

$$\|u - \hat{u}\|_{H^\sigma_0(\Omega)} \le \frac{C_*}{a_*} \|f_\# - \hat{f}_\#\|_{L^{2^\#}(\Omega)} + \frac{1}{a_*} \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_{\boldsymbol{L}^2(\mathbb{R}^N)}. \tag{2.10}$$

It is well known (see [8, p. 203], for instance) that to solve (2.9) is equivalent to finding $u \in H^\sigma_0(\Omega)$, such that

$$\Gamma \equiv f' - \mathcal{L}^\sigma_A u \in \partial I_{\mathbb{K}^\sigma_g}(u) \quad \text{in } H^{-\sigma}(\Omega), \tag{2.11}$$

where $\mathcal{L}^\sigma_A : H^\sigma_0(\Omega) \to H^{-\sigma}(\Omega)$ is the linear continuous operator defined by

$$\langle \mathcal{L}^\sigma_A w, v\rangle_\sigma = \int_{\mathbb{R}^N} A D^\sigma w \cdot D^\sigma v, \quad \forall v, w \in H^\sigma_0(\Omega),$$

and $\Gamma = \Gamma(u) \in H^{-\sigma}(\Omega)$ is an element of the sub-gradient of the indicatrix function $I_{\mathbb{K}^\sigma_g}$ of the convex set $\mathbb{K}^\sigma_g$ at u:

$$I_{\mathbb{K}^\sigma_g}(v) = \begin{cases} 0 & \text{if } v \in \mathbb{K}^\sigma_g, \\ +\infty & \text{if } v \in H^\sigma_0(\Omega) \setminus \mathbb{K}^\sigma_g. \end{cases}$$

By the Riesz theorem, there exists a unique $\gamma = \gamma(u) \in H^\sigma_0(\Omega)$ corresponding to $\Gamma = \Gamma(u)$ given by (2.11) (recall (2.5)) and the couple $(u, \gamma) \in \mathbb{K}^\sigma_g \times H^\sigma_0(\Omega)$ solves the problem

$$\int_{\mathbb{R}^N} (A D^\sigma u + D^\sigma \gamma) \cdot D^\sigma v = \int_\Omega f_\# v + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma v, \quad \forall v \in H^\sigma_0(\Omega). \tag{2.12}$$

If we denote $\widehat{\gamma} = \gamma(\hat{u})$, with $\hat{u}$ solving (2.9) with $\hat{f}_\#$ and $\hat{\boldsymbol{f}}$ given in (2.3), using (2.10) and (2.2), we easily obtain, by the Riesz isometry $\|\Gamma\|_{H^{-\sigma}(\Omega)} = \|\gamma\|_{H^\sigma_0(\Omega)}$,

$$\begin{aligned} &\|\gamma - \widehat{\gamma}\|_{H^\sigma_0(\Omega)} \\ &\quad \le C_* \left(1 + \frac{a^*}{a_*}\right) \|f_\# - \hat{f}_\#\|_{L^{2^\#}(\Omega)} + \left(1 + \frac{a^*}{a_*}\right) \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|_{\boldsymbol{L}^2(\mathbb{R}^N)}. \end{aligned} \tag{2.13}$$

We have then proven the following result.

Theorem 2.1. *Under the previous assumptions, namely,* (2.1), (2.2), (2.3), *and* (2.8), *there exists a unique solution of* (2.9), *which also satisfies* (2.12) *with a unique* $\gamma = \gamma(u) \in H^\sigma_0(\Omega)$, *obtained through* (2.11) *and depending on the data through* (2.13).

Remark 2.2. This result extends to the Riesz fractional gradient the limit case $\sigma = 1$, where the classical gradients of u and γ are extended by zero in $\mathbb{R}^N \setminus \Omega$. A natural and important question is to find a more direct relation of the potential γ with the solution u through the existence of a Lagrange multiplier λ, such that

$$D^\sigma \gamma = \lambda D^\sigma u. \tag{2.14}$$

In the classical case $\sigma = 1$, with $A = Id$, $\Omega \subseteq \mathbb{R}^2$ simply connected, and f' and g given by positive constants, corresponding to the elasto-plastic torsion problem, Brézis has proven the existence and uniqueness of a bounded function

$$\lambda \geq 0 \quad \text{such that} \quad \lambda\big(|Du| - g\big) = 0 \text{ a.e. in } \Omega,$$

which is even continuous if Ω is convex (see [11] for references). Although (2.14) is an open question in the general case of Theorem 2.1, for strictly positive bounded threshold g, it has been shown to hold in the sense of finite additive measures in [10], following the case $\sigma = 1$ of [3].

Using a variant of a classical penalisation method proposed in [8, p. 376] with $\varepsilon \in (0, 1)$ and

$$k_\varepsilon(t) = 0,\ t \leq 0, \quad k_\varepsilon(t) = \frac{t}{\varepsilon},\ 0 \leq t \leq \frac{1}{\varepsilon}, \quad k_\varepsilon(t) = \frac{1}{\varepsilon^2},\ t \geq \frac{1}{\varepsilon}, \tag{2.15}$$

we may consider the approximating quasi-linear problem: find $u_\varepsilon \in H^\sigma_0(\Omega)$, such that

$$\begin{aligned}\int_{\mathbb{R}^N} &\big(AD^\sigma u_\varepsilon + \widehat{k}_\varepsilon(u_\varepsilon)\, D^\sigma u_\varepsilon\big) \cdot D^\sigma v \\ &= \int_\Omega f_\# v + \int_{\mathbb{R}^N} f \cdot D^\sigma v,\ \forall v \in H^\sigma_0(\Omega),\end{aligned} \tag{2.16}$$

where we set

$$\widehat{k}_\varepsilon = \widehat{k}_\varepsilon(u_\varepsilon) = k_\varepsilon\big(|D^\sigma u_\varepsilon|^2 - g^2\big) \quad \text{with } k_\varepsilon \text{ given by (2.15).}$$

In the proof of the approximation theorem, we shall require the following assumption: for each $R > 0$, there exists a g_R, such that

$$g(x) \geq g_R > 0, \quad \text{for a.e. } x \in B_R = \big\{x \in \mathbb{R}^N : |x| < R\big\}. \tag{2.17}$$

Theorem 2.3. *Under the assumptions of Theorem 2.1, let also (2.17) hold. Then, the unique solution $u_\varepsilon \in H^\sigma_0(\Omega)$ of (2.16), as $\varepsilon \to 0$, is such that*

$$u_\varepsilon \underset{\varepsilon\to 0}{\rightharpoonup} u \qquad \text{in } H^\sigma_0(\Omega)\text{-weak}, \tag{2.18}$$

$$\widehat{k}_\varepsilon D^\sigma u_\varepsilon \underset{\varepsilon\to 0}{\rightharpoonup} D^\sigma \gamma \quad \text{in } \Psi'_\sigma\text{-weak}, \tag{2.19}$$

where $(u,\gamma)\in\mathbb{K}^\sigma_g\times H^\sigma_0(\Omega)$ *is the unique couple given in Theorem* 2.1 *and satisfying* (2.12) *and* Ψ_σ *is the vector space defined in* (2.6).

Proof. Since the quasi-linear operator $\widehat{A}_\varepsilon : H^\sigma_0(\Omega)\to H^{-\sigma}(\Omega)$ defined by the left-hand side of (2.16) is bounded, strongly monotone, coercive, and hemicontinuous, the existence and uniqueness of u_ε solution to (2.16) is classical (see [8], for instance).

Taking $v=u_\varepsilon$ in (2.16) and recalling that $\widehat{\kappa}_\varepsilon(u_\varepsilon)\ge 0$, it is clear that we have, with $C_\sigma>0$ independent of ε, $0<\varepsilon<1$:

$$\|u_\varepsilon\|_{H^\sigma_0(\Omega)} \le \frac{C_*}{a_*}\|f_\#\|_{L^{2^\#}(\Omega)} + \frac{1}{a_*}\|\boldsymbol{f}\|_{\boldsymbol{L}^2(\mathbb{R}^N)} \equiv C_\sigma, \tag{2.20}$$

so that we have (2.18) at least for a generalised subsequence and some $u\in H^\sigma_0(\Omega)$. Consequently, from (2.16), we also obtain

$$\|\widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon\|_{\Psi'_\sigma} = \sup_{\substack{v\in H^\sigma_0(\Omega)\\ \|v\|_{H^\sigma_0(\Omega)}=1}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon(u_\varepsilon) D^\sigma u_\varepsilon\cdot Dv \le (a_*+a^*)C_\sigma,$$

for all ε, $0<\varepsilon<1$, by using (2.20) and recalling (2.2). Here we use the definition (2.5) and we consider $\boldsymbol{L}^2(\mathbb{R}^N)$, identified to its dual, as a subspace of Ψ'_σ, the dual of $\Psi_\sigma\subseteq\boldsymbol{L}^2(\mathbb{R}^N)$. Hence, for a generalised subsequence $\varepsilon\to0$, we also have

$$\widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \underset{\varepsilon\to0}{\rightharpoonup} \Lambda \quad \text{in } \Psi'_\sigma\text{-weak.} \tag{2.21}$$

In order to prove that $u\in\mathbb{K}^\sigma_g$, i.e., $|D^\sigma u|\le g$ a.e. in $\mathbb{R}^N$, we consider, for $R>0$,

$$\begin{aligned} U_{\varepsilon,R} &= \left\{x\in B_R : 0\le \left|D^\sigma u_\varepsilon(x)\right|^2 - g^2(x)\le\sqrt{\varepsilon}\right\},\\ V_{\varepsilon,R} &= \left\{x\in B_R : \left|D^\sigma u_\varepsilon(x)\right|^2 - g^2(x) > \sqrt{\varepsilon}\right\}\end{aligned}$$

and we observe that, using the assumptions (2.17), (2.20), and $\widehat{\kappa}_\varepsilon(|D^\sigma u^\varepsilon|^2-g^2)\ge0$, from (2.16) it follows that

$$g_R^2\int_{B_R}\widehat{\kappa}_\varepsilon \le \int_{\mathbb{R}^N}\widehat{\kappa}_\varepsilon g^2 \le \int_{\mathbb{R}^N}\widehat{\kappa}_\varepsilon|D^\sigma u_\varepsilon|^2 \le \frac{a_*}{2}C_\sigma^2, \quad 0<\varepsilon<1. \tag{2.22}$$

Consequently, for all $R>0$, we conclude that $|D^\sigma u|\le g$ in B_R from

$$\begin{aligned}\int_{B_R}\big(|D^\sigma u|-g\big)^+ &\le \varliminf_{\varepsilon\to0}\int_{B_R}\big(|D^\sigma u_\varepsilon|-g\big)^+\\ &= \varliminf_{\varepsilon\to0}\left[\int_{U_{\varepsilon,R}}\big(|D^\sigma u_\varepsilon|-g\big)+\int_{V_{\varepsilon,R}}\big(|D^\sigma u_\varepsilon|-g\big)\right]\end{aligned}$$

since

$$\int_{U_{\varepsilon,R}} \big(|D^\sigma u_\varepsilon| - g\big) \leq \frac{1}{g_R}\int_{U_{\varepsilon,R}} \big(|D^\sigma u_\varepsilon|^2 - g^2\big) \leq \frac{|B_R|\sqrt{\varepsilon}}{g_R},$$

$$\begin{aligned}\int_{V_{\varepsilon,R}} \big(|D^\sigma u_\varepsilon| - g\big) &\leq |V_{\varepsilon,R}|^{\frac12}\big(\|D^\sigma u_\varepsilon\|_{\boldsymbol{L}^2(B_R)} + \|g\|_{\boldsymbol{L}^2(B_R)}\big)\\ &\leq \big(C_\sigma + \|g\|_{\boldsymbol{L}^2(\mathbb{R}^N)}\big)|V_{\varepsilon,R}|^{\frac12}\end{aligned}$$

with

$$|V_{\varepsilon,R}| = \int_{V_{\varepsilon,R}} 1 \leq \int_{V_{\varepsilon,R}} \frac{\widehat{\kappa}_\varepsilon}{k_\varepsilon(\sqrt{\varepsilon})} \leq \sqrt{\varepsilon}\int_{B_R} \widehat{\kappa}_\varepsilon \leq \frac{a_* C_\sigma^2}{2g_R^2}\sqrt{\varepsilon}.$$

Now, observing that for arbitrary $v \in \mathbb{K}_g^\sigma$ we have

$$\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \cdot D^\sigma(v - u_\varepsilon) \leq \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|\big(|D^\sigma v| - |D^\sigma u_\varepsilon|\big) \leq 0$$

(since $\widehat{\kappa}_\varepsilon > 0$ if $|D^\sigma u_\varepsilon| > g \geq |D^\sigma v|$), from (2.16) we obtain

$$\int_{\mathbb{R}^N} AD^\sigma u_\varepsilon \cdot D^\sigma(v - u_\varepsilon) \geq \int_\Omega f_\#(v - u_\varepsilon) + \int_{\mathbb{R}^N} \boldsymbol{f}\cdot D^\sigma(v-u_\varepsilon), \quad \forall v \in \mathbb{K}_g^\sigma,$$

and, passing to the limit as $\varepsilon \to 0$, we conclude that u solves (2.9), by using (2.18) and the lower semi-continuity

$$\varliminf_{\varepsilon\to 0} \int_{\mathbb{R}^N} AD^\sigma u_\varepsilon \cdot D^\sigma u_\varepsilon \geq \int_{\mathbb{R}^N} AD^\sigma u \cdot D^\sigma u. \tag{2.23}$$

Finally, taking an arbitrary $\boldsymbol{G} = D^\sigma v \in \Psi_\sigma$ and taking $\varepsilon \to 0$ in (2.16), by recalling (2.21), (2.12), and (2.5) we find

$$\begin{aligned}\langle \Lambda, \boldsymbol{G}\rangle_{\Psi_\sigma} &= \lim_{\varepsilon\to 0}\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \cdot D^\sigma v = \int_{\mathbb{R}^N} (D^\sigma\phi - AD^\sigma u)\cdot D^\sigma v\\ &= \int_{\mathbb{R}^N} D^\sigma\gamma \cdot D^\sigma v,\end{aligned}$$

yielding the conclusion (2.19), by the uniqueness of u and γ. ■

3. The charges approach with a σ-gradient constraint in L^∞

In the framework of the previous section, we consider now the convex set $\mathbb{K}_g^\sigma$ defined by (2.7) with the assumption

$$g \in L^\infty(\mathbb{R}^N), \quad 0 < g_* \leq g(x) \leq g^* \text{ a.e. } x \text{ in } \mathbb{R}^N, \tag{3.1}$$

for some constants g_* and g^*. It is clear that $\mathbb{K}_g^\sigma$ is still closed for the topology of $H_0^\sigma(\Omega)$ in the space

$$\Upsilon_\infty^\sigma(\Omega) = \big\{v \in H_0^\sigma(\Omega) : D^\sigma v \in \boldsymbol{L}^\infty(\mathbb{R}^N)\big\}, \quad 0 < \sigma \le 1, \tag{3.2}$$

and therefore, by the fractional Morrey–Sobolev inequality (1.7) for $\sigma > \frac{N}{p}$, we have, for all $0 < \beta < \sigma$,

$$\mathbb{K}_g^\sigma \subset \Upsilon_\infty^\sigma(\Omega) \subset \mathcal{C}^{0,\beta}(\bar{\Omega}) \subset L^\infty(\Omega). \tag{3.3}$$

Here $\mathcal{C}^{0,\beta}(\bar{\Omega})$ is the space of the Hölder continuous functions with exponent β. As observed in [10], (3.3) is a consequence of Theorem 7.63 of [1] (see also [12, Theorem 2.2]), which yields

$$\begin{aligned}\|u\|_{L^\infty(\Omega)} &\le C_p \|D^\sigma u\|_{\boldsymbol{L}^p(\mathbb{R}^N)} \\ &\le C_p \|D^\sigma u\|^{1-\frac{2}{p}}_{\boldsymbol{L}^\infty(\mathbb{R}^N)} \|D^\sigma u\|^{\frac{2}{p}}_{\boldsymbol{L}^2(\mathbb{R}^N)}, \quad \forall u \in \Upsilon_\infty^\sigma(\Omega),\end{aligned} \tag{3.4}$$

where $C_p > 0$ is the Sobolev constant corresponding to any $p > \frac{N}{\sigma} \vee 2$.

Therefore, in this case, we can extend the result of the solvability of the variational inequality (2.9) with data in L^1:

$$f_\# \in L^1(\Omega) \quad \text{and} \quad \boldsymbol{f} \in \boldsymbol{L}^1(\mathbb{R}^N). \tag{3.5}$$

Theorem 3.1. *Under the assumptions* (2.1), (2.2), (2.3), *and* (3.1), *the unique solution u to* (2.9) *also satisfies the continuous dependence estimates* (2.10). *Moreover, if in addition $(\boldsymbol{f}, f_\#)$ and $(\hat{\boldsymbol{f}}, \hat{f}_\#)$ also satisfy* (3.5), *the following estimate holds:*

$$\|u - \hat{u}\|_{H_0^\sigma(\Omega)} \le a_p \|f_\# - \hat{f}_\#\|^{\frac{1}{2-\frac{2}{p}}}_{L^1(\Omega)} + b_1 \|\boldsymbol{f} - \hat{\boldsymbol{f}}\|^{\frac{1}{2}}_{\boldsymbol{L}^1(\mathbb{R}^N)}, \tag{3.6}$$

where $p > \frac{N}{\sigma} \vee 2$ as in (3.4) *and $a_p, b_1 > 0$ are constants.*

Consequently, the variational inequality (2.9) *is also uniquely solvable with the assumption* (2.3) *replaced by* (3.5) *and the estimate* (3.6) *still holds in this case.*

Proof. While the first part of this theorem is also a direct consequence of the Stampacchia theorem, the estimate (3.6) follows easily from (2.9). Indeed, if we set $\bar{u} = u - \hat{u}$, $\bar{f}_\# = f_\# - \hat{f}_\#$, and $\bar{\boldsymbol{f}} = \boldsymbol{f} - \hat{\boldsymbol{f}}$, we have

$$\begin{aligned}a_* \|\bar{u}\|^2_{H_0^\sigma(\Omega)} &= a_* \int_{\mathbb{R}^N} \|D^\sigma \bar{u}\|^2 \\ &\le \|\bar{u}\|_{L^\infty(\Omega)} \|\bar{f}_\#\|_{L^1(\Omega)} + \|D^\sigma \bar{u}\|_{\boldsymbol{L}^\infty(\Omega)} \|\bar{\boldsymbol{f}}\|_{\boldsymbol{L}^1(\Omega)} \\ &\le C_p (2g^*)^{1-\frac{2}{p}} \|D^\sigma \bar{u}\|^{\frac{2}{p}}_{\boldsymbol{L}^2(\Omega)} \|\bar{f}_\#\|_{L^1(\Omega)} + 2g^* \|\bar{\boldsymbol{f}}\|_{\boldsymbol{L}^1(\Omega)},\end{aligned} \tag{3.7}$$

by (3.4) and the assumption (3.1). Hence, (3.6) follows easily by applying Young's inequality and $\sqrt{\phi+\psi}\leq\sqrt{\phi}+\sqrt{\psi}$ to the right-hand side of (3.7), where we obtain the constants a_p and b_1 depending on C_p, a_*, g^*, and $p>\frac{N}{\sigma}\vee 2$. The solvability of (2.9) under the assumption (3.5) can be easily obtained using (3.6), approximating the solution by a Cauchy sequence in $H^\sigma_0(\Omega)$ of solutions $u_\nu \xrightarrow[\nu\to 0]{} u$, where u_ν solves (2.9) with approximating sequences

$$f_{\#_\nu} \xrightarrow[\nu\to 0]{} f_\# \text{ in } L^1(\Omega) \quad\text{and}\quad \boldsymbol{f}_\nu \xrightarrow[\nu\to 0]{} \boldsymbol{f} \text{ in } L^1(\mathbb{R}^N) \tag{3.8}$$

with $f_{\#_\nu}\in L^2(\Omega)$ and $\boldsymbol{f}_\nu\in \boldsymbol{L}^2(\mathbb{R}^N)$, for instance, with $f_\nu=(f\wedge\frac{1}{\nu})\vee(-\frac{1}{\nu})$ by truncation. ■

Remark 3.2. This result with L^1-data extends Theorem 2.1 of [10] which considered only the case $\boldsymbol{f}\equiv 0$. If the data $f_\#\in L^{2^\#}(\Omega)$ and $\boldsymbol{f}\in\boldsymbol{L}^2(\mathbb{R}^N)\cap\boldsymbol{L}^1(\mathbb{R}^N)$ hold, our approximation Theorem 2.3 also holds for the solution (u,γ) to (2.11)-(2.12) under the assumption (3.1), which implies $g\in L^2(B_R)$ for all $R>0$, since the proof is the same.

It is also possible to obtain with L^1-data the $\frac{1}{2}$-Hölder continuity of the map $L^\infty(\mathbb{R}^N)\ni g\mapsto u\in H^\sigma_0(\Omega)$ with g satisfying (3.1) and u solution to (2.9), extending Theorem 2.2 of [10].

Theorem 3.3. *Under the assumptions* (2.1), (2.2), *and* (3.5), *let u and $\hat{u}$ be the solutions to* (2.9) *corresponding to g and $\hat{g}$ satisfying* (3.1). *Then, there exists a constant $C_*>0$, depending on g_* and the data, but independent of the solutions, such that*

$$\|u-\hat{u}\|_{H^\sigma_0(\Omega)}\leq C_*\|g-\hat{g}\|^{\frac{1}{2}}_{L^\infty(\mathbb{R}^N)}. \tag{3.9}$$

Proof. Denote $\delta=\|g-\hat{g}\|_{L^\infty(\mathbb{R}^N)}$, and take as test functions in (2.9), respectively,

$$w=\frac{g_*}{g_*+\delta}\hat{u}\in\mathbb{K}^\sigma_g \quad\text{and}\quad \hat{w}=\frac{g_*}{g_*+\delta}u\in\mathbb{K}^\sigma_{\hat{g}}$$

for the variational inequality for u and for $\hat{u}$.

Observing that

$$|u-\hat{w}|\leq\frac{\delta}{g_*}|u| \quad\text{and}\quad \left|D^\sigma(u-\hat{w})\right|\leq\frac{\delta}{g_*}|D^\sigma u|$$

and similarly for $\hat{u}-w$, we obtain (3.9) from

$$\begin{aligned}a_*\|u-\hat{u}\|^2_{H^\sigma_0(\Omega)} &\leq \int_{\mathbb{R}^N} AD^\sigma(u-\hat{u})\cdot D^\sigma(u-\hat{u})\\ &= \int_{\mathbb{R}^N} AD^\sigma u\cdot D^\sigma(u-w)+\int_{\mathbb{R}^N} AD^\sigma u\cdot D^\sigma(w-\hat{u})\\ &\quad+\int_{\mathbb{R}^N} AD^\sigma\hat{u}\cdot D^\sigma(\hat{u}-\hat{w})+\int_{\mathbb{R}^N} AD^\sigma\hat{u}\cdot D^\sigma(\hat{w}-u)\end{aligned}$$

$$
\begin{aligned}
&\leq \int_\Omega f_\#\big((u-w)+(\hat{u}-\hat{w})\big)+\int_{\mathbb{R}^N} \boldsymbol{f}\cdot D^\sigma\big((u-w)+(\hat{u}-\hat{w})\big)\\
&\quad+\frac{2\delta}{g_*}\int_{\mathbb{R}^N}|AD^\sigma u\cdot D^\sigma\hat{u}|\\
&=\int_\Omega f_\#\big((u-\hat{w})+(\hat{u}-w)\big)+\int_{\mathbb{R}^N} \boldsymbol{f}\cdot D^\sigma\big((u-\hat{w})+(\hat{u}-w)\big)\\
&\quad+\frac{2\delta}{g_*}\int_{\mathbb{R}^N}|AD^\sigma u\cdot D^\sigma\hat{u}|\\
&\leq \frac{2\delta}{g_*}\big(C_p {g^*}^{1-\frac{2}{p}}\eta_p^{\frac{2}{p}}\|f_\#\|_{L^1(\Omega)}+g^*\|\boldsymbol{f}\|_{\boldsymbol{L}^1(\mathbb{R}^N)}+a^*\eta_p^2\big),
\end{aligned}
$$

by using (3.4) and $\eta_p = a_p\|f_\#\|_{L^1(\Omega)}^{\frac{1}{2-\frac{2}{p}}} + b_1\|\boldsymbol{f}\|_{\boldsymbol{L}^2(\mathbb{R}^N)}^{\frac{1}{2}}$, which is a general upper bound for $\|D^\sigma u\|_{L^2(\mathbb{R}^N)}$ and $\|D^\sigma \hat{u}\|_{L^2(\mathbb{R}^N)}$, just by taking $v\equiv 0$ in (2.9) and calculating as in (3.6). ■

Remark 3.4. This theorem allows to obtain solutions to quasi-variational inequalities of the type (2.9), with the solution dependent on the convex sets $\mathbb{K}^\sigma_{G[u]}$ as in (2.7) with $g = G[u]$, where $G : L^{2^*}(\Omega)\to L^\infty_{g_*}(\mathbb{R}^N)$, being $L^\infty_{g_*}(\mathbb{R}^N)=\{h\in L^\infty(\mathbb{R}^N) : h(x)\geq g_*>0 \text{ a.e. } x\in\mathbb{R}^N\}$, or $G : \mathcal{C}(\overline{\Omega})\to L^\infty_{g_*}(\mathbb{R}^N)$ are continuous and bounded operators, as in [10, Section 4], where only the case $f_\#\in L^2(\Omega)$ and $\boldsymbol{f}\equiv 0$ was considered.

As we observed in Remark 3.2, the solution u to the variational inequality with bounded σ-gradient constraint and data satisfying (2.3) also solves (2.12), but the extra terms involving γ can be interpreted with a Lagrange multiplier λ in a generalised sense extending Theorem 3.1 of [10] to L^1-data. Here we use the duality in $L^\infty(\mathbb{R}^N)$ and in $\boldsymbol{L}^\infty(\mathbb{R}^N)$ with the notation

$$
\langle\!\langle \lambda\boldsymbol{\alpha},\boldsymbol{\beta}\rangle\!\rangle=\langle\lambda,\boldsymbol{\alpha}\cdot\boldsymbol{\beta}\rangle,\quad \forall\lambda\in L^\infty(\mathbb{R}^N)'\ \forall\boldsymbol{\alpha},\boldsymbol{\beta}\in\boldsymbol{L}^\infty(\mathbb{R}^N). \tag{3.10}
$$

Theorem 3.5. *Under the assumptions* (2.1), (2.2), (3.1), *and* (2.3) *or* (3.5), *there exists* $(u,\lambda)\in\Upsilon^\sigma_\infty(\Omega)\times L^\infty(\mathbb{R}^N)'$, *such that*

$$
\begin{aligned}
&\int_{\mathbb{R}^N} AD^\sigma u\cdot D^\sigma w+\langle\!\langle\lambda D^\sigma u, D^\sigma w\rangle\!\rangle\\
&\qquad=\int_\Omega f_\# w+\int_{\mathbb{R}^N}\boldsymbol{f}\cdot D^\sigma w,\quad \forall w\in\Upsilon^\sigma_\infty(\Omega),
\end{aligned}
\tag{3.11}
$$

$$
|D^\sigma u|\leq g \text{ a.e. in } \mathbb{R}^N,\quad \lambda\geq 0\quad\text{and}\quad \lambda\big(|D^\sigma u|-g\big)=0 \text{ in } L^\infty(\mathbb{R}^N)'. \tag{3.12}
$$

Moreover, u is the unique solution to the variational inequality (2.9).

Proof. (i) First we suppose (2.3), i.e., $f_\# \in L^2(\Omega)$ and $\boldsymbol{f} \in \boldsymbol{L}^2(\mathbb{R}^N)$, and, from the approximation problem (2.16), in addition to (2.20), we obtain the *a priori* estimates independent of $0 < \varepsilon < 1$:

$$\|\widehat{\kappa}_\varepsilon\|_{L^1(\mathbb{R}^N)} \leq \frac{a_*}{2g_*^2} C_\sigma^2 \equiv \frac{C_1}{g_*^2}, \tag{3.13}$$

$$\|\widehat{\kappa}_\varepsilon\|_{L^\infty(\mathbb{R}^N)'} \leq \frac{C_1}{g_*^2}, \tag{3.14}$$

$$\|\widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)'} \leq \frac{C_1}{g_*}. \tag{3.15}$$

Indeed, (3.13) follows from (2.22) with the assumption (3.1), which implies (3.14), by definition of the dual norm, as well as (3.15), by using (3.13) and again (2.22):

$$\begin{aligned} \|\widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)'} &= \sup_{\substack{\boldsymbol{\beta} \in \boldsymbol{L}^\infty(\mathbb{R}^N) \\ \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)}=1}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \cdot \boldsymbol{\beta} \\ &\leq \left(\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|^2 \right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon \right)^{\frac{1}{2}} \leq \frac{C_1}{g_*}. \end{aligned}$$

By the estimates (3.14), (3.15), and the Banach–Alaoglu–Bourbaki theorem, at least for some generalised subsequence $u_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} u$ in $H_0^\sigma(\Omega)$ also

$$\widehat{\kappa}_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} \lambda \text{ weakly in } L^\infty(\mathbb{R}^N)' \quad \text{and} \quad \widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \underset{\varepsilon\to 0}{\longrightarrow} \Lambda \text{ weakly in } \boldsymbol{L}^\infty(\mathbb{R}^N)'.$$

Since $\widehat{\kappa}_\varepsilon \geq 0$ a.e., $\lambda \geq 0$ in $L^\infty(\mathbb{R}^N)'$, and letting $\varepsilon \to 0$ in (2.16) with $w \in \Upsilon_\infty^\sigma(\Omega)$, u and Λ satisfy

$$\begin{aligned} &\int_{\mathbb{R}^N} A D^\sigma u \cdot D^\sigma w + \langle\!\langle \Lambda, D^\sigma w \rangle\!\rangle \\ &\qquad = \int_\Omega f_\# w + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma w, \quad \forall w \in \Upsilon_\infty^\sigma(\Omega). \end{aligned} \tag{3.16}$$

Letting $\varepsilon \to 0$ in (2.16) with $v = u_\varepsilon$ and using (2.23), we easily find that

$$\overline{\lim_{\varepsilon\to 0}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|^2 \leq \langle\!\langle \Lambda, D^\sigma u \rangle\!\rangle.$$

Recalling that $(|D^\sigma u_\varepsilon|^2 - g^2)\widehat{\kappa}_\varepsilon \geq 0$ and $|D^\sigma u| \leq g$ a.e. $x \in \mathbb{R}^N$, we obtain

$$\left\langle \lambda, |D^\sigma u|^2 \right\rangle \leq \langle \lambda, g^2 \rangle = \lim_{\varepsilon\to 0} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon g^2 \leq \overline{\lim_{\varepsilon\to 0}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|^2 \leq \langle\!\langle \Lambda, D^\sigma u \rangle\!\rangle.$$

Since we get the opposite inequality from

$$\begin{aligned}0 &\le \overline{\lim_{\varepsilon\to 0}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon \big|D^\sigma(u_\varepsilon - u)\big|^2 \\ &= \overline{\lim_{\varepsilon\to 0}} \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|^2 - 2\lim_{\varepsilon\to 0}\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma u_\varepsilon \cdot D^\sigma u + \lim_{\varepsilon\to 0}\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u|^2 \\ &\le \langle\!\langle \Lambda, D^\sigma u\rangle\!\rangle - 2\langle\!\langle \Lambda, D^\sigma u\rangle\!\rangle + \big\langle \lambda, |D^\sigma u|^2\big\rangle = -\langle\!\langle \Lambda, D^\sigma u\rangle\!\rangle + \big\langle \lambda, |D^\sigma u|^2\big\rangle,\end{aligned}$$

we conclude $\langle\!\langle \Lambda, D^\sigma u\rangle\!\rangle = \langle \lambda, |D^\sigma u|^2\rangle$ and

$$\lim_{\varepsilon\to 0}\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon \big|D^\sigma(u_\varepsilon - u)\big|^2 = 0. \tag{3.17}$$

Hence, for any $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(\mathbb{R}^N)$, we have

$$\begin{aligned}\big|\langle\!\langle \Lambda - \lambda D^\sigma u, \boldsymbol{\beta}\rangle\!\rangle\big| &= \lim_{\varepsilon\to 0}\left|\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma(u_\varepsilon - u)\cdot \boldsymbol{\beta}\right| \\ &\le \lim_{\varepsilon\to 0}\left[\left(\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon \big|D^\sigma(u_\varepsilon - u)\big|^2\right)^{\frac{1}{2}} \|\widehat{\kappa}_\varepsilon\|_{L^1(\mathbb{R}^N)} \|\boldsymbol{\beta}\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)}\right] = 0,\end{aligned}$$

showing that

$$\Lambda = \lambda D^\sigma u \quad \text{in } \boldsymbol{L}^\infty(\mathbb{R}^N)'$$

and that, in fact, (3.16) is equivalent to (3.11).

It remains to show the last equation of (3.12) which follows easily from (recall (3.1))

$$\begin{aligned}0 &= \big\langle \lambda, (g^2 - |D^\sigma u|^2)\varphi\big\rangle = \big\langle \lambda, \big(g - |D^\sigma u|\big)\big(g + |D^\sigma u|\big)\varphi\big\rangle \\ &\ge g_*\big\langle \lambda, \big(g - |D^\sigma u|\big)\varphi\big\rangle = g_*\big\langle \lambda\big(g - |D^\sigma u|\big), \varphi\big\rangle \ge 0\end{aligned}$$

for arbitrarily $\varphi \in L^\infty(\Omega)$, $\varphi \ge 0$, which holds provided that we show

$$\big\langle \lambda, (g^2 - |D^\sigma u|^2)\varphi\big\rangle = 0. \tag{3.18}$$

As above, using (3.17), we have first

$$\begin{aligned}\langle \lambda, g^2\varphi\rangle &\le \lim_{\varepsilon\to 0}\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u_\varepsilon|^2 \varphi \\ &= \lim_{\varepsilon\to 0}\bigg(\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon \big|D^\sigma(u_\varepsilon - u)\big|^2\varphi \\ &\qquad + 2\int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon D^\sigma(u_\varepsilon - u)\cdot D^\sigma u\varphi + \int_{\mathbb{R}^N} \widehat{\kappa}_\varepsilon |D^\sigma u|^2\varphi\bigg) \\ &= \big\langle \lambda, |D^\sigma u|^2\varphi\big\rangle\end{aligned}$$

and, since $u \in \mathbb{K}^\sigma_g$ and $\varphi, \lambda \geq 0$, it also holds that

$$\big\langle \lambda, \big(g^2 - |D^\sigma u|^2\big)\varphi \big\rangle \geq 0.$$

To show that u is the unique solution to (2.9), it suffices to take $w = u - v$, with an arbitrary $v \in \mathbb{K}^\sigma_g$, and observe that, by (3.18),

$$\begin{aligned} \langle\!\langle \lambda D^\sigma u, D^\sigma(v-u) \rangle\!\rangle &\leq \big\langle \lambda, |D^\sigma u|\big(|D^\sigma v| - |D^\sigma u|\big)\big\rangle \\ &\leq \big\langle \lambda, |D^\sigma u|\big(g - |D^\sigma u|\big)\big\rangle \\ &= \left\langle \lambda\big(g^2 - |D^\sigma u|^2\big), \frac{|D^\sigma u|}{g + |D^\sigma u|} \right\rangle = 0. \end{aligned}$$

(ii) In the second case, if (3.5) holds, we can use approximation by solutions (u_ν, λ_ν) of (3.11)-(3.12) corresponding to data $f_{\#\nu} \in L^{2^\#}(\Omega)$ and $\boldsymbol{f}_\nu \in \boldsymbol{L}^2(\mathbb{R}^N)$ satisfying (3.8), as in Theorem 3.1.

Using the estimate (3.6), it is clear that

$$u_\nu \xrightarrow[\nu \to 0]{} u \quad \text{in } H^\sigma_0(\Omega) \tag{3.19}$$

and u solves (2.9).

For $\varphi \in L^\infty(\mathbb{R}^N)$, setting $b = \frac{\|\varphi\|_{L^\infty(\mathbb{R}^N)}}{g_*^2}$, recalling (3.1), and using (3.11) and (3.12) for λ_ν, which also implies that $\langle \lambda_\nu, g^2 - |D^\sigma u_\nu|^2 \rangle = 0$, we have

$$\begin{aligned} \big|\langle \lambda_\nu, \varphi \rangle\big| &\leq \langle \lambda_\nu, bg^2 \rangle \\ &= b\big\langle \lambda_\nu, |D^\sigma u_\nu|^2 \big\rangle = b\langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma u_\nu \rangle\!\rangle \\ &\leq b\bigg(\int_\Omega f_\# u_\nu + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma u_\nu \bigg) \leq C\,\frac{\|\varphi\|_{L^\infty(\mathbb{R}^N)}}{g_*^2}, \end{aligned} \tag{3.20}$$

where the constant $C > 0$ depends only on the L^1-norms of $f_\#$ and $\boldsymbol{f}$ and on the constants a_p and b_1 of (3.6), being consequently independent of ν. Then, λ_ν is uniformly bounded in $L^\infty(\mathbb{R}^N)'$ and we may assume, for some generalised subsequence,

$$\lambda_\nu \underset{\nu \to 0}{\rightharpoonup} \lambda \text{ in } L^\infty(\mathbb{R}^N)'\text{-weakly}^*, \quad \text{with } \lambda \geq 0, \tag{3.21}$$

and, since $\Lambda_\nu = \lambda_\nu D^\sigma u_\nu$ is also bounded in $\boldsymbol{L}^\infty(\mathbb{R}^N)'$ (recall $\|D^\sigma u_\nu\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)} \leq g^*$), also

$$\Lambda_\nu \underset{\nu \to 0}{\rightharpoonup} \Lambda \quad \text{in } \boldsymbol{L}^\infty(\mathbb{R}^N)'\text{-weakly}^*. \tag{3.22}$$

Therefore, taking the limit $\nu \to 0$ in (3.11), we find that (u, λ) solves

$$\begin{aligned} &\int_{\mathbb{R}^N} AD^\sigma u \cdot D^\sigma w + \langle\!\langle \Lambda, D^\sigma w \rangle\!\rangle \\ &\qquad = \int_\Omega f_\# w + \int_{\mathbb{R}^N} \boldsymbol{f} \cdot D^\sigma w, \quad \forall w \in \Upsilon^\sigma_\infty(\Omega). \end{aligned} \tag{3.23}$$

Recalling (3.18) with $\varphi = 1$, we have

$$\left\langle \lambda_\nu, |D^\sigma u|^2 \right\rangle \leq \langle \lambda_\nu, g^2 \rangle = \left\langle \lambda_\nu, |D^\sigma u_\nu|^2 \right\rangle. \tag{3.24}$$

Using the equalities (3.24) and (3.19), we have

$$\begin{aligned} 0 &\leq \frac{1}{2}\left\langle \lambda_\nu, \left|D^\sigma (u_\nu - u)\right|^2 \right\rangle \\ &= \frac{1}{2}\left(\left\langle \lambda_\nu, |D^\sigma u_\nu|^2 \right\rangle - 2\langle \lambda_\nu, D^\sigma u_\nu \cdot D^\sigma u\rangle + \left\langle \lambda_\nu, |D^\sigma u|^2 \right\rangle\right) \\ &\leq \left\langle \lambda_\nu, |D^\sigma u_\nu|^2 \right\rangle - \left\langle \lambda_\nu, D^\sigma u_\nu \cdot D^\sigma u \right\rangle = \langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma (u_\nu - u) \rangle\!\rangle \\ &= \int_\Omega f_{\#\nu}(u_\nu - u) + \int_{\mathbb{R}^N} \boldsymbol{f}_\nu \cdot D^\sigma (u_\nu - u) \\ &\quad - \int_{\mathbb{R}^N} A D^\sigma u_\nu \cdot D^\sigma (u_\nu - u) \xrightarrow[\nu \to 0]{} 0, \end{aligned} \tag{3.25}$$

being the last equality satisfied because (u_ν, λ_ν) solves problem (3.11)-(3.12) with data $f_{\#\nu}$ and $\boldsymbol{f}_\nu$.

Then, from (3.23) we can conclude that u in fact solves (3.11) from the equality

$$\begin{aligned} \langle\!\langle \Lambda, D^\sigma w \rangle\!\rangle &= \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma w \rangle\!\rangle \\ &= \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma u, D^\sigma w \rangle\!\rangle + \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma (u_\nu - u), D^\sigma w \rangle\!\rangle \\ &= \lim_{\nu \to 0} \langle \lambda_\nu, D^\sigma u \cdot D^\sigma w \rangle = \langle \lambda, D^\sigma u \cdot D^\sigma w \rangle = \langle\!\langle \lambda D^\sigma u, D^\sigma w \rangle\!\rangle, \end{aligned} \tag{3.26}$$

which is valid for all $w \in \Upsilon^\sigma_\infty(\Omega)$ since (3.25) implies that

$$\begin{aligned} \left|\langle\!\langle \lambda_\nu D^\sigma (u_\nu - u), D^\sigma w \rangle\!\rangle\right| &= \left|\left\langle \lambda_\nu, D^\sigma (u_\nu - u) \cdot D^\sigma w \right\rangle\right| \\ &\leq \left\langle \lambda_\nu, \left|D^\sigma (u_\nu - u)\right| |D^\sigma w| \right\rangle \\ &\leq \left(\left\langle \lambda_\nu, \left|D^\sigma (u_\nu - u)\right|^2 \right\rangle\right)^{\frac{1}{2}} \left(\left\langle \lambda_\nu, |D^\sigma w|^2 \right\rangle\right)^{\frac{1}{2}} \xrightarrow[\nu \to 0]{} 0, \end{aligned}$$

where we have used the Hölder inequality for charges in the last inequality.

From (3.26), we find $\langle\!\langle \Lambda, D^\sigma u \rangle\!\rangle = \langle \lambda, |D^\sigma u|^2 \rangle$ and

$$\begin{aligned} \langle \lambda, g^2 \rangle &= \lim_{\nu \to 0} \langle \lambda_\nu, g^2 \rangle = \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma u_\nu \rangle\!\rangle \\ &= \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma u \rangle\!\rangle + \lim_{\nu \to 0} \langle\!\langle \lambda_\nu D^\sigma u_\nu, D^\sigma (u_\nu - u) \rangle\!\rangle \\ &= \lim_{\nu \to 0} \langle\!\langle \Lambda_\nu, D^\sigma u \rangle\!\rangle = \langle\!\langle \Lambda, D^\sigma u \rangle\!\rangle = \left\langle \lambda, |D^\sigma u|^2 \right\rangle. \end{aligned}$$

Finally, we can now complete the proof of the theorem by using this equality in the form $\langle \lambda(g^2 - |D^\sigma u|^2), 1 \rangle = 0$ and again the Hölder inequality to conclude the third condition in (3.12) with an arbitrarily $\varphi \in L^\infty(\mathbb{R}^N)$,

$$\begin{aligned}\big|\langle\lambda(g-|D^\sigma u|),\varphi\rangle\big| &\le \langle\lambda(g-|D^\sigma u|),|\varphi|\rangle\\ &= \left\langle\lambda\big(g^2-|D^\sigma u|^2\big), \frac{|\varphi|}{g+|D^\sigma u|}\right\rangle\\ &\le \langle\lambda\big(g^2-|D^\sigma u|^2\big),1\rangle^{\frac12}\left\langle\lambda\big(g^2-|D^\sigma u|^2\big), \frac{|\varphi|^2}{\big(g+|D^\sigma u|\big)^2}\right\rangle^{\frac12}\\ &= 0. \qquad\blacksquare\end{aligned}$$

The second part of this proof actually shows a generalised continuous dependence of the solution and of the Lagrange multiplier with respect to the L^1-data.

Corollary. *Under the assumptions* (2.1)*,* (2.2)*,* (3.1)*, and* (3.5)*, if* $(u_\nu,\lambda_\nu)\in\Upsilon^\sigma_\infty(\Omega)\times L^\infty(\mathbb{R}^N)'$ *are the solutions to* (3.11) *and* (3.12) *corresponding to* L^1*-data satisfying* (3.8)*, as* $\nu\to 0$*, we have the convergence, for some generalised subsequence or net,*

$$u_\nu \xrightarrow[\nu\to 0]{} u \text{ in } H^\sigma_0(\Omega) \quad \text{and} \quad \lambda_\nu \underset{\nu\to 0}{\rightharpoonup} \lambda \text{ in } L^\infty(\mathbb{R}^N)'\text{-weakly}^*,$$

where $(u,\lambda)\in\Upsilon^\sigma_\infty(\Omega)\times L^\infty(\mathbb{R}^N)'$ *also solves* (3.11)*-*(3.12)*.*

4. Convergence to the local problem as $\sigma \nearrow 1$

It is easy to check that all the theorems of the preceding two sections hold in the limit case $\sigma = 1$, when $D^\sigma = D$ is the classical gradient and the data $f_\#$ and $\boldsymbol{f}$ satisfy (2.3) (with $f_\# \in L^{\frac{2N}{N+2}}(\Omega)$, if $N>2$, $f_\#\in L^q(\Omega)$, $\forall q<\infty$ if $N=2$ and $q=\infty$ if $N=1$) or (3.5), and g satisfies (2.8), (2.17) or (3.1), respectively.

In this section, we show a continuous dependence of the solution u^σ and of the Lagrange multiplier λ^σ when $\sigma\nearrow 1$. For the sake of simplicity, we take $f_\#=0$ and $\boldsymbol{f}\in \boldsymbol{L}^1(\mathbb{R}^N)$, so that the limit variational inequality reads

$$u\in\mathbb{K}_g=\big\{v\in H^1_0(\Omega): |Dv|\le g \text{ a.e. in } \Omega\big\}, \tag{4.1}$$

$$\int_\Omega ADu\cdot D(v-u) \ge \int_\Omega \boldsymbol{f}\cdot D(v-u), \quad \forall v\in\mathbb{K}_g. \tag{4.2}$$

Likewise, observing that setting $\sigma=1$ in (3.2) we have $\Upsilon_\infty(\Omega)=W^{1,\infty}_0(\Omega)$, we can write the limit Lagrange multiplier problem in the following form: find $(u,\lambda)\in W^{1,\infty}_0(\Omega)\times L^\infty(\Omega)'$

$$\int_\Omega ADu\cdot Dw + \langle\!\langle\lambda Du, Dw\rangle\!\rangle = \int_\Omega \boldsymbol{f}\cdot Dw, \quad \forall w\in W^{1,\infty}_0(\Omega), \tag{4.3}$$

$$|Du|\le g \text{ a.e. in } \Omega, \quad \lambda\ge 0 \quad \text{and} \quad \lambda\big(|Du|-g\big)=0 \text{ in } L^\infty(\Omega)'. \tag{4.4}$$

In (4.3), we denote the duality in $\boldsymbol{L}^\infty(\Omega)$ similarly to (3.10), as we can always consider the solution and the test functions extended by zero in $\mathbb{R}^N \setminus \Omega$, since $\partial\Omega$ is $\mathcal{C}^{0,1}$.

We first recall an important consequence of the fact that the Riesz kernel is an approximation of the identity, as remarked by Kurokawa in [7].

Proposition 4.1. *If $h \in L^p(\mathbb{R}^N) \cap \mathcal{C}(\mathbb{R}^N)$, for some $p \geq 1$, is bounded and uniformly continuous in $\mathbb{R}^N$, then*

$$\lim_{\alpha \to 0} \|I_\alpha * h - h\|_{L^\infty(\mathbb{R}^N)} = 0.$$

As a consequence, we have

$$D^\sigma w \xrightarrow[\sigma \nearrow 1]{} Dw \quad \text{in } \boldsymbol{L}^\infty(\mathbb{R}^N), \text{ for all } w \in \mathcal{C}^1_c(\mathbb{R}^N). \tag{4.5}$$

Proof. In [7, Proposition 2.10], it is proved that

$$I_\alpha * h(x) \xrightarrow[\alpha \to 0]{} h(x)$$

at each point of continuity of any function $h \in L^p(\mathbb{R}^N)$, $1 \leq p < \infty$, and it is not difficult to check that this convergence is uniform in $x \in \mathbb{R}^N$ for bounded and uniformly continuous functions (see [2]). Then, (4.5) is an immediate consequence of Theorem 1.2 of [12], which established that $D^s w = I_{1-s} * Dw$ for all $w \in \mathcal{C}^\infty_c(\mathbb{R}^N)$, being the proof equally valid for functions only in $\mathcal{C}^1_c(\mathbb{R}^N)$. ∎

Remark 4.2. The convergence (4.5), as well as in $\boldsymbol{L}^p(\mathbb{R}^N)$ for $p \geq 1$, has been shown in [6, Proposition 4.4] for functions of $\mathcal{C}^2_c(\mathbb{R}^N)$. By density of $\mathcal{C}^\infty_c(\mathbb{R}^N)$ in $L^p(\mathbb{R}^N)$ for $p \geq 1$, in [4] it was shown that the convergence $D^\sigma h \xrightarrow[\sigma \nearrow 1]{} Dh$ holds in $L^p(\mathbb{R}^N)$, for $1 < p < \infty$, if $h \in W^{1,p}(\mathbb{R}^N)$.

For $\chi \in L^\infty(\mathbb{R}^N)'$, we denote its restriction to $\Omega \subset \mathbb{R}^N$ by $\chi_\Omega \in L^\infty(\Omega)'$, defined by

$$\langle \chi_\Omega, \varphi \rangle = \langle \chi, \tilde{\varphi} \rangle, \quad \forall \varphi \in L^\infty(\Omega),$$

where $\tilde{\varphi}$ is the extension of φ by zero to $\mathbb{R}^N \setminus \Omega$,

Theorem 4.3. *Let $f \in \boldsymbol{L}^1(\mathbb{R}^N)$ ($f_\# = 0$) and let g be given as in (3.1). Then, if $(u^\sigma, \lambda^\sigma) \in \Upsilon^\sigma_\infty(\Omega) \times L^\infty(\mathbb{R}^N)'$ are the solutions to (3.11)-(3.12), we have, for a generalised subsequence, the convergences, for any s, $0 < s < \sigma < 1$:*

$$u^\sigma \xrightarrow[\sigma \nearrow 1]{} u \text{ in } H^s_0(\Omega) \quad \text{and} \quad \lambda^\sigma_\Omega \xrightarrow[\sigma \nearrow 1]{} \lambda \text{ in } L^\infty(\Omega)'\text{-weakly*}, \tag{4.6}$$

where $(u, \lambda) \in W^{1,\infty}_0(\Omega) \times L^\infty(\Omega)'$ is a solution to (4.3)-(4.4) and u is the unique solution to (4.1)-(4.2).

Proof. Setting $v = 0$ in (2.9), or $w = u^\sigma$ in (3.11), we immediately obtain

$$\|u^\sigma\|_{H_0^\sigma(\Omega)} = \|D^\sigma u^\sigma\|_{\boldsymbol{L}^2(\mathbb{R}^N)} \leq \left(\frac{g^*}{a_*}\|\boldsymbol{f}\|_{L^1(\mathbb{R}^N)}\right)^{\frac{1}{2}} \equiv C_1, \tag{4.7}$$

where C_1 is independent of σ, $0 < \sigma < 1$. Hence, arguing as in (3.20), using (3.11)-(3.12), it also follows easily that

$$\|\lambda^\sigma\|_{L^\infty(\mathbb{R}^N)'} = \sup_{\substack{\varphi\in L^\infty(\mathbb{R}^N)\\ \|\varphi\|_{L^\infty(\mathbb{R}^N)}=1}} \langle\lambda^\sigma,\varphi\rangle \leq \frac{\|\boldsymbol{f}\|_{\boldsymbol{L}^1(\mathbb{R}^N)}}{g_*^2}. \tag{4.8}$$

Then, using $\Lambda^\sigma = \lambda^\sigma D^\sigma u^\sigma$ and recalling $\|D^\sigma u^\sigma\|_{\boldsymbol{L}^\infty(\mathbb{R}^N)} \leq g^*$, from the estimates (4.7) and (4.8), we may take a generalised subsequence $\sigma \nearrow 1$ such that, by the compactness of $H_0^\sigma(\Omega) \hookrightarrow H_0^s(\Omega)$, $0 < s < \sigma \leq 1$,

$$\begin{cases} u^\sigma \xrightarrow[\sigma\nearrow 1]{} u & \text{in } H_0^s(\Omega),\\ D^\sigma u^\sigma \underset{\sigma\nearrow 1}{\rightharpoonup} \boldsymbol{\chi} & \text{in } \boldsymbol{L}^2(\mathbb{R}^N)'\text{-weak and } \boldsymbol{L}^\infty(\mathbb{R}^N)'\text{-weak}^*, \end{cases} \tag{4.9}$$

$$\lambda^\sigma \underset{\sigma\nearrow 1}{\rightharpoonup} \widetilde{\lambda} \text{ in } L^\infty(\mathbb{R}^N)'\text{-weak}^*, \quad \Lambda^\sigma \underset{\sigma\nearrow 1}{\rightharpoonup} \widetilde{\Lambda} \text{ in } \boldsymbol{L}^\infty(\mathbb{R}^N)'\text{-weak}. \tag{4.10}$$

Denoting by $\tilde{u}^\sigma$ the extension of u^σ by zero to $\mathbb{R}^N \setminus \Omega$, from (4.9) we conclude that $\boldsymbol{\chi} = D\tilde{u}$ and in fact $u \in H_0^1(\Omega)$, and then $D\tilde{u} = \widetilde{Du}$. Indeed, recalling the convergence (4.5), we have

$$\begin{aligned}\int_{\mathbb{R}^N} \boldsymbol{\chi}\cdot\boldsymbol{\varphi} &= \lim_{\sigma\nearrow 1}\int_{\mathbb{R}^N} D^\sigma u^\sigma\cdot\boldsymbol{\varphi} = -\lim_{\sigma\nearrow 1}\int_{\mathbb{R}^N} \tilde{u}^\sigma(D^\sigma\cdot\boldsymbol{\varphi})\\ &= -\int_{\mathbb{R}^N}\tilde{u}(D\cdot\boldsymbol{\varphi}) = \int_{\mathbb{R}^N} D\tilde{u}\cdot\boldsymbol{\varphi},\end{aligned}$$

with an arbitrary $\boldsymbol{\varphi} \in \mathcal{C}_c^\infty(\mathbb{R}^N)$.

On the other hand, given any measurable set $\omega \subset \Omega$, we have now

$$\int_\omega |Du|^2 \leq \varliminf_{\sigma\nearrow 1}\int_\omega |D^\sigma u^\sigma|^2 \leq \int_\omega g^2$$

and therefore $|Du| \leq g$ a.e. in Ω, which yields $u \in \mathbb{K}_g \subset W_0^{1,\infty}(\Omega)$.

Passing to the limit $\sigma \nearrow 1$ in (3.11), first with $w \in \mathcal{C}_c^\infty(\Omega)$

$$\int_{\mathbb{R}^N} AD^\sigma u^\sigma\cdot D^\sigma w + \langle\!\langle\Lambda^\sigma, D^\sigma w\rangle\!\rangle = \int_{\mathbb{R}^N}\boldsymbol{f}\cdot D^\sigma w$$

and using (4.5), (4.9), and (4.10), since $\chi = \widetilde{Du}$ and $D\tilde{w} = \widetilde{Dw}$, we obtain

$$\int_\Omega ADu \cdot Dw + \langle\!\langle \Lambda, Dw \rangle\!\rangle = \int_\Omega f \cdot Dw, \tag{4.11}$$

by setting $\Lambda = \widetilde{\Lambda}_\Omega$ and $\langle\!\langle \Lambda, Dw \rangle\!\rangle = \langle\!\langle \widetilde{\Lambda}, D\tilde{w} \rangle\!\rangle$.

Note that for each $w \in W_0^{1,\infty}(\Omega)$ we may choose $w_\nu \in \mathcal{C}_c^\infty(\Omega)$ such that $w_\nu \xrightarrow[\nu\to\infty]{} w$ in $H_0^1(\Omega)$ and $Dw_\nu \underset{\nu\to\infty}{\rightharpoonup} Dw$ in $L^\infty(\Omega)$-weak* in (4.11) and we may pass to the generalised limit $\nu \to \infty$, concluding that (4.11) also holds for all $w \in W_0^{1,\infty}(\Omega)$. So, in order to see that u and $\lambda = \widetilde{\lambda}_{|\Omega}$, i.e., the restriction to Ω of the limit charge $\widetilde{\lambda}$ in (4.10), solve (4.3), we need to show that

$$\langle\!\langle \Lambda, Dw \rangle\!\rangle = \langle\!\langle \lambda Du, Dw \rangle\!\rangle = \langle \lambda, Du \cdot Dw \rangle, \quad \forall w \in W_0^{1,\infty}(\Omega). \tag{4.12}$$

We show first (4.12) for $w = u$, i.e., $\langle\!\langle \Lambda, Du \rangle\!\rangle = \langle \lambda, |Du|^2 \rangle$, in two steps.

Observing that $\widetilde{\lambda} \geq 0$ and $|Du| \leq g$, we have $\langle \lambda, |Du|^2 \rangle \leq \langle\!\langle \Lambda, Du \rangle\!\rangle$ from

$$\begin{aligned}\langle \lambda, |Du|^2 \rangle \leq \langle \widetilde{\lambda}, g^2 \rangle &= \lim_{\sigma\nearrow 1} \langle \lambda^\sigma, g^2 \rangle = \lim_{\sigma\nearrow 1} \langle \lambda^\sigma, |D^\sigma u^\sigma|^2 \rangle \\ &= \lim_{\sigma\nearrow 1} \langle\!\langle \lambda^\sigma D^\sigma u^\sigma, D^\sigma u^\sigma \rangle\!\rangle \\ &= \overline{\lim_{\sigma\nearrow 1}} \int_{\mathbb{R}^N} (f - AD^\sigma u^\sigma) \cdot D^\sigma u^\sigma \\ &\leq \int_{\mathbb{R}^N} (f - AD\tilde{u}) \cdot D\tilde{u} = \langle\!\langle \widetilde{\Lambda}, D\tilde{u} \rangle\!\rangle = \langle\!\langle \Lambda, Du \rangle\!\rangle. \end{aligned} \tag{4.13}$$

Note that $D^\sigma u^\sigma \underset{\sigma\nearrow 1}{\rightharpoonup} D\tilde{u}$ in $L^2(\mathbb{R}^N)$-weak and hence

$$\underline{\lim_{\sigma\nearrow 1}} \int_{\mathbb{R}^N} AD^\sigma u^\sigma \cdot D^\sigma u^\sigma \geq \int_{\mathbb{R}^N} AD\tilde{u} \cdot D\tilde{u} = \int_\Omega ADu \cdot Du.$$

On the other hand, we find $\langle\!\langle \Lambda, Du \rangle\!\rangle \leq \langle \lambda, |Du|^2 \rangle$ by noting that $\Lambda^\sigma = \lambda^\sigma D^\sigma u^\sigma$ and, similarly,

$$0 \leq \langle \lambda^\sigma, |D^\sigma u^\sigma - D\tilde{u}|^2 \rangle = \langle\!\langle \lambda^\sigma D^\sigma u^\sigma, D^\sigma u^\sigma \rangle\!\rangle - 2\langle\!\langle \Lambda^\sigma, D\tilde{u} \rangle\!\rangle + \langle \lambda^\sigma, |D\tilde{u}|^2 \rangle \tag{4.14}$$

yields

$$\begin{aligned} 2\langle\!\langle \widetilde{\Lambda}, D\tilde{u} \rangle\!\rangle = 2 \lim_{\sigma\nearrow 1} \langle\!\langle \Lambda^\sigma, D\tilde{u} \rangle\!\rangle &\leq \overline{\lim_{\sigma\nearrow 1}} \int_{\mathbb{R}^N} (f - AD^\sigma u^\sigma) \cdot D^\sigma u^\sigma + \lim_{\sigma\nearrow 1} \langle \lambda^\sigma, |D\tilde{u}|^2 \rangle \\ &\leq \int_{\mathbb{R}^N} (f - AD\tilde{u}) \cdot D\tilde{u} + \langle \lambda, |Du|^2 \rangle = \langle\!\langle \widetilde{\Lambda}, D\tilde{u} \rangle\!\rangle + \langle \lambda, |Du|^2 \rangle. \end{aligned}$$

As a consequence of $\langle\!\langle \Lambda, Du\rangle\!\rangle = \langle \lambda, |Du|^2\rangle$, from (4.14) we deduce

$$\lim_{\sigma\nearrow 1} \big\langle \lambda^\sigma, |D^\sigma u^\sigma - D\tilde{u}|^2\big\rangle = 0, \tag{4.15}$$

which by the Hölder inequality yields, for any $\boldsymbol{\beta} \in \boldsymbol{L}^\infty(\mathbb{R}^N)$,

$$\begin{aligned}\big|\langle\!\langle \widetilde{\Lambda} - \widetilde{\lambda} D\tilde{u}, \boldsymbol{\beta}\rangle\!\rangle\big| &= \lim_{\sigma\nearrow 1} \big|\langle\!\langle \Lambda^\sigma - \lambda^\sigma D\tilde{u}, \boldsymbol{\beta}\rangle\!\rangle\big| = \lim_{\sigma\nearrow 1} \big|\langle\!\langle \lambda^\sigma (D^\sigma u^\sigma - D\tilde{u}), \boldsymbol{\beta}\rangle\!\rangle\big| \\ &\leq \lim_{\sigma\nearrow 1} \big\langle \lambda^\sigma, |D^\sigma u^\sigma - D\tilde{u}|\, |\boldsymbol{\beta}|\big\rangle \\ &\leq \lim_{\sigma\nearrow 1} \big\langle \lambda^\sigma, |D^\sigma u^\sigma - D\tilde{u}|^2\big\rangle^{\frac{1}{2}} \big\langle \lambda^\sigma, |\boldsymbol{\beta}|^2\big\rangle^{\frac{1}{2}} = 0,\end{aligned}$$

and, consequently, (4.12) follows from

$$\Lambda = \lambda Du \quad \text{in } \boldsymbol{L}^\infty(\Omega)'.$$

This equality in (4.12) with $g > 0$ implies that

$$\big\langle \lambda, |Du|^2\big\rangle = \langle \widetilde{\lambda}, g^2\rangle \geq \langle \lambda, g^2_{|\Omega}\rangle \geq \big\langle \lambda, |Du|^2\big\rangle,$$

and $\langle \lambda, |Du|^2 - g^2\rangle = 0$ (here $g = g_{|\Omega}$). Then, exactly the same argument as at the end of the proof of Theorem 3.4 shows that λ and u satisfy the third condition of (4.4).

Finally, since we also have

$$\langle\!\langle \lambda Du, D(v-u)\rangle\!\rangle \leq 0, \quad \forall v \in \mathbb{K}_g,$$

(4.3) implies (4.2) and this concludes the proof of the theorem. ■

Remark 4.4. In the Hilbertian case of $g \in L^2(\Omega)$, $g \geq 0$, and $\boldsymbol{f} \in \boldsymbol{L}^2(\mathbb{R}^N)$, it is easy to show the convergence of the solutions $(u^\sigma, \gamma^\sigma) \in \Upsilon^\sigma_\infty(\Omega) \times H^\sigma_0(\Omega)$ given by Theorem 2.1, also in the case $f_\# = 0$ to simplify, as $\sigma \nearrow 1$ to the local problem for $(u, \gamma) \in W^{1,\infty}_0(\Omega) \times H^1_0(\Omega)$, satisfying (2.11) with $\sigma = 1$ and

$$\int_\Omega (ADu + D\sigma) \cdot Dv = \int_\Omega \boldsymbol{f} \cdot Dv, \quad \forall v \in H^1_0(\Omega). \tag{4.16}$$

Indeed, as in (2.10) and (2.13), the a priori estimates

$$\|u^\sigma\|_{H^\sigma_0(\Omega)} \leq \frac{1}{a_*} \|\boldsymbol{f}\|_{\boldsymbol{L}^2(\mathbb{R}^N)} \quad \text{and} \quad \|\gamma^\sigma\|_{H^\sigma_0(\Omega)} \leq \left(1 + \frac{a^*}{a_*}\right) \|\boldsymbol{f}\|_{\boldsymbol{L}^2(\mathbb{R}^N)}$$

allow us to take sequences

$$u^\sigma \xrightarrow[\sigma\nearrow 1]{} u \text{ and } \gamma^\sigma \xrightarrow[\sigma\nearrow 1]{} \gamma \text{ in } H^s_0(\Omega), \quad 0 < s < 1,$$

in (2.12) with $v \in H^1_0(\Omega) \subset H^\sigma_0(\Omega)$, in order to obtain (4.16) and, using (2.18), the $\Gamma = \Gamma(u) \in H^{-\sigma}(\Omega)$ corresponding to γ satisfies (2.11) with $\sigma = 1$.

Funding. The research of José-Francisco Rodrigues was partially done under the framework of the Project PTDC/MATPUR/28686/2017 at CMAFcIO/ULisboa. The research of Assis Azevedo and Lisa Santos was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020.

References

[1] R. A. Adams, *Sobolev Spaces*. Pure Appl. Math. 65, Academic Press, New York, 1975 Zbl 0314.46030 MR 0450957

[2] A. Azevedo, J.-F. Rodrigues, and L. Santos, Nonlocal Lagrange multipliers and transport densities. 2022, arXiv:2208.14274

[3] A. Azevedo and L. Santos, Lagrange multipliers and transport densities. *J. Math. Pures Appl. (9)* **108** (2017), no. 4, 592–611 Zbl 1386.35486 MR 3698170

[4] J. C. Bellido, J. Cueto, and C. Mora-Corral, Γ-convergence of polyconvex functionals involving s-fractional gradients to their local counterparts. *Calc. Var. Partial Differential Equations* **60** (2021), no. 1, Paper No. 7 Zbl 1455.49008 MR 4179861

[5] G. E. Comi and G. Stefani, A distributional approach to fractional Sobolev spaces and fractional variation: existence of blow-up. *J. Funct. Anal.* **277** (2019), no. 10, 3373–3435 Zbl 1437.46039 MR 4001075

[6] G. E. Comi and G. Stefani, A distributional approach to fractional Sobolev spaces and fractional variation: asymptotics I. *Rev. Mat. Complut.* **36** (2023), no. 2, 491–569 Zbl 07683439 MR 4581759

[7] T. Kurokawa, On the Riesz and Bessel kernels as approximations of the identity. *Sci. Rep. Kagoshima Univ.* (1981), no. 30, 31–45 Zbl 0531.40007 MR 643223

[8] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod, Paris, 1969 Zbl 0189.40603 MR 0259693

[9] J.-F. Rodrigues, *Obstacle Problems in Mathematical Physics*. North-Holland Math. Stud. 134, North-Holland, Amsterdam, 1987 Zbl 0606.73017 MR 880369

[10] J.-F. Rodrigues and L. Santos, On nonlocal variational and quasi-variational inequalities with fractional gradient. *Appl. Math. Optim.* **80** (2019), no. 3, 835–852; correction in *Appl. Math. Optim.* **84** (2021), 3565–3567 Zbl 1429.49011 MR 4026601

[11] J.-F. Rodrigues and L. Santos, Variational and quasi-variational inequalities with gradient type constraints. In *Topics in Applied Analysis and Optimisation—Partial Differential Equations, Stochastic and Numerical Analysis*, pp. 319–361, CIM Ser. Math. Sci., Springer, Cham, 2019 Zbl 1442.49012 MR 4410580

[12] T.-T. Shieh and D. E. Spector, On a new class of fractional partial differential equations. *Adv. Calc. Var.* **8** (2015), no. 4, 321–336 Zbl 1330.35510 MR 3403430

[13] T.-T. Shieh and D. E. Spector, On a new class of fractional partial differential equations II. *Adv. Calc. Var.* **11** (2018), no. 3, 289–307 Zbl 1451.35257 MR 3819528

[14] M. Šilhavý, Fractional vector analysis based on invariance requirements (critique of coordinate approaches). *Contin. Mech. Thermodyn.* **32** (2020), no. 1, 207–228 Zbl 1443.26004 MR 4048032

[15] K. Yosida, *Functional Analysis*. 6th edn., Grundlehren Math. Wiss. 123, Springer, Berlin, 1980 Zbl 0435.46002 MR 617913

Assis Azevedo
CMAT and Departamento de Matemática, Escola de Ciências, Universidade do Minho,
Campus de Gualtar, 4710-057 Braga, Portugal; assis@math.uminho.pt

José-Francisco Rodrigues
CMAFcIO and Departamento de Matemática, Faculdade de Ciências, Universidade de Lisboa,
1749-016 Lisboa, Portugal; jfrodrigues@ciencias.ulisboa.pt

Lisa Santos
CMAT and Departamento de Matemática, Escola de Ciências, Universidade do Minho,
Campus de Gualtar, 4710-057 Braga, Portugal; lisa@math.uminho.pt

DOI 10.4171/8ECM/09

The topology of dissipative systems

Héctor Barge and José M. R. Sanjurjo

Abstract. This expository article is dedicated to the study of some topological features of dissipative flows defined in locally compact metric spaces, especially in manifolds and in the Euclidean space. We show that they exhibit a host of interesting topological properties in areas as diverse as Conley's index theory, population dynamics, and the dynamics of planar systems.

1. Introduction

In the study of flows in non-compact spaces, the dissipative ones play an important role. Their interest lies in the fact that it is possible to reduce the fundamental part of the flow and its asymptotic behavior to a compact set. The concept of dissipativity was introduced by Levinson in 1944 [28] for flows in the Euclidean space in his study of the periodically forced van der Pol equation.

This expository article is dedicated to the study of some topological features of dissipative flows defined in locally compact metric spaces, especially in manifolds and in the Euclidean space. We show that they exhibit a host of interesting topological properties in areas as diverse as Conley's index theory, population dynamics, and the dynamics of planar systems.

Through the paper we shall consider continuous dynamical systems (or flows) $\varphi : M \times \mathbb{R} \to M$, where M is a locally compact metric space. If M is not compact, then every flow can be extended to the Alexandrov compactification $M \cup \{\infty\}$ of M by leaving fixed the point ∞.

The main reference for the elementary concepts of dynamical systems is [10] but we also recommend [34, 35, 37]. We use the notation $\gamma(x)$ for the *trajectory* of the point x, i.e.,

$$\gamma(x) = \{xt \mid t \in \mathbb{R}\}.$$

Similarly, for the *positive semi-trajectory* and the *negative semi-trajectory*

$$\gamma^+(x) = \{xt \mid t \in \mathbb{R}_+\}, \quad \gamma^-(x) = \{xt \mid t \in \mathbb{R}_-\}.$$

2020 Mathematics Subject Classification. Primary 34D23; Secondary 37G35.
Keywords. Dissipative flow, global attractor, Borsuk homotopy, Conley index, continuation.

By the *omega-limit* of a point x we understand the set

$$\omega(x) = \bigcap_{t>0} \overline{x[t,\infty)}.$$

In an analogous way, the *negative omega-limit* is the set

$$\omega^*(x) = \bigcap_{t<0} \overline{x(-\infty,t]}.$$

An invariant compactum K is *stable* if every neighborhood U of K contains a neighborhood V of K such that $V[0,\infty) \subset U$.

We recall that an *attractor* is a stable invariant compactum K satisfying that there exists a neighborhood U of K such that $\emptyset \neq \omega(x) \subset K$, for every $x \in U$. A *repeller* is just an attractor for the reverse flow given by $\overline{\varphi}(x,t) = \varphi(x,-t)$.

If K is an attractor, its region (or basin) of attraction of K is the set

$$\mathcal{A}(K) = \{x \in M \mid \emptyset \neq \omega(x) \subset K\}.$$

It is well known that $\mathcal{A}(K)$ is an open invariant set. If in particular $\mathcal{A}(K)$ is the whole phase space, we say that K is a *global attractor*.

We use some topological notions through this paper. We recommend the books of Hatcher and Spanier [24, 44] to cover this material. We use the notation H^* for the singular cohomology. We consider cohomology taking coefficients in $\mathbb{Z}$.

If a pair of spaces (X, A) satisfies that its cohomology $H^k(X, A)$ is finitely generated for each k and is non-zero only for a finite number of values of k (as it happens if (X, A) is a pair of compact manifolds), its *Poincaré polynomial* is defined as

$$P_t(X, A) = \sum_{k\geq 0} \operatorname{rk} H^k(X, A) t^k.$$

There is a form of homotopy theory which has proved to be the most convenient for the study of the global topological properties of the invariant spaces involved in dynamics, namely *Borsuk's homotopy theory* or *shape theory*, introduced and studied by Karol Borsuk. We present here a short introduction based on the presentation given by Kapitanski and Rodnianski in [27].

A metric space X is said to be an *absolute neighborhood retract* or, shortly, an *ANR* if it satisfies that whenever there exists an embedding $f : X \to Y$ of X into a metric space Y such that $f(X)$ is closed in Y, there exists a neighborhood U of $f(X)$ such that $f(X)$ is a retract of U. Some examples of ANRs are manifolds, CW complexes, and polyhedra. Besides, an open subset of an ANR is an ANR and a retract of ANR is also an ANR. For more information about ANRs we recommend [26]. Notice that by Kuratowski–Wojdyslawski theorem, every metric space can be embedded in an ANR as a closed subspace.

Let X be a closed subset of an ANR M and Y a closed subset of an ANR N. Denote by $\mathbb{U}(X; M)$ (resp. $\mathbb{U}(Y; N)$) the set of all open neighborhoods of X in M (resp. Y in N).

Let $\mathbf{f} = \{f : U \to V\}$ be a collection of continuous maps from the neighborhoods $U \in \mathbb{U}(X; M)$ to $V \in \mathbb{U}(Y; N)$. We say that $\mathbf{f}$ is a *mutation* from X to Y if it satisfies

(1) for every $V \in \mathbb{U}(Y; N)$ there exists at least a map $f : U \to V$ in $\mathbf{f}$;

(2) if $f : U \to V$ is in $\mathbf{f}$, then the restriction $f|_{U_1} : U_1 \to V_1$ is also in $\mathbf{f}$ for every neighborhood $U_1 \subset U$ and every neighborhood $V_1 \supset V$;

(3) if two maps $f, f' : U \to V$ are in $\mathbf{f}$, there exists a neighborhood $U_1 \subset U$ such that the restrictions $f|_{U_1}$ and $f'|_{U_1}$ are homotopic.

An example of mutation is the *identity mutation* $\mathrm{id}_{\mathbb{U}(X;M)}$ consisting of the identity maps $\mathrm{id} : U \to U$.

Composition of mutations $\mathbf{f} = \{f : U \to V\}$, $\mathbf{g} = \{g : V \to W\}$ from X to Y and from Y to Z, respectively, is defined in the straightforward way. Two mutations $\mathbf{f} = \{f : U \to V\}$ and $\mathbf{f}' = \{f' : U' \to V'\}$ (both from X to Y) are said to be *homotopic* if for every pair of maps $f : U \to V$ and $f' : U' \to V$ belonging to $\mathbf{f}$ and $\mathbf{f}'$, respectively, there exists a neighborhood $U_0 \in \mathbb{U}(X; M)$, $U_0 \subset U \cap U'$ such that $f|_{U_0}$ is homotopic to $f'|_{U_0}$. It is easy to see that homotopy of mutations is an equivalence relation.

Two metric spaces X and Y have the same *Borsuk homotopy type* or *shape*, denoted by $\mathrm{Sh}(X) = \mathrm{Sh}(Y)$, if they can be embedded as closed sets in ANRs M and N in such a way that there exist mutations $\mathbf{f} = \{f : U \to V\}$ and $\mathbf{g} = \{g : V \to U\}$ such that the compositions $\mathbf{gf}$ and $\mathbf{fg}$ are homotopic to the identity mutations $\mathrm{id}_{\mathbb{U}(X;M)}$ and $\mathrm{id}_{\mathbb{U}(Y;N)}$, respectively. In this case, the mutation $\mathbf{f}$ (resp. $\mathbf{g}$) is said to be a *shape equivalence*.

We stress the following basic features whose proofs can be found in [11].

(1) The notion of shape of sets depends neither on the ANRs they are embedded in nor on the particular embeddings.

(2) Spaces belonging to the same homotopy type have the same shape.

(3) ANRs have the same shape if and only if they have the same homtopy type.

In the case of plane continua, the relation of having the same Borsuk homotopy type has an easy visualization as it establishes the following result.

Theorem 1.1 (Borsuk [11]). *Two continua K and L contained in $\mathbb{R}^2$ have the same Borsuk homotopy type if and only if they disconnect $\mathbb{R}^2$ in the same number of connected components. In particular, a continuum has the Borsuk homotopy type of a point if and only if it does not disconnect $\mathbb{R}^2$. A continuum has the Borsuk homotopy type of a circle if and only if it disconnects $\mathbb{R}^2$ into two connected components.*

Every continuum has the Borsuk homotopy type of a wedge of circles, finite or infinite (Hawaiian earring).

For more information about Borsuk homotopy theory we recommend the books [11, 17, 32]. The papers [3, 9, 19–21, 23, 27, 39, 40, 42] illustrate some applications of this theory to the study of dynamical systems.

An important class of invariant compacta is the so-called *isolated invariant sets* (see [15, 16, 18] for details). These are compact invariant sets K which possess an *isolating neighborhood*, i.e., a compact neighborhood N such that K is the maximal invariant set in N.

To introduce the Conley index, that plays an essential role in this paper, we use a special kind of isolating neighborhoods, the so-called *isolating blocks*. More precisely, an isolating block N is an isolating neighborhood such that there are compact sets $N^i, N^o \subset \partial N$, called the entrance and the exit sets, satisfying

(1) $\partial N = N^i \cup N^o$;

(2) for each $x \in N^i$ there exists $\varepsilon > 0$ such that $x[-\varepsilon, 0) \subset M \setminus N$ and for each $x \in N^o$ there exists $\delta > 0$ such that $x(0, \delta] \subset M \setminus N$;

(3) for each $x \in \partial N \setminus N^i$ there exists $\varepsilon > 0$ such that $x[-\varepsilon, 0) \subset \mathring{N}$ and for every $x \in \partial N \setminus N^o$ there exists $\delta > 0$ such that $x(0, \delta] \subset \mathring{N}$.

These blocks form a neighborhood basis of K in M.

Let K be an isolated invariant set. Its *Conley index* $h(K)$ is defined as the pointed homotopy type of the topological space $(N/N^o, [N^o])$, where N is an isolating block of K. A weak version of the Conley index which will be useful for us is the *cohomological index* defined as $CH^*(K) = H^*(h(K))$. It can be proved that $CH^*(K) \cong H^*(N, N^o)$. Our main references for the Conley index theory are [15, 38]. An exhaustive study of the Conley index in the case of two-dimensional flows can be found in [2, 4] and some applications of this theory to the evolution of the Lorenz strange set are contained in [8]. In addition, the Conley index has recently been used to find counterexamples to the triangulation conjecture (see [30, 31]).

The Conley index allows us to establish some conections between local and global dynamics via Morse decompositions. We recall that if K is a compact invariant set, a finite collection $\{M_1, \ldots, M_n\}$ of pairwise disjoint invariant subcompacta of K is a *Morse decomposition* if it satisfies that

$$\text{for each } x \in \left(K \setminus \bigcup_{i=1}^{n} M_i\right), \quad \omega(x) \subset M_j \text{ and } \omega^*(x) \subset M_k \text{ with } j < k.$$

Each set M_i is said to be a *Morse set*.

Given a Morse decomposition $\{M_1, M_2, \dots, M_k\}$ of an isolated invariant set K, there exists a polynomial $Q(t)$ whose coefficients are non-negative integers such that

$$\sum_{i=1}^{n} P_t\big(h(M_i)\big) = P_t\big(h(K)\big) + (1+t)Q(t).$$

This formula, which relates the Conley indices of the Morse sets with the Conley index of the isolated invariant set, is known as the *Morse equation* of the Morse decomposition and it generalizes the classical Morse inequalities.

Another central concept of the Conley index theory that plays a crucial role in this paper is that of continuation of isolated invariant sets. Let M be a locally compact metric space, and let $\varphi_\lambda : M \times \mathbb{R} \to M$ be a parametrized family of flows (parametrized by $\lambda \in [0,1]$, the unit interval). The family $(K_\lambda)_{\lambda \in J}$, where $J \subset [0,1]$ is a closed (non-degenerate) subinterval and, for each $\lambda \in J$, K_λ is an isolated invariant set for φ_λ, is said to be a *continuation* if for each $\lambda_0 \in J$ and each N_{λ_0} isolating neighborhood for K_{λ_0}, there exists $\delta > 0$ such that N_{λ_0} is an isolating neighborhood for K_λ for every $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta) \cap J$. We say that the family $(K_\lambda)_{\lambda \in J}$ is a continuation of K_{λ_0} for each $\lambda_0 \in J$.

Notice that [38, Lemma 6.1] ensures that if K_{λ_0} is an isolated invariant set for φ_{λ_0}, there always exists a continuation $(K_\lambda)_{\lambda \in J_{\lambda_0}}$ of K_{λ_0} for some closed (non-degenerate) subinterval $\lambda_0 \in J_{\lambda_0} \subset [0,1]$.

There is a simpler definition of continuation based on [38, Lemma 6.2]. There, it is proved that if $\varphi_\lambda : M \times \mathbb{R} \to M$ is a parametrized family of flows and if N_1 and N_2 are isolating neighborhoods of the same isolated invariant set for φ_{λ_0}, then there exists $\delta > 0$ such that N_1 and N_2 are isolating neighborhoods for φ_λ, for every $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta) \cap [0,1]$, with the property that, for every λ, the isolated invariant subsets in N_1 and N_2 which have N_1 and N_2 as isolating neighborhoods coincide.

Therefore, the family $(K_\lambda)_{\lambda \in J}$, with K_λ an isolated invariant set for φ_λ, is a continuation if for every $\lambda_0 \in J$ there are an isolating neighborhood N_{λ_0} for K_{λ_0} and a $\delta > 0$ such that N_{λ_0} is an isolating neighborhood for K_λ, for every $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta) \cap J$.

Notice that, since this should not lead to any confusion, sometimes we will only say that K_λ is a continuation of K_{λ_0} without specifying the subinterval $J \subset [0,1]$ to which the parameters belong.

In the particular case that K_{λ_0} is an attractor for $\lambda_0 \in J$, there exists $\delta > 0$ such that K_λ is attractor with $\mathrm{Sh}(K_\lambda) = \mathrm{Sh}(K_{\lambda_0})$ for $\lambda \in (\lambda_0 - \delta, \lambda_0 + \delta) \cap J$ (see [41, Theorem 4]).

The paper is structured as follows. In Section 2 the concept of dissipative flow is introduced and some of the basic properties of this class of flows are presented. In particular, we see that dissipative flows coincide with those that have a global attractor.

We also present some characterizations of the global attractor of a dissipative flow in the Euclidean space. Section 3 is devoted to study parametrized families of dissipative flows. We see that the property of being a global attractor is not robust and introduce a characterization of continuations that consist entirely of global attractors. We also survey some results regarding the bifurcation global to non-global. In Section 4 we study connections between dissipative flows and populations dynamics and present some results about uniform persistence, a central concept in population dynamics. Finally, in Section 5, we present some results that ensure that the global attractor of a dissipative flow defined on the non-negative orthant of the plane is contained in the boundary.

2. Dissipative flows

We start by recalling the definition of dissipative flow and some of its basic properties. We assume that M is a locally compact, non-compact metric space.

Definition 2.1 (Levinson 1944). A flow $\varphi : M \times \mathbb{R} \to M$ is said to be *dissipative* provided that, for each $x \in M$, the omega limit $\omega(x) \neq \emptyset$ and the closure of the set

$$\Omega(\varphi) = \bigcup_{x \in M} \omega(x)$$

is compact.

The following characterization of dissipative flows, which gives a very clear interpretation of their dynamics, was provided by Pliss.

Proposition 2.2 (Pliss 1966 [36]). *A flow $\varphi : M \times \mathbb{R} \to M$ is dissipative if and only it has a global attractor.*

It should be noted that, in general, the global attractor does not necessarily coincide with the closure of $\Omega(\varphi)$. On the other hand, it can be seen that the flow φ is dissipative if and only if $\{\infty\}$ is a repeller.

The following result gives a characterization of the global attractor of a dissipative flow. It relies heavily on the non-existence of bounded orbits outside the attractor.

Proposition 2.3. *Let φ be a dissipative flow in $\mathbb{R}^n$ and K a compact invariant set. Then K is the global attractor if and only if $\mathbb{R}^n \setminus K$ does not contain bounded orbits.*

In the case of flows on the two-dimensional Euclidean space it is possible to obtain a simpler characterization of global attractors of dissipative flows.

Theorem 2.4 (Barge–Sanjurjo [5]). *Let K be an isolated invariant continuum of a dissipative flow φ in $\mathbb{R}^2$. The following conditions are equivalent:*

(i) *K is a global attractor;*

(ii) *there are no fixed points in* $\mathbb{R}^2 \setminus K$ *and there exists an orbit* γ *connecting* ∞ *and* K *(i.e., such that* $\|\gamma(t)\| \to \infty$ *when* $t \to -\infty$ *and* $\omega(\gamma) \subset K$*).*

This result is inspired by the following result that gives a relation between global asymptotic stability of a fixed point and the non-existence of additional fixed points in the case of dissipative discrete dynamical systems.

Theorem 2.5 (Alarcón–Guíñez–Gutiérrez [1], Ortega–Ruiz del Portal [33]). *Assume that* $h \in \mathcal{H}_+$ *(orientation preserving homeomorphisms of* $\mathbb{R}^2$*) is dissipative and* p *is an asymptotically stable fixed point of* h*. The following conditions are equivalent:*

(i) p *is globally asymptotically stable;*

(ii) $\mathrm{fix}(h) = p$ *and there exists an arc* $\gamma \subset S^2$ *with end points at* p *and* ∞ *such that* $h(\gamma) = \gamma$*.*

The proof in [1] is based on Brouwer's theory of fixed point free homeomorphisms of the plane. Ortega and Ruiz del Portal give in [33] an alternative proof based on the theory of prime ends.

The previous results suggest that if K is an attractor of a dissipative flow, then $\mathcal{A}(K)$ being bounded is in the sharpest contrast to K being global.

3. Robustness of global attractors

This section is dedicated to the presentation of some results related to properties of dissipative systems that concern Conley's index theory.

We give a simple example which shows that the property of being global is not a robust property for an attractor since small perturbations of the flow can create bounded orbits in its region of attraction.

Example 3.1. Consider the family of ordinary differential equations defined on the plane in polar coordinates:

$$\begin{cases} r' = -r^3\left(\frac{1}{r} - \lambda\right)^2, \\ \theta' = 1, \end{cases} \qquad \lambda \in [0, 1].$$

The phase portraits of this family of differential equations are depicted in Figure 1. The picture on the left describes the phase portrait for the parameter $\lambda = 0$. We see that in this case the origin is a globally attracting fixed point and the orbit of any other point spirals towards it. The picture on the right describes the phase portrait when $\lambda > 0$. In this case we see that the origin is still an asymptotically stable fixed point but it is not a global attractor anymore since, for each $\lambda > 0$, the circle centered at the origin and radius $1/\lambda$ is a periodic trajectory which attracts uniformly all the points of the unbounded component of its complement and repels all the points of the bounded one except the origin.

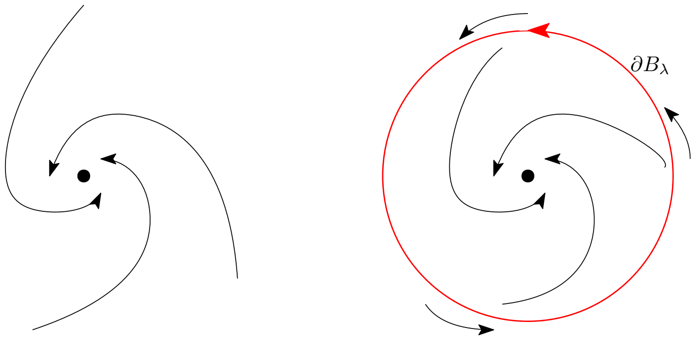


Figure 1. Phase portraits of the family of ordinary differential equations from Example 3.1 for $\lambda = 0$ (left) and $\lambda > 0$ (right).

Example 3.1 motivates the following definition.

Definition 3.2. A parametrized family of dissipative flows $\varphi_\lambda : M \times \mathbb{R} \to M$ is said to be *coercive* if for any continuation K_λ of the global attractor K_0 of φ_0 there exists a λ_0 such that $\mathcal{A}_\lambda(K_\lambda)$ is bounded for every λ with $0 < \lambda < \lambda_0$.

However, note that in this situation, since all the flows are dissipative, then each φ_λ still has a global attractor $\widehat{K}_\lambda$ but the family of global attractors is not a continuation of K_0.

The following definition introduces a notion which is, in some sense, the opposite of the previous one.

Definition 3.3. A parametrized family of dissipative flows $\varphi_\lambda : M \times \mathbb{R} \to M$ is said to be *uniformly dissipative* provided that for each $x \in M$ and $\lambda \in [0, 1]$ we have that $\omega_\lambda(x) \neq \emptyset$ and the closure of the set

$$\Omega = \bigcup_{\lambda \in [0,1]} \Omega(\varphi_\lambda)$$

is compact.

The importance of the above definition is that it can be used to provide a characterization of continuations that consist entirely of global attractors.

Theorem 3.4 (Barge–Sanjurjo [7]). *Let $\varphi_\lambda : M \times \mathbb{R} \to M$ be a parametrized family of dissipative flows with $\lambda \in [0, 1]$. Let K_λ denote the global attractor of φ_λ. Then the family $(K_\lambda)_{\lambda \in [0,1]}$ is a continuation of K_0 if and only if the family $(\varphi_\lambda)_{\lambda \in [0,1]}$ is uniformly dissipative.*

We see a nice application of the previous result.

Example 3.5. An important example of global attractor is provided by the Lorenz equations

$$\begin{cases} x' = \sigma(y - x), \\ y' = rx - y - xz, \\ z' = xy - bz, \end{cases}$$

where σ, r, and b are three real positive parameters. If we fix σ and b, we obtain a family of flows

$$\varphi_r : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}^3$$

corresponding to the Lorenz equations for the different values of r.

E. N. Lorenz proved that for every value of r there exists a global attractor of zero volume for the flow associated to these equations. This attractor should not be confused with the famous Lorenz attractor, which is a proper subset of the global attractor.

The family φ_r is uniformly dissipative and, as a consequence, it defines a continuation of global attractors K_r. The proof of this fact uses the function

$$V = rx^2 + \sigma y^2 + \sigma(z - 2r)^2.$$

C. Sparrow studied in [45] this function and showed that it is a Lyapunov function for the flow φ_r. By using this function he was able to prove that K_r lies in a ball B_r centered at 0 and with radius $O(r)$, such that $O(r)$ depends continuously on r. Hence, if we consider an arbitrary r_0 and an interval $[c, d]$ containing r_0, we have that the set $C = \overline{\bigcup_{c \leq r \leq d} B_r}$ is compact and that $\emptyset \neq \omega_r(x) \subset C$ for every $x \in \mathbb{R}^3$ and every $r \in [c, d]$. Therefore, the family of Lorenz flows φ_r is uniformly dissipative and the corresponding family K_r of global attractors is a continuation.

The coercive families of flows are in sharp contrast with the uniformly dissipative families. For coercive families, the continuations of global attractors are never global. The study of coercive families of flows has some topological interest. The following result provides a graphic characterization of this kind of families.

Theorem 3.6 (Barge–Sanjurjo [7]). *Let φ_λ, with $\lambda \in [0, 1]$, be a coercive family of flows in $\mathbb{R}^n$. We denote by K_0 the global attractor of φ_0 and by K_λ, with $\lambda \in [0, 1]$, a continuation of K_0. Then there exists $\lambda_0 > 0$ such that for every λ with $0 < \lambda < \lambda_0$ there is an isolated invariant compactum C_λ in $\mathbb{R}^n \setminus K_\lambda$ such that*

(i) *C_λ separates $\mathbb{R}^n$ into two components and K_λ lies in the bounded component;*

(ii) *C_λ has the Borsuk homotopy type (shape) of S^{n-1};*

(iii) *C_λ attracts uniformly all the points of the unbounded component and repels all the points of the bounded one which are not in K_λ;*

(iv) $\operatorname{diam} C_\lambda \to \infty$ *when* $\lambda \to 0$, *where* $\operatorname{diam} C_\lambda$ *denotes the diameter of* C_λ.

Moreover, the existence of such a C_λ *for* $0 < \lambda < \lambda_0$ *is sufficient for the family to be coercive.*

In view of the previous results, it is interesting to study in all its generality the mechanism which produces the global to non-global bifurcation in families of dissipative flows. With this objective we introduce the following definition.

Definition 3.7. Let $\varphi_\lambda : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, with $\lambda \in [0, 1]$, be a parametrized family of dissipative flows. The family is said to be *polar* if it has arbitrarily large bounded trajectories. More precisely, for every $L > 0$ (arbitrarily large) there is a $\lambda_0 > 0$ such that for every λ with $0 < \lambda < \lambda_0$ there is a bounded trajectory γ_λ of φ_λ and a $t_\lambda < 0$ such that $\|\gamma_\lambda(t)\| > L$ for every t with $-\infty < t < t_\lambda$.

Obviously, if K_λ is a continuation of the global attractor K_0 of φ_0, then for L sufficiently large, γ_λ lies in $\mathbb{R}^n \setminus K_\lambda$.

The following proposition makes it clear that polarity is a key notion regarding the transition from global to non-global.

Proposition 3.8 (Barge–Sanjurjo [7]). *Let* $\varphi_\lambda : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, *with* $\lambda \in [0, 1]$, *be a parametrized family of dissipative flows. Then the family is polar if and only if for every continuation* K_λ *of the global attractor* K_0 *of* φ_0 *there exists a* $\lambda_0 > 0$ *such that* K_λ *is a non-global attractor for every* λ *with* $0 < \lambda < \lambda_0$.

The following result describes the general picture of the polar families of dissipative flows.

Theorem 3.9 (Barge–Sanjurjo [7]). *If* $\varphi_\lambda : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, *with* $\lambda \in [0, 1]$, *is a polar family of dissipative flows, then there exists a* $\lambda_0 > 0$ *such that for every* λ *with* $0 < \lambda < \lambda_0$ *the maximal invariant compactum lying in* $\mathbb{R}^n \setminus K_\lambda$ *for the flow* φ_λ, *which we denote by* C_λ, *is non-empty and isolated, and its cohomological Conley index is trivial in every dimension. Moreover, the family is coercive if and only if* C_λ *has the Borsuk homotopy type of* S^{n-1}.

The isolated invariant compactum C_λ can be seen as the obstruction for the existence of a continuation of global attractors. An interesting feature of the above proposition is that it provides an equivalence between a topological property (having the Borsuk homotopy type of S^{n-1}) and a dynamical property (coercivity).

4. Dissipative flows and populations dynamics

Another area in which dissipative systems play a fundamental role is population dynamics. We shall suppose here that M is a closed subset of a larger locally compact metric space X and denote by ∂M the boundary of M in X.

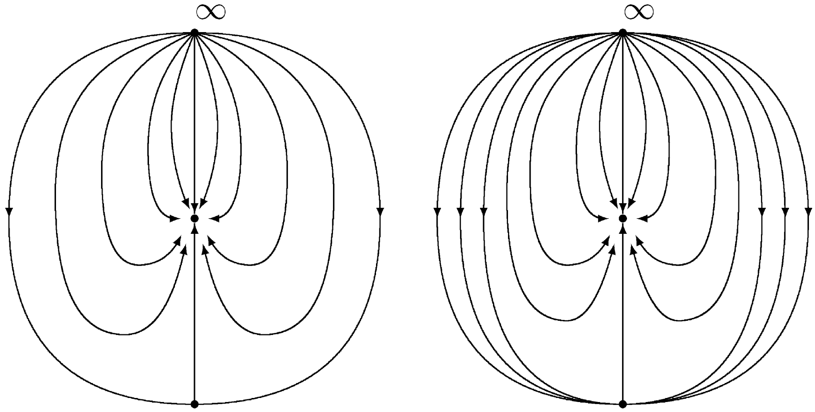


Figure 2. A small perturbation of a uniformly persistent flow that is not uniformly persistent.

Definition 4.1. We will say that the dissipative flow $\varphi : M \times \mathbb{R} \to M$ is *uniformly persistent* if there exists $\beta > 0$ such that for every $x \in \overset{\circ}{M}$

$$\liminf\left\{d\left(\varphi(x,t), \partial M\right) \mid t \to \infty\right\} \geq \beta.$$

If M is compact, then φ is persistent if and only if ∂M is a repeller of φ. If M is not compact, then φ is persistent if and only if $\partial M \cup \{\infty\}$ is a repeller for the flow extended to $M \cup \{\infty\}$. As a consequence, there exists a compactum K which attracts all points $x \notin \partial M$. This compactum is called the *global attractor* of the system and represents a state of coexistence of all the species that make up the population. The most significant case is when $X = \mathbb{R}^n$, $M = \mathbb{R}^n_+$ (the non-negative orthant). The boundary $\partial\mathbb{R}^n_+$ represents populations some of whose components have become extinct.

It is easy to see that, in a general context, uniform persistence is not a robust property. For instance, the illustration in Figure 2 shows that small perturbations of a uniformly persistent flow can destroy this property.

Despite this fact, we see in our next result that all uniformly persistent flows have *weak continuation properties*, meaning by this that small perturbations of the flow never drive to extinction populations within a certain range (which can be arbitrarily chosen).

Theorem 4.2 (Weak continuation of uniform persistence, Sanjurjo [43]). *Let X be a locally compact metric space and let M be a closed subset of X. Suppose we are given a (continuous) parametrized family of dissipative dynamical systems φ_λ, with $\lambda \in I$, on M, for which ∂M is invariant. Further, assume that φ_0 is uniformly persistent. Then there exists $\beta > 0$ such that for every compact set $K \subset \overset{\circ}{M}$ there exists $\lambda_0 > 0$*

such that

$$\liminf\left\{d\left(\varphi_\lambda(x,t),\partial M\right)\mid t\to\infty\right\}\geq\beta$$

for every $\lambda\leq\lambda_0$ and for every $x\in K$.

When M is the non-negative orthant, some nice topological conclusions can be reached about special regions of the flow. In particular, there is a contractible region where populations are guaranteed their survival and another region of spherical shape where populations have their survival compromised.

Corollary 4.3 (Sanjurjo [43]). *Let φ_λ, with $\lambda\in I$, be a (continuous) parametrized family of dissipative flows on the non-negative orthant $\mathbb{R}^n_+$. Further, assume that φ_0 is uniformly persistent. Then there exists $\alpha>0$ such that for every ε and every L with $0<\varepsilon<L$ there exists $\lambda_0>0$ such that*

(i) $\liminf\{d(\varphi_\lambda(x,t),\partial\mathbb{R}^n_+)|t\to\infty\}>\alpha$ *for every x with $d(x,\partial\mathbb{R}^n_+)\geq\varepsilon$ and $\|x\|\leq L$ and for every $\lambda\leq\lambda_0$,*

(ii) *the set*

$$W_\lambda=\left\{x\in\mathbb{R}^n_+\mid\liminf\left\{d\left(\varphi_\lambda(x,t),\partial\mathbb{R}^n_+\right)\mid t\to\infty\right\}>\alpha\right\}$$

is contractible and the set

$$R_\lambda=\left\{x\in\mathbb{R}^n_+\mid\liminf\left\{d\left(\varphi_\lambda(x,t),\partial\mathbb{R}^n_+\right)\mid t\to\infty\right\}\leq\alpha\right\}\cup\{\infty\}$$

has the Borsuk homotopy type (shape) of S^{n-1} for every $\lambda\leq\lambda_0$.

It would be of interest to study the implications of these results in some particular situations. Theorem 4.2 suggests that permanence does not vanish completely in an abrupt way. Even if it does not continue, permanence still remains when we limit ourselves to populations within a certain range. As an interesting case, S. Cano-Casanova and J. López-Gómez prove in [14] (see also [29]) that permanence of two species is possible under strong mutual aggression. In other words, they prove that if the birth rates are high enough, then the species are permanent irrespective of the competition strength in the regions where competition occurs. They actually measure how large the birth rate must be.

An interesting problem would be to study to what extent permanence remains for populations within a certain range despite their reproduction rate being below the limit threshold.

As we said before, uniformly persistent flows have a global attractor towards which all the states of the interior evolve. The following results concern the fine structure of this global attractor of the flow and some of its topological properties. We recall that a continuum K is *point-like* in $\mathbb{R}^n$ provided $\mathbb{R}^n\setminus K$ is homeomorphic to $\mathbb{R}^n\setminus\{p\}$, where p is a point.

Theorem 4.4 (Sanjurjo [43]). *Let $\varphi : \mathbb{R}^n_+ \times \mathbb{R} \to \mathbb{R}^n_+$ be a dissipative flow. If φ is uniformly persistent then:*

(i) *Suppose L is a point-like repeller (in particular a repelling point) in the interior of $\mathbb{R}^n_+$, then there exists an attractor K_0 with the Borsuk homotopy type (shape) of S^{n-1} contained in the global attractor K and whose basin of attraction is int $\mathbb{R}^n_+ \setminus L$.*

(ii) *Suppose L is a repeller with the Borsuk homotopy type (shape) of S^{n-1} in the interior of $\mathbb{R}^n_+$. Then L decomposes $\mathring{\mathbb{R}}^n_+$ into two connected components. Moreover, if the bounded component is simply connected, then there exists an attractor with the Borsuk homotopy type (shape) of a point contained (together with its basin of attraction) in the interior of the global attractor K.*

In our next result we see that the Morse theory of uniformly persistent flows with an attracting cycle can be described in a simple way, irrespective of the complexity of the flow in the boundary. Suppose $\varphi : \mathbb{R}^n_+ \times \mathbb{R} \to \mathbb{R}^n_+$ is a uniformly persistent flow. We say that $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$ is a natural Morse decomposition of the flow if

(a) $\{M_1, M_2\}$ is an attractor-repeller decomposition of the global attractor K,

(b) $M_i \subset \partial\mathbb{R}^n_+$ for $i \geq 3$, and

(c) $\{M_1, M_2, \ldots, M_k, \infty\}$ is a Morse decomposition of $\mathbb{R}^n_+ \cup \{\infty\}$.

By the Morse equation of $\mathcal{M}$ we mean the Morse equation of $\{M_1, M_2, \ldots, M_k, \infty\}$. The next theorem shows that if M_1 is an attracting cycle or, more generally, an attractor with the Borsuk homotopy type (shape) of S^1, then the Morse equation of $\mathcal{M}$ takes a simple form. On the opposite direction we see that using this equation we can recognize the existence of attractors with the Borsuk homotopy type (shape) of S^1 in the plane or attractors whose suspension has the Borsuk homotopy type (shape) of S^2 for higher dimensions.

Theorem 4.5 (Sanjurjo [43]). *Let $\varphi : \mathbb{R}^n_+ \times \mathbb{R} \to \mathbb{R}^n_+$ be a dissipative flow. Suppose φ is uniformly persistent and $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$ is a natural Morse decomposition of $\mathbb{R}^n_+$ for φ. Then:*

(i) *If M_1 has the Borsuk homotopy type of S^1, then the Morse equation of the decomposition $\mathcal{M}$ with coefficients in $\mathbb{Z}$ or a field is*

$$1 + t + t^2 = 1 + (1 + t)t. \tag{4.1}$$

(ii) *Conversely, if the Morse equation of $\mathcal{M}$ is (4.1), then* $\mathrm{Sh}(M_1) = \mathrm{Sh}(S^1)$ *for $n = 2$ and* $\mathrm{Sh}(\Sigma M_1) = \mathrm{Sh}(S^2)$ *for $n \geq 2$, where ΣM_1 is the suspension of M_1.*

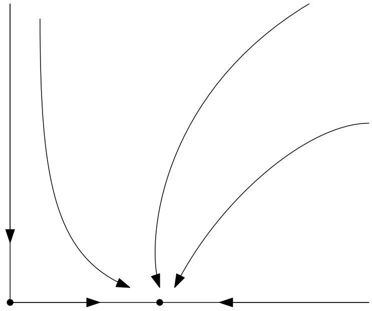


Figure 3. Phase portrait of the Lotka–Volterra system for $a/\lambda \leq c/d$.

5. Planar dissipative flows

In this section, we present some results regarding dissipative flows defined on the non-negative orthant of the plane.

Example 5.1. Consider the Lotka–Volterra equation in $\mathbb{R}^2_+$ (see [25] for more information):

$$\begin{cases} \dot{x} = x(a - by - \lambda x), \\ \dot{y} = y(-c + dx - \mu y), \end{cases} \qquad a, b, c, d, \lambda > 0 \text{ and } \mu \geq 0.$$

This equation, which plays a central role in population dynamics, induces a family of dissipative flows depending on the parameters. The point $(a/\lambda, 0) \in \partial\mathbb{R}^2_+$ is a fixed point (regardless of the parameter value) which is a sink for $a/\lambda \leq c/d$ (Figure 3). In this case, there are no fixed points in $\mathring{\mathbb{R}}^2_+$ and the global attractor of the flow is the closed interval $[0, a/\lambda] \times \{0\}$ contained in $\partial\mathbb{R}^2_+$. As a consequence, the extinction of one of the populations takes place. This situation is, in a certain sense, the opposite of that described for uniformly persistent flows.

Motivated by the situation just described, we present some results that ensure that the global attractor of a dissipative flow defined on $\mathbb{R}^2_+$ is contained in $\partial\mathbb{R}^2_+$.

Theorem 5.2 (Barge–Sanjurjo [6]). *Suppose that $\varphi : \mathbb{R}^2_+ \times \mathbb{R} \to \mathbb{R}^2_+$ is a flow without equilibria in $\mathring{\mathbb{R}}^2_+$. Then, the ω-limit (resp. the ω^*-limit) of any point, when non-empty, is entirely composed of fixed points and, hence, it is contained in $\partial\mathbb{R}^2_+$. If, in addition, the fixed point set is bounded and totally disconnected, then the ω-limit (resp. the ω^*-limit) of each trajectory, when non-empty, is a singleton. Moreover, if the flow is dissipative, the following hold.*

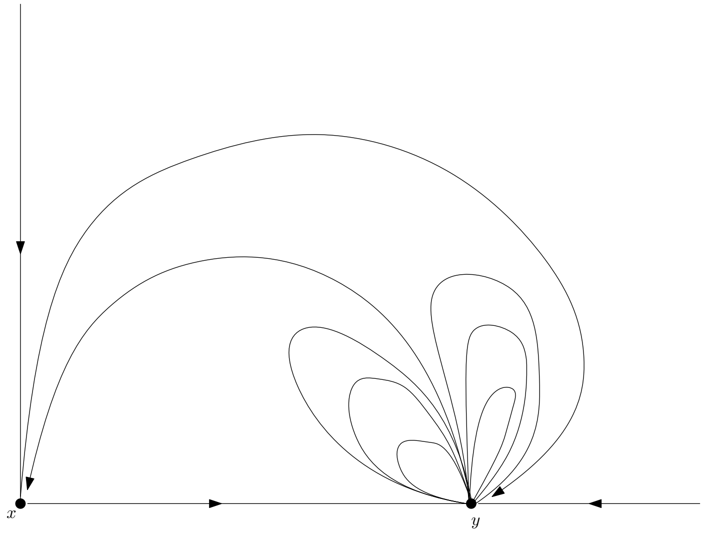


Figure 4. Dissipative flow in $\mathbb{R}^2_+$ with only two fixed points x and y, both contained in $\partial\mathbb{R}^2_+$ and such that $I(x) = 4$ and $I(y) = +\infty$.

(i) *Given $x \in \mathbb{R}^2_+$, if $\gamma^-(x)$ is bounded so is $\gamma(x)$. Hence, both $\omega(x)$ and $\omega^*(x)$ are non-empty and entirely composed of fixed points.*

(ii) *Let K be the global attractor of $\varphi|_{\partial\mathbb{R}^2_+}$. Then, K is the global attractor of φ if and only if K is isolated for φ or, equivalently, $\gamma^-(x)$ is unbounded for each $x \in \mathring{\mathbb{R}}^2_+$.*

Remark. The fact that $\omega(x)$ is composed of fixed points for discrete systems of the disc having all the fixed points in the boundary was proved by Campos, Ortega, and Tineo in [13] by using some ideas of Brown [12] and a classical result of Brouwer (see [22]) on homeomorphisms of the plane. The proof of the previous result, that can be seen in [6], makes use of the Poincaré–Bendixson theorem.

Let $\Gamma_B(\varphi)$ be the set of bounded trajectories of φ and let x be an equilibrium point. We define

$$\Gamma(x) := \big\{\gamma \in \Gamma_B \mid x \in \omega(\gamma) \cup \omega^*(\gamma)\big\}.$$

Definition 5.3. Let x be an equilibrium point. We define the index $I(x) \in \mathbb{N} \cup \{+\infty\}$ to be $k \in \mathbb{N}$ if the cardinal of $\Gamma(x)$ is k and $I(x) = +\infty$ if the cardinal of $\Gamma(x)$ is not finite.

Remark. For each $k \in \mathbb{N} \cup \{+\infty\}$ there exists a flow on $\mathbb{R}^2_+$ with all its equilibria contained in $\partial\mathbb{R}^2_+$ and having a fixed point x such that $I(x) = k$. In Figure 4, a flow having a fixed point of index 4 is depicted.

Theorem 5.4 (Barge–Sanjurjo [6]). *Suppose that $\varphi : \mathbb{R}^2_+ \times \mathbb{R} \to \mathbb{R}^2_+$ is a dissipative flow having a countable amount of fixed points, all of them contained in $\partial\mathbb{R}^2_+$. Then, the global attractor of the flow is in the boundary if and only if all the fixed points have finite index. In such a case, for each fixed point x, $I(x)$ is either 1, 2 or 3. Moreover, if the index of an isolated fixed point x takes the value 1, then $\{x\}$ is the global attractor.*

Acknowledgments. The authors thank the referee for their careful reading of the manuscript and helpful remarks.

Funding. Research supported by the Spanish Ministerio de Ciencia e Innovación grants PGC2018-098321-B-I00 and PID2021-126124NB-I00.

References

[1] B. Alarcón, V. Guíñez, and C. Gutierrez, Planar embeddings with a globally attracting fixed point. *Nonlinear Anal.* **69** (2008), no. 1, 140–150 Zbl 1178.37028 MR 2417859

[2] H. Barge, Regular blocks and Conley index of isolated invariant continua in surfaces. *Nonlinear Anal.* **146** (2016), 100–119 Zbl 1376.37041 MR 3556331

[3] H. Barge, A. Giraldo, and J. M. R. Sanjurjo, Bifurcations, robustness and shape of attractors of discrete dynamical systems. *J. Fixed Point Theory Appl.* **22** (2020), no. 2, Paper No. 29 Zbl 1434.37015 MR 4077285

[4] H. Barge and J. M. R. Sanjurjo, Unstable manifold, Conley index and fixed points of flows. *J. Math. Anal. Appl.* **420** (2014), no. 1, 835–851 Zbl 1304.37017 MR 3229857

[5] H. Barge and J. M. R. Sanjurjo, Fixed points, bounded orbits and attractors of planar flows. In *A Mathematical Tribute to Professor José María Montesinos Amilibia*, pp. 125–132, Dep. Geom. Topol. Fac. Cien. Mat. UCM, Madrid, 2016 Zbl 1371.37043 MR 3525929

[6] H. Barge and J. M. R. Sanjurjo, Flows in $\mathbb{R}^2_+$ without interior fixed points, global attractors and bifurcations. *Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM* **112** (2018), no. 3, 671–683 Zbl 1395.37014 MR 3819723

[7] H. Barge and J. M. R. Sanjurjo, Dissipative flows, global attractors and shape theory. *Topology Appl.* **258** (2019), 392–401 Zbl 1414.37013 MR 3926436

[8] H. Barge and J. M. R. Sanjurjo, A Conley index study of the evolution of the Lorenz strange set. *Phys. D* **401** (2020), 132162, 11 Zbl 1453.37018 MR 4034683

[9] H. Barge and J. M. R. Sanjurjo, Higher dimensional topology and generalized Hopf bifurcations for discrete dynamical systems. *Discrete Contin. Dyn. Syst.* **42** (2022), no. 6, 2585–2601 Zbl 07528588 MR 4421505

[10] N. P. Bhatia and G. P. Szegő, *Stability Theory of Dynamical Systems*. Classics Math., Springer, Berlin, 2002 Zbl 0993.37001 MR 1887295

[11] K. Borsuk, *Theory of Shape*. Monografie Matematyczne 59, PWN—Polish Scientific Publishers, Warsaw, 1975 Zbl 0317.55006 MR 0418088

[12] M. Brown, Homeomorphisms of two-dimensional manifolds. *Houston J. Math.* **11** (1985), no. 4, 455–469 Zbl 0605.57005 MR 837985

[13] J. Campos, R. Ortega, and A. Tineo, Homeomorphisms of the disk with trivial dynamics and extinction of competitive systems. *J. Differential Equations* **138** (1997), no. 1, 157–170 Zbl 0886.34025 MR 1458459

[14] S. Cano-Casanova and J. López-Gómez, Permanence under strong aggressions is possible. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **20** (2003), no. 6, 999–1041 Zbl 1086.35054 MR 2008687

[15] C. Conley, *Isolated Invariant Sets and the Morse Index*. CBMS Reg. Conf. Ser. Math. 38, American Mathematical Society, Providence, RI, 1978 Zbl 0397.34056 MR 511133

[16] C. Conley and R. Easton, Isolated invariant sets and isolating blocks. *Trans. Amer. Math. Soc.* **158** (1971), 35–61 Zbl 0223.58011 MR 279830

[17] J. Dydak and J. Segal, *Shape Theory. An Introduction*. Lecture Notes in Math. 688, Springer, Berlin, 1978 Zbl 0401.54028 MR 520227

[18] R. W. Easton, Isolating blocks and symbolic dynamics. *J. Differential Equations* **17** (1975), 96–118 Zbl 0293.58011 MR 370663

[19] A. Giraldo, V. F. Laguna, and J. M. R. Sanjurjo, Uniform persistence and Hopf bifurcations in $\mathbb{R}^n_+$. *J. Differential Equations* **256** (2014), no. 8, 2949–2964 Zbl 1332.37015 MR 3199752

[20] A. Giraldo, M. A. Morón, F. R. Ruiz del Portal, and J. M. R. Sanjurjo, Some duality properties of non-saddle sets. *Topology Appl.* **113** (2001), no. 1-3, 51–59 Zbl 1003.37009 MR 1821846

[21] A. Giraldo, M. A. Morón, F. R. Ruiz Del Portal, and J. M. R. Sanjurjo, Shape of global attractors in topological spaces. *Nonlinear Anal.* **60** (2005), no. 5, 837–847 Zbl 1060.37010 MR 2113160

[22] L. Guillou, Théorème de translation plane de Brouwer et généralisations du théorème de Poincaré–Birkhoff. *Topology* **33** (1994), no. 2, 331–351 Zbl 0924.55001 MR 1273787

[23] B. Günther and J. Segal, Every attractor of a flow on a manifold has the shape of a finite polyhedron. *Proc. Amer. Math. Soc.* **119** (1993), no. 1, 321–329 Zbl 0822.54014 MR 1170545

[24] A. Hatcher, *Algebraic Topology*. Cambridge University Press, Cambridge, 2002 Zbl 1044.55001 MR 1867354

[25] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998 Zbl 0914.90287 MR 1635735

[26] S.-t. Hu, *Theory of Retracts*. Wayne State University Press, Detroit, 1965 Zbl 0145.43003 MR 0181977

[27] L. Kapitanski and I. Rodnianski, Shape and Morse theory of attractors. *Comm. Pure Appl. Math.* **53** (2000), no. 2, 218–242 Zbl 1026.37007 MR 1721374

[28] N. Levinson, Transformation theory of non-linear differential equations of the second order. *Ann. of Math. (2)* **45** (1944), 723–737 Zbl 0061.18910 MR 11505

[29] J. López-Gómez, Strong competition with refuges. *Nonlinear Anal.* **30** (1997), no. 8, 5167–5178 Zbl 0910.92030 MR 1726019

[30] C. Manolescu, The Conley index, gauge theory, and triangulations. *J. Fixed Point Theory Appl.* **13** (2013), no. 2, 431–457 Zbl 1282.57001 MR 3122335

[31] C. Manolescu, Pin(2)-equivariant Seiberg–Witten Floer homology and the triangulation conjecture. *J. Amer. Math. Soc.* **29** (2016), no. 1, 147–176 Zbl 1343.57015 MR 3402697

[32] S. Mardešić and J. Segal, *Shape Theory. The Inverse System Approach*. North-Holland Math. Libr. 26, North-Holland, Amsterdam, 1982 Zbl 0495.55001 MR 676973

[33] R. Ortega and F. R. Ruiz del Portal, Attractors with vanishing rotation number. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 6, 1569–1590 Zbl 1253.37048 MR 2835324

[34] J. Palis Jr. and W. de Melo, *Geometric Theory of Dynamical Systems. An Introduction*. Springer, New York, 1982 Zbl 0491.58001 MR 669541

[35] S. Y. Pilyugin, *Introduction to Structurally Stable Systems of Differential Equations*. Birkhäuser, Basel, 1992 Zbl 0747.34031 MR 1158874

[36] V. A. Pliss, *Nonlocal Problems of the Theory of Oscillations*. Academic Press, New York, 1966 Zbl 0151.12104 MR 0196199

[37] J. C. Robinson, *Infinite-Dimensional Dynamical Systems. an Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*. Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, 2001 Zbl 0980.35001 MR 1881888

[38] D. Salamon, Connected simple systems and the Conley index of isolated invariant sets. *Trans. Amer. Math. Soc.* **291** (1985), no. 1, 1–41 Zbl 0573.58020 MR 797044

[39] J. J. Sánchez-Gabites, Dynamical systems and shapes. *Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Mat. RACSAM* **102** (2008), no. 1, 127–159 Zbl 1151.37017 MR 2416242

[40] J. M. R. Sanjurjo, Multihomotopy, Čech spaces of loops and shape groups. *Proc. London Math. Soc. (3)* **69** (1994), no. 2, 330–344 Zbl 0826.55004 MR 1281968

[41] J. M. R. Sanjurjo, On the structure of uniform attractors. *J. Math. Anal. Appl.* **192** (1995), no. 2, 519–528 Zbl 0823.58033 MR 1332224

[42] J. M. R. Sanjurjo, Global topological properties of the Hopf bifurcation. *J. Differential Equations* **243** (2007), no. 2, 238–255 Zbl 1126.37036 MR 2371787

[43] J. M. R. Sanjurjo, On the fine structure of the global attractor of a uniformly persistent flow. *J. Differential Equations* **252** (2012), no. 9, 4886–4897 Zbl 1263.37045 MR 2891350

[44] E. H. Spanier, *Algebraic Topology*. McGraw-Hill, New York, 1966 Zbl 0145.43303 MR 0210112

[45] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Appl. Math. Sci. 41, Springer, New York, 1982 Zbl 0504.58001 MR 681294

Héctor Barge
E.T.S. Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Madrid, Spain;
h.barge@upm.es

José M. R. Sanjurjo
Facultad de Ciencias Matemáticas and Instituto de Matemática Interdisciplinar (IMI),
Universidad Complutense de Madrid, 28040 Madrid, Spain; jose_sanjurjo@mat.ucm.es

DOI 10.4171/8ECM/32

Onset of fracture in random heterogeneous particle chains

Laura Lauerbach, Stefan Neukamm, Mathias Schäffner, and
Anja Schlömerkemper

Abstract. In mechanical systems, it is of interest to know the onset of fracture in dependence of the boundary conditions. Here we study a one-dimensional model which allows for an underlying heterogeneous structure in the discrete setting. Such models have recently been studied in the passage to the continuum by means of variational convergence (Γ-convergence). The Γ-limit results determine thresholds of the boundary condition, which mark a transition from purely elastic behavior to the occurrence of a crack. In this article, we provide a notion of fracture in the discrete setting and show that its continuum limit yields the same threshold as that obtained from the Γ-limit. Since the calculation of the fracture threshold is much easier with the new method, we see a good chance that this new approach will turn out useful in applications.

1. Introduction

The mechanical behavior of one-dimensional systems has been of interest for decades. Such systems serve as toy models for higher-dimensional theoretical investigations and are of interest with respect to one-dimensional structures; see, e.g., [8,9,11,12,21]. In order to understand the effective behavior of materials, the systems are studied as the number of particles tends to infinity.

In this article, we focus on the occurrence of cracks and continue a mathematical analysis of the effective behavior of one-dimensional discrete systems in the passage to the continuum. In particular, we strive for insight into the threshold for the overall prescribed length ℓ of a chain. If ℓ is smaller than the threshold, the system will show elastic behavior, whereas cracks are energetically favored if ℓ is larger than the threshold. The interaction potentials between the particles or atoms of the discrete chain are allowed to be in a large class of convex-concave potentials, which include for instance the classical Lennard-Jones potentials. The system is then modeled with

2020 Mathematics Subject Classification. Primary 74Q05; Secondary 74R10, 74A45, 41A60, 74G65.
Keywords. Fracture, discrete system, stochastic homogenization, Γ-convergence, Lennard-Jones potentials.

the help of an energy functional that is the sum of all the interaction potentials; see (2.1). Here we restrict to the interactions of nearest neighbors; for related studies with interactions beyond nearest neighbors we refer to [4, 5, 20].

In view of misplaced atoms or of chains consisting of several different kind of particles, we allow for a random distribution of the interaction potentials; see Assumption 2.1 and (2.2) for details. The limit passage is then also referred to as stochastic homogenization; cf., e.g., [1, 7, 10, 17]. As a special case, also materials with a periodic heterostructure are included; cf. also [15].

An appropriate mathematical technique for the passage of energy functionals from discrete to continuous systems is based on the notion of Γ-convergence, which is a notion of a variational convergence and (under coercivity assumptions) ensures that minimizers of the discrete system converge to minimizers of the system in the continuum limit; see, e.g., [2, 3, 19] and references cited therein. As the number of particles tends to infinity, the energy functional converges to a functional that allows for describing cracks. In particular, it is shown that cracks in the continuum limit emerge if a critical stretch is exceeded. On the other hand, on the discrete level a similar notion of a "critical stretch" or a notion for the onset of a crack has not been introduced so far.

In this article, which is partially based on the PhD thesis [13, Chapter 7] of L. Lauerbach, we focus on the emergence of cracks in atomistic chains. On the level of the continuum limiting model of the chain, "crack" has a clear meaning – it is the point where the continuum deformation features a jump and there is no interaction between the different segments separated by the jump. In contrary, on the level of a discrete chain with $n + 1$ particles, the notion of "crack" cannot be unambiguously defined, since always neighboring particles interact. In the present paper, we introduce a notion of "onset of a crack" at the discrete level for a chain with $n + 1$ particles. For simplicity, we discuss the key idea in the case of a chain with $n + 1$ particles that is composed of (random) potentials that are convex around its ground state and otherwise concave, i.e., for deformations larger than an inflection point z_{frac}. We call a deformation u *elastic* if the individual interaction potentials along the chain are only evaluated in their convex region. In contrary, a deformation that is *not* elastic invokes at least one bond that "lives" in the concave region of the corresponding potential. Next, we consider the energy minimizers u_n of the chain with $n + 1$ particles and prescribed total length $\ell > 0$. If the minimizers u_n are elastic for all $n \in \mathbb{N}$, then we do not expect the occurrence of crack in the continuum limit; while in the other case, we expect that minimizing sequences show a concentration of strain on a finite number of weak bonds and thus a "crack" emerges in the continuum limit. Based on these heuristics, we introduce a "critical stretch" ℓ_n^* for random chains with $n + 1$ particles. Firstly, we prove that it converges, for $n \to \infty$, to the jump-threshold predicted by the zeroth-order Γ-limit of the discrete energy, which has been obtained earlier in [14].

Secondly, we establish a first-order expansion of the critical stretch and show that the coefficients of the expansion term agree with the values predicted by the first-order Γ-limit of the discrete energy derived in [14]. Since the proofs in [14] are technically quite involved, it is interesting to learn that there is a much simpler method for the derivation of the jump threshold in the continuum limit. We expect that the new notions of a fracture point and of a jump threshold in the discrete setting turn out to be useful also in a wider class of applications. They might be compared to the Γ-convergence analysis of weak-membrane and Blake–Zisserman models in [6, 18], which invoke a combination of piecewise affine and piecewise constant interpolations that require the identification of strain concentration on the discrete level as well.

The outline of this article is as follows: in Section 2, we introduce the model in the discrete setting, including the assumptions on the large class of interaction potentials in the random setting. Further, we provide the definition of a critical stretch (Definition 2.1), which corresponds to the jump threshold. We assert the asymptotic behavior of the critical stretch as the number of particles tends to infinity (Theorem 2.1) and compare the limit to the corresponding Γ-convergence results. Moreover, we consider a rescaled setting, define the rescaled jump threshold, and assert its asymptotic behavior as n tends to infinity (Theorem 2.2). Finally, we compare also this result with the corresponding Γ-convergences result. All proofs are provided in Section 3.

2. Setup and main results

We consider a chain of $n+1$ atoms that in a reference configuration are placed at the sites in $\frac{1}{n}\mathbb{Z}\cap[0,1]$; see Figure 1. The deformation of the atoms is referred to as

$$u_n : \frac{1}{n}\mathbb{Z}\cap[0,1]\to\mathbb{R}.$$

For the passage from discrete systems to their continuous counterparts, it is useful to identify the discrete functions with their piecewise affine interpolations, more precisely, with the functions in

$$\mathcal{A}_n := \Big\{u\in C\big([0,1]\big) : u \text{ is affine on } (i,i+1)\frac{1}{n},\ i\in\{0,1,\dots,n-1\}, \\ \text{and monotonically increasing}\Big\}.$$

We shall also consider clamped boundary conditions for the chain and thus introduce for $\ell>0$ the set

$$\mathcal{A}_{n,\ell} := \big\{u\in\mathcal{A}_n : u(0)=0,\ u(1)=\ell\big\}.$$

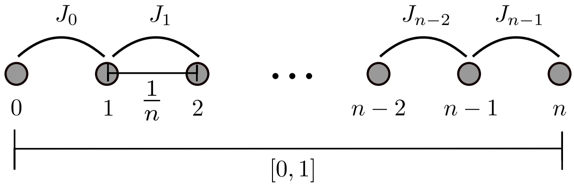


Figure 1. Chain of $n+1$ atoms with reference position $\frac{i}{n}$. The potential J_i describes the nearest neighbor interaction of atom i and $i+1$, $i=0,\dots,n-1$. The characteristic length scale is $\frac{1}{n}$ and the interval is $[0,1]$.

We consider a discrete energy functional of the form

$$\begin{aligned}\mathcal{A}_{n,\ell} \ni u \mapsto E_n(u) &:= \sum_{i=0}^{n-1} \frac{1}{n} J_i\left(\frac{u\left(\frac{i+1}{n}\right)-u\left(\frac{i}{n}\right)}{\frac{1}{n}}\right)\\ &= \sum_{i=0}^{n-1} \frac{1}{n} J_i\left(n\left(u\left(\frac{i+1}{n}\right)-u\left(\frac{i}{n}\right)\right)\right), \qquad (2.1)\end{aligned}$$

where $J_i : (0,\infty) \to \mathbb{R}$ is a potential describing the interaction between the ith atom and its neighbor to the right. We are interested in random heterogeneous chains of atoms, and thus assume that the potentials $\{J_i\}_{i\in\mathbb{Z}}$ are random with a distribution that is stationary and ergodic. We appeal to the following standard setup: let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space and $(\tau_i)_{i\in\mathbb{Z}}$ a family of measurable maps $\tau_i : \Omega \to \Omega$ such that

- (Group property) $\tau_0\omega = \omega$ for all $\omega \in \Omega$ and $\tau_{i_1+i_2} = \tau_{i_1}\tau_{i_2}$ for all $i_1, i_2 \in \mathbb{Z}$,
- (Stationarity) $\mathbb{P}(\tau_i B) = \mathbb{P}(B)$ for every $B \in \mathcal{F}$, $i \in \mathbb{Z}$,
- (Ergodicity) For all $B \in \mathcal{F}$, it holds that $(\tau_i(B) = B \ \forall i \in \mathbb{Z}) \Rightarrow \mathbb{P}(B) = 0$ or $\mathbb{P}(B) = 1$.

We then consider the energy functional

$$E_n : \Omega \times \mathcal{A}_n \to \mathbb{R} \cup \{+\infty\}$$

with

$$E_n(\omega, u) := \sum_{i=0}^{n-1} \frac{1}{n} J\left(\tau_i\omega, n\left(u\left(\frac{i+1}{n}\right)-u\left(\frac{i}{n}\right)\right)\right), \qquad (2.2)$$

where the random potential satisfies the following assumptions:

Assumption 2.1. Let $J : \Omega \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be jointly measurable with $J(\cdot, z) = \infty$ if $z \le 0$. For $\mathbb{P}$-a.e. $\omega \in \Omega$, the following conditions hold true:

(A1) (Regularity) $J(\omega, \cdot) \in C^3(0,\infty)$.

(A2) (Behavior at 0 and ∞) There exist functions $\psi^+, \psi^- \in C(0,\infty)$, independent of ω, such that

$$\lim_{z\to 0+} \psi^-(z) = \infty \quad \text{and} \quad \lim_{z\to\infty} \psi^+(z) = 0,$$

and

$$J(\omega,z) \geq \psi^-(z) \text{ for all } 0 < z \leq 1 \quad \text{and} \quad \left|J(\omega,z)\right| \leq \psi^+(z) \text{ for all } z \geq 1.$$

(A3) (Convex-monotone structure) Suppose strict convexity close to 0 in form of

$$z_{\text{frac}}(\omega) := \sup\left\{z > 0 \,:\, J''(\omega,s) := \partial_s^2 J(\omega,s) > 0 \text{ for all } s \in (0,z)\right\} > 0,$$

and assume that $J(\omega,\cdot)$ is monotonically increasing on $[z_{\text{frac}}(\omega),\infty)$.

(A4) (Non-degenerate ground state) Suppose that $J(\omega,\cdot)$ admits a unique minimizer $\delta(\omega) \in (0, z_{\text{frac}}(\omega)]$, called the ground state of $J(\omega,\cdot)$. There exists a constant $c > 0$, independent of ω, such that $\frac{1}{c} > \delta(\omega) > c$ and

$$\forall z \in \delta(\omega) + (-c,c) : c \leq J''(\omega,z) \leq \frac{1}{c} \quad \text{and} \quad \left|J'''(\omega,z)\right| \leq \frac{1}{c}.$$

Next, we introduce the following central quantities for a random heterogeneous chain with $n+1$ particles:

Definition 2.1 (Critical stretch of a chain with $n+1$ particles). Consider the situation of Assumption 2.1. Let $n \in \mathbb{N}$ and $\omega \in \Omega$. The critical stretch $\ell_n^*(\omega)$ is defined as the largest number such that

$$\inf_{\mathcal{A}_n^{\text{el}}(\omega)\cap\mathcal{A}_{n,\ell}} E_n(\omega,\cdot) = \inf_{\mathcal{A}_{n,\ell}} E_n(\omega,\cdot) \quad \text{for all } 0 \leq \ell < \ell_n^*(\omega),$$

where we denote by

$$\mathcal{A}_n^{\text{el}}(\omega) := \left\{u \in \mathcal{A}_n \,:\, \frac{u\left(\frac{i+1}{n}\right) - u\left(\frac{i}{n}\right)}{\frac{1}{n}} \leq z_{\text{frac}}(\tau_i\omega) \text{ for all } i = 0,\ldots,n-1\right\}$$

the set of purely elastic deformations.

The idea behind the above definition is the following: a deformation $u \in \mathcal{A}_n^{\text{el}}(\omega)$ only sees the strictly convex region of the interaction potentials. Thus, we could replace the potentials $J(\tau_i\omega, z)$ in the definition of the energy function E_n by (globally) convex potentials with superlinear growth without changing the energy for deformations in $\mathcal{A}_n^{\text{el}}(\omega)$. As it is well known, such energies do not allow for fracture in the continuum limit. The definition of the critical stretch implies that a prescribed macroscopic stretch (or compression) $\ell < \ell_n^*(\omega)$ can be realized by a deformation in

$\mathcal{A}_n^{\mathrm{el}}(\omega)$ and thus prohibits the formation of a jump, while, for $\ell > \ell_n^*(\omega)$, deformations with minimal energy are required to explore the non-convex region of at least one of the interaction potentials. We may refer to the bonds $[i, i+1]$ that are evaluated outside the convex region as "weak" bonds. If a jump occurs in the limit, then the minimizing sequence shows a concentration of strain in the weak bonds. We thus expect that $\ell_n^*(\omega)$ almost surely converges in the limit $n \to \infty$ to the continuum fracture threshold that can be defined on the level of the continuum Γ-limit; see below. In our first result, we prove that ℓ_n^* indeed converges and we identify its limit, which is the statistical mean of the ground states:

Theorem 2.1. *Let Assumption 2.1 be fulfilled. Then,*

$$\lim_{n\to\infty} \ell_n^*(\omega) = \mathbb{E}[\delta] \quad \textit{for } \mathbb{P}\textit{-a.e. } \omega \in \Omega.$$

(The proof of Theorem 2.1 can be found in Section 3.1.)

Next, we consider the special case when $\delta(\omega)$ is deterministic, say $\delta(\omega) = 1$ for $\mathbb{P}$-a.e. In that case, we establish a first-order expansion of $\ell_n^*(\omega)$ around its limit $\mathbb{E}[\delta] = 1$ of the form

$$\ell_n^*(\omega) \approx 1 + \sqrt{\frac{1}{n}}\sqrt{\frac{\beta}{\underline{\alpha}}},$$

where β is related to the maximal energy barrier among the random potentials J, and $1/\underline{\alpha}$ is the statistical mean of the curvatures of the random potentials at the ground state.

Theorem 2.2. *Let Assumption 2.1 be satisfied and assume that $\delta(\omega) = 1$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. Consider the rescaled jump threshold $\gamma_n^*(\omega) := \frac{\ell_n^*(\omega)-1}{\sqrt{\frac{1}{n}}}$. Then*

$$\lim_{n\to\infty} \gamma_n^*(\omega) = \lim_{n\to\infty} \frac{\ell_n^*(\omega) - 1}{\sqrt{\frac{1}{n}}} = \sqrt{\frac{\beta}{\underline{\alpha}}} \quad \textit{for } \mathbb{P}\textit{-a.e. } \omega \in \Omega,$$

where

$$\underline{\alpha} := \left(\mathbb{E}\left[\left(\frac{1}{2}J''(\omega,1)\right)^{-1}\right]\right)^{-1} \quad \textit{and} \quad \beta := \operatorname*{ess\,inf}_{\omega\in\Omega}\big(-J(\omega,1)\big). \tag{2.3}$$

(The proof of Theorem 2.2 can be found in Section 3.2.)

We finally relate the above results to the zeroth- and first-order Γ-limits of E_n subject to clamped boundary conditions, i.e.,

$$E_n^\ell(\omega,\cdot) : L^1(0,1) \to \mathbb{R} \cup \{+\infty\}, \quad E_n^\ell(\omega, u) := \begin{cases} E_n(\omega, u) & \text{if } u \in \mathcal{A}_{n,\ell}, \\ +\infty & \text{else.} \end{cases}$$

The zeroth-order Γ-limit of the discrete energy yields a homogenized energy functional. In the present setting of nearest-neighbor interactions, [14] allows to characterize the homogenized energy functional by

$$E^{\ell}_{\mathrm{hom}}(u) = \int_0^1 J_{\mathrm{hom}}\big(u'(x)\big)\,\mathrm{d}x,$$

where the homogenized energy density map $z \mapsto J_{\mathrm{hom}}(z)$ is convex, lower semicontinuous, monotonically decreasing and satisfies

$$\lim_{z\to 0+} J_{\mathrm{hom}}(z) = +\infty. \tag{2.4}$$

Moreover, the minimum values of $E^{\ell}_{n}(\omega,\cdot)$ and E^{ℓ}_{hom} satisfy

$$\lim_{n\to\infty} \inf_{u} E^{\ell}_{n}(\omega,u) = \min_{u} E^{\ell}_{\mathrm{hom}}(u) = J_{\mathrm{hom}}(\ell),$$

and therefore can be calculated as

$$\min_{u} E^{\ell}_{\mathrm{hom}}(u) = J_{\mathrm{hom}}(\ell) = \begin{cases} J_{\mathrm{hom}}(\ell) & \text{for } \ell < \mathbb{E}[\delta], \\ J_{\mathrm{hom}}\big(\mathbb{E}[\delta]\big) & \text{for } \ell \geq \mathbb{E}[\delta]. \end{cases}$$

Hence, the threshold between the elastic and the jump regimes is $\mathbb{E}[\delta]$, which is identical to the limit of $\ell^{*}_{n}(\omega)$; see Theorem 2.1. Secondly, we recall a Γ-limit result from [16] for the rescaled energy functional

$$H^{\gamma_n}_{n}(\omega,v) = \begin{cases} H_n(\omega,v) & \text{if } v \in \mathcal{A}_{n,\gamma_n}, \\ +\infty & \text{otherwise,} \end{cases}$$

where $(\gamma_n)_n$ is a sequence of non-negative numbers with $\gamma_n \to \gamma \geq 0$ and

$$H_n(\omega,v) := \sum_{i=0}^{n-1}\left(J\left(\tau_i\omega, \frac{v\big(\frac{i+1}{n}\big) - v\big(\frac{i}{n}\big)}{\sqrt{\frac{1}{n}}} + \delta(\tau_i\omega)\right) - J\big(\tau_i\omega, \delta(\tau_i\omega)\big)\right).$$

The Γ-limit is shown to be given as

$$H^{\gamma}(v) = \underline{\alpha}\int_0^1 \big|v'(x)\big|^2\,\mathrm{d}x + \beta\#S_v,$$

with homogenized elastic coefficient $\underline{\alpha}$, jump parameter β, $\#S_v$ being the number of jumps of v, and v satisfying boundary conditions which depend on γ. Moreover, it holds true that

$$\lim_{n\to\infty} \inf_{v} H^{\gamma_n}_{n}(\omega,v) = \min_{v} H^{\gamma}(v) = \min\{\underline{\alpha}\gamma^2, \beta\},$$

which yields that the minima of the energy are given by

$$\min_v H^\gamma(v) = \min\{\underline{\alpha}\gamma^2, \beta\} = \begin{cases} \underline{\alpha}\gamma^2 & \text{if } \gamma < \sqrt{\frac{\beta}{\underline{\alpha}}}, \\ \beta & \text{if } \gamma \geq \sqrt{\frac{\beta}{\underline{\alpha}}}. \end{cases}$$

Hence the threshold between elasticity and fracture in the rescaled case is $\sqrt{\frac{\beta}{\underline{\alpha}}}$, which equals the limit of the jump threshold γ_n^* in Theorem 2.2.

In summary, although the techniques by which the results are calculated are completely different, they yield the same result regarding the jump threshold in the continuum setting. The derivation of the limiting jump threshold with help of the newly defined jump threshold in the discrete setting is, however, much easier and thus is of interest for applications. It remains an open problem to analyze corresponding questions in higher dimensional settings. In the following section, we provide the proofs of the above theorems.

3. Proofs

For the upcoming analysis, it is convenient to introduce the notation

$$M_n(\omega, \ell) := \min\left\{\frac{1}{n}\sum_{i=0}^{n-1} J(\tau_i\omega, z^i) : \frac{1}{n}\sum_{i=0}^{n-1} z^i = \ell\right\}$$

to denote the minimum energy of a discrete chain of length ℓ. We begin with an elementary (yet, convenient) reformulation of the critical stretch ℓ_n^* (cf. Definition 2.1).

Lemma 3.1. *Consider the situation of Assumption 2.1. Let $n \in \mathbb{N}$ and $\omega \in \Omega$. Then, it holds*

$$M_n(\omega, \ell) = \min_{u \in \mathcal{A}_{n,\ell}} E_n(\omega, u). \tag{3.1}$$

Moreover, $\ell_n^*(\omega)$ is the largest number such that for all $0 < \ell < \ell_n^*(\omega)$ there exists $\bar{z} \in \mathbb{R}^n$ satisfying

$$M_n(\omega,\ell) = \frac{1}{n}\sum_{i=0}^{n-1} J(\tau_i\omega, \bar{z}^i), \quad \frac{1}{n}\sum_{i=0}^{n-1}\bar{z}^i = \ell, \quad \bar{z}^i \leq z_{\text{frac}}(\tau_i\omega) \quad \forall i \in \{0,\dots,n-1\}. \tag{3.2}$$

Proof of Lemma 3.1. The identity (3.1) follows by a simple change of variables, that is by setting

$$z^i = n\left(u\left(\frac{i+1}{n}\right) - u\left(\frac{i}{n}\right)\right),$$

and the direct method of the calculus of variations.

Next, we give an argument regarding the characterization of ℓ_n^*. The definition of $\ell_n^*(\omega)$, see Definition 2.1, and (3.1) imply that

$$\inf_{\mathcal{A}_n^{\mathrm{el}}(\omega)\cap\mathcal{A}_{n,\ell}} E_n(\omega,\cdot) = M_n(\omega,\ell) < \infty \quad \forall \ell \in \big(0, \ell_n^*(\omega)\big).$$

Since $\mathcal{A}_n^{\mathrm{el}}(\omega) \cap \mathcal{A}_{n,\ell}$ is compact, there exists $\bar{u} \in \mathcal{A}_n^{\mathrm{el}}(\omega) \cap \mathcal{A}_{n,\ell}$ such that

$$E_n(\omega,\bar{u}) = \inf_{\mathcal{A}_n^{\mathrm{el}}(\omega)\cap\mathcal{A}_{n,\ell}} E_n(\omega,\cdot).$$

Clearly, $\bar{z} \in \mathbb{R}^n$ defined as $\bar{z}^i = n(\bar{u}(\frac{i+1}{n}) - \bar{u}(\frac{i}{n}))$ satisfies (3.2).

Now we suppose that for some $\ell \geq \ell_n^*$ there exists $\bar{z} \in \mathbb{R}^n$ satisfying (3.2). With help of the same change of variables as above, we find $\bar{u} \in \mathcal{A}_n^{\mathrm{el}}(\omega) \cap \mathcal{A}_{n,\ell}$ satisfying $E_n(\omega,\bar{u}) = M_n(\omega,\ell)$ which contradicts the definition of ℓ_n^*. ■

Lemma 3.2. *Let Assumption 2.1 be satisfied. Then, $J(\omega,\cdot)$ is increasing on $[\delta(\omega),\infty)$ and it holds that*

$$z_{\mathrm{frac}}^{\sup} := \sup\big\{z_{\mathrm{frac}}(\omega) : \omega \in \Omega\big\} < \infty. \tag{3.3}$$

Proof of Lemma 3.2. For convenience we drop the dependence on ω in our notation and simply write $J(z)$, δ, and z_{frac} instead of $J(\omega,z)$, $\delta(\omega)$, and $z_{\mathrm{frac}}(\omega)$, respectively. We first prove that J is increasing on $[\delta,\infty)$. On $[z_{\mathrm{frac}},\infty)$ this directly follows from (A3). On $[\delta, z_{\mathrm{frac}})$ this follows from the convexity of J on $(0, z_{\mathrm{frac}})$ and the fact that δ minimizes J. Next, we prove (3.3). We first note that (A2) and (A3) imply that

$$\forall z \in (\delta,\infty) : J(\delta) \leq J(z) \leq 0. \tag{3.4}$$

Moreover, (A4) implies that $z_{\mathrm{frac}} \geq \delta + c$. Thus, for all $\eta \in (0,c)$ we obtain

$$\begin{aligned} 0 \geq J(z_{\mathrm{frac}}) &= J(\delta+\eta) + \int_{\delta+\eta}^{z_{\mathrm{frac}}} J'(t)\,\mathrm{d}t \\ &\geq J(\delta+\eta) + J'(\delta+\eta)\big(z_{\mathrm{frac}} - (\delta+\eta)\big), \end{aligned} \tag{3.5}$$

where the second inequality holds, since J' is increasing on $(\delta+\eta, z_{\mathrm{frac}})$ thanks to (A3). (A4) yields

$$J'(\delta+\eta) = J'(\delta+\eta) - J'(\delta) = \int_{\delta}^{\delta+\eta} J''(s)\,ds \geq c\eta.$$

Thus, by rearranging terms in (3.5) and appealing to (3.4) and the previous estimate we get

$$z_{\mathrm{frac}} \leq \delta + \eta - \frac{J(\delta+\eta)}{J'(\delta+\eta)} \leq \delta + \eta - \frac{J(\delta)}{c\eta}. \tag{3.6}$$

It remains to bound $\delta = \delta(\omega)$ and $-J(\delta) = -J(\omega, \delta(\omega))$ by a constant that is independent of ω. From (A4) and (A2), we get

$$\delta \in \left(c, \frac{1}{c}\right) \quad \text{and} \quad -J(\delta) \leq \max_{z\in[c,\frac{1}{c}+\eta]} \max\left\{-\psi^-(z),\, \left|\psi^+(z)\right|\right\} =: d < \infty, \tag{3.7}$$

and thus, (3.6) yields $z_{\text{frac}} \leq \frac{1}{c} + \eta + \frac{d}{c\eta}$. ■

3.1. Proof of Theorem 2.1

Proof of Theorem 2.1. Note that $\omega \mapsto \delta(\omega)$ is (as a minimizer of a measurable function) measurable. Moreover, by (3.7) δ is a non-negative and bounded and thus an L^1-random variable. Thus the ergodic theorem yields

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \delta(\tau_i\omega) = \mathbb{E}[\delta], \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} J\big(\tau_i\omega, \delta(\tau_i\omega)\big) = \mathbb{E}\big[J(\delta)\big] \tag{3.8}$$

for $\mathbb{P}$-a.e. $\omega \in \Omega$. For the rest of the proof, we consider $\omega \in \Omega$ such that (3.8) is valid and drop the dependence on ω. In particular, we set

$$\delta_i := \delta(\tau_i\omega), \quad z^i_{\text{frac}} := z_{\text{frac}}(\tau_i\omega), \quad \text{and} \quad J_i(z) := J(\tau_i\omega, z).$$

Step 1. We show that $\bar{A} := \limsup_{n\to\infty} \ell^*_n \leq \mathbb{E}[\delta]$.

Without loss of generality, we suppose that $\bar{A} = \lim_{n\to\infty} \ell^*_n$ and prove $\bar{A} \leq \mathbb{E}[\delta]$ by contradiction. Assume that there exists $\varepsilon \in (0, c)$ such that $\bar{A} > \mathbb{E}[\delta] + 3\varepsilon$. By (3.8), we find that $\bar{N} \in \mathbb{N}$ such that

$$\ell^*_n > \frac{1}{n} \sum_{i=0}^{n-1} \delta_i + 2\varepsilon =: k_n \quad \text{for } n > \bar{N}. \tag{3.9}$$

In view of Lemma 3.1, there exists a sequence $(\bar{z}_n)_n$ satisfying for $n \geq \bar{N}$

$$\frac{1}{n}\sum_{i=0}^{n-1} \bar{z}^i_n = k_n, \quad \frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}^i_n) = M_n(k_n), \quad \bar{z}^i_n \leq z^i_{\text{frac}} \quad \forall i \in \{0, \dots, n-1\}. \tag{3.10}$$

We claim that

$$\limsup_{n\to\infty} M_n(k_n) \leq \mathbb{E}\big[J(\delta)\big], \tag{3.11}$$

$$\liminf_{n\to\infty} M_n(k_n) \geq \mathbb{E}\big[J(\delta)\big] + c_\varepsilon, \tag{3.12}$$

for some $c_\varepsilon > 0$. Clearly, (3.11) and (3.12) yield a contradiction. Hence the assumption $\bar{A} > \mathbb{E}[\delta] + 3\varepsilon$ is wrong and $\bar{A} \leq \mathbb{E}[\delta]$ follows by the arbitrariness of $\varepsilon > 0$.

Substep 1.1. Proof of (3.11). Let $z_n \in \mathbb{R}^n$ be given by $z_n^i := \delta_i$ for $i \geq 1$ and $z_n^0 := \delta_0 + 2n\varepsilon$. Since $\frac{1}{n}\sum_{i=0}^{n-1} z_n^i = k_n$, we have

$$\begin{aligned} M_n(k_n) &\leq \frac{1}{n}\sum_{i=1}^{n-1} J_i(\delta_i) + \frac{1}{n}J_0(\delta_0 + 2n\varepsilon) \\ &= \frac{1}{n}\sum_{i=0}^{n-1} J_i(\delta_i) + \frac{1}{n}\big(J_0(\delta_0 + 2n\varepsilon) - J_0(\delta_0)\big). \end{aligned}$$

Hence, (3.11) follows by (A2) and (3.8).

Substep 1.2. Proof of (3.12). Let $\bar{z}_n$ be as in (3.10) and set

$$I_n := \big\{i \in \{0, \dots, n-1\} : \bar{z}_n^i > \delta_i + \varepsilon\big\}.$$

Obviously, it holds that $0 \leq |I_n|/n \leq 1$ and we claim

$$\frac{|I_n|}{n} \geq \frac{\varepsilon}{z_{\mathrm{frac}}^{\sup}} > 0 \quad \text{for all } n \in \mathbb{N}, \tag{3.13}$$

where $z_{\mathrm{frac}}^{\sup} \in (0, \infty)$ is as in Lemma 3.2. Indeed,

$$\begin{aligned} \frac{1}{n}\sum_{i=0}^{n-1}\delta_i + 2\varepsilon = k_n &= \frac{1}{n}\sum_{i=0}^{n-1}\bar{z}_n^i = \frac{1}{n}\sum_{i\in I_n}\bar{z}_n^i + \frac{1}{n}\sum_{i\notin I_n}\bar{z}_n^i \\ &\overset{(3.10)}{\leq} \frac{|I_n|}{n}z_{\mathrm{frac}}^{\sup} + \frac{1}{n}\sum_{i=0}^{n-1}(\delta_i + \varepsilon) \end{aligned}$$

implies (3.13). Finally, using the monotonicity of J_i on (δ_i, ∞) (see Lemma 3.2) and (A4), we obtain

$$\begin{aligned} \frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) &= \frac{1}{n}\sum_{i\in I_n} J_i(\bar{z}_n^i) + \frac{1}{n}\sum_{i\notin I_n} J_i(\bar{z}_n^i) \geq \frac{1}{n}\sum_{i\in I_n} J_i(\delta_i + \varepsilon) + \frac{1}{n}\sum_{i\notin I_n} J_i(\delta_i) \\ &\geq \frac{1}{n}\sum_{i\in I_n}\Big(J_i(\delta_i) + \frac{1}{2}c\varepsilon^2\Big) + \frac{1}{n}\sum_{i\notin I_n} J_i(\delta_i) = \frac{1}{n}\sum_{i=0}^{n-1} J_i(\delta_i) + \frac{|I_n|}{n}\frac{1}{2}c\varepsilon^2, \end{aligned}$$

where $c > 0$ is as in (A4). Sending $n \to \infty$, we obtain with help of (3.8) and (3.13) the claim (3.12).

Step 2. We claim $\underline{A} := \liminf_{n\to\infty} \ell_n^* \geq \mathbb{E}[\delta]$.

For all $\varepsilon > 0$, we show that

$$\ell_n^* \geq \frac{1}{n}\sum_{i=0}^{n-1}\delta_i - \varepsilon =: k_n \quad \forall n \in \mathbb{N}, \tag{3.14}$$

which in combination with (3.8) implies that $\underline{A} := \liminf_{n\to\infty} \ell_n^* \geq \mathbb{E}[\delta]$ by the arbitrariness of $\varepsilon > 0$.

Let $\bar{z}_n$ be such that

$$\frac{1}{n}\sum_{i=0}^{n-1} \bar{z}_n^i = k_n, \quad \frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) = M_n(k_n).$$

We show that $\bar{z}_n^i \leq \delta_i < z_{\text{frac}}^i$ $\forall i \in \{0,\dots,n-1\}$, which obviously implies (3.14). Indeed, the optimality condition for $\bar{z}_n$ implies that there exists a Lagrange multiplier $\Lambda \in \mathbb{R}$ such that $\Lambda = J_i'(\bar{z}_n^i)$ for all $i \in \{0,\dots,n-1\}$. Since

$$\frac{1}{n}\sum_{i=0}^{n-1}(\bar{z}_n^i - \delta_i) \leq -\varepsilon,$$

there exists $\hat{i} \in \{0,\dots,n-1\}$ such that $\bar{z}_n^{\hat{i}} \in (0,\delta_i)$ and thus $J_{\hat{i}}'(\bar{z}_n^{\hat{i}}) < 0$. Hence $J_i'(\bar{z}_n^i) < 0$ for all $i \in \{0,\dots,n-1\}$. Since $J_i' \geq 0$ on (δ_i,∞) by Lemma 3.2, we conclude that $\bar{z}_n^i \leq \delta_i \leq z_{\text{frac}}^i$ and thus $\ell_n^* \geq k_n$ by Lemma 3.1. ■

3.2. Proof of Theorem 2.2

We begin with a preliminary structure result for minimizers of the minimum problem in the definition of $M_n(\omega, 1+n^{-\frac{1}{2}}D)$ for some $D > 0$; see (3.1).

Proposition 3.3. *Let Assumption 2.1 be satisfied and assume that $\delta(\omega) = 1$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. Fix $D > 0$. There exist $\bar{N} \in \mathbb{N}$ and a sequence (N_n) satisfying $N_n \to \infty$ such that the following statements hold true for $\mathbb{P}$-a.e. $\omega \in \Omega$ and $n \geq \bar{N}$.*

Let $\bar{z}_n \in \mathbb{R}^n$ be such that

$$\frac{1}{n}\sum_{i=0}^{n-1} \bar{z}_n^i = 1 + n^{-\frac{1}{2}}D \quad \text{and} \quad \frac{1}{n}\sum_{i=0}^{n-1} J(\tau_i\omega, \bar{z}_n^i) = M_n(\omega, 1+n^{-\frac{1}{2}}D). \tag{3.15}$$

Then, it holds that

$$\bar{z}_n^i \in [1, 1 + c^{-2}n^{-\frac{1}{2}}D] \cup [N_n, \infty) \quad \text{for all } i \in \{0,\dots,n-1\}, \tag{3.16}$$

where $c > 0$ is as in (A4).

Proof of Proposition 3.3. We consider $\omega \in \Omega$ such that $\delta(\tau_i\omega) = 1$ $\forall i \in \mathbb{N}$ and drop the dependence on ω. Moreover, we use the shorthand notation $z_{\text{frac}}^i := z_{\text{frac}}(\tau_i\omega)$ and $J_i(z) := J(\tau_i\omega, z)$.

Step 1. We show that

$$0 \leq J'(\bar{z}_n^i) \leq \frac{1}{c}Dn^{-\frac{1}{2}} \quad \text{for all } i \in \{0,\dots,n-1\}, \tag{3.17}$$

where $c > 0$ is as in (A4).

By the optimality condition for $\bar{z}_n$, there exists a Lagrange multiplier $\Lambda \in \mathbb{R}$ such that $\Lambda = J_i'(\bar{z}_n^i)$ for all $i \in \{0, \dots, n-1\}$. Since

$$\frac{1}{n}\sum_{i=0}^{n-1} \bar{z}_n^i = 1 + n^{-\frac{1}{2}} D,$$

there exists $i_1 \in \{0, \dots, n-1\}$ such that $\bar{z}_n^{i_1} \geq 1 + n^{-\frac{1}{2}} D > 1$. Lemma 3.2 and the assumption $\delta(\tau_i \omega) = 1$ imply that J_i is increasing on $(1, \infty)$ and thus we have $\Lambda \geq 0$. Moreover, there exists $i_2 \in \{0, \dots, n-1\}$ such that $\bar{z}_n^{i_2} \leq 1 + n^{-\frac{1}{2}} D$. For n sufficiently large such that $n^{-\frac{1}{2}} D < c$, where $c > 0$ as in (A4), we have (using that $J_i'(1) = 0$)

$$0 \leq \Lambda = J_{i_2}'(\bar{z}_n^{i_2}) = \int_1^{\bar{z}_n^{i_2}} J_{i_2}''(t)\, dt \overset{(A4)}{\leq} \frac{1}{c} n^{-\frac{1}{2}} D.$$

Since $\Lambda = J_i'(\bar{z}_n^i)$ for all $i \in \{0, \dots, n-1\}$, the claim (3.17) follows.

Step 2. Argument for (3.16).

We firstly observe that (3.17) implies that $1 \leq \bar{z}_n^i$ for all $i \in \{0, \dots, n-1\}$ (recall $J_i'(z) < 0$ on $(0, 1)$). The remaining estimates of (3.16) are proven in three steps.

Substep 2.1. We claim that for n sufficiently large, $\bar{z}_n^i \leq z_{\mathrm{frac}}^i$ implies that

$$\bar{z}_n^i \leq 1 + c^{-2} n^{-\frac{1}{2}} D,$$

where $c > 0$ is as in (A4). Indeed, using $J_i''(s) > 0$ on $(0, z_{\mathrm{frac}}^i)$ and (A4), we deduce from $\bar{z}_n^i \leq z_{\mathrm{frac}}^i$ and n sufficiently large that

$$c^{-1} D n^{-\frac{1}{2}} \overset{(3.17)}{\geq} J_i'(\bar{z}_n^i) = \int_1^{\bar{z}_n^i} J_i''(t)\, dt \overset{(A4)}{\geq} c \min\{\bar{z}_n^i - 1, c\}.$$

From the above inequality, we deduce that $\bar{z}_n^i - 1 \geq c$ implies that $n \leq D^2/c^6$. Hence, $\bar{z}_n^i - 1 < c$ and thus $1 \leq \bar{z}_n^i \leq 1 + c^{-2} D n^{-\frac{1}{2}}$ for $n > D^2/c^6$.

Substep 2.2. There exists $M < \infty$, depending only on $\psi^-(1)$ from (A2) and $c > 0$ from (A4), such that

$$\sup_{n \in \mathbb{N}} |I_n^w| \leq M, \quad \text{where } I_n^w := \left\{ i \in \{0, \dots, n-1\} : \bar{z}_n^i \geq z_{\mathrm{frac}}^i \right\}. \tag{3.18}$$

Suppose that $|I_n^w| \geq 2$ and consider some $i_n \in I_n^w$. Define

$$\hat{z}_n^i := \begin{cases} \bar{z}_n^i & \text{if } i \notin I_n^w, \\ 1 & \text{if } i \in I_n^w \setminus \{i_n\}, \\ 1 + \sum_{i \in I_n^w} (\bar{z}_n^i - 1) & \text{if } i = i_n. \end{cases} \tag{3.19}$$

By construction, we have $\sum_{i=0}^{n-1}\bar{z}_n^i = \sum_{i=0}^{n-1}\hat{z}_n^i$ and thus by (3.15)

$$\begin{aligned} 0 &\geq \sum_{i=0}^{n-1}\big(J_i(\bar{z}_n^i) - J_i(\hat{z}_n^i)\big) \\ &= \sum_{i\in I_n^w\setminus\{i_n\}}\big(J_i(\bar{z}_n^i) - J_i(1)\big) + J_{i_n}(\bar{z}_n^{i_n}) - J_{i_n}(\hat{z}_n^{i_n}). \end{aligned} \tag{3.20}$$

By the monotonicity of J_i on $(1,\infty)$, (A3), and (A4) in the form

$$J_i(z_{\text{frac}}^i) - J_i(1) \geq J_i(1+c) - J_i(1) = \int_1^{1+c}\int_1^s J_i''(t)\,dt\,ds \geq \frac{1}{2}c^3$$

(where $c > 0$ is as in (A4)), we find

$$J_i(\bar{z}_n^i) - J_i(1) \geq J_i(z_{\text{frac}}^i) - J_i(1) \geq \frac{1}{2}c^3 := \eta \quad \forall i \in I_n^w. \tag{3.21}$$

Moreover, using $\hat{z}_n^{i_n} \geq 1$ and thus $J_{i_n}(\hat{z}_n^{i_n}) \leq 0$ (which follows from the monotonicity of J_i on $(1,\infty)$ and (A2)), we obtain

$$J_{i_n}(\bar{z}_n^{i_n}) - J_{i_n}(\hat{z}_n^{i_n}) \geq J_{i_n}(1) \overset{(A2)}{\geq} \psi^-(1). \tag{3.22}$$

Combining (3.20)–(3.22), we deduce the uniform bound $|I_n^w| \leq 1 - \eta^{-1}\psi^-(1)$.

Substep 2.3. We show that there exists (N_n) satisfying $N_n \to \infty$ as $n \to \infty$ such that $\bar{z}_n^i \geq N_n$ for all $i \in I_n^w$, where I_n^w is defined in (3.18).

We argue by contradiction and assume that there exists $A \in [1,\infty)$ and an index $\hat{i} \in I_n^w$ such that $\bar{z}_n^{\hat{i}} \leq A$. For n sufficiently large, we show that this contradicts (3.15). Define

$$\tilde{z}_n^i := \begin{cases} 1 & \text{if } i = \hat{i}, \\ \bar{z}_n^i + \big(n - |I_n^w|\big)^{-1}(\bar{z}_n^{\hat{i}} - 1) & \text{if } i \notin I_n^w, \\ \bar{z}_n^i & \text{if } i \in I_n^w \setminus \{\hat{i}\}. \end{cases} \tag{3.23}$$

By construction, we have $\sum_{i=0}^{n-1}\tilde{z}_n^i = \sum_{i=0}^{n-1}\bar{z}_n^i$. Since $\bar{z}_n$ is a minimizer (see (3.15)),

$$0 \geq \sum_{i=0}^{n-1}\big(J_i(\bar{z}_n^i) - J_i(\tilde{z}_n^i)\big) = J_{\hat{i}}(\bar{z}_n^{\hat{i}}) - J_{\hat{i}}(1) + \sum_{i\notin I_n^w}\big(J_i(\bar{z}_n^i) - J_i(\tilde{z}_n^i)\big).$$

By (3.21), we have $J_{\hat{i}}(\bar{z}_n^{\hat{i}}) - J_{\hat{i}}(1) \geq \eta(c) > 0$. To obtain a contradiction, it suffices to show that the second term on the right-hand side vanishes as n tends to infinity. This can be seen as follows: on the one hand, we have $\bar{z}_n^i \in [1, 1 + c^{-2}n^{-\frac{1}{2}}D]$ for all $i \notin I_n^w$ by Substep 2.1, and on the other hand, we have

$$(n - |I_n^w|)^{-1}(\bar{z}_n^{\hat{i}} - 1) \leq (n - M)^{-1}(A - 1),$$

thanks to $|I_n^w| \leq M$. Hence, $\bar{z}_n^i, \tilde{z}_n^i \in [1, 1 + \frac{c}{2}]$ for n sufficiently large (depending only on c, D, M, and A). Now, a quadratic Taylor expansion of J_i at $\bar{z}_n^i$ yields (using $|J''(z)| \leq c^{-1}$ for $z \in [1, 1 + c)$; see (A4))

$$\begin{aligned}\sum_{\substack{i=1\\ i\notin I_n^w}}^{n} \big|J_i(\bar{z}_n^i) - J_i(\tilde{z}_n^i)\big| &\leq \sum_{i=0}^{n-1} \Big(\big|J_i'(\bar{z}_n^i)\big|(n-M)^{-1}(A-1) + c^{-1}(n-M)^{-2}(A-1)^2\Big)\\ &\overset{(3.17)}{\leq} n(n-M)^{-1}c^{-1}(A-1)\big(n^{-\frac{1}{2}}D + (A-1)(n-M)^{-1}\big)\\ &\leq Cn^{-\frac{1}{2}},\end{aligned}$$

where $C < \infty$ depends only on A, c, D, and M. ■

Proof of Theorem 2.2. By the ergodic theorem, it holds that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} J''(\tau_i\omega, 1)^{-1} = \mathbb{E}\big[J''(1)^{-1}\big], \qquad \lim_{n\to\infty} \beta_n(\omega) = \beta \tag{3.24}$$

for $\mathbb{P}$-a.e. $\omega \in \Omega$, where β is defined in (2.3) and

$$\beta_n(\omega) := \min\big\{-J(\tau_i\omega, 1) : i \in \{0, \dots, n-1\}\big\}. \tag{3.25}$$

In Step 3 below, we provide an argument for the limit $\beta_n \to \beta$.

In Steps 1 and 2, we consider $\omega \in \Omega$ such that (3.24) and the conclusion of Proposition 3.3 are valid. Moreover, we drop the dependence on ω and use the shorthand notation $z_{\mathrm{frac}}^i := z_{\mathrm{frac}}(\tau_i\omega)$ and $J_i(z) := J(\tau_i\omega, z)$.

Step 1. We prove $\bar{A} := \limsup_{n\to\infty} \gamma_n^* \leq \sqrt{\frac{\beta}{\underline{\alpha}}}$ by contradiction: assume that there exists $\varepsilon > 0$ and $\bar{N} \in \mathbb{N}$ such that

$$\ell_n^* > 1 + n^{-\frac{1}{2}} \sqrt{\frac{\beta}{\underline{\alpha}}}(1 + \varepsilon) =: k_n \quad \text{for } n > \bar{N}. \tag{3.26}$$

In view of Lemma 3.1, there exists $(\bar{z}_n)_n$ satisfying

$$\frac{1}{n}\sum_{i=0}^{n-1} \bar{z}_n^i = k_n, \quad \frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) = M_n(k_n), \quad \bar{z}_n^i \leq z_{\mathrm{frac}}^i \quad \forall i \in \{0, \dots, n-1\}. \tag{3.27}$$

We show that

$$\limsup_{n\to\infty} n\left(M_n(k_n) - \frac{1}{n}\sum_{i=0}^{n-1} J_i(1)\right) \leq \beta, \tag{3.28}$$

$$\liminf_{n\to\infty} n\left(\frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) - \frac{1}{n}\sum_{i=0}^{n-1} J_i(1)\right) \geq \beta(1 + \varepsilon)^2. \tag{3.29}$$

Clearly, (3.28) and (3.29) contradict (3.27) for n sufficiently large.

Substep 1.1. Argument for (3.29).

We claim that there exists $K < \infty$ such that for all n sufficiently large

$$n\left(\frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) - \frac{1}{n}\sum_{i=0}^{n-1} J_i(1)\right) \geq \left(\frac{1}{n}\sum_{i=0}^{n-1}\left(\frac{1}{2}J_i''(1)\right)^{-1}\right)^{-1}\frac{\beta}{\underline{\alpha}}(1+\varepsilon)^2 - \frac{K}{\sqrt{n}}, \tag{3.30}$$

where $\overline{\alpha}$ and β are defined in (2.3). Note that (3.24) and (3.30) imply (3.29).

We prove (3.30). By (3.26), (3.27), and Proposition 3.3 (applied with $D = \sqrt{\frac{\beta}{\underline{\alpha}}(1+\varepsilon)^2}$), we get

$$1 \leq z_n^i \leq 1 + n^{-\frac{1}{2}}C \tag{3.31}$$

for some $C < \infty$ independent of n. Hence, a Taylor expansion yields

$$\sum_{i=0}^{n-1} J_i(\bar{z}_n^i) = \sum_{i=0}^{n-1} J_i(1) + \frac{1}{2}\sum_{i=0}^{n-1} J_i''(1)(\bar{z}_n^i - 1)^2 + \frac{1}{6}\sum_{i=0}^{n-1} J_i'''(\xi_n^i)(\bar{z}_n^i - 1)^3, \tag{3.32}$$

where $\xi_n^i \in [1, \bar{z}_n^i]$. To estimate the second term on the right-hand side, note that Cauchy–Schwarz' inequality yields

$$\left(\sum_{i=0}^{n-1}(\bar{z}_n^i - 1)\right)^2 \leq \left(\frac{1}{2}\sum_{i=0}^{n-1} J_i''(1)(\bar{z}_n^i - 1)^2\right)\left(\sum_{i=0}^{n-1}\left(\frac{1}{2}J_i''(1)\right)^{-1}\right).$$

Combined with the identity $\sum_{i=0}^{n-1}(\bar{z}_n^i - 1) = n(k_n - 1) = \sqrt{n}\sqrt{\frac{\beta}{\underline{\alpha}}}(1+\varepsilon)$, we get

$$\left(\frac{1}{n}\sum_{i=0}^{n-1}\left(\frac{1}{2}J_i''(1)\right)^{-1}\right)^{-1}\frac{\beta}{\underline{\alpha}}(1+\varepsilon)^2 \leq \frac{1}{2}\sum_{i=0}^{n-1} J_i''(1)(\bar{z}_n^i - 1)^2. \tag{3.33}$$

Moreover, (3.31) and (A4) imply for n sufficiently large that

$$\frac{1}{6}\sum_{i=0}^{n-1} J_i'''(\xi_n^i)(\bar{z}_n^i - 1)^3 \geq -\frac{C^3}{6c\sqrt{n}}. \tag{3.34}$$

Clearly, (3.32)–(3.34) imply (3.30) (with $K = \frac{C^3}{6c}$).

Substep 1.2. Argument for (3.28).

For every $n \in \mathbb{N}$, we choose $\hat{i}_n \in \{0, \dots, n-1\}$ such that $-J_{\hat{i}_n}(1) = \beta_n$ (see (3.25)) and define $z_n \in \mathbb{R}^n$ as

$$z_n^i = \begin{cases} 1 & \text{if } i \in \{0, \dots, n-1\} \setminus \{\hat{i}_n\}, \\ 1 + n(k_n - 1) & \text{if } i = \hat{i}_n. \end{cases}$$

Since $\frac{1}{n}\sum_{i=0}^{n-1} z_n^i = k_n = 1 + n^{-\frac{1}{2}}\sqrt{\frac{\beta}{\underline{\alpha}}}(1+\varepsilon)$, we have

$$n\left(M_n(k_n) - \frac{1}{n}\sum_{i=0}^{n-1} J_i(1)\right) \leq J_{\hat{\imath}_n}\big(1 + n(k_n - 1)\big) - J_{\hat{\imath}_n}(1)$$
$$\leq \psi^+\left(1 + \sqrt{n}\sqrt{\frac{\beta}{\underline{\alpha}}}(1+\varepsilon)\right) + \beta_n,$$

where the second inequality holds by (A2) and the choice of $\hat{\imath}_n$. Now, (3.28) follows from (3.24) and assumption (A2).

Step 2. Proof of $\underline{A} := \liminf_{n\to\infty} \gamma_n^* \geq \sqrt{\frac{\beta}{\underline{\alpha}}}$.

We show that, for every $\varepsilon > 0$, there exists $\bar{N} \in \mathbb{N}$ such that

$$\ell_n^* \geq 1 + n^{-\frac{1}{2}}\sqrt{\frac{\beta}{\underline{\alpha}}}(1-\varepsilon) =: k_n \quad \text{for } n > \bar{N}. \tag{3.35}$$

Note that (3.35) implies that $\liminf_{n\to\infty}\gamma_n^* \geq \sqrt{\frac{\beta}{\underline{\alpha}}}(1-\varepsilon)$ for all $\varepsilon > 0$, and thus the claim.

Let $(\bar{z}_n)_n$ be a sequence satisfying for all $n \in \mathbb{N}$,

$$\frac{1}{n}\sum_{i=0}^{n-1}\bar{z}_n^i = k_n, \quad M_n(k_n) = \frac{1}{n}\sum_{i=0}^{n-1} J_i(\bar{z}_n^i). \tag{3.36}$$

To prove (3.35), we only need to show that

$$z_n^i \leq z_{\text{frac}}^i \quad \text{for all } i \in \{0,\dots,n-1\} \text{ for } n \text{ sufficiently large}, \tag{3.37}$$

depending only on $\underline{\alpha}$ β, c, and $\varepsilon > 0$.

Substep 2.1. We show that

$$\limsup_{n\to\infty} n\left(M_n(k_n) - \frac{1}{n}\sum_{i=0}^{n-1} J_i(1)\right) \leq \beta(1-\varepsilon). \tag{3.38}$$

Set

$$\hat{z}_n^i := 1 + n^{-\frac{1}{2}}\sqrt{\frac{\beta}{\underline{\alpha}}}(1-\varepsilon)\left(\frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{\alpha_i}\right)^{-1}\frac{1}{\alpha_i},$$

where $\alpha_i := \frac{1}{2}J_i''(1)$. By construction, we have

$$\frac{1}{n}\sum_{i=0}^{n-1}\hat{z}_n^i = k_n, \quad 0 \leq \hat{z}_n^i - 1 \leq n^{-\frac{1}{2}}C, \tag{3.39}$$

where $C < \infty$ depends only on $\underline{\alpha}$, β, and $c > 0$ from (A4) (note that (A4) implies that $\alpha_i \leq \frac{1}{2c}$ and $\frac{1}{\alpha_i} \leq \frac{2}{c}$). Hence, a Taylor expansion of J_i at 1 and (A4) yield for n

sufficiently large

$$\begin{aligned}\sum_{i=0}^{n-1}\big(J_i(\hat{z}_n^i)-J_i(1)\big) &\le \sum_{i=0}^{n-1}\alpha_i(\hat{z}_n^i-1)^2+\frac{1}{6c}\sum_{i=0}^{n-1}(\hat{z}_n^i-1)^3\\ &\le \frac{\beta}{\underline{\alpha}}(1-\varepsilon)^2\bigg(\frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{\alpha_i}\bigg)^{-1}+\frac{C^3}{6c}n^{-\frac{1}{2}},\end{aligned}$$

where $C<\infty$ is the same as in (3.39). Finally, (3.24) implies that $(\frac{1}{n}\sum_{i=0}^{n-1}\frac{1}{\alpha_i})^{-1}\le \underline{\alpha}(1+\varepsilon)$ for n sufficiently large and thus (3.38) follows.

Substep 2.2. We now prove (3.37) by contraposition. Suppose that $\bar{z}_n^{\hat{i}} > z_{\text{frac}}^{\hat{i}}$ for some $\hat{i}\in\{0,\dots,n-1\}$. Then, Proposition 3.3 yields $\bar{z}_n^{\hat{i}}\ge N_n$ for some (N_n) with $N_n\to\infty$, and thus $J_{\hat{i}}(\bar{z}_n^{\hat{i}})\ge -\sup_{s\ge N_n}\psi^+(s)$ by (A2). Hence, with $J_i(\bar{z}_n^i)\ge J_i(1)$ and $-J_{\hat{i}}(1)\ge\beta$, we therefore get

$$\sum_{i=0}^{n-1}\big(J_i(\bar{z}_n^i)-J_i(1)\big)\ge J_{\hat{i}}(\bar{z}_n^{\hat{i}})-J_{\hat{i}}(1)\ge\beta-\sup_{s\ge N_n}\psi^+(s).$$

Since $\sup_{s\ge N_n}\psi^+(s)\to 0$ for $n\to\infty$, the above lower bound combined with the upper bound (3.38) and (3.36) yields a contradiction for n sufficiently large, and thus (3.37) follows.

Step 3. Argument for $\beta_n\to\beta$ almost surely in (3.24).

The sequence $(\beta_n(\omega))_n\subset\mathbb{R}$ is decreasing and it holds that $\beta_n(\omega)\ge\beta$ for all $n\in\mathbb{N}$. Hence, there exists $\widehat{\beta}(\omega)\ge\beta$ such that

$$\lim_{n\to\infty}\beta_n(\omega)=\widehat{\beta}(\omega)\ge\beta.$$

It remains to show that $\widehat{\beta}(\omega)=\beta$ for $\mathbb{P}$-a.e. $\omega\in\Omega$. We argue by contradiction and therefore suppose that there exist $\varepsilon>0$ and a set $\Omega'\subset\Omega$ with positive measure such that $\widehat{\beta}(\omega)\ge\beta+\varepsilon$ for all $\omega\in\Omega'$. Then we obtain for all $\omega\in\Omega'$ that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\chi_{\{-J(\tau_i\omega,1)\le\beta+\frac{1}{2}\varepsilon\}}(\tau_i\omega)=0,$$

where χ_A denotes the indicator function. Clearly, this contradicts the ergodic theorem and the definition of β in the form

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n-1}\chi_{\{-J(\tau_i\omega,1)\le\beta+\frac{1}{2}\varepsilon\}}=\mathbb{E}[\chi_{\{-J(1)\le\beta+\frac{1}{2}\varepsilon\}}]>0\quad\text{for }\mathbb{P}\text{-a.e. }\omega\in\Omega.$$

Hence the theorem is proven. ■

Acknowledgements. During the work on this project, LL was affiliated most of the time with the Institute of Mathematics at the University of Würzburg, Germany. LL gratefully acknowledges the kind hospitality of the Technische Universität Dresden during her research visits.

Funding. Research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within project 405009441 and TU Dresden's Institutional Strategy "The Synergetic University".

References

[1] R. Alicandro, M. Cicalese, and A. Gloria, Integral representation results for energies defined on stochastic lattices and application to nonlinear elasticity. *Arch. Ration. Mech. Anal.* **200** (2011), no. 3, 881–943 Zbl 1294.74056 MR 2796134

[2] A. Braides, G. Dal Maso, and A. Garroni, Variational formulation of softening phenomena in fracture mechanics: the one-dimensional case. *Arch. Ration. Mech. Anal.* **146** (1999), no. 1, 23–58 Zbl 0945.74006 MR 1682660

[3] A. Braides and M. S. Gelli, Continuum limits of discrete systems without convexity hypotheses. *Math. Mech. Solids* **7** (2002), no. 1, 41–66 Zbl 1024.74004 MR 1900933

[4] A. Braides, A. J. Lew, and M. Ortiz, Effective cohesive behavior of layers of interatomic planes. *Arch. Ration. Mech. Anal.* **180** (2006), no. 2, 151–182 Zbl 1093.74013 MR 2210908

[5] A. Braides and M. Solci, Asymptotic analysis of Lennard-Jones systems beyond the nearest-neighbour setting: a one-dimensional prototypical case. *Math. Mech. Solids* **21** (2016), no. 8, 915–930 Zbl 1370.74013 MR 3538510

[6] A. Chambolle, Finite-differences discretizations of the Mumford–Shah functional. *M2AN Math. Model. Numer. Anal.* **33** (1999), no. 2, 261–288 Zbl 0947.65076 MR 1700035

[7] G. Dal Maso and L. Modica, Nonlinear stochastic homogenization and ergodic theory. *J. Reine Angew. Math.* **368** (1986), 28–42 Zbl 0582.60034 MR 850613

[8] M. Friedrich and U. Stefanelli, Crystallization in a one-dimensional periodic landscape. *J. Stat. Phys.* **179** (2020), no. 2, 485–501 Zbl 1437.82024 MR 4091566

[9] C. L. Hall, T. Hudson, and P. van Meurs, Asymptotic analysis of boundary layers in a repulsive particle system. *Acta Appl. Math.* **153** (2018), 1–54 Zbl 1387.74093 MR 3745729

[10] O. Iosifescu, C. Licht, and G. Michaille, Variational limit of a one dimensional discrete and statistically homogeneous system of material points. *Asymptot. Anal.* **28** (2001), no. 3-4, 309–329 Zbl 1031.74041 MR 1878798

[11] S. Jansen, W. König, B. Schmidt, and F. Theil, Surface energy and boundary layers for a chain of atoms at low temperature. *Arch. Ration. Mech. Anal.* **239** (2021), no. 2, 915–980 Zbl 1456.82666 MR 4201619

[12] M. Kimura and P. van Meurs, Quantitative estimate of the continuum approximations of interacting particle systems in one dimension. *SIAM J. Math. Anal.* **53** (2021), no. 1, 681–709 Zbl 1459.82181 MR 4205668

[13] L. Lauerbach, *Stochastic homogenization in the passage from discrete to continuous systems—fracture in composite materials*. Ph.D. thesis, University of Würzburg, 2020

[14] L. Lauerbach, S. Neukamm, M. Schäffner, and A. Schlömerkemper, Mechanical behaviour of heterogeneous nanochains in the Γ-limit of stochastic particle systems. 2019, arXiv:1909.06607

[15] L. Lauerbach, M. Schäffner, and A. Schlömerkemper, On continuum limits of heterogeneous discrete systems modelling cracks in composite materials. *GAMM-Mitt.* **40** (2018), no. 3, 184–206 MR 3796762

[16] L. Lauerbach and A. Schlömerkemper, Derivation of a variational model for brittle fracture from a random heterogeneous particle chain. 2021, arXiv:2104.08607

[17] S. Neukamm, M. Schäffner, and A. Schlömerkemper, Stochastic homogenization of nonconvex discrete energies with degenerate growth. *SIAM J. Math. Anal.* **49** (2017), no. 3, 1761–1809 Zbl 1362.74028 MR 3650427

[18] M. Ruf, Discrete stochastic approximations of the Mumford–Shah functional. *Ann. Inst. H. Poincaré C Anal. Non Linéaire* **36** (2019), 887–937 Zbl 1417.49041 MR 3955107

[19] L. Scardia, A. Schlömerkemper, and C. Zanini, Boundary layer energies for nonconvex discrete systems. *Math. Models Methods Appl. Sci.* **21** (2011), no. 4, 777–817 Zbl 1268.49016 MR 2795506

[20] M. Schäffner and A. Schlömerkemper, On Lennard-Jones systems with finite range interactions and their asymptotic analysis. *Netw. Heterog. Media* **13** (2018), no. 1, 95–118 Zbl 1407.49015 MR 3811556

[21] L. Truskinovsky, Fracture as a phase transition. In *Contemporary Research in the Mechanics and Mathematics of Materials*, pp. 322–332, CIMNE, Barcelona, 1996

Laura Lauerbach
Institute of Mathematics, University of Kassel, Heinrich-Plett-Straße 40, 34132 Kassel, Germany; lauerbach@mathematik.uni-kassel.de

Stefan Neukamm
Faculty of Mathematics, Technische Universität Dresden, 01069 Dresden, Germany; stefan.neukamm@tu-dresden.de

Mathias Schäffner
Institute of Mathematics, Martin Luther University Halle-Wittenberg, 06099 Halle (Saale), Germany; mathias.schaeffner@tu-dortmund.de

Anja Schlömerkemper
Institute of Mathematics, University of Würzburg, Emil-Fischer-Straße 40, 97074 Würzburg, Germany; anja.schloemerkemper@mathematik.uni-wuerzburg.de

DOI 10.4171/8ECM/29

Siu's lemma: Generalizations and applications

Xiangyu Zhou and Langfeng Zhu

Abstract. In this survey paper, we present some generalizations of Siu's lemma related to multiplier ideal sheaves and discuss their applications in some problems related to optimal L^2 extension, comparison between singular metrics on exceptional fibers of twisted relative pluricanonical bundles, and subadditivity of Kodaira–Iitaka dimensions with multiplier ideal sheaves. We also discuss some ideas in the proofs.

1. Introduction

L^2 extensions with precise estimates and multiplier ideal sheaves related to plurisubharmonic (psh) functions and their singularities have been playing a useful role in recent progress in several complex variables and complex geometry. Siu's lemma deals with multiplier ideal sheaves which is an important invariant of the singularities of the psh functions and quite closely related to L^2 extensions. In the present paper, we will outline recent progress on the generalizations and applications of Siu's lemma. Before we present our main results, let us first recall some notions and notations (see [7–9, 18, 20, 25]), which will be used in this paper.

Let X be a complex manifold. A function $\varphi : X \to [-\infty, +\infty)$ is said to be *quasi-plurisubharmonic* (quasi-psh) if φ is locally the sum of a psh function and a smooth function.

A *singular (Hermitian) metric* h of a holomorphic line bundle L over X is simply a Hermitian metric which is expressed locally as $e^{-\varphi}$ with respect to local holomorphic trivialization of L such that $\varphi \in L^1_{\mathrm{loc}}$. The curvature current $\sqrt{-1}\Theta_{L,h} := \sqrt{-1}\partial\overline{\partial}\varphi$ is well defined on X. A holomorphic line bundle L is called pseudoeffective if it is endowed with a singular Hermitian metric h with positive or semipositive curvature current (i.e., φ is psh in the sense of distribution). In particular, L is called a positive line bundle if φ is smooth strictly psh; L is called a big line bundle if the

2020 Mathematics Subject Classification. Primary 32D15; Secondary 32L10, 32W05, 14C30, 14F18, 32U05.
Keywords. Optimal L^2 extension, multiplier ideal sheaf, singularities of plurisubharmonic functions, strong openness, Kodaira–Iitaka dimension, twisted relative pluricanonical bundles.

curvature current is a Kähler current, i.e., $\Theta \geq \varepsilon\omega$ for some $\varepsilon > 0$, where ω is the $(1,1)$ form associated to a Kähler metric.

A quasi-psh function φ on X is said to have *(neat) analytic singularities* if every point $x \in X$ possesses an open neighborhood U on which φ can be written as

$$\varphi = c \log \sum_{1 \leq j \leq j_0} |g_j|^2 + u,$$

where c is a nonnegative number, $g_j \in \mathcal{O}_X(U)$, and u is bounded on U ($u \in C^\infty(U)$).

1.1. Multiplier ideal sheaf

For any quasi-psh function φ on X, the L^p *multiplier ideal sheaf* ($0 < p < +\infty$) is defined by

$$\mathcal{I}_{L^p}(\varphi)_x = \left\{ f \in \mathcal{O}_{X,x};\ \exists U \ni x \text{ such that } \int_U |f|^p e^{-\varphi}\, d\lambda < +\infty \right\},$$

where $U \subset X$ is a coordinate chart, and $d\lambda$ is the Lebesgue measure. $\mathcal{I}_{L^2}(\varphi)$ is just the usual multiplier ideal sheaf $\mathcal{I}(\varphi)$.

We will write the $L^{\frac{2}{m}}$ multiplier ideal sheaf $\mathcal{I}_{L^{\frac{2}{m}}}(\varphi)$ (m is a positive integer) by $\mathcal{I}_m(\varphi)$ for simplicity.

We list some basic properties of the multiplier ideal sheaves as follows.

(1) Nadel's theorem: $\mathcal{I}(\varphi)$ is coherent.

Consequently, the nonlocally integrable point set of $e^{-\varphi}$ (= the zero set of $\mathcal{I}(\varphi) = \operatorname{supp} \mathcal{O}/\mathcal{I}(\varphi)$) is an analytic set, since the support of a coherent analytic sheaf is an analytic set.

(2) Theorem: A multiplier ideal sheaf is integrally closed, i.e., the integral closure of $\mathcal{I}(\varphi)$ is itself.

(3) Nadel's vanishing theorem: Let $(L, e^{-\varphi})$ be a big line bundle on a compact Kähler manifold X. Then

$$H^q\big(X, K_X \otimes L \otimes \mathcal{I}(\varphi)\big) = 0,$$

for any $q \geq 1$.

Recently, a new property of the multiplier ideal sheaves, i.e., the strong openness of the multiplier ideal sheaves, is established by the solution of Demailly's strong openness conjecture [13]. The solution and its applications are based on the above basic properties of the multiplier ideal sheaves.

Demailly's strong openness conjecture. For any psh function φ on X, one has

$$\mathcal{I}(\varphi) = \mathcal{I}_+(\varphi) := \bigcup_{\varepsilon > 0} \mathcal{I}\big((1+\varepsilon)\varphi\big) = \mathcal{I}\big((1+\varepsilon_0)\varphi\big).$$

The last equality is well known by the Noetherian property of the coherent analytic sheaves. This conjecture was also stated by Y. T. Siu [25], Demailly–Kollár [10] and many others. For a reformulation of the conjecture, see Theorem 3.2.

It should be noted that $(1+\varepsilon)\varphi$ in the conjecture could be replaced by any increasing sequence of psh functions which converges to φ [13]; more generally, stability of the multiplier ideal sheaves holds by Guan–Li–Zhou [15] based on [12]. Strong openness also holds for L^p multiplier ideal sheaves for $0<p<\infty$; it follows from the strong openness for L^2 multiplier ideal sheaves and Hölder's inequality; see Fornaess [11]. By the strong openness, it follows that L^p multiplier ideal sheaves are coherent [6, 25].

1.2. Optimal L^2 extension

In [21], Ohsawa and Takegoshi obtained the following L^2 extension theorem with psh weights.

Theorem 1.1 ([21]). *Let $\Omega \subset \mathbb{C}^n$ be a bounded pseudoconvex domain, φ a psh function on Ω, and s a holomorphic function on Ω. Let*

$$H := \{x \in \Omega;\ s(x)=0\}.$$

Assume that $|s| \leq 1$ on Ω and $ds \not\equiv 0$ on H. Then there exists an absolute constant C such that, for every holomorphic function f on H satisfying

$$\int_H \frac{|f|^2 e^{-\varphi}}{|ds|^2}\, dV_H < +\infty,$$

there exists a holomorphic function F on Ω satisfying $F=f$ on H and

$$\int_\Omega |F|^2 e^{-\varphi}\, d\lambda_n \leq C \int_H \frac{|f|^2 e^{-\varphi}}{|ds|^2}\, dV_H,$$

where $d\lambda_n$ is the $2n$-dimensional Lebesgue measure, and dV_H is the $2(n-1)$-dimensional Hausdorff measure on H.

Unifying various L^2 extension theorems, a general L^2 extension theorem with precise estimate for the almost Stein case and its geometric meaning was established and discovered in [14]. Later on, the optimal L^2 extension theorem with singular metrics for the Kähler case was obtained in [6, 30] by using Guan–Zhou's work on optimal L^2 extension and strong openness of multiplier ideal sheaves, and the generalized Siu's lemma stated below plays also a key role in [30].

1.3. Siu's lemma

In the study of algebraic geometry problems such as Fujita's conjecture [1, 24], Siu obtained the semi-continuity of multiplier ideal sheaves and the following lower limit property about integrals with psh weights having trivial multiplier ideal sheaves by using Theorem 1.1.

Theorem 1.2 (Siu's lemma; see [22]). *Let $\varphi(z', z'')$ be a nonpositive psh function on $\mathbb{B}^1_r \times \mathbb{B}^{n-1}_r$ such that*

$$\int_{z''\in\mathbb{B}^{n-1}_r} e^{-\varphi(0,z'')}\, d\lambda_{n-1} < +\infty, \tag{1.1}$$

where $\mathbb{B}^{n-1}_r$ denotes the open ball in $\mathbb{C}^{n-1}$ centered at 0 with radius $r > 0$, and $d\lambda_{n-1}$ denotes the $2(n-1)$-dimensional Lebesgue measure. Assume that $r_1 \in (0, r)$. Then there exists a positive number C independent of φ, such that

$$\varliminf_{z'\to 0} \int_{z''\in\mathbb{B}^{n-1}_{r_1}} e^{-\varphi(z',z'')}\, d\lambda_{n-1} \le C \int_{z''\in\mathbb{B}^{n-1}_r} e^{-\varphi(0,z'')}\, d\lambda_{n-1}. \tag{1.2}$$

The inequality (1.1) means that φ restricted to the center fiber has a trivial multiplier ideal sheaf.

Siu's lemma was also used by Phong and Sturm [22] to obtain a holomorphic stability result for 1-parameter deformations; i.e., for any nonpositive psh function $\varphi(z', z'')$ which has neat analytic singularities and satisfies (1.1), the stronger equality

$$\lim_{z'\to 0} \int_{z''\in\mathbb{B}^{n-1}_{r_1}} e^{-\varphi(z',z'')}\, d\lambda_{n-1} = \int_{z''\in\mathbb{B}^{n-1}_{r_1}} e^{-\varphi(0,z'')}\, d\lambda_{n-1} \tag{1.3}$$

holds.

In general, one could not expect that (1.3) holds for a general nonpositive psh function φ which satisfies (1.1) but does not have neat analytic singularities (one can see [29] for a simple counterexample).

Theorem 1.3 (Lemma on the semi-continuity of multiplier ideal sheaves; [1, 24]). *The limit of the zero-sets of the multiplier ideal sheaves defined by a holomorphic family of multivalued holomorphic sections contains the zero-set of the limit.*

For a concrete explanation of the above result, the reader is referred to [1, Lemma 6.1] and [24, Section 3]. The above two lemmas follow from the L^2 extension theorem.

An equivalent version of the above semi-continuity is as follows. Let $\varphi(z', z'')$ be a psh function on $\Delta^n \times \Delta^m$. If $e^{-\varphi(z',z'')}$ is not integrable at $z' = 0$ for almost all $z'' \in \Delta^m \setminus 0$, then $e^{-\varphi(z',0)}$ is not integrable at $z' = 0$.

The generalized Siu lemma stated below implies both Siu's lemma and Siu's semi-continuity of multiplier ideal sheaves which seem to not imply each other, and could be regarded as a unified version of both properties.

1.4. Main content of the present paper

In [29], we generalized Siu's lemma by proving a limit property, which implies Siu's lemma with an optimal estimate. In [32], we further generalized Siu's lemma to the case that the multiplier ideal sheaf of φ is not necessarily trivial when restricted to the center fiber.

Moreover, we used in [32] the generalization of Siu's lemma with nontrivial multiplier ideal sheaves to prove a refined optimal L^2 extension theorem with singular metrics in the Kähler case. As another application, we gave in [32] a positive answer to a comparison question posed by Berndtsson–Păun [5] and Păun–Takayama [23] about singular metrics on exceptional fibers of twisted relative pluricanonical bundles.

By using optimal L^2 extension theorem with singular metrics in the Kähler case, one can prove the positivity or pseudoeffectivity of twisted relative pluricanonical bundles with singular metrics in the Kähler case (see [6, 30, 32] for the Kähler case, and see also [3, 5, 23] for the projective case). This positivity can be used to study the subadditivity of Kodaira–Iitaka dimensions for Kähler fibrations (see [27, 33]).

We also proved in [33] a more general version of Siu's lemma with nontrivial multiplier ideal sheaves near a subvariety, which generalizes the submanifold case obtained in [32]. This result gives a relation between two measures used in previous L^2 extension theorems in [9, 14, 19, 31, 34].

In the rest sections, we will discuss the above results explicitly.

2. A generalization of Siu's lemma with trivial multiplier ideal sheaves

Under similar assumptions as in Siu's lemma (Theorem 1.2), we proved the following limit property, which is a generalization of Siu's lemma.

Theorem 2.1 ([29]). *Let $\varphi(z',z'')$ be a psh function on $\mathbb{B}_r^k \times \mathbb{B}_r^{n-k}$ $(1 \le k \le n)$ such that*

$$\int_{z''\in\mathbb{B}_r^{n-k}} e^{-\varphi(0,z'')}\,d\lambda_{n-k} < +\infty.$$

Let P be a nonnegative continuous function on $\mathbb{B}_r^k \times \mathbb{B}_r^{n-k}$. Assume that $r_1 \in (0,r)$. Then

$$\lim_{\varepsilon\to 0}\frac{1}{\lambda_k(\mathbb{B}_\varepsilon^k)}\int_{\mathbb{B}_\varepsilon^k\times\mathbb{B}_{r_1}^{n-k}} P(z',z'')e^{-\varphi(z',z'')}\,d\lambda_n = \int_{z''\in\mathbb{B}_{r_1}^{n-k}} P(0,z'')e^{-\varphi(0,z'')}\,d\lambda_{n-k}, \tag{2.1}$$

where $\lambda_k(\mathbb{B}_\varepsilon^k) :=$ the $2k$-dimensional Lebesgue measure of $\mathbb{B}_\varepsilon^k$.

It is easy to see that (2.1) implies that (1.2) holds with $C = 1$.

The following two results are used in the proof of Theorem 2.1.

Lemma 2.2 ([22, 29]). *Let $\varphi(z', z'')$ be a negative psh function on $\mathbb{B}^k_\delta \times \mathbb{B}^{n-k}_\delta$ $(1 \le k \le n, \delta > 0)$ such that*

$$I_\varphi := \int_{z'' \in \mathbb{B}^{n-k}_\delta} e^{-\varphi(0,z'')}\, d\lambda_{n-k} < +\infty \quad (I_\varphi := e^{-\varphi(0)} \text{ if } k = n).$$

Assume that $r_1 \in (0, \delta)$. Then there exist two positive numbers C and $\varepsilon_\varphi \in (0, r_1]$ (C is independent of φ), such that

$$\frac{1}{\varepsilon^{2k}} \int_{\mathbb{B}^k_\varepsilon \times \mathbb{B}^{n-k}_{r_1}} e^{-\varphi(z',z'')}\, d\lambda_n \le C I_\varphi^{k+1}$$

for all $\varepsilon \in (0, \varepsilon_\varphi]$ (z'' and $\mathbb{B}^{n-k}_{r_1}$ will disappear if $k = n$).

Theorem 2.3 ([2]). *Let φ be a psh function on the ball $\mathbb{B}^n_r$ of $\mathbb{C}^n$ centered at 0 with radius r. Assume that*

$$\int_{\mathbb{B}^n_r} e^{-\varphi}\, d\lambda_n < +\infty.$$

Let $\delta \in (0, r)$. Then there exists $\beta > 0$ such that

$$\int_{\mathbb{B}^n_\delta} e^{-(1+\beta)\varphi}\, d\lambda_n < +\infty.$$

The idea in our proof of Theorem 2.1 consists of two steps.

The first step is to control the integral near the set $\{\varphi = -\infty\}$, which can be completed by using the openness property of multiplier ideal sheaves (Theorem 2.3) and a variation of Siu's lemma (Lemma 2.2).

The second step is to prove that Theorem 2.1 holds for φ which is bounded, which can be completed by mainly using Lebesgue's dominated convergence theorem.

To be more precise, let r, r_1 be as in Theorem 2.1 and denote $\frac{r+r_1}{2}$ by δ. Then $r_1 < \delta < r$ and we obtain from Theorem 2.3 that

$$\int_{z'' \in \mathbb{B}^{n-k}_\delta} e^{-(1+\beta)\varphi(0,z'')}\, d\lambda_{n-k} < +\infty$$

for some positive number β.

Then applying Lemma 2.2 to the psh function $(1 + \beta)\varphi$, we have

$$\frac{1}{\varepsilon^{2k}} \int_{\mathbb{B}^k_\varepsilon \times \mathbb{B}^{n-k}_{r_1}} e^{-(1+\beta)\varphi(z',z'')}\, d\lambda_n \le C \tag{2.2}$$

for all small enough ε, where C is a positive constant independent of ε.

Let v be a positive integer. Then (2.2) implies that

$$\frac{1}{\lambda(\mathbb{B}^k_\varepsilon)}\int_{\{\varphi\le -v\}\cap(\mathbb{B}^k_\varepsilon\times\mathbb{B}^{n-k}_{r_1})} P(z',z'')e^{-\varphi(z',z'')}\,d\lambda_n \le C_1 e^{-\beta v}$$

for all small enough ε, where C_1 is a positive constant independent of ε.

Hence the integral near $\{\varphi=-\infty\}$ is uniformly small if v is sufficiently large, and we complete the first step.

Set $\varphi_v=\max\{\varphi,-v\}$. The second step is to prove

$$\begin{aligned}\lim_{\varepsilon\to 0}\frac{1}{\lambda(\mathbb{B}^k_\varepsilon)}\int_{\mathbb{B}^k_\varepsilon\times\mathbb{B}^{n-k}_{r_1}} & P(z',z'')e^{-\varphi_v(z',z'')}\,d\lambda_n \\ &= \int_{z''\in\mathbb{B}^{n-k}_{r_1}} P(0,z'')e^{-\varphi_v(0,z'')}\,d\lambda_{n-k},\end{aligned}$$

which can be obtained by mainly using Lebesgue's dominated convergence theorem (see [29] for the details).

3. A generalization of Siu's lemma with nontrivial multiplier ideal sheaves

In both Theorem 1.2 and Theorem 2.1, the multiplier ideal sheaf of φ is trivial when restricted to the center fiber. It is natural to consider the case when the multiplier ideal sheaf is nontrivial.

In [32], we obtained the following generalization of Siu's lemma for psh functions having nontrivial multiplier ideal sheaves when restricted to the center fiber.

Theorem 3.1 ([32,35]). *Let $p\in(0,2]$. Let $\varphi(z',z'')$ be a psh function on $\mathbb{B}^k_r\times\mathbb{B}^{n-k}_r$ $(1\le k\le n)$, let $P(z',z'')$ be a nonnegative continuous function on $\mathbb{B}^k_r\times\mathbb{B}^{n-k}_r$, let $M(z')$ be a bounded nonnegative measurable function on $\mathbb{C}^k$ with compact support, and let $f(z'')$ be a holomorphic function on $\mathbb{B}^{n-k}_r$ satisfying*

$$\int_{z''\in\mathbb{B}^{n-k}_r}\left|f(z'')\right|^p e^{-\varphi(0,z'')}\,d\lambda_{n-k}<+\infty.$$

Assume that $r_1,r_2\in(0,r)$ and $r_1<r_2$. Let β be a positive number such that

$$I_\beta:=\int_{z''\in\mathbb{B}^{n-k}_{r_2}}\left|f(z'')\right|^p e^{-(1+\beta)\varphi(0,z'')}\,d\lambda_{n-k}<+\infty \tag{3.1}$$

and $\alpha\in(1-\frac{p}{2k}\beta,1)\cap[0,1)$. Then there exists a holomorphic function $F(z',z'')$ on

$\mathbb{B}^k_r \times \mathbb{B}^{n-k}_{r_2}$ such that $F(0, z'') = f(z'')$ on $\mathbb{B}^{n-k}_{r_2}$,

$$\int_{(z',z'')\in\mathbb{B}^k_r\times\mathbb{B}^{n-k}_{r_2}} \frac{\big|F(z',z'')\big|^p e^{-(1+\beta)\varphi(z',z'')}}{|z'|^{2k\alpha}}\, d\lambda_n < +\infty, \tag{3.2}$$

and

$$\lim_{\varepsilon\to 0+} \int_{(z',z'')\in\mathbb{C}^k\times\mathbb{B}^{n-k}_{r_1}} \frac{1}{\varepsilon^{2k}} M\left(\frac{z'}{\varepsilon}\right) P(z',z'')\big|F(z',z'')\big|^p e^{-\varphi(z',z'')}\, d\lambda_n$$
$$= \int_{z'\in\mathbb{C}^k} M(z')\, d\lambda_k \int_{z''\in\mathbb{B}^{n-k}_{r_1}} P(0,z'')\big|f(z'')\big|^p e^{-\varphi(0,z'')}\, d\lambda_{n-k}. \tag{3.3}$$

Moreover, any holomorphic extension F of f satisfying (3.2) has the property (3.3).

The existence of β in Theorem 3.1 is guaranteed by the strong openness property of multiplier ideal sheaves, i.e., Theorem 3.2 below.

Theorem 3.2 ([13]). *Let $p \in (0, +\infty)$. Let φ be a psh function on the unit ball $\mathbb{B}^n_1$ of $\mathbb{C}^n$. Assume that F is a holomorphic function on $\mathbb{B}^n_1$ satisfying*

$$\int_{\mathbb{B}^n_1} |F|^p e^{-\varphi}\, d\lambda_n < +\infty.$$

Then there exists $r \in (0, 1)$ and $\beta \in (0, +\infty)$ such that

$$\int_{\mathbb{B}^n_r} |F|^p e^{-(1+\beta)\varphi}\, d\lambda_n < +\infty.$$

In [32], we proved Theorem 3.1 by developing the method established in [30] and using the iteration method in [4] or [5].

The existence of a holomorphic extension F satisfying (3.2) can be obtained by using L^2 extension theorems. The main property that needs to be proved is (3.3).

The key step in our proof of (3.3) is to construct holomorphic functions F_ε on $\mathbb{B}^k_{r_2} \times \mathbb{B}^{n-k}_{r_2}$ such that $F_\varepsilon = f$ on $\{0\} \times \mathbb{B}^{n-k}_{r_2}$,

$$\int_{\mathbb{B}^k_\varepsilon\times\mathbb{B}^{n-k}_{r_2}} |F_\varepsilon|^p e^{-(1+\beta)\varphi}\, d\lambda_n \le C_1\varepsilon^{2k}, \tag{3.4}$$

and

$$\int_{\mathbb{B}^k_{r_2}\times\mathbb{B}^{n-k}_{r_2}} |F_\varepsilon|^p e^{-(1+\beta)\varphi}\, d\lambda_n \le C_2\varepsilon^{-2\beta_1} \tag{3.5}$$

for any $\varepsilon \in (0, r_2)$, where $\beta_1 \in (0, \frac{p}{2})$ is a small enough positive number.

Then by some calculation, we can get

$$\int_{\mathbb{B}_\varepsilon^k\times\mathbb{B}_{r_2}^{n-k}} \frac{|F-F_\varepsilon|^p e^{-(1+\beta_2)\varphi}}{\varepsilon^{2k}}\,d\lambda_n \le C_3 \tag{3.6}$$

for all ε small enough, where $\beta_2\in(0,\beta)$ is a small enough positive number, and C_3 is a positive number independent of ε.

Then we get the desired property (3.3) by using (3.4), (3.6), and the following proposition.

Proposition 3.3 ([30]). *Let $\varphi(z',z'')$ be a psh function on $\mathbb{B}_r^k\times\mathbb{B}_r^{n-k}$, $f(z'')$ a holomorphic function on $\mathbb{B}_r^{n-k}$, $P(z',z'')$ a nonnegative continuous function on $\mathbb{B}_r^k\times\mathbb{B}_r^{n-k}$, and $M(z')$ a bounded nonnegative measurable function on $\mathbb{C}^k$ with compact support. Let C, β_2, r_1, and r_2 be positive numbers. Assume that $r_1<r_2<r$. Suppose that F is a holomorphic function on $\mathbb{B}_{r_2}^k\times\mathbb{B}_{r_2}^{n-k}$ satisfying $F(0,z'')=f(z'')$ $(\forall z''\in\mathbb{B}_{r_2}^{n-k})$,*

$$\sup_{\mathbb{B}_{r_2}^k\times\mathbb{B}_{r_2}^{n-k}}|F|\le C,$$

and

$$\varlimsup_{\varepsilon\to0+}\frac{1}{\varepsilon^{2k}}\int_{(z',z'')\in\mathbb{B}_\varepsilon^k\times\mathbb{B}_{r_2}^{n-k}}\left|F(z',z'')\right|^p e^{-(1+\beta_2)\varphi(z',z'')}\,d\lambda_n\le C.$$

Then

$$\begin{aligned}\lim_{\varepsilon\to0+}&\int_{(z',z'')\in\mathbb{C}^k\times\mathbb{B}_{r_1}^{n-k}}\frac{1}{\varepsilon^{2k}}M\left(\frac{z'}{\varepsilon}\right)P(z',z'')\left|F(z',z'')\right|^p e^{-\varphi(z',z'')}\,d\lambda_n\\&=\int_{z'\in\mathbb{C}^k}M(z')\,d\lambda_k\int_{z''\in\mathbb{B}_{r_1}^{n-k}}P(0,z'')\left|f(z'')\right|^p e^{-\varphi(0,z'')}\,d\lambda_{n-k}.\end{aligned}$$

4. A refined optimal L^2 extension theorem with singular metrics on Kähler manifolds

Now we regard Theorem 3.1 as a local property on coordinate charts, and discuss a global version of it on complex manifolds.

Let $\mathfrak{R}$ be the class of functions defined by

$$\Big\{R\in C^\infty(-\infty,0];\ R>0,\ R\text{ is decreasing},\ C_R:=\int_{-\infty}^0\frac{1}{R(t)}\,dt<+\infty$$
$$\text{and } e^tR(t)\text{ is bounded above on }(-\infty,0]\Big\}.$$

By using Theorems 3.1 and 3.2, we obtained the following refined optimal L^2 extension theorem with singular metrics on Kähler manifolds as a global version of Theorem 3.1.

Theorem 4.1 ([32]). *Let $R \in \mathfrak{R}$ and let $\Pi : X \to S$ be a surjective proper holomorphic map from a Kähler manifold (X, ω) of dimension n to a Stein domain S contained in the unit ball $\mathbb{B}^k \subset \mathbb{C}^k$ $(1 \le k \le n)$, where ω is a Kähler metric on X. With respect to the coordinate functions on $\mathbb{C}^k$, we regard Π as a k-dimensional vector of holomorphic functions $s = (s_1, \dots, s_k)$ on X. Assume that $0 \in S$ and $ds := ds_1 \wedge \cdots \wedge ds_k$ is nonvanishing on*

$$X_0 := \{x \in X;\ s(x) = 0\}.$$

Let (L, h) be a holomorphic line bundle over X equipped with a singular Hermitian metric h such that the curvature current $\sqrt{-1}\Theta_{L,h} \ge 0$. Write h as $h_1 e^{-\phi}$, where h_1 is any fixed smooth metric of L and ϕ is a global quasi-psh function on X. Assume that $f \in H^0(X_0, K_X|_{X_0} + L|_{X_0})$ satisfies

$$\int_{X_0} \frac{|f|^2_{\omega,h}}{|ds|^2_\omega}\, dV_{X_0,\omega_0} < +\infty,$$

where $dV_{X_0,\omega_0} := \frac{\omega_0^{n-k}}{(n-k)!}$ and ω_0 is the Kähler metric on X_0 induced from ω. Let β be a positive number such that

$$\mathcal{I}\big((1+\beta)\phi + \log|s|^{2k}\big)_x = \mathcal{I}\big(\phi + \log|s|^{2k}\big)_x$$

and

$$f \in \mathcal{I}\big((1+\beta)\phi|_{X_0}\big)_x$$

hold for any $x \in X_0$ (the existence of β is guaranteed by Theorem 3.2). Then there exists $F \in H^0(X, K_X + L)$ such that $F = f$ on X_0,

$$F \in \mathcal{I}\big((1+\beta)\phi + \alpha \log|s|^{2k}\big)_x \quad \big(\forall \alpha \in [0,1),\ \forall x \in X_0\big), \tag{4.1}$$

$$\int_X \frac{|F|^2_{\omega,h}}{|s|^{2k} R\big(\log|s|^{2k}\big)}\, dV_{X,\omega} \le C_R \frac{(2\pi)^k}{k!} \int_{X_0} \frac{|f|^2_{\omega,h}}{|ds|^2_\omega}\, dV_{X_0,\omega_0}, \tag{4.2}$$

and

$$\lim_{\varepsilon\to 0+} \int_X \frac{1}{\varepsilon^{2k}} M\Big(\frac{s}{\varepsilon}\Big) P|F|^2_{\omega,h}\, dV_{X,\omega} = 2^k \int_{z'\in\mathbb{C}^k} M(z')\, d\lambda_k \int_{X_0} \frac{P|f|^2_{\omega,h}}{|ds|^2_\omega}\, dV_{X_0,\omega_0}, \tag{4.3}$$

where M is as in Theorem 3.1, and P is any nonnegative continuous function on X.

One can see [6, 9, 28, 30] for many L^2 extension theorems with singular metrics on Kähler manifolds without the properties (4.1) and (4.3).

The properties (4.1) and (4.3) refine the optimal L^2 extension theorem with singular metrics on Kähler manifolds discussed in [30]. In our proof of Theorem 4.1 (even without (4.1) and (4.3)), a key step is to use some construction essentially used in the proof of Theorem 3.1 in the case $p = 2$ (see (3.4), (3.5), and [30]).

Furthermore, applying the iteration method in [4] or [5], the following refined optimal $L^{\frac{2}{m}}$ extension theorem with singular metrics on Kähler manifolds can be obtained from Theorem 4.1.

Theorem 4.2 ([32]). *Let R, Π, (X,ω), S, (L,h), (X_0,ω_0), s, ds, h_1, ϕ, M, and P be the same as in Theorem 4.1. Let $h_\omega := (dV_{X,\omega})^{-1}$ (it defines a smooth Hermitian metric on K_X). Assume that $f \in H^0(X_0, mK_X|_{X_0} + L|_{X_0})$ (m is a positive integer) satisfies*

$$C_f := \int_{X_0} \frac{|f|^{\frac{2}{m}}_{h_\omega^{\otimes m}\otimes h}}{|ds|^2_\omega}\, dV_{X_0,\omega_0} < +\infty,$$

and assume that there exists a holomorphic extension F_1 of f to an open neighborhood of X_0 in X. Then there exists a positive number β and a holomorphic section $F \in H^0(X, mK_X + L)$ such that $F = f$ on X_0,

$$F \in \mathcal{I}_m\Big((1+\beta)\frac{\phi}{m} + \alpha \log|s|^{2k}\Big)_x \quad \big(\forall \alpha \in [0,1),\ \forall x \in X_0\big), \tag{4.4}$$

$$\int_X \frac{|F|^{\frac{2}{m}}_{h_\omega^{\otimes m}\otimes h}}{|s|^{2k} R\big(\log|s|^{2k}\big)}\, dV_{X,\omega} \le C_R \frac{(2\pi)^k}{k!} C_f, \tag{4.5}$$

and

$$\begin{aligned}\lim_{\varepsilon\to 0+} \int_X \frac{1}{\varepsilon^{2k}} M\Big(\frac{s}{\varepsilon}\Big) P |F|^{\frac{2}{m}}_{h_\omega^{\otimes m}\otimes h}\, dV_{X,\omega} \\ = 2^k \int_{z'\in\mathbb{C}^k} M(z')\, d\lambda_k \int_{X_0} \frac{P|f|^{\frac{2}{m}}_{h_\omega^{\otimes m}\otimes h}}{|ds|^2_\omega}\, dV_{X_0,\omega_0}.\end{aligned} \tag{4.6}$$

5. Comparison of singular metrics on exceptional fibers of twisted relative pluricanonical bundles

Theorem 4.2 can be used to prove the positivity of the twisted relative pluricanonical bundles (Theorem 5.1), and it can also be used to obtain a comparison result of singular metrics on exceptional fibers of twisted relative pluricanonical bundles (Theorem 5.2).

Let X be an n-dimensional Kähler manifold, Y a k-dimensional connected complex manifold ($1 \leq k \leq n$), and (L, h) a pseudoeffective holomorphic line bundle over X.

Let $\Pi : X \to Y$ be a surjective proper holomorphic map. Denote by Y_0 the set of all points in Y which are regular values of Π. Let $X_y := \Pi^{-1}(y)$, $L_y := L|_{X_y}$, $h_y := h|_{X_y}$, $Y_h := \{y \in Y_0;\ h_y \not\equiv +\infty\}$ and

$$Y_{m,\mathrm{ext}} := \left\{y \in Y_0;\ \dim H^0(X_y, mK_{X_y} + L_y) = \operatorname{rank} \Pi_*(mK_{X/Y} + L)\right\}.$$

Then $Y_{m,\mathrm{ext}}$ is the Zariski open subset of Y consisting of all $y \in Y_0$ such that every section in $H^0(X_y, mK_{X_y} + L_y)$ has a holomorphic extension to some open neighborhood of X_y in X. Denote $\Pi^{-1}(Y_{m,\mathrm{ext}})$ by $X_{m,\mathrm{ext}}$.

Denote by ω_y the Kähler metric on X_y induced by ω. Let $dV_{X_y,\omega_y} := \frac{\omega_y^{n-k}}{(n-k)!}$ and let $h_{\omega_y} := (dV_{X_y,\omega_y})^{-1}$ (it defines a smooth metric on K_{X_y}).

For every $y \in Y_{m,\mathrm{ext}}$ and every $x \in X_y$, by the isomorphism

$$(mK_{X/Y} + L)|_{X_y} \simeq mK_{X_y} + L_y,$$

the *relative m-Bergman kernel* $B^o_{m,X/Y}$ of the line bundles $(mK_{X/Y} + L)|_{X_{m,\mathrm{ext}}}$ is defined as

$$B^o_{m,X/Y}(x) := \sup\left\{u(x) \otimes \overline{u(x)};\ u \in H^0(X_y, mK_{X_y} + L_y)\right.$$
$$\left.\text{and} \int_{X_y} |u|^{\frac{2}{m}}_{h^{\otimes m}_{\omega_y} \otimes h_y}\, dV_{X_y,\omega_y} \leq 1\right\}.$$

Assume that $B^o_{m,X/Y} \not\equiv 0$. Then the following positivity of the twisted relative pluricanonical bundles $mK_{X/Y} + L$ holds. The projective case was proved in [3, 5, 23]. We give a proof of the Kähler case in [32] by using Theorem 4.2 (see also [6, 30] for the Kähler case).

Theorem 5.1 ([3, 5, 6, 23, 30, 32]). *The metric $(B^o_{m,X/Y})^{-1}$ is a singular metric on $(mK_{X/Y} + L)|_{X_{m,\mathrm{ext}}}$ with semipositive curvature current. Moreover, $(B^o_{m,X/Y})^{-1}$ extends across $X \setminus X_{m,\mathrm{ext}}$ uniquely to a singular metric $(B_{m,X/Y})^{-1}$ on $mK_{X/Y} + L$ with semipositive curvature current on all of X.*

Theorems 4.1 and 5.1 can be used to prove the positivity of the direct images of twisted relative pluricanonical bundles with singular metrics for Kähler fibrations (see [32], and see also [3, 16, 23] for the projective case).

Now we discuss comparison of singular metrics on exceptional fibers of twisted relative pluricanonical bundles.

For $y \in Y_0 \setminus Y_{m,\mathrm{ext}}$, the fiber X_y is called an exceptional fiber. The metric $(B_{m,X/Y})^{-1}$ over the exceptional fibers X_y is defined as the unique extension of $(B^o_{m,X/Y})^{-1}$.

There is an extremal metric $(B_{m,y})^{-1}$ on the bundle $mK_{X_y} + L_y$ over the exceptional fibers X_y defined as

$$B_{m,y}(x) := \sup \Big\{ u(x) \otimes \overline{u(x)};\ u \in H^0(X_y, mK_{X_y} + L_y), \int_{X_y} |u|^{\frac{2}{m}}_{h_{\omega_y}^{\otimes m} \otimes h_y} dV_{X_y,\omega_y} \leq 1 \text{ and } u \text{ has a holomorphic extension to some open neighborhood of } X_y \text{ in } X \Big\}.$$

When the metric h on L is continuous, the inequality

$$(B_{m,X/Y})^{-1}|_{X_y} \leq (B_{m,y})^{-1} \quad (\forall y \in Y_0 \setminus Y_{m,\mathrm{ext}}) \tag{5.1}$$

was obtained in [5,23] in the projective case. When h has arbitrary singularities, (5.1) was guessed in [5,23].

By using Theorem 4.2, we obtained the following result, which shows that (5.1) is actually an equality in the Kähler case for those h with arbitrary singularities.

Theorem 5.2 ([32]). $(B_{m,X/Y})^{-1}|_{X_y} = (B_{m,y})^{-1}$ *holds for any* $y \in Y_0 \setminus Y_{m,\mathrm{ext}}$.

The key point in the proof of Theorem 5.2 is to obtain a holomorphic extension whose $L^{\frac{2}{m}}$ integral with singular metrics on nearby fibers has a lower limit property with an optimal estimate. This property can be implied by (4.6).

6. Subadditivity of generalized Kodaira–Iitaka dimensions

Theorem 5.1 can be used to prove the subadditivity of the generalized Kodaira–Iitaka dimensions with multiplier ideal sheaves for certain Kähler fibrations (Theorem 6.2).

Let X be a connected compact complex manifold, and (L, h_L) a holomorphic $\mathbb{Q}$-line bundle on X with a singular metric h_L. Let k_0 be the smallest positive integer such that k_0L is a holomorphic line bundle.

In terms of the singular metric h_L of L, we will denote the $L^{\frac{2}{m}}$ multiplier ideal sheaf $\mathcal{I}_{L^{\frac{2}{m}}}(\varphi)$ (m is a positive integer) by the global notation $\mathcal{I}_m(h_L)$, where φ is the local weight of h_L. We also write $\mathcal{I}_1(h_L)$ as $\mathcal{I}(h_L)$ for simplicity.

The notion of the generalized Kodaira–Iitaka dimension with multiplier ideal sheaves is defined below.

Definition 6.1 ([33]). The generalized Kodaira–Iitaka dimension $\kappa(X, K_X + L, h_L)$ is defined to be

$$\sup \left\{ v \in \mathbb{Z};\ \varlimsup_{k \to +\infty} \frac{h^0\big(X, (kk_0K_X + kk_0L) \otimes \mathcal{I}_{kk_0}(h_L)\big)}{k^v} > 0 \right\}$$

if $\varlimsup_{k\to+\infty} h^0(X, (kk_0K_X + kk_0L) \otimes \mathcal{I}_{kk_0}(h_L)) \neq 0$. Otherwise, $\kappa(X, K_X + L, h_L)$ is defined to be $-\infty$.

Then the following subadditivity of the generalized Kodaira–Iitaka dimensions for certain Kähler fibrations holds.

Theorem 6.2 ([33]). *Let $\Pi : X \to Y$ be a surjective holomorphic map with connected fibers from a compact Kähler manifold X to a compact connected complex manifold Y; $\dim X = n$ and $\dim Y = m$. Let L be a holomorphic $\mathbb{Q}$-line bundle on X possessing a singular metric h_L such that the curvature current $\sqrt{-1}\Theta_{L,h_L} \geq 0$ on X. Assume that the canonical bundle K_Y of Y possesses a singular metric h such that*

(a) $\sqrt{-1}\Theta_{K_Y,h} \geq 0$ *on Y in the sense of currents,*

(b) *there exists an open subset U of Y and a continuous positive $(1,1)$-form γ on U such that $\sqrt{-1}\Theta_{K_Y,h} \geq \gamma$ on U in the sense of currents.*

Then

$$\kappa(X, K_X + L, h_L) \geq \kappa(Z, K_Z + L|_Z, h_L|_Z) + m, \tag{6.1}$$

where Z denotes a general fiber of Π.

If X and Y are projective, and (L, h_L) is trivial, Theorem 6.2 is just Kawamata–Viehweg's result; that is,

$$\kappa(X) \geq \kappa(Z) + \dim Y$$

holds for Y of general type [17, 26].

If $\mathcal{I}(h_L) = \mathcal{O}_X$ (for example, (X, L) is Kawamata log terminal), (6.1) becomes

$$\kappa(X, K_X + L) \geq \kappa(Z, K_Z + L|_Z) + \dim Y.$$

One can also see [27] for some related results in the case when (X, L) is Kawamata log terminal.

Our proof of Theorem 6.2 is analytic and relies on Theorem 5.1 and a general L^2 extension theorem with singular metrics on weakly pseudoconvex Kähler manifolds [9, Theorem 2.8 and Remark 2.9 (b)]. They are used to prove the following $L^{\frac{2}{k}}$ extension theorem (Theorem 6.3 below) for twisted pluricanonical sections on compact Kähler manifolds, which is the crucial step in the proof of Theorem 6.2.

For $y \in Y$, denote $\Pi^{-1}(y)$ by X_y. Denote $L|_{X_y}$ simply by L_y. Denote by Y_0 the set of all points in Y which are regular values of Π.

Let k_0 be the smallest positive integer such that $k_0 L$ is a holomorphic line bundle, and let k be a positive integer such that $k_0|k$. Let

$$\begin{aligned} Y_{k,\mathrm{ext}} &:= \big\{y \in Y_0;\ h^0(X_y, kK_{X_y} + kL_y) = \operatorname{rank} \Pi_*(kK_{X/Y} + kL)\big\}, \\ \widetilde{Y}_{k,h_L,\mathrm{ext}} &:= \big\{y \in Y_0;\ h^0\big(X_y, (kK_{X_y} + kL_y) \otimes \mathcal{I}_k(h_L)|_{X_y}\big) \\ &\qquad\qquad = \operatorname{rank} \Pi_*\big((kK_{X/Y} + kL) \otimes \mathcal{I}_k(h_L)\big)\big\}, \end{aligned}$$

and

$$Y_{k,h_L,\mathrm{ext}} := \big\{y \in Y_{k,\mathrm{ext}} \cap \widetilde{Y}_{k,h_L,\mathrm{ext}};\ \mathcal{I}_k(h_L|_{X_y}) = \mathcal{I}_k(h_L)|_{X_y}\big\}.$$

Denote $\bigcap_{k\in\mathbb{Z}^+,\, k_0|k} Y_{k,h_L,\mathrm{ext}}$ simply by $Y_{h_L,\mathrm{ext}}$. It is not hard to see that the $2m$-dimensional Lebesgue measure of $Y \setminus Y_{h_L,\mathrm{ext}}$ is zero and $\kappa(X_y, K_{X_y} + L_y, h_L|_{X_y})$ is independent of y when $y \in Y_{h_L,\mathrm{ext}}$.

Theorem 6.3 ([33]). *Let Π, X, Y, (L, h_L) be the same as in Theorem 6.2. Let X_y, L_y, k, Y_0, $Y_{k,\mathrm{ext}}$, and $Y_{k,h_L,\mathrm{ext}}$ be the notations defined above. Let $\varphi \le 0$ be a quasi-psh function on Y, which is smooth outside q distinct points $\{y_j\}_{j=1}^q$. For each y_j $(1 \le j \le q)$, assume that there exists a coordinate ball around y_j with coordinate functions $z = (z_1, z_2, \ldots, z_m)$ such that $z(y_j) = 0$ and $\varphi(z) - \log|z|^{2m}$ is smooth. Moreover, assume that the canonical bundle K_Y of Y possesses a singular Hermitian metric h such that*

$$(k-1)\sqrt{-1}\Theta_{K_Y,h} + \alpha\sqrt{-1}\partial\bar{\partial}\varphi \ge 0 \quad \text{on } Y \text{ for all } \alpha \in [1, 1+\varepsilon], \tag{6.2}$$

where ε is a positive number. Denote the pluripolar set $\{h = +\infty\}$ by Σ_h. Assume that $\{y_j\}_{j=1}^q \subset Y_{k,\mathrm{ext}} \setminus \Sigma_h$. Then there exists a positive constant C such that, for any

$$f \in H^0\Big(\bigcup_{j=1}^q X_{y_j}, (kK_X + kL)|_{\bigcup_{j=1}^q X_{y_j}} \otimes \mathcal{I}_k(h_L|_{\bigcup_{j=1}^q X_{y_j}})\Big),$$

there exists $F \in H^0(X, (kK_X + kL) \otimes \mathcal{I}_k(h_L))$ such that

$$F|_{\bigcup_{j=1}^q X_{y_j}} = f$$

and

$$\int_X \big(|F|^2_{h_\omega^k \otimes h_L^k}\big)^{\frac{1}{k}}\, dV_{X,\omega} \le C \sum_{j=1}^q \int_{X_{y_j}} \big(|f|^2_{h_\omega^k \otimes h_L^k}\big)^{\frac{1}{k}}\, dV_{X_{y_j},\omega_{y_j}}, \tag{6.3}$$

where ω is the Kähler metric on X, ω_{y_j} is the Kähler metric on X_{y_j} induced from ω, $dV_{X,\omega} := \frac{\omega^n}{n!}$ is the volume form on X, and $h_\omega := (dV_{X,\omega})^{-1}$ (it defines a smooth Hermitian metric on K_X).

The idea in our proof of Theorem 6.2 is sketched below.

Let k be a positive integer sufficiently divisible, and let

$$p := h^0\big(X_y, (kK_{X_y} + kL_y) \otimes \mathcal{I}_k(h_L|_{X_y})\big).$$

Then

$$p \geq \alpha k^{\kappa(Z, K_Z + L|_Z, h_L|_Z)}$$

for some positive number α independent of k.

Let $q := \beta k^m$ and let $\{y_j\}_{j=1}^q$ be q distinct points in U in general position, where β is some positive number independent of k. We will make some explanation on the degree m $(= \dim Y)$ in the end of this section.

Assume U is small enough and let $\eta \in H^0(U, K_Y)$ be a holomorphic frame of $K_Y|_U$. For each $j = 1, 2, \dots, q$, let

$$\{e_{ij}\}_{i=1}^p \subset H^0\big(X_{y_j}, (kK_{X_{y_j}} + kL_{y_j}) \otimes \mathcal{I}_k(h_L|_{X_{y_j}})\big)$$

be a basis.

Let δ_{lj} be the Kronecker delta function, where $1 \leq l \leq q$ and $1 \leq j \leq q$.

For each j,

$$f_{i,l,j} := \delta_{lj} e_{ij} \otimes (\Pi^* \eta)^k|_{X_{y_j}} \quad (1 \leq i \leq p,\ 1 \leq l \leq q)$$

belongs to $H^0(X_{y_j}, (kK_X|_{X_{y_j}} + kL|_{X_{y_j}}) \otimes \mathcal{I}_k(h_L|_{X_{y_j}}))$. Let

$$f_{i,l} \in H^0\Bigg(\bigcup_{j=1}^q X_{y_j}, (kK_X + kL)|_{\bigcup_{j=1}^q X_{y_j}} \otimes \mathcal{I}_k(h_L|_{\bigcup_{j=1}^q X_{y_j}})\Bigg)$$

be defined by $f_{i,l}|_{X_{y_j}} = f_{i,l,j}$ for all $1 \leq i \leq p$, $1 \leq l \leq q$, and $1 \leq j \leq q$.

Then by using Theorem 6.3, we can obtain that there exist holomorphic sections

$$\{F_{i,l}\}_{1 \leq i \leq p,\, 1 \leq l \leq q} \subset H^0\big(X, (kK_X + kL) \otimes \mathcal{I}_k(h_L)\big)$$

such that

$$F_{i,l}|_{\bigcup_{j=1}^q X_{y_j}} = f_{i,l} \quad \text{for all } i \text{ and } l.$$

It is obvious that $\{F_{i,l}\}_{1 \leq i \leq p,\, 1 \leq l \leq q}$ is linearly independent. Therefore,

$$h^0\big(X, (kK_X + kL) \otimes \mathcal{I}_k(h_L)\big) \geq pq \geq \alpha\beta k^{\kappa(Z, K_Z + L|_Z, h_L|_Z) + m}.$$

Hence

$$\kappa(X, K_X + L, h_L) \geq \kappa(Z, K_Z + L|_Z, h_L|_Z) + m.$$

Now we make some explanation on the degree m ($= \dim Y$) in the definition of q. In the above proof, we need to construct a quasi-psh function φ on Y such that (6.2) holds on Y when we use Theorem 6.3. In fact, we construct the function φ by splicing the polar functions $\log|z(y) - z(y_j)|^{2m}$ around the points $\{y_j\}_{j=1}^{q}$, and we need to control the negative part of $\sqrt{-1}\partial\bar{\partial}\varphi$ such that (6.2) holds on Y. In this way, $\sqrt{-1}\partial\bar{\partial}\varphi$ may get more negativity when q becomes larger. The integer m is just the largest degree in the definition of q such that (6.2) holds on Y.

7. A generalization of Siu's lemma with nontrivial multiplier ideal sheaves near a subvariety

Let (X, ω) be an n-dimensional Kähler manifold with a Kähler metric ω, let (L, h) be a holomorphic line bundle over X equipped with a singular metric h, and let ψ be a quasi-psh function on X.

Let $S := V(\mathcal{I}(\psi))$ be the zero variety of the multiplier ideal sheaf $\mathcal{I}(\psi)$. Then ψ is said to have *log canonical singularities* along S if $\mathcal{I}((1-\varepsilon)\psi)|_S = \mathcal{O}_X|_S$ for every $\varepsilon > 0$.

Denote by S^0 the set of regular points of S.

Definition 7.1 (see [9, 14, 19]). Assume that ψ has neat analytic singularities and has log canonical singularities along $S = V(\mathcal{I}(\psi))$. The positive measure $dV_{X,\omega}[\psi]$ on S^0 (the set of regular points of S) is defined by

$$\int_{S^0} g\, dV_{X,\omega}[\psi] = \lim_{t\to-\infty} \int_{\{x\in X: t<\psi(x)<t+1\}} \tilde{g} e^{-\psi}\, dV_{X,\omega} \tag{7.1}$$

for any compactly supported nonnegative continuous function g on S^0, where $\tilde{g}$ is a compactly supported nonnegative continuous extension of g to X such that $(\operatorname{supp}\tilde{g}) \cap S \subset S^0$.

If $f \in H^0(S^0, (K_X \otimes L)|_{S^0})$, then $|f|^2_{\omega,h} dV_{X,\omega}[\psi]$ is a positive measure on S^0 which depends on the property of h on S^0. There is another way to define a positive measure which is defined not only on the property of h on S^0 but also on the property of h near $\{\psi = -\infty\}$ (see Definition 7.3).

Note that ψ is neither assumed to have neat analytic singularities nor assumed to have log canonical singularities along S in Definition 7.2 and Definition 7.3.

Definition 7.2 ([9]). The *restricted multiplier ideal sheaf* $\mathcal{I}'_{\psi}(h)$ is defined to be the set of germs $f \in \mathcal{I}(h)_x \subset \mathcal{O}_{X,x}$ such that there exists a coordinate neighborhood U of x satisfying

$$\varlimsup_{t\to-\infty} \int_{\{y\in U: t<\psi(y)<t+1\}} |f|^2 e^{-\varphi-\psi}\, d\lambda < +\infty,$$

where U is small enough such that h can be written as $e^{-\varphi}$ with respect to a local holomorphic trivialization of L on a neighborhood of $\overline{U}$, and $d\lambda$ is the $2n$-dimensional Lebesgue measure on U.

Denote by S' the zero set of the ideal sheaf

$$\mathcal{J} := \{g \in \mathcal{O}_X;\ g \cdot \mathcal{I}(h) \subset \mathcal{I}(he^{-\psi})\}.$$

Let f be an element in

$$H^0\big(X, \mathcal{O}_X(K_X \otimes L) \otimes \mathcal{I}'_{\psi}(h)/\mathcal{I}(he^{-\psi})\big).$$

Then f is actually supported on S'.

Definition 7.3 ([9]). The positive measure $|f|^2_{\omega,h} dV'_{X,\omega}[\psi]$ (a purely formal notation) on S' is defined as the minimum element of the partially ordered set of positive measures $d\mu$ satisfying

$$\int_{S'} g d\mu \geq \overline{\lim_{t\to-\infty}} \int_{\{x\in X: t<\psi(x)<t+1\}} g|\tilde{f}|^2_{\omega,h} e^{-\psi}\, dV_{X,\omega}$$

for any nonnegative continuous function g on X with $\operatorname{supp} g \subset\subset X$, where $\tilde{f}$ is a smooth extension of f to X such that

$$\tilde{f} - \hat{f} \in \mathcal{O}_X(K_X \otimes L) \otimes_{\mathcal{O}_X} \mathcal{I}(he^{-\psi}) \otimes_{\mathcal{O}_X} \mathcal{C}^{\infty}$$

locally for any local holomorphic representation $\hat{f}$ of f.

The following generalization of Siu's lemma with nontrivial multiplier ideal sheaves near a subvariety gives a relation between the two measures in Definitions 7.1 and 7.3. It can be proved by using Hironaka's desingularization theorem and the method in the proof of Theorem 3.1.

Theorem 7.4 ([33]). *Let $\Omega_0 \subset \mathbb{C}^n$ be a bounded domain, ψ a negative quasi-psh function on Ω_0 with neat analytic singularities, and φ a negative psh function on Ω_0. Denote by S the zero variety of $\mathcal{I}(\psi)$ and assume that ψ has log canonical singularities along S. Let Ω be a pseudoconvex domain such that $\Omega \subset\subset \Omega_0$. Suppose that f is a holomorphic function on S^0 satisfying*

$$\int_{S^0} |f|^2 e^{-\varphi}\, d\lambda[\psi] < +\infty,$$

where S^0 is the set of regular points of S, $d\lambda$ is the Lebesgue measure on Ω_0, and $d\lambda[\psi]$ is the positive measure on S^0 defined as in Definition 7.1. Then there exists a positive number β such that

$$\int_{\Omega\cap S^0} |f|^2 e^{-(1+\beta)\varphi}\, d\lambda[\psi] < +\infty,$$

and there exists a holomorphic function F on Ω such that

$$F = f \quad \text{on } \Omega \cap S^0 \tag{7.2}$$

and

$$\int_\Omega \frac{|F|^2 e^{-(1+\beta)\varphi}}{e^\psi(\psi^2+1)}\, d\lambda < +\infty. \tag{7.3}$$

Moreover, any holomorphic function F on Ω satisfying (7.2) *and* (7.3) *has the property*

$$\lim_{t\to-\infty}\int_{\Omega\cap\{t<\psi<t+1\}} v|F|^2 e^{-\varphi-\psi}\, d\lambda = \int_{\Omega\cap S^0} v|f|^2 e^{-\varphi}\, d\lambda[\psi]$$

for any compactly supported nonnegative continuous function v on Ω.

Theorem 7.4 shows that the two ways to define the measures are the same when ψ has neat analytic singularities and has log canonical singularities along S.

By using Theorem 7.4, we proved in [33] that the L^2 extension theorem in [31] can be regarded as a corollary of the L^2 extension theorem in [34].

Funding. Xiangyu Zhou was partially supported by the National Natural Science Foundation of China (No. 11688101 and No. 11431013). Langfeng Zhu was partially supported by the National Natural Science Foundation of China (No. 12022110).

References

[1] U. Angehrn and Y. T. Siu, Effective freeness and point separation for adjoint bundles. *Invent. Math.* **122** (1995), no. 2, 291–308 Zbl 0847.32035 MR 1358978

[2] B. Berndtsson, The openness conjecture and complex Brunn–Minkowski inequalities. In *Complex Geometry and Dynamics*, pp. 29–44, Abel Symp. 10, Springer, Cham, 2015 Zbl 1337.32001 MR 3587460

[3] B. Berndtsson and M. Păun, Bergman kernels and the pseudoeffectivity of relative canonical bundles. *Duke Math. J.* **145** (2008), no. 2, 341–378 Zbl 1181.32025 MR 2449950

[4] B. Berndtsson and M. Păun, A Bergman kernel proof of the Kawamata subadjunction theorem. 2008, arXiv:0804.3884

[5] B. Berndtsson and M. Păun, Bergman kernels and subadjunction. 2010, arXiv:1002.4145

[6] J. Cao, Ohsawa–Takegoshi extension theorem for compact Kähler manifolds and applications. In *Complex and Symplectic Geometry*, pp. 19–38, Springer INdAM Ser. 21, Springer, Cham, 2017 Zbl 1405.32029 MR 3645303

[7] J.-P. Demailly, Singular Hermitian metrics on positive line bundles. In *Complex Algebraic Varieties (Bayreuth, 1990)*, pp. 87–104, Lecture Notes in Math. 1507, Springer, Berlin, 1992 Zbl 0784.32024 MR 1178721

[8] J.-P. Demailly, *Analytic Methods in Algebraic Geometry*. Surv. Mod. Math. 1, International Press, Somerville, MA; Higher Education Press, Beijing, 2012 Zbl 1271.14001 MR 2978333

[9] J.-P. Demailly, Extension of holomorphic functions defined on non reduced analytic subvarieties. In *The Legacy of Bernhard Riemann After One Hundred and Fifty Years. Vol. I*, pp. 191–222, Adv. Lect. Math. (ALM) 35, Int. Press, Somerville, MA, 2016 Zbl 1360.14025 MR 3525916

[10] J.-P. Demailly and J. Kollár, Semi-continuity of complex singularity exponents and Kähler–Einstein metrics on Fano orbifolds. *Ann. Sci. École Norm. Sup. (4)* **34** (2001), no. 4, 525–556 Zbl 0994.32021 MR 1852009

[11] J. E. Fornaess, Several complex variables. 2015, arXiv:1507.00562

[12] Q. Guan and X. Zhou, Effectiveness of Demailly's strong openness conjecture and related problems. *Invent. Math.* **202** (2015), no. 2, 635–676 Zbl 1333.32014 MR 3418242

[13] Q. Guan and X. Zhou, A proof of Demailly's strong openness conjecture. *Ann. of Math. (2)* **182** (2015), no. 2, 605–616 Zbl 1329.32016 MR 3418526

[14] Q. Guan and X. Zhou, A solution of an L^2 extension problem with an optimal estimate and applications. *Ann. of Math. (2)* **181** (2015), no. 3, 1139–1208 Zbl 1348.32008 MR 3296822

[15] Q. A. Guan, Z. Q. Li, and X. Y. Zhou, Estimation of weighted L^2 norm related to Demailly's Strong Openness Conjecture. 2016, arXiv:1603.05733

[16] C. Hacon, M. Popa, and C. Schnell, Algebraic fiber spaces over abelian varieties: around a recent theorem by Cao and Păun. In *Local and Global Methods in Algebraic Geometry*, pp. 143–195, Contemp. Math. 712, Amer. Math. Soc., Providence, RI, 2018 Zbl 1398.14018 MR 3832403

[17] Y. Kawamata, Characterization of abelian varieties. *Compositio Math.* **43** (1981), no. 2, 253–276 Zbl 0471.14022 MR 622451

[18] A. M. Nadel, Multiplier ideal sheaves and Kähler–Einstein metrics of positive scalar curvature. *Ann. of Math. (2)* **132** (1990), no. 3, 549–596 Zbl 0731.53063 MR 1078269

[19] T. Ohsawa, On the extension of L^2 holomorphic functions. V. Effects of generalization. *Nagoya Math. J.* **161** (2001), 1–21 Zbl 0986.32002 MR 1820210

[20] T. Ohsawa, *L^2 Approaches in Several Complex Variables, Towards the Oka–Cartan Theory with Precise Bounds*. Springer Monogr. Math., Springer, Tokyo, 2018 Zbl 1439.32003 MR 3887636

[21] T. Ohsawa and K. Takegoshi, On the extension of L^2 holomorphic functions. *Math. Z.* **195** (1987), no. 2, 197–204 Zbl 0625.32011 MR 892051

[22] D. H. Phong and J. Sturm, Algebraic estimates, stability of local zeta functions, and uniform estimates for distribution functions. *Ann. of Math. (2)* **152** (2000), no. 1, 277–329 Zbl 0995.11065 MR 1792297

[23] M. Păun and S. Takayama, Positivity of twisted relative pluricanonical bundles and their direct images. *J. Algebraic Geom.* **27** (2018), no. 2, 211–272 Zbl 1430.14017 MR 3764276

[24] Y.-T. Siu, The Fujita conjecture and the extension theorem of Ohsawa–Takegoshi. In *Geometric Complex Analysis (Hayama, 1995)*, pp. 577–592, World Sci. Publ., River Edge, NJ, 1996 Zbl 0941.32021 MR 1453639

[25] Y.-T. Siu, Invariance of plurigenera and torsion-freeness of direct image sheaves of pluricanonical bundles. In *Finite or Infinite Dimensional Complex Analysis and Applications*, pp. 45–83, Adv. Complex Anal. Appl. 2, Kluwer Acad. Publ., Dordrecht, 2004 Zbl 1044.32016 MR 2058399

[26] E. Viehweg, Die Additivität der Kodaira Dimension für projektive Faserräume über Varietäten des allgemeinen Typs. *J. Reine Angew. Math.* **330** (1982), 132–142 Zbl 0466.14009 MR 641815

[27] J. Wang, On the Iitaka conjecture $C_{n,m}$ for Kähler fibre spaces. *Ann. Fac. Sci. Toulouse Math. (6)* **30** (2021), no. 4, 813–897 Zbl 07469482 MR 4350100

[28] L. Yi, An Ohsawa–Takegoshi theorem on compact Kähler manifolds. *Sci. China Math.* **57** (2014), no. 1, 9–30 Zbl 1302.32035 MR 3146512

[29] X. Zhou and L. Zhu, A generalized Siu's lemma. *Math. Res. Lett.* **24** (2017), no. 6, 1897–1913 Zbl 1394.32027 MR 3762700

[30] X. Zhou and L. Zhu, An optimal L^2 extension theorem on weakly pseudoconvex Kähler manifolds. *J. Differential Geom.* **110** (2018), no. 1, 135–186 Zbl 1426.53082 MR 3851746

[31] X. Zhou and L. Zhu, Optimal L^2 extension of sections from subvarieties in weakly pseudoconvex manifolds. *Pacific J. Math.* **309** (2020), no. 2, 475–510 Zbl 1458.32013 MR 4202022

[32] X. Zhou and L. Zhu, Siu's lemma, optimal L^2 extension and applications to twisted pluricanonical sheaves. *Math. Ann.* **377** (2020), no. 1-2, 675–722 Zbl 1452.32014 MR 4099619

[33] X. Zhou and L. Zhu, Subadditivity of generalized Kodaira dimensions and extension theorems. *Internat. J. Math.* **31** (2020), no. 12, 2050098, 36 Zbl 1457.32059 MR 4184430

[34] X. Zhou and L. Zhu, Extension of cohomology classes and holomorphic sections defined on subvarieties. *J. Algebraic Geom.* **31** (2022), no. 1, 137–179 Zbl 07459642 MR 4372411

[35] X. Y. Zhou and L. F. Zhu, Optimal L^2 extension and Siu's lemma. *Acta Math. Sin. (Engl. Ser.)* **34** (2018), no. 8, 1289–1296 Zbl 1402.32013 MR 3843436

Xiangyu Zhou
Institute of Mathematics, AMSS, Chinese Academy of Sciences, Beijing 100190, P. R. China;
xyzhou@math.ac.cn

Langfeng Zhu
School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P. R. China;
zhulangfeng@amss.ac.cn

European Congress of Mathematics

The European Congress of Mathematics, held every four years, is a well-established major international mathematical event. Following those in Paris (1992), Budapest (1996), Barcelona (2000), Stockholm (2004), Amsterdam (2008), Kraków (2012), and Berlin (2016), the Eighth European Congress of Mathematics (8ECM) took place in Portorož, Slovenia, June 20–26, 2021, with about 1700 participants from all over the world, mostly online due to Covid pandemic.

Ten plenary and thirty invited lectures along with the special Abel and Hirzebruch lectures formed the core of the program. As in all the previous EMS congresses, ten outstanding young mathematicians received the EMS prizes in recognition of their research achievements. In addition, two more prizes were awarded: The Felix Klein Prize for a remarkable solution of an industrial problem and the Otto Neugebauer Prize for a highly original and influential piece of work in the history of mathematics. The program was complemented by five public lectures, several exhibitions, and 62 minisymposia with about 1000 contributions, spread over all areas of mathematics. A number of panel discussions and meetings were organized, covering a variety of issues ranging from the future of mathematical publishing and the role of the ERC to public awareness of mathematics.

These proceedings provide a permanent record of current mathematics of highest quality by presenting extended versions of seven plenary, six prize, and fourteen invited lectures as well as eleven lectures from minisymposia keynote speakers, all of which were delivered during the congress.

https://ems.press
ISBN 978-3-98547-051-8

EMS PRESS