European Congress of Mathematics

ECM Sevilla2024

Sevilla, 15–19 July, 2024

Edited by
Guillermo P. Curbera
María Luz Puertas
Alfonso Suárez-Llorens



EMS PRESS


EMS
PRESS

European Congress of Mathematics



Sevilla, 15–19 July, 2024

Edited by
Guillermo P. Curbera
María Luz Puertas
Alfonso Suárez-Llorens

EMS PRESS

Editors

Guillermo P. Curbera
Facultad de Matemáticas
Universidad de Sevilla
Calle Tarfia s/n
41012 Sevilla, Spain

Email: curbera@us.es

María Luz Puertas
Departamento de Matemáticas
Universidad de Almería
Carretera Sacramento s/n
04120 Almería, Spain

Email: mpuertas@ual.es

Alfonso Suárez-Llorens
Departamento de Estadística
e Investigación Operativa
Facultad de Ciencias
Universidad de Cádiz
Calle República Saharaui
11519 Puerto Real (Cádiz), Spain

Email: alfonso.suarez@uca.es

2020 Mathematics Subject Classification: 00B25

ISBN 978-3-98547-105-8, eISBN 978-3-98547-605-3, DOI 10.4171/9ECM

The content of this volume is licensed under the CC BY 4.0 license, with the exception of the logo and branding of the European Mathematical Society and EMS Press, and where otherwise noted.

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

Email: info@ems.press

https://ems.press

© 2026 EMS Press

Cover photograph by Antonio Rojas, kindly donated for this book.
Typesetting by Compuscript Ltd, Shannon, Ireland
Printed in Germany
♾ Printed on acid free paper

Preface

These proceedings aim at presenting a faithful account of all that occurred at the 9th European Congress of Mathematics, held in Seville from July 15 to 19, 2024. They contain polished and extended versions of most of the plenary, invited, prize and public lectures that were delivered at the congress. These, together with the large number of minisymposia, contributed communications, posters and satellite conferences organized, provide a lively snapshot of the current situation and the recent developments of mathematical research in Europe.

The proceedings also include a detailed account of the organization and development of the congress. In particular, of the opening and closing ceremonies, including the presentation of the prizes of the European Mathematical Society, as well as the Felix Klein, Otto Neugebauer, Paul Lévy and EMS/ECMI Lánczos prizes.

The congress would not have been possible without the work of a large number of committees, collaborators, and public and private sponsors, which are promptly indicated.

We would like to express our warm thanks to the authors of the articles for their effort in the preparation of their contributions. We also would like to express our gratitude to EMS Press, the publishing house of the European Mathematical Society, for its support.

Guillermo P. Curbera, 9ECM organizing committee member
María Luz Puertas, 9ECM organizing committee member
Alfonso Suárez-Llorens, 9ECM organizing committee member



Señor Don
Juan González-Meneses López
Presidente del Comité Organizador del IX Congreso Europeo de Matemáticas

De Su Majestad el Rey

Quisiera enviarle mi felicitación más cordial por el éxito de la novena edición del Congreso Europeo de Matemáticas, considerado el segundo más importante del mundo, que ha batido récord histórico en asistencia de expertos.

Lo he seguido con interés y me consta que los conferenciantes que se han dado cita en Sevilla en tan significativo encuentro desarrollan una extraordinaria labor de investigación y divulgación de mucho valor para la sociedad.

Enhorabuena a todo el comité organizador y también a los matemáticos galardonados con los cuatro premios especiales "Felix Klein", "Paul Lévy", "Otto Neugebauer" y "Cornelius Lanczos EMS/ECMI de Software Matemático" y con los diez premios EMS por el reconocimiento obtenido.

Muy afectuosamente,

Felipe R.



EUROPEAN MATHEMATICAL SOCIETY

ECM Sevilla2024

9TH EUROPEAN CONGRESS OF MATHEMATICS

15-19 July 2024, Sevilla, Spain

EMS Prizes
Invited Talks
Mini-symposia

Thematic sessions
Public Lectures
Panel discussions

Hirzebruch Lecture
Étienne Ghys

Abel Lecture
Avi Wigderson

Plenary Speakers

Martin Bridson
University of Oxford

Annalisa Buffa
EPF Lausanne

Maxim Kontsevich
IHES

André Neves
University of Chicago

Eero Saksman
University of Helsinki

Benny Sudakov
ETH Zurich

Fabio Toninelli
Technische Universität Wien

Vlad Vicol
Courant Institute

Anna Wienhard
Max Planck Institute for Mathematics in the Sciences

Tamar Ziegler
Hebrew University of Jerusalem

www.ecm2024sevilla.com

Contents

Preface v

Contents xi

Past congresses xv

Organization of the congress xvii

9ECM committees xxv

Prize winners xxxi

Program of the congress xxxiii

Sponsors and exhibitors xlix

Opening ceremony li

Plenary lectures

Chasing finite shadows of infinite groups through geometry 1
by Martin R. Bridson

Generalized Riemann–Hilbert correspondence, wall-crossing and resurgence . 45
by Maxim Kontsevich

Restricted subgraphs of edge-colored graphs and applications 73
by Benny Sudakov

Large-scale behavior of out-of-equilibrium systems at and above the critical dimension 103
by Giuseppe Cannizzaro and Fabio Toninelli

Anomalous diffusion via iterative quantitative homogenization: An overview of the main ideas 125
by Scott Armstrong and Vlad Vicol

Positivity and non-commutativity 169
by Anna Wienhard

Sign patterns of the Liouville function 203
by Tamar Ziegler

Invited lectures

A discontinuous Galerkin method for the three-dimensional heterodimer model with application to prion-like proteins' dynamics 217
by Paola F. Antonietti, Mattia Corti, and Giacomo Lorenzon

Dualities in convex geometric analysis and beyond 241
by Shiri Artstein-Avidan

The continuum problem: Hilbert's first problem revisited 259
by David Asperó and Ralf Schindler

Self-affine sets and measures – A survey 277
by Balázs Bárány

Applications of the renormalisation group 303
by Roland Bauerschmidt

A free boundary approach of spectral optimization problems: Extremal Neumann eigenvalues 317
by Dorin Bucur

Courant, Bézout and persistence 333
by Lev Buhovsky

Unstable vortices and non-uniqueness for 2D Euler and $\alpha-$gSQG 351
by Ángel Castro

Graph product structure: Theory and applications 377
by Vida Dujmović and David R. Wood

Weak mixing in polygonal billiards 389
by Giovanni Forni

Partition regularity of homogeneous quadratics: Current trends and challenges 411
by Nikos Frantzikinakis

Symmetries of Grothendieck rings in representation theory 467
by David Hernandez

Forward and inverse problems in nonlinear acoustics 489
by Barbara Kaltenbacher

Introduction to inverse problems for non-linear partial differential equations .. 515
by Matti Lassas

Cheeger–Müller and Bismut–Zhang theorem on singular spaces 541
by Ursula Ludwig

Strong convergence of unitary and permutation representations of discrete groups 581
by Michael Magee

The Poisson–Dirichlet problem in domains with Ahlfors regular boundary 601
by Ariel Barton, Svitlana Mayboroda, and Alberto Pacati

Cohomology rings of character varieties 637
by Anton Mellit

Moduli of Galois representations and generalized class groups 665
by James Newton

Differential calculus and optimization in persistence module categories 677
by Steve Oudot

Spectral gaps in quantum computing 713
by David Pérez-García

The matrix $\mathbf{A}_2$ conjecture 737
by Stefanie Petermichl

Proofs, computational complexity, and games 761
by Pavel Pudlák

The $K(\pi, 1)$-conjecture for Artin groups 785
by Mario Salvetti

Universal symmetries: Global equivariant homotopy theory 807
by Stefan Schwede

New tools in hierarchical hyperbolicity: A survey 829
by Alessandro Sisto

Noncommutative singularity theory 843
by Gavin Brown and Michael Wemyss

EMS prize lectures

Sketches of nonuniformly elliptic Schauder theory 865
by Cristiana De Filippis

An introduction to representations of p-adic groups 903
by Jessica Fintzen

Recent progress on the problem of soliton resolution 921
by Jacek Jendrej

Recent progress on sparse equidistribution 941
by Adam M. Kanigowski

Recent progress in the study of Latin squares 963
by Richard Montgomery

Fourier uniqueness pairs and interpolation formulas 975
by Danylo Radchenko

Other prize lectures

Learning physics simulations for improving design processes in the industry . . 991
by Fabien Casenave

On the 125th anniversary of his birth: My “encounters” with Otto Neugebauer 1013
by Reinhard Siegmund-Schultze

Integrable PDE’s for random growth 1039
by Jeremy Quastel

MUMPS: MUltifrontal Massively Parallel Solver for the direct solution of sparse linear equations 1063
by Patrick R. Amestoy, Jean-Yves L’Excellent, Theo Mary, and Chiara Puglisi

Public lecture

Tales of (mathematical) paradise 1077
by José Ferreirós

Past congresses

1992	Paris
1996	Budapest
2000	Barcelona
2004	Stockholm
2008	Amsterdam
2012	Kraków
2016	Berlin
2020	Portorož (held in 2021)

Organization of the congress

Between July 15 and 19, 2024, the 9th edition of the European Congress of Mathematics took place in Seville. Since the council of the European Mathematical Society, meeting in Bled (Slovenia) in July 2020, chose Seville's candidacy as the venue for this edition, the organizing committee had worked diligently, culminating its work during the five days of the congress.

With a total of 1,400 registered participants, in addition to exhibitors, press, volunteers, and companions, the congress was a complete success in terms of attendance, surpassing the previous in-person edition (the 7th ECM, held in Berlin in 2016). The multiple lectures, minisymposia, thematic, and poster sessions, supplemented with other activities, brought together the most prominent researchers in mathematics, both European and from the entire world. For one week, Seville was the world's mathematical hub.

Even though Seville was the main venue of the congress, the participation in the organizing committee of mathematicians from the five Andalusian universities offering degrees in mathematics has given the congress a distinct Andalusian character. The organizing committee consisted of 18 members from the universities of Almería, Cádiz, Granada, Málaga and Sevilla (the list of its members is presented in section Organizing committee below). Additionally, numerous collaborators from all these universities helped the success of the congress by participating in different roles in its organization (their names are listed in section Collaborators).

From the very beginning, the organizers have had the support of the main national scientific institution, the Consejo Superior de Investigaciones Científicas (CSIC), and the main mathematical institutions in the country: the Real Sociedad Matemática Española (RSME), the Sociedad Española de Matemática Aplicada (SeMA), the Societat Catalana de Matemàtiques (SCM), the Sociedad de Estadística e Investigación Operativa (SEIO), the Comité Español de Matemáticas (CEMAT), and the major research institutes: the Instituto de Ciencias Matemáticas (ICMAT), the Centre de Recerca Matemàtica (CRM), the Basque Center for Applied Mathematics (BCAM), the Instituto de Matemáticas de la Universidad de Granada (IMAG), the Instituto de Matemáticas de la Universidad de Sevilla (IMUS), and the Centro de Investigación y Tecnología Matemática de Galicia (CITMAga). His Majesty the King Felipe VI agreed to preside over the honorary committee (for the full list of members, see section Honorary committee), which also included the president of the government, the minister of science, innovation and universities, the president of the Junta de Andalucía, the president of the Diputación de Sevilla, the mayor of the city of Seville, the rectors of the universities involved in the organization, the presidents of the Span-

ish mathematical societies, and other authorities. This long list reflects the great institutional support received by this project.

We have also received an invaluable and generous help from a large number of sponsors (listed in section Sponsors and exhibitors below).

Thanks to the collaboration of the Comité Español de Matemáticas (CEMAT) and many Spanish faculties and mathematical institutes, we were able to set up an excellent volunteer program through which a total of 100 volunteers, undergraduate, master, and first-year PhD students in mathematics from universities across the country contibuted to the organization during the development of the congress, greatly helping it to run smoothly, while also experiencing the unique opportunity to participate in a congress of this magnitude.

It is tradition at European Mathematical Society congresses to support mathematicians from developing countries and young researchers with fewer resources. On this occasion, the congress organization and the European Mathematical Society granted a total of 120 scholarships that allowed as many mathematicians to finance their participation in the congress. The grant committee (see the members' list in section Grant committee), responsible for selecting the beneficiaries, had the invaluable collaboration of the Centre International de Mathématiques Pures et Appliquées (CIMPA), an organization dedicated to promoting mathematical research in developing countries. CIMPA also collaborated in the organization of the minisymposium *Shared horizons: Collaborations between Europe and the global South*, in which several of the scholarship recipients participated.

In the months leading up to the congress, the scientific committee appointed by the European Mathematical Society, presided by Leonid Polterovich and consisting of 17 members (see their names listed in section Scientific committee), selected the plenary and invited speakers who would give talks at the congress. In parallel, the local scientific committee, chaired by Juan González-Meneses and consisting of 17 mathematicians from different Spanish universities (see their names in section Local scientific committee) received various proposals for minisymposia, talks for special sessions and posters, and decided on their acceptance.

A total of 123 people collaborated in one way or another in the organization of the congress, including organizers, members of the scientific committees, satellite committee, grant committee, and prize committees.

During the weekend prior to the congress, the biennial meeting of the council of the European Mathematical Society was held in Granada, organized by the Vicerrectorado de Investigación y Transferencia and the Instituto de Matemáticas de la Universidad de Granada (IMAG), and coordinated by Magdalena Rodríguez from the Universidad de Granada. At this meeting, among other matters, new vice-presidents (Jorge Buescu and Victoria Gould), secretary (Jiří Rákosník), and executive committee members of the EMS for the 2025–28 period were elected. The vote to

choose the city that would hold the venue for the next European Congress of Mathematics also took place. London and Bologna were the two candidate cities, with Bologna ultimately being selected. Participants in the meeting also had the opportunity to enjoy a mathematical visit to the Alhambra palace.

The day before the official opening of the congress, a conference organized by the European Women in Mathematics association was held at the Faculty of Fine Arts of the Universidad de Sevilla. The conference, attended by a large group of researchers participating in the congress, featured five talks (listed in section EWM event), followed by a debate and social activities.

As is customary in congresses of this magnitude, several satellite conferences took place during the preceding weeks. On this occasion, a total of 27 congresses were held throughout Spain and Portugal, covering a wide variety of topics (these conferences are listed in section Satellite conferences), which were selected by the satellite committee (see the list of members in section Satellite committee).

The congress officially began with the openning ceremony held at the Teatro de la Maestranza in Seville, with Professor Jesús Soto as master of ceremonies, attended by over a thousand participants. Juan González-Meneses and Isabel Fernández, president and vice-president of the organizing committee, spoke first to welcome attendees and express their gratitude to all the people and institutions who collaborated on making the congress a success. Next, the president of the European Mathematical Society, Jan Philip Solovej, took the floor, offering a brief overview of the scientific program of the congress and thanking the organizing committee for their work. Following speeches by the vice-minister for university, research and innovation of the Junta de Andalucía, Lorena Garrido, the deputy mayor of Seville, Álvaro Pimentel, the vice-rector for research of the Universidad de Sevilla, Julián Martínez, finally declared the congress officially inaugurated.

The ceremony continued with a musical interlude by the renowned flamenco pianist David Dorantes, who delighted the audience with a selection of his best compositions, accompanied by a dance performance.

The climax of the ceremony followed: the presentation of the European Mathematical Society prizes, which are awarded at each European Congress of Mathematics to ten mathematicians under 35 years of age in recognition of outstanding contributions in their respective fields. The president of the EMS prize committee (members listed in section EMS prize committee), Pavel Exner, took the stage to announce the winners (the list of the awardees is presented in section EMS prize below). All of them (except Maria Colombo and Nina Holden, absent due to their recent maternity) went on stage to receive their award from the president of the European Mathematical Society: an accrediting diploma and an Andalusian-themed tile.

Next, Peregrina Quintela, president of the committee for the Felix Klein prize in applied mathematics (for the list of members, see section Felix Klein prize commit-

tee), announced the winner (whose name is given in section Felix Klein prize). Karine Chemla, president of the committee for the Otto Neugebauer prize in the history of mathematics (the members of the committee are listed in section Otto Neugebauer prize committee), announced the winner of this edition (the name is given in section Otto Neugebauer prize). Barbara Kaltenbacher, a member of the committee for the EMS/ECMI Lánczos prize in mathematical software (see section EMS/ECMI Lánczos prize committee), spoke next to announce its award (find the name in section EMS/ECMI Lánczos prize). Finally, Konstantin Khanin, a member of the committee for the Paul Lévy prize in probability (see section Paul Lévy prize committee), accompanied by Vincent Piron, Lévy's own grandson, took the stage to present the award to the winner (named in section Paul Lévy prize).

The congress venue was the Escuela Técnica Superior de Ingeniería of the Universidad de Sevilla, located in the Cartuja island on the banks of the Guadalquivir river. Plenary and special lectures were delivered in the nearby Barceló convention center, where a lecture hall with more than one thousand seats was available.

Including plenary, invited, award winners', special and public lectures, minisymposia, and thematic sessions, the scientific program of the congress featured over 950 talks, which gives an idea of the magnitude achieved by the event.

The highlight, undoubtedly, were the plenary lectures (listed in section Plenary lectures): ten top-level speakers selected by the scientific committee. Other highlights were the Abel lecture (see section Abel lecture), given at each edition by a recipient of the namesake prize, and the Hirzebruch lecture (see section Hirzebruch lecture), given by renown mathematicians whose work has an important facet of cooperation. All of them were livestreamed and their recordings can be viewed on the online channel of the Universidad de Sevilla: the opening ceremony on July 15,[1] and the plenary sessions on July 16,[2] July 17,[3] July 18,[4] and July 19.[5]

All 14 prize winners (except Maria Colombo, absent due to her recent maternity) also gave lectures (find the list of the talks in section Prize lectures).

[1] https://tv.us.es/media/OPENING+CEREMONY.+9TH+EUROPEAN+CONGRESS+OF+MATHEMATICS.+15+DE+JULIO+DE+2024/1_ixpcdbt1/231298773

[2] https://tv.us.es/media/9TH+EUROPEAN+CONGRESS+OF+MATHEMATICS.+16+DE+JULIO+DE+2024/1_r386h4wf/231298773

[3] https://tv.us.es/media/9TH+EUROPEAN+CONGRESS+OF+MATHEMATICS.+17+DE+JULIO+DE+2024/1_hat837lr/232310953

[4] https://tv.us.es/media/9TH+EUROPEAN+CONGRESS+OF+MATHEMATICS.+18+DE+JULIO+DE+2024/1_eeqd4lh1/232310953

[5] https://tv.us.es/media/9TH+EUROPEAN+CONGRESS+OF+MATHEMATICS.+19+DE+JULIO+DE+2024/1_t1g1lr5y/232310953

Additionally, selected by the scientific committee, a total of 31 invited lectures covered the most important recent advances in various fields of mathematics (find the list in section Invited lectures).

The congress hosted 64 minisymposia, with a total of 592 oral presentations, far exceeding the largest number of minisymposia organized to date at a European Congress of Mathematics (find the list of minisymposia in section Minisymposia). These are special sessions proposed and organized by mathematicians from around the world, and accepted by the local scientific committee of the congress. Each organizing team of a minisymposium (a maximum of three people) chooses the speakers, who give talks on an specific topic.

Another important activity of the congress were the contributed sessions: consisting of oral contributions that are not invited or part of a minisymposium. Proposed by the speakers themselves, the local scientific committee decided on their suitability for the congress. A total of 200 oral communications were accepted, distributed in 18 contributed sessions (find the ist of contributed sessions in section Contributed sessions).

A special contributed sessions was organized by the European Mathematical Society Young Academy (EMYA), an association created within the European Mathematical Society to support young researchers in mathematics and promote their participation in activities. This session consisted of 21 lightning talks by young researchers. In addition, EMYA organized an "ice breaking" session, specially aimed at young people who had never participated in a major event like the 9ECM.

Panel discussions also played a major role in the congress: eight in total, all with top-level international speakers (the list of panel discussions can be found in section Panel discussions).

Finally, we highlight the 44 posters presented by attendees at the congress and exhibited throughout the week, complementing the scientific program.

One of the objectives of the congress was the public dissemination of mathematics and related topics. For this, several exhibitions, talks, and events beyond the scientific program were organized. Four exhibitions were organized focusing on mathematics and its cultural role (their list and descriptions are presented in section Exhibitions). During the afternoons of the congress, and in parallel to the minisymposia and contributed sessions, four public lectures brought to the congress attendees to learn about examples of the excellence of science carried out at the Universidad de Sevilla. The lecturers were researchers from departments of logic and philosophy of science, physiology, atomic, molecular and nuclear physics, and ecology (see section Public lectures: Beyond mathematics).

During the congress, an outreach lecture (see section Outreach lecture) was also given by one of the invited speakers. It was attended by several high school students participating in a program to foster mathematical talent, who enjoyed an interactive

lecture based on the *Quintessence puzzle*, a new generation of 3D puzzles based on the geometry and combinatorics of the 120-cell polytope. There was also a special lecture (see section London Mathematical Society lecture) organized by the London Mathematical Society.

A cultural visit to the Real Alcázar de Sevilla was organized for the congress participants. The Alcázar is the oldest royal palace still in use in Europe. Its history stretches back more than a thousand years, featuring key moments in Spanish history. The Congress dinner was held on Thursday, July 18, at the *Restaurante "El 29"*, located within the María Luisa Park, a historical garden in the center of the city of Seville.

There were 17 mathematical societies, publishers, and companies that set up their stands at the congress venue to offer participants information about their activities and products (see section Exhibitors).

The closing ceremony of the 9th European Congress of Mathematics took place in the afternoon of Friday, July 19, at the auditorium of the Escuela Técnica Superior de Ingeniería of the Universidad de Sevilla, main venue of the congress. The ceremony was presided over by María Isabel Hartillo Hermoso, vice-rector for digital transformation of the Universidad de Sevilla, with interventions by Juan González-Meneses, president of the organizing committee of the 9ECM, María Ángeles Japón, vice-director of the IMUS, Jan Philip Solovej, president of the EMS, Andrés Sáez Pérez, director of the Escuela Técnica Superior de Ingeniería of the Universidad de Sevilla and Eva Ortega Paíno, secretary-general for research of the ministry for science, innovation and universities. The ceremony was officially closed by the vice-rector, María Isabel Hartillo Hermoso.

The press office of the 9ECM was handled by the agency "Divulga", headed by Ignacio Fernández Bayo. This agency was in charge of the press offices of previous important mathematical events in Spain: the ICM 2006, the IMO 2008 and the ICIAM 2019. They took care of the national and international press coverage of the Congress as well as the communication during the congress. Scientific advisement to the press office was done by Francisco Torres de Lizaur, from the Universidad de Sevilla, and social media communications were in the hands of Ramón Flores, from the Universidad de Sevilla. The official agency of the Congress was "Pacífico", with a long experience in large scientific events and a magnificent and experienced team.

Weather conditions in Sevilla suggested providing meals and coffee breaks at the conference venue. This was a challenge in view of the large number of participants. The task was perfectly done by a local provider, delivering a total number of 5,000 meals during the conference week.

Despite the bad weather forecasts (given the location and dates of the congress), the expected financial fluctuations, and other organizational setbacks, it turned out that the 9th European Congress of Mathematics was a complete success.

Organizing a congress of this magnitude is a colossal task, requiring the work and involvement of a large number of people. On behalf of the organizing committee, we want to express our deep gratitude to all collaborators, volunteers, and organizations that have supported this project from the beginning, and to the sponsors and public bodies that have provided funding, without whom we could never have successfully carried out this task.

The congress has been a window through which Europe and the world have been able to appreciate the level and organizational capacity of the Spanish, and particularly Andalusian, mathematical community.

9ECM committees

Honorary committee
His Majesty King Felipe VI of Spain (president)
Pedro Sánchez Pérez-Castejón, prime minister of Spain
Diana Morant Ripoll, minister for science, innovation and universities
Juan Manuel Moreno Bonilla, president of the regional government of Andalusia
José Luis Sanz Ruiz, mayor of Seville
José Carlos Gómez Villamandos, counsellor of university, research and innovation of the regional government of Andalusia
Francisco Javier Fernández de los Ríos Torres, president of the provincial government of Seville
Miguel Ángel Castro Arroyo, rector, University of Seville
Pedro Mercado Pacheco, rector, University of Granada
Casimiro Mantell Serrano, rector, University of Cádiz
José J. Céspedes Lorente, rector, University of Almería
Jan Philip Solovej, president of the European Mathematical Society
Alfonso Gordaliza Ramos, president of the Spanish Committee for Mathematics
Eva Gallardo Gutiérrez, president of the Royal Spanish Mathematical Society
Manuel J. Castro Díaz, president of the Spanish Society of Applied Mathematics
Montserrat Alsina Aubach, president of the Catalan Mathematical Society
Begoña Vitoriano Villanueva, president of the Spanish Society of Statistics, Operations Research and Data Science
Antonio Pulido Gutiérrez, president of the Cajasol Foundation
Raimundo Pérez-Hernández y Torra, director of the Ramón Areces Foundation

Scientific committee
Leonid Polterovich (chair), Tel Aviv University
Kari Astala, University of Helsinki
Karine Beauchard, ENS Rennes
Tara Brendle, University of Glasgow
Barbara Fantechi, SISSA
Isabel Fernández, University of Seville
Francisco Gancedo, University of Seville
María Ángeles Gil, University of Oviedo
Bernhard Keller, Paris Cité University
Pär Kurlberg, KTH Royal Institute of Technology

Mariusz Lemańczyk, Nicolaus Copernicus University in Toruń
George Marinescu, University of Cologne,
"Simion Stoilow" Institute of Mathematics of the Romanian Academy
Martin Markl, Institute of Mathematics of the Czech Academy of Sciences
Bojan Mohar, Simon Fraser University, University of Ljubljana
Ilaria Perugia, University of Vienna
Jonathan Pila, University of Oxford
Bálint Tóth, Alfréd Rényi Institute of Mathematics and University of Bristol

Local scientific committee
Juan González-Meneses (chair), University of Seville
Sara Arias de Reyna, University of Seville
Guillermo P. Curbera, University of Seville
Elena Fernández, University of Cádiz
José Antonio Gálvez, University of Granada
María Isabel Garrido, Complutense University of Madrid
María Ángeles Japón, University of Seville
Mariano Mateos, University of Oviedo
Rosa María Miró, University of Barcelona
Tomás Morales de Luna, University of Málaga
Juan José Moreno, University of Almería
David Orden, University of Alcalá
Victoria Otero, University of Santiago de Compostela
María del Carmen Reguera, University of Málaga
David Ruiz, University of Granada
Enric Ventura i Capell, Polytechnic University of Catalonia
Begoña Vitoriano, Complutense University of Madrid

EMS prize committee
Paver Exner (chair), Institute of Nuclear Physics, Czech Academy of Sciences, Czech Technical University
José Antonio Carrillo de la Plata, University of Oxford
Heike Faßbender, TU Braunschweig
Elena Fernández Areizaga, Universidad de Cádiz
Ben Green, University of Oxford
Kathryn Hess Bellwald, EPFL
Juliette Kennedy, University of Helsinki
Tomasz Łuczak, Adam Mickiewicz University in Poznań
Tere Martinez Seara, Universitat Politècnica de Catalunya

Volodymyr Mazorchuk, Uppsala University
Pierre Pansu, Paris-Saclay University
Giuseppe Savaré, Bocconi University
Kristian Seip, Norwegian University of Science and Technology
Ofer Zeitouni, Weizmann Institute of Science

Felix Klein prize committee
Peregrina Quintela Estévez (chair), University of Santiago de Compostela
Krzysztof Burnecki, Wrocław University of Science and Technology
Patrice Hauret, Michelin
Ralf Korn, RPTU Kaiserslautern–Landau
Karl Kunisch, Johann Radon Institute for Computational and Applied Mathematics

Otto Neugebauer prize committee
Karine Chemla (chair), CNRS, Paris Cité University
Erica Luciano, University of Turin
Antoni Malet, Pompeu Fabra University, CNRS, Paris Cité University
Roman Murawski, Adam Mickiewicz University in Poznań
Peter Ullrich, University of Koblenz

EMS/ECMI Lánczos prize committee
Patrick Farrell (chair), University of Oxford
Jennifer Balakrishnan, Boston University
Elena Celledoni, Norwegian University of Science and Technology
Sharon Glotzer, University of Michigan
William Stein, SageMath
Lawrence Mitchell, NVIDIA
Marc Pfetsch, Technical University of Darmstadt

Paul Lévy prize committee
Wendelin Werner (chair), University of Cambridge
Sourav Chatterjee, Stanford University
Patricia Gonçalves, University of Lisbon
Alice Guionnet, ENS Lyon
Martin Hairer, EPFL
Allan Sly, Princeton University

Organizing committee
Juan González-Meneses (president), University of Seville
Isabel Fernández (vice president), University of Seville
Mirta Castro Smirnova, University of Seville
Guillermo P. Curbera, University of Seville
María Ángeles Rodríguez-Bellido, University of Seville
Antonio Rojas-León, University of Seville
José Alfredo Cañizo, University of Granada
Enrique D. Fernández-Nieto, University of Seville
Fernando Fernandez Sánchez, University of Seville
Francisco Gancedo, University of Seville
Andrés Jiménez-Losada, University of Seville
Victoria Martín-Márquez, University of Seville
María Luz Muñoz-Ruiz, University of Málaga
María Luz Puertas, University of Almería
Magdalena Rodríguez, University of Granada
Alfonso Suárez-Llorens, University of Cádiz
Francisco Torres de Lizaur, University of Seville
Inmaculada Ventura Molina, University of Seville

Satellite committee
José Alfredo Cañizo (chair), University of Granada
Jorge Buescu, University of Lisbon
Isabel Fernández, University of Seville
Francisco Gancedo, University of Seville
María Luz Muñoz-Ruiz, University of Málaga
Beatrice Pelloni, Heriot-Watt University
María Luz Puertas, University of Almería
Alfonso Suárez-Llorens, University of Cádiz

Grant committee
Mirta María Castro Smirnova (chair), University of Seville
Christophe Ritzenthaler, University of Rennes, CIMPA
Lidia Fernández, University of Granada, CIMPA
Martina Magliocca, University of Seville
María Ángeles García Ferrero, Institute of Mathematical Sciences (ICMAT)

Collaborators
Antonio Jesús Arriaza Gómez, University of Cádiz

Alfonso José Bello Espina, University of Cádiz
Alberto Castaño Domínguez, University of Seville
Ramón Flores, University of Seville
Eduardo García Juárez, University of Seville
Alfonso Márquez Martínez, University of Seville
Cristian Morales Rodrigo, University of Seville
Ana Delia Olvera Cervantes, University of Cádiz, Technological University of the Mixteca
Miguel Ortega Titos, University of Granada
Gema Pigueiras Voces, University of Cádiz
José A. Prado Bassas, University of Seville
Marta Sánchez Sánchez, University of Cádiz
José María Tornero Sánchez, University of Seville

Prize winners

EMS prize
Maria Colombo, EPFL
Cristiana De Filippis, University of Parma
Jessica Fintzen, University of Bonn
Nina Holden, New York University
Tom Hutchcroft, California Institute of Technology
Jacek Jendrej, Sorbonne Paris North University
Adam M. Kanigowski, University of Maryland
Frederick Manners, University of California San Diego
Richard Montgomery, University of Warwick
Danylo Radchenko, CNRS, University of Lille

Felix Klein prize
Fabien Casenave, Safran Tech, Digital Sciences & Technologies

Otto Neugebauer prize
Reinhard Siegmund-Schultze, University of Agder

EMS/ECMI Lánczos prize
Patrick Amestoy, Mumps Technologies
Jean-Yves L'excellent, Mumps Technologies
Theo Mary, Mumps Technologies

Paul Lévy prize
Jeremy Quastel, University of Toronto

Program of the congress

Plenary lectures
Martin Bridson, *Chasing finite shadows of infinite groups through geometry*
Annalisa Buffa, *On the impact of geometric simplification in the numerical simulation of partial differential equations*
Maxim Kontsevich, *Generalized Riemann–Hilbert correspondence, wall-crossing and resurgence*
André Neves, *Recent progress in the theory of minimal surfaces*
Eero Saksman, *Gaussian multiplicative chaos in random geometry and analytic number theory*
Benny Sudakov, *Restricted subgraphs of edge-colored graphs and applications*
Fabio Toninelli, *Large-scale behavior of (super-)critical stochastic PDEs*
Vlad Vicol, *Anomalous diffusion*
Anna Wienhard, *A beautiful world beyond hyperbolic geometry: Anosov representations and higher Teichmüller spaces*
Tamar Ziegler, *Sign patterns of the Möbius function*

Abel lecture
Avi Wigderson, *The value of errors in proofs*

Hirzebruch lecture
Étienne Ghys, *Homeomorphisms of the sphere*

Invited lectures
Paola Antonietti, *Mathematical and numerical modeling of neurodegenerative diseases*
Shiri Artstein-Avidan, *Dualities in asymptotic geometric analysis and beyond*
David Asperó and Ralf Schindler, *The continuum problem: Hilbert's first problem revisited*
Balázs Bárány, *Self-affine sets and measures*
Roland Bauerschmidt, *Applications of the renormalisation group*
Dorin Bucur, *A free boundary approach of spectral optimization problems: Extremal Neumann eigenvalues*
Lev Buhovsky, *Courant, Bézout and persistence*
Ángel Castro, *Unstable vortices and non-uniqueness for 2D Euler and α-gSQG*
Vida Dujmovic and David Wood, *Graph product structure: Theory and applications*
Giovanni Forni, *Weakly mixing polygonal billiards*
Nikos Frantzikinakis, *Partition regularity of Pythagorean pairs and more*

Adam Harper, *Low moments of character sums*
David Hernandez, *Symmetries of Grothendieck rings in representation theory*
Barbara Kaltenbacher, *Forward and inverse problems in nonlinear acoustics*
Matti Lassas, *Inverse problems for non-linear partial differential equations*
Ursula Ludwig, *Cheeger–Müller theorem for singular spaces*
Michael Magee, *Optimal spectral gaps*
Svitlana Mayboroda, *Free boundary problems for the partially reflected Brownian motion*
Anton Mellit, *Cohomology rings of character varieties*
Pablo Mira, *Free boundary minimal annuli in the unit ball*
James Newton, *Modularity of Galois representations and generalized class groups*
Steve Oudot, *Differential calculus and optimization in persistence module categories*
Stefanie Petermichl, *The matrix A2 conjecture*
David Pérez-García, *Spectral gaps in quantum computing*
Pavel Pudlak, *Proofs, computational complexity, and games*
Mario Salvetti, *The $K(\pi, 1)$-conjecture for Artin groups*
Ralf Schindler and David Asperó, *The continuum problem: Hilbert's first problem revisited*
Saul Schleimer, *Solving the word problem in the mapping class group in quasi-linear time*
Stefan Schwede, *Universal symmetries: Global-equivariant homotopy theory*
Alessandro Sisto, *An invitation to hierarchical hyperbolicity*
Yilin Wang, *Recent developments on the Loewner energy for Jordan curves*
Michael Wemyss, *Normal forms of noncommutative functions*
David Wood and Vida Dujmovic, *Graph product structure: theory and applications*

Prize lectures
Cristiana De Filippis, *Nonuniformly elliptic Schauder theory*
Jessica Fintzen, *An introduction to representations of p-adic groups*
Nina Holden (online speaker), *Random surfaces*
Tom Hutchcroft, *Probability and the structure of groups*
Jacek Jendrej, *Recent progress on the problem of soliton resolution*
Adam M. Kanigowski, *Ergodic and spectral properties of smooth flows on surfaces*
Frederick Manners, *Rough structure theorems in additive combinatorics and their applications*
Richard Montgomery, *Transversals in Latin squares*
Danylo Radchenko, *Fourier uniqueness pairs and interpolation formulas*
Fabien Casenave, *Learning physics simulations for design processes in the industry*
Reinhard Siegmund-Schultze, *On the 125th anniversary of his birth: My 'encounter' with Otto Neugebauer*

Patrick Amestoy, Jean-Yves L'Excellent and Theo Mary, *MUMPS: Multifrontal Massively Parallel Solver for the direct solution of sparse linear equations*
Jeremy Quastel, *Integrable PDE's for random growth*

Public lectures: Beyond mathematics
José Ferreirós, University of Seville, *Tales from (mathematical) paradise*
José López Barneo, University of Seville, *Search for a neuroprotective therapy for Parkinson's disease*
Eleonora Viezzer, University of Seville, *Wave-particle interactions in tokamak fusion plasmas*
Montserrat Vilá, Doñana Biological Station (EBD-CSIC), *Biotic globalization by invasive species*

Special lectures

London Mathematical Society lecture
Heather Harrington, *Application driven topological data analysis*

Outreach lecture
Saul Schleimer, *Puzzling the 120-cell*

Minisymposia

Analysis, approximation and control of chemotaxis models
María Ángeles Rodríguez-Bellido, University of Seville
Antonio Suárez Fernández, University of Seville
Cristian Morales-Rodrigo, University of Seville

Analytical, computational and geometrical approaches to inverse problems
Pedro Caro, Basque Center for Applied Mathematics
Tony Liimatainen, University of Helsinki
Samuli Siltanen, University of Helsinki

Applied and combinatorial topology
José Antonio Vilches, University of Seville
Rocío González Díaz, University of Seville
Enrique Macías-Virgós, University of Santiago de Compostela

Asymptotic convex analysis
Bernardo González Merino, University of Murcia
Carlos Hugo Jiménez Gómez, University of Seville
Julián Haddad, University of Seville

Calculus of variations problems in materials science
Arghir-Dani Zarnescu, Basque Center for Applied Mathematics
Maria-Carme Calderer, University of Minnesota

Collective phenomena of Fermionic systems
Jinyeop Lee, University of Basel
Emanuela L. Giacomelli, LMU Munich

Combinatorial designs
Raúl Manuel Falcón Ganfornina, University of Seville
Félix Gudiel Rodríguez, University of Seville
Víctor Álvarez Solano, University of Seville

Complex, quaternionic and exceptional geometry
Fabio Vlacci, University of Trieste
Simon Salamon, King's College London
Chiara De Fabritiis, Marche Polytechnic University

Computational group theory: Algorithmic problems and cryptographic applications
María Cumplido, University of Seville
Marialaura Noce, University of Salerno

Computer algebra and mathematical interactions
Francesca Cioffi, University of Naples Federico II
Cristina Bertone, University of Turin

Current developments in mathematical fluid dynamics
Daniel Lear, University of Cantabria
Diego Alonso-Oran, University of La Laguna
Rafael Granero-Belinchon, University of Cantabria

D-modules and applications
Alberto Castaño Domínguez, University of Seville
María Cruz Fernández Fernández, University of Seville

Dynamical aspects of nonlinear problems in reaction-diffusion equations, Hamiltonian systems, and ODEs
Andrea Tellini, Technical University of Madrid
Guglielmo Feltrin, University of Udine
Elisa Sovrano, University of Modena and Reggio Emilia

EU-MATHS-IN: Success stories of mathematical technologies in societal challenges and industry
Manuel Cruz, LEMA-ISEP-IPP and EU-MATHS-IN

Barbara Kaltenbacher, European Mathematical Society and University of Klagenfurt
Matthias Ehrhardt, University of Wuppertal and EU-MATHS-IN

Extremal and probabilistic combinatorics
Richard Lang, University of Hamburg
Nina Kamčev, University of Zagreb

Financial and insurance risks
Marta Sánchez Sánchez, University of Cádiz
Alfonso José Bello Espina, University of Cádiz
Gema Pigueiras Voces, University of Cádiz

Fixed point theory and applications to fractional ordinary and partial difference and differential equations
Erdal Karapinar, Çankaya University
Adrian Petrusel, Babeș-Bolyai University
Marija Cvetković, University of Niš

Function spaces and related topics
Luboš Pick, Charles University
Lenka Slavíková, Charles University
Andrea Cianchi, University of Florence

Generalizations of complex analysis and applications
Eckhard Hitzer, International Christian University
Dmitrii Legatiuk, University of Erfurt
Sebastià Xambó, Polytechnic University of Catalonia

Geometric and regularity aspects of sub-elliptic PDEs: From kinetic equations to Hormander's vector fields
Francesca Anceschi, Marche Polytechnic University
Mirco Piccinini, University of Parma
Annalaura Rebucci, Max Planck Institute for Mathematics in the Sciences

Geometric data science presenting days
Vitaliy Kurin, University of Liverpool

Geometry of Banach and metric spaces and fixed point theory
Łukasz Piasecki, Maria Curie-Skłodowska University in Lublin
Jeimer Villada, Maria Curie-Skłodowska University in Lublin
María Japón Pineda, University of Seville

Geometry, algebra and asymptotic analysis of differential equations and dynamical systems
Fernando Sanz Sánchez, University of Valladolid

Javier Sanz Gil, University of Valladolid
Alberto Fernández Boix, University of Valladolid

Groups, skew braces and the Yang–Baxter equation
Marco Trombetti, University of Naples Federico II
Vincent Pérez-Calabuig, University of Valencia
Marzia Mazzotta, University of Salento

Interfaces between interacting particle systems and PDEs: From analysis to applications
Hui Huang, University of Graz
Alexandra Holzinger, University of Oxford

Inverse Galois theory and arithmetic
Tony Shaska, Oakland University
Sara Arias De Reyna, University of Seville

Knot theory and low dimensional manifolds
Marithania Silvero, University of Seville
Lukas Lewark, ETH Zurich
Mark Powell, University of Glasgow

Linear algebra and matrix analysis
Fernando De Terán, Charles III University of Madrid
Alicia Roca, Technical University of Valencia
Ángeles Carmona, Polytechnic University of Catalonia

Mathematical aspects on reduced order modeling
Enrique Delgado Ávila, University of Seville
Alejandro Bandera Moreno, University of Seville
Tomás Chacón Rebollo, University of Seville

Mathematical challenges in inverse problems
Barbara Kaltenbacher, University of Klagenfurt
Eric Bonnetier, Grenoble Alpes University
Thorsten Hohage, University of Göttingen

Mathematical methods for biological problems
Ana Carpio, Complutense University of Madrid

Mathematical methods in reliability analysis
Nuria Torrado, Autonomous University of Madrid
Antonio Arriaza, University of Cádiz
Julio Mulero, University of Alicante

Mathematical modeling in life sciences
Adérito Araújo, University of Coimbra
Andrea Pugliese, University of Trento

MIDAS – Mathematical information in the digital age of science
Olaf Teschke, FIZ Karlsruhe, EMS Committee on Publications and Electronic Dissemination, zbMATH Open, EOSC
Evelyne Miot, CNRS, EMS Committee on Publications and Electronic Dissemination, Institut Fourier, Grenoble Alpes University, MathDoc
Moritz Schubotz, FIZ Karlsruhe, zbMATH Open, EOSC

Geometry of submanifolds, celebrating Bang-Yen Chen's 80th anniversary
Bogdan Suceavă, California State University
Mihaela Vajiac, Chapman University
Alfonso Carriazo, University of Seville

New mathematical approaches in developmental biology
Elena Camacho Aguilar, Andalusian Centre for Developmental Biology
Fernando Casares, Andalusian Centre for Developmental Biology
Luciano Marcon, Andalusian Centre for Developmental Biology

New trends in calculus of variations
Virginia De Cicco, Sapienza University of Rome
Anderson Melchor Hernandez, Polytechnic University of Turin
Valeria Chiadò Piat, Polytechnic University of Turin

New trends in control and inverse problems
Anna Doubova, University of Seville

New trends in many particle and mean field models
Idriss Mazari-Fouquer, CEREMADE, Paris Dauphine University – PSL
Domènec Ruiz-I-Balet, Imperial College London
Laura Kanzler, CEREMADE, Paris Dauphine University – PSL

Nonlinear and nonlocal PDE in biology and collective behavior
José Cañizo, University of Granada, IMAG
Bruno Volzone, Parthenope University of Naples

Non-local operators and stochastic processes: Fractional processes, random fields, and applications
Enrica Pirozzi, University of Naples Federico II
Enrico Scalas, Sapienza University of Rome

Nonlocality in hyperbolic PDEs
Felisia Angela Chiarello, University of L'Aquila

Lukas Pflug, University of Erlangen–Nuremberg
Alexander Keimer, University of Erlangen–Nuremberg

Nonstandard dualities, associated dynamics, and algorithms
Patrick Combettes, North Carolina State University
Michel De Lara, ENPC

Operator algebras
Joan Bosa Puigredon, University of Zaragoza
Antonio Peralta, University of Granada
Francesc Perera, Autonomous University of Barcelona

Operator semigroups and evolution equations: Theory and applications
Christian Budde, University of the Free State

Orthogonal polynomials and special functions
Mirta María Castro Smirnova, University of Seville
Ignacio Nahuel Zurrián, University of Seville
Manuel Domínguez De La Iglesia, National Autonomous University of Mexico

Particles unleashed: From the theory of interacting particle systems to applications in optimization, data science, and machine learning
Konstantin Riedl, Technical University of Munich
Tim Roith, Helmholtz Imaging, DESY
Michael Herty, RWTH Aachen University

Qualitative theory and applications of boundary value problems for nonlinear ODEs and PDEs
Paola Rubbioni, University of Perugia
Serena Matucci, University of Florence
Radu Precup, Babeş-Bolyai University

Reducing the computational complexity of eigenvalue problems arising from partial differential equations
Daniele Boffi, King Abdullah University of Science and Technology
Fleurianne Bertrand, Chemnitz University of Technology

Regularisation by noise
Mate Gerencser, TU Wien

Rings, modules, and factorizations
Laura Cossu, University of Graz
Alfred Geroldinger, University of Graz
Mara Pompili, University of Graz

Shared horizons: Collaborations between Europe and the global South
Balazs Szendrői, University of Vienna
Sophie Dabo-Niang, University of Lille
Christophe Ritzenthaler, University of Rennes, CIMPA

Special functions, orthogonal polynomials, Q-series and applications
Howard Cohl, National Institute of Standards and Technology
Robert Maier, University of Arizona
Roberto Costas-Santos, Loyola University Andalusia

Spectral geometry and related topics
Luigi Provenzano, Sapienza University of Rome
Davide Buoso, University of Eastern Piedmont
Katie Gittins, Durham University

Spectral theory of differential operators in mathematical physics
Pavel Exner, Nuclear Physics Institute, Czech Academy of Sciences, Czech Technical University
Konstantin Pankrashkin, Carl von Ossietzky University of Oldenburg

Stochastic and harmonic analysis
Li Chen, Aarhus University
Fabrice Baudoin, Aarhus University

Stochastic dynamical systems
Ting Gao, Huazhong University of Science and Technology
Jinqiao Duan, Great Bay University

Stochastic processes in cell biology
Yann Lanoiselée, Basque Center for Applied Mathematics
Jakub Ślęzak, Wrocław University of Science and Technology

Symmetry methods and other approaches for nonlinear partial differential equations with applications
Tamara M. Garrido Letrán, University of Cádiz
Maria Luz Gandarias, University of Cádiz
Almudena P. Marquez, University of Cádiz

The mathematics of data-driven reduced-order modeling for dynamical systems
Zoran Tomljanovic, University of Osijek
Ion Victor Gosea, Max Planck Institute for Dynamics of Complex Technical Systems

Towards the regularity theory, PDEs free boundaries and interfaces
Edgard Pimentel, University of Coimbra
Makson Santos, Instituto Superior Técnico

Variational and topological methods in PDEs
Jacopo Schino, University of Warsaw
Bartosz Bieganowski, University of Warsaw

Variational methods and nonlinear analysis
Ana Cristina Barroso, University of Lisbon
Elvira Zappale, Sapienza University of Rome

Women in control
Lorena Bociu, North Carolina State University

Contributed sessions
Algebra
Algebraic and complex geometry
Analysis and its applications
Combinatorics and discrete mathematics
Decision, optimization and control
Differential geometry and applications
Dynamical systems, ordinary differential equations and applications
EMYA lightning talks by young participants
General topics
History of mathematics, mathematics education, popularization of mathematics
Logic and mathematical aspects of computer science
Mathematical physics
Mathematics in science and technology
Number theory
Numerical analysis, scientific computing
Partial differential equations and applications
Probability, statistics and financial mathematics
Topology

Panel discussions

How to enhance mathematical competence? Lessons from great-scale assessments
Organized by the European Mathematical Society and the Royal Spanish Mathematical Society
Participants:
Luis J. Rodríguez-Muñiz (chair), vice-president of the Royal Spanish Mathematical Society, University of Oviedo
Nuria Climent, president of the Spanish Society of Research on Mathematics Education, University of Huelva
Susana Carreira, University of Algarve

Giorgio Bolondi, Free University of Bozen-Bolzano
Andreas Eichler, University of Kassel

Mathematics and artificial intelligence
Organized by the Basque Center for Applied Mathematics
Participants:
Luis Vega (chair), Basque Center for Applied Mathematics
Rosa Crujeiras, University of Santiago de Compostela
Matti Lassas, University of Helsinki
Justo Puerto, University of Seville
David Ríos, AXA Chair in Adversarial Risk Analysis, Institute of Mathematical Sciences (ICMAT)
Begoña Vitoriano, Complutense University of Madrid

Translating European mathematics
Organized by the Institute of Mathematical Sciences (ICMAT) and the International Commission on the History of Mathematics (IMU)
Participants
Norbert Schappacher (chair), IRMA
June Barrow-Green, The Open University and London School of Economics, president of the International Commission on the History of Mathematics (IMU)
Frances Goldman, English editor in the LMS of the Russian translation journals
François Lê, member of the "The Brill–Noether report on the theory of algebraic functions (1894): translation and analysis project"

European Research Council panel
Organized by the European Research Council
Participants:
Maria Gonzalez Cerveron (chair), ERCEA Mathematics Panel coordinator, European Research Council
Annalisa Buffa, AdG panel member, StG and AdG grantee
Giovanni Forni, AdG grantee
Stefanie Petermichl, Consolidator grantee
Jan Philip Solovej, AdG panel member, panel chair and AdG grantee

New scenarios in Open Science
Organized by the Centre de Recerca Matemàtica
Participants:
Miguel Benítez (chair), Basque Center for Applied Mathematics
André Gaul, EMS Press
Ignasi Labastida i Juan, University of Barcelona
Pilar Rico, Spanish Foundation for Science and Technology (FECYT)

Celebrating and supporting women in Mathematics
Organized by the European Women in Mathematics
Participants:
Tara Brendle, vice-president of the Edinburgh Mathematical Society, University of Glasgow
Eva Gallardo Gutiérrez, president of Royal Spanish Mathematical Society, Complutense University of Madrid
Maria Gonzalez Cerveron, European Research Council

Sustainability panel and group discussion
Organized by the European Mathematical Society Young Academy

The role of institutes of mathematics in the advancement of mathematical research
Organized by the Institute of Mathematics of the University of Seville
Participants:
Justo Puerto (chair), director of the Institute of Mathematics of the University of Seville
Lluís Alsedà, director of the Centre de Recerca Matemàtica
Martin Bridson, president of the Clay Mathematics Institute
Rosa Crujeiras, director of the Galician Centre for Mathematical Research and Technology
Miguel Sánchez, director of the Institute of Mathematics of the University of Granada
Anna Wienhard, director of the Max Planck Institute for Mathematics in the Sciences
José A. Lozano, director of the Basque Center for Applied Mathematics
Javier Aramayona, director of the Institute of Mathematical Sciences (ICMAT)

Activities by the European Mathematical Society Young Academy

EMYA lighting talks
EMYA ice breaking session

Activities by the European Women in Mathematics

EWM event
Speakers:
Tara Brendle, University of Glasgow
Jessica Fintzen, University of Bonn
Elena Gaburro, University of Bordeaux
Chiara Saffirio, University of Basel
Maria Yakerson, University of Oxford

EWM panel discussion

EWM exhibition
Exhibitions

Maps: Cartographic heritage in Seville from the 15th to the 18th century[6]
Coordinator: Guillermo Curbera, University of Seville
Cajasol Foundation, June 27 to July 20, 2024
The exhibition featured a selection of maps and books from the General Archive of the Indies, the Columbina Library, and the old collection of the University of Seville Library. To complement this exhibition, three conferences were held, given by Antonio Durán (Professor of mathematical analysis at the University of Seville), Consuelo Varela (former director of the School of Spanish-American Studies and the Alcázar of Seville), and Antonio Sánchez Martínez (Professor of history and philosophy of science at the Autonomous University of Madrid). More than three thousand visitors enjoyed the exhibition during those days.

Nudos: A journey into knot theory[7]
Coordinators: Juan González-Meneses, Marithania Silvero and Francisco Vega, University of Seville
Higher Technical School of Engineering of the University of Seville (congress venue)
A collection of 3D models of knots and interactive links to guide the visitor through the amazing world of knot theory, from basic concepts to some of the most intriguing features. Created by Juan González-Meneses, Marithania Silvero, and Francisco Vega.

EWM exhibition: Celebrating and supporting women in mathematics
Coordinator: European Women in Mathematics
Higher Technical School of Engineering of the University of Seville (congress venue)
Biographies and photographs of 39 women researchers in mathematics from around the world.

A geometric walk through the Alhambra
Coordinator: Miguel Ortega Titos, University of Granada
Higher Technical School of Engineering of the University of Seville (congress venue)
This exhibition analyzes the decoration of the Alhambra palaces from a geometric perspective, attempting to discover some of the mathematical secrets used by Nasrid artisans: the use of proportions, Nasrid decorative elements, and examples of crystallographic groups.

[6] https://expobus.us.es/s/mapas/page/mapas-index
[7] https://institucional.us.es/exponudos/

Satellite conferences

Towards infinite dimension and beyond in quantum information
May 5 to May 10, 2024, Granada, Spain

Generalized geometry meets string theory
May 12 to May 17, 2024, Granada, Spain

Positive solutions of polynomial systems arising from real-life applications
May 19 to May 24, 2024, Granada, Spain

Bridging the gap: The importance of communicating mathematical research to laypeople and in education
May 21 to May 24, 2024, Mallorca, Spain

Climate-inclusive ecosystem modeling: Understanding the dynamics of ecosystems in a changing world
May 27 to May 29, 2024, Barcelona, Spain

From evolution to bioengineering of biological patterning mechanisms. Mathematical advances and challenges
June 2 to June 7, 2024, Granada, Spain

BAC24 Barcelona analysis conference
June 3 to June 7, 2024, Barcelona, Spain

Functor and tensor categories, models and systems
June 3 to June 7, 2024, Santiago de Compostela, Spain

PDE methods in machine learning: From continuum dynamics to algorithms
June 9 to June 14, 2024, Granada, Spain

Methods in Banach spaces
June 10th to June 14th, 2024, Badajoz, Spain

Geometry of field theories
June 12 to June 14, 2024, Madrid, Spain

From classical to modern analysis: In memory of professor José Carlos Petronilho
June 24 to June 28, 2024, Sanlúcar de Barrameda, Spain

EACA2024
June 24 to June 26, 2024, San Lorenzo de El Escorial, Spain

Topology, algebra, and categories in logic (TACL)
June 24 to July 5, 2024, Barcelona, Spain

European school of differential geometry (ESDG)
June 24 to July 5, 2024, Granada, Spain

Interactions between geometry, algebra and topology
June 25 to July 5, 2024, Barcelona, Spain

Perspectives in representation theory
July 1 to July 3, 2024, Coimbra, Portugal

Applied nonlinear dynamical systems and chaos. Celebrating Stephen Wiggins's 65th birthday
July 1 to July 3, 2024, Madrid, Spain

Discrete mathematics days 2024
July 3 to July 5, 2024, Alcalá de Henares, Spain

Alhambra PDE days
July 8 to July 12, 2024, Granada, Spain

International meeting on numerical semigroup 2024 (IMNS2024)
July 8 to July 12, 2024, Jerez de la Frontera, Spain

Recent advances in harmonic analysis
July 8 to July 12, 2024, Málaga, Spain

Combinatorial designs and codes (CODESCO 2024)
July 8 to July 12, 2024, Sevilla, Spain

Integrability and moduli, a conference in honour of Leon Takhtajan
July 8 to July 12, 2024, Lisbon, Portugal

2nd international workshop on mathematics and physical sciences
July 11 to July 12, 2024, Évora, Portugal

Regularity theory and free boundary problems: From PDE to interfaces
July 22 to July 26, 2024, Coimbra, Portugal

32nd international fall workshop on geometry and physics
September 2 to September 5, 2024, Coimbra, Portugal

Sponsors and exhibitors

Main sponsors
Fundación Ramón Areces
Foundation Compositio Mathematica
Fundación Cajasol

Gold sponsor
American Mathematical Society

Exhibitors
American Mathematical Society
Basque Center for Applied Mathematics
Cambridge University Press
Centre de Recerca Matemàtica
De Gruyter
Elsevier
European Mathematical Society
EMS Press
Heidelberg Laureate Forum Foundation
Instituto de Matemáticas de la Universidad de Sevilla
London Mathematical Society
Red Española Matemática-Industria (math-in)
Mathematics in France
Oxford University Press
Princeton University Press
Project Euclid
Sociedad Española de Matemática Aplicada
Societat Catalana de Matemàtiques
Springer Nature
The Royal Society Publishing

Other sponsors
Basque Center for Applied Mathematics
Casio Educational Division
Centre de Recerca Matemàtica
Instituto de Ciencias Matemáticas
Consejo Superior de Investigaciones Científicas

Ministerio de Ciencia, Innovación y Universidades
Fundación Española de Ciencia y Tecnología

Institutional support beta
Universidad Pública de Navarra - Departamento de Estadística, Informática y Matemáticas
Universidad de Cádiz - Departamento de Estadística e Investigación Operativa
Universidad de Sevilla – Departamento de Álgebra

Institutional support gamma
Centro de Investigación y Tecnología Matemática de Galicia (CITMAga)
Real Sociedad Matemática Española
Universidad de Almería, Facultad de Ciencias Experimentales
Universidad de Almería, Departamento de Matemáticas
Universidad Carlos III de Madrid, Departamento de Matemáticas
Universitat Politècnica de Catalunya, Departamento de Matemáticas
Universidad de Granada, Departamento de Geometría y Topología
Universidad de Sevilla, Departamento de Matemática Aplicada I
Universidad de Sevilla, Departamento de Matemática Aplicada II

Institutional support
Universidad de Extremadura, Departamento de Matemáticas
Universidad de Santiago de Compostela, Departamento de Matemática Aplicada

Opening ceremony

Juan González-Meneses, president of the 9ECM
Isabel Fernández, vice-president of the 9ECM

President of the European Mathematical Society,
vice-chansellor of university, research and innovation of the government of Andalusia,
general secretary of research and innovation of the government of Andalusia,
deputy mayor of Seville,
vice rector of research of the University of Seville,
presidents and directors of mathematical societies, faculties and institutes of research,
chairs and members of committees,
dear participants,

The day arrived at last! Welcome to the 9th European Congress of Mathematics! With the collaboration of many people and institutions, we have put together a program for the conference that serves as a showcase for European mathematics. In the mornings, you will enjoy the plenary lectures, together with the Abel and the Hirzebruch lectures. After lunch, prize lectures and invited lectures, plus the minisymposia, contributed sessions and poster sessions. But there is more, each afternoon we will have two panel discussions, and a public lecture about a scientific topic closely related with mathematics. Last, but not least, the outreach lecture, and four exhibitions. Three of them in the main venue of the congress, and one of them, with cartographic gems from the 15th to 18th centuries, in the city center.

We would like to thank the European Mathematical Society for the trust placed in us. Four years ago, the EMS Council decided to bring the European Congress of Mathematics to our city, to Seville. And we are extremely grateful for that. In return, we also want to bring Seville into the congress. You will experience the historic heritage of Seville during the cultural visit to the Real Alcázar. During the congress dinner, you will be immersed in the María Luisa park, a historic garden full of stunning buildings, which is also one of the green lungs of our city. You will see Seville even in the typography of your badges, which mimics the lettering of the street signs in the city.

We would like to thank all the people that collaborated with us to prepare the program for the congress: speakers, members of the committees, organizers of minisymposia and panel discussions, chairs, and of course, our volunteers. Thank you so much to all of them.

All this would not have been possible without the support of many people and institutions. The King of Spain presides the honorary committee. The government of Spain, the government of Andalusia, the council of the province of Seville, and the town hall of Seville, have all contributed to the success of this congress. The univer-

sities of Granada, Cádiz, Málaga and Almería configure the presence in this congress of the Andalusian universities. Let me personally thank the impressive work of Mirta Castro, Guillermo Curbera, María Ángeles Rodríguez Bellido, Antonio Rojas, and each and every member of the organizing committee. The help of sponsors like the Spanish Foundation for Science and Technology (FECYT), the American Mathematical Society, Foundation Compositio Mathematica, Ramón Areces Foundation and Cajasol Foundation, has been invaluable.[8] And, of course, the support that the organizers have received from our own university, the University of Seville, is worth mentioning.

But we have been overwhelmed by the response of the mathematical community:

- Mathematical societies from Spain, RSME, SeMA, SCM, SEIO, backed our candidacy and support the congress in many ways, as well as institutions like the Spanish National Research Council (CSIC).
- The Institute of Mathematics of the University of Seville, and other Spanish Research Institutes, IMAG, ICMAT, BCAM, CRM are responsible for the organization of really interesting panel discussions. Special thanks go to Magdalena Rodríguez, IMAG and the University of Granada for the stunning organization of the EMS Council in Granada.
- 27 satellite events have been celebrated, all around Spain and Portugal.
- The Spanish Committee for Mathematics, together with faculties from all over Spain, starting by the faculty of mathematics of Seville, made it possible that we can count with 100 volunteers, from many different places in Spain, to help all of you during this week.
- The solidarity of departments, faculties, and individual mathematicians, together with the EMS and CIMPA, allowed us to offer 120 grants for people having difficulties to attend the congress. We are very proud of that.
- And finally, all of you, more than 1,300 participants who are here in Seville attending the congress, with nearly 1,000 talks and posters. Those of you who accepted the invitation to give a lecture, and those who will participate in any other way, you are the reason why this congress exists.

We have no words to express our gratitude to all of you for this support. The best we can do is to say a huge "thank you", and assure you that we worked really, really hard to make this congress a success.

Right in front of where we are, there is a replica of the Nao Victoria, the ship that in the 16th century set sail from the port of Seville and completed the first circumnavigation of the world, the greatest maritime feat of all time. Just as that ship

[8] All sponsors are listed in section Sponsors and exhibitors

opened new paths, we hope this congress will forge new connections and significant advancements.

We hope this will not be your last visit to Seville, we hope you come back to our university, to our Institute of Mathematics, the IMUS, to work together, to reinforce the bonds that will be formed during this congress. We welcome you to our city. We hope that, from today, it will become a bit of your own too. Thank you very much.

Jan Philip Solovej, president of the European Mathematical Society

Honored guests,
dear colleagues,

It is my great honor and immense pleasure to welcome you today to the 9th ECM.

This congress offers a fantastic opportunity, as has already been mentioned, to celebrate mathematics in all its beauty, all its depth, and all its diversity. We have a fantastic and fascinating program ahead of us.

Soon, we will learn the names of the prize winners of the ten European Mathematical Society prizes. We will also learn the recipient of the Felix Klein prize for mathematical applications, as well as the winner of the Otto Neugebauer prize in the history of mathematics. In addition, we will announce the prize winners of the EMS/ECMI Lánczos prize in mathematical software and the recipient of the Paul Lévy prize in probability. The last two prizes are new, and we are looking forward to celebrating their winners.

The scientific program offers a rich and diverse mix of plenary lectures, invited lectures, contributed talks, minisymposia, and panel discussions. I would also like to highlight events organized by the association of European Women in Mathematics and by the recently founded European Mathematical Society Young Academy, EMYA.

Of course, we should not forget the social events, which provide valuable opportunities for interaction and exchange. I would like to thank the scientific committee and all the prize committees for their tremendous efforts and hard work. I also wish to thank all the speakers for accepting our invitations. I am equally grateful to all the prize winners for accepting their awards. I would like to thank the University of Seville for hosting us, and I wish to express a very special thanks to the local organizing committee—Juan, Isabel, and their entire team. They have carried out an enormous amount of work, and they truly deserve all the credit. Last but not least, I want to thank all of you for participating in this congress. It is wonderful to see such a large and vibrant audience. Although I regret that I cannot attend all the events myself, I am delighted that you are all here, contributing to what I am confident will be a truly outstanding congress.

Lorena Garrido Serrano, deputy minister of the department of university, research, and innovation of the regional government of Andalusia

It is an honor to take part in the opening ceremony of this 9th European Congress of Mathematics on behalf of the regional government of Andalusia. I would like to take this opportunity to welcome all the participants and organizers of the congress to Andalusia. This land is friendly and welcoming, so I am sure it will leave you with a satisfying memory when your stay comes to an end. Apart from being a welcoming land, it is a community with a deep commitment to science and knowledge, as evidenced daily by the work of our entire university system.

We are proud that this congress returns to Spain and is held for the first time in Andalusia. Our community contributes significantly to the scientific production in Spain, representing over an 11% of the one of the European Union. Andalusia enjoys a reputation for excellence in mathematics research throughout most of its territory, centered around the Andalusian Interuniversity Institute of Mathematics, which is now a reality, as a result of the commitment of our department of university, research, and innovation and the universities of Seville and Granada. Specifically, the mathematics institutes of the universities of Seville and Granada are our two flagships in scientific production in this field.

In the case of Seville, I hope that it will soon be awarded the seal of excellence for which it has been striving for years, accumulating merits in a growing list of achievements. Evidence of this is the position of both institutes, Granada and Seville, in international rankings. From Andalusia, we are already working on the creation of new units of excellence that will serve as stepping stones towards obtaining the national seals of excellence named María de Maeztu or Severo Ochoa. This is our commitment to the well-developed research strength of this land in the field of mathematics.

Our responsibility towards the sciences does not end at the peak of the research career; rather, we are convinced that success is built from the ground up, with an investment of time and dedication that integrates with the future research community. Therefore, we have updated our offering of university degrees after remaining unchanged for 14 years. And we have done this with a deep commitment to sciences, resulting in a variety of degrees where 50% of the undergraduate, master's, and doctoral programs are in STEM fields.

Nevertheless, there is still much to do. We need to work on fostering scientific vocations among our schoolchildren, not just teenagers, but even at earlier stages. We need to work towards increasing the presence of women in science, whose day we celebrate each year with the commitment to continue addressing that gender gap. I'll keep it brief. I just want to emphasize my hope that this congress will be productive

for all, and that you end up convinced that Andalusia is a major laboratory in Southern Europe. Thank you very much.

Álvaro Jesús Pimentel Siles, third deputy mayor of Seville

President of the organizing committee and members of the various committees of the Congress,
president of the European Mathematical Society,
deputy minister of university, research, and innovation of the regional government of Andalusia,
vice-chancellors of the University of Seville,
researchers, congress participants, ladies, and gentlemen,

I would like to start by conveying greetings on behalf of the mayor, José Luis Sanz, and extending a warm welcome from the city council of Seville to all participants in this 9th European Congress of Mathematics, hosted by the capital of Andalusia until July 19.

It has already been mentioned in this opening ceremony that this is the second most important congress in the world in this discipline, and that more than 1,300 mathematicians from across Europe will have the opportunity to present and discuss the latest research advances in their respective fields. Therefore, I can only say that from the city council of Seville, we are very proud that our city was chosen for this event which, four years later, follows in the footsteps of Portorož, Slovenia.

Seville aims to be, during this week, the best setting to welcome you and to promote the dissemination of a science that extends beyond the walls of universities and research centers. History, architecture, technology, and craftsmanship are also mathematics.

I am strongly convinced that mathematics is fundamental for the comprehensive development of the individual. It stimulates logic, reasoning, and order. It gives us the capacity to face complex and decisive situations. Therefore, mathematics is more than mere operations with numbers: its presence and impact in everyday life are undeniable.

Discovering the importance of this science for our daily lives is also another goal of this congress. Among the 12 plenary lectures, 64 symposia, or the 270 communications that form the main body of the research, allow me to highlight the lectures, activities for youth, and thematic exhibitions that are part of the congress agenda.

I know that the most important seed you are going to sow during these days goes beyond the impeccable academic dossiers that will be part of the congress records. If even one child participates in a lecture, or a young person takes interest in your exhibitions... this congress will have been a resounding success. The future of society hinges on having well-educated, prepared citizens who contribute to building a

better world through their work. To achieve this, we must not only foster academic understanding but also problem-solving skills, critical thinking, and self-confidence.

In Seville, we know a lot about all I'm talking about. We have two public universities with researchers who are leaders in research and dissemination. We have the most important scientific and technological park in Southern Europe, PCT Cartuja, where all of you will have the opportunity to spend most of these days of the congress. Moreover, you will have the fortune of being in a cutting-edge hub in Europe for technological development, academic training, R&D, and where the society of the future is being shaped.

I want to conclude by congratulating the organizing committee because I know it hasn't been easy to shape such a large-scale event, but they have worked diligently to ensure everything goes perfectly. Also, to the University of Seville and the School of Engineering for being outstanding hosts of this event. And of course, to the awardees in recognition of their excellent contributions in the field of mathematics.

I reiterate that it is a pride to have you in Seville. I hope that your experience in the city during these days will also leave you with the best memories and impressions of a city that, thanks to you, is already considered in the finest history of research and innovation.

Bienvenidos a Sevilla.
Welcome to Seville.
Willkommen in Sevilla.
Vilkamen in Sevilla.
Bienvenue à Séville.
Vianveniú a Sevilla.
Muchas gracias.

Julián Martínez Fernández, vice-rector for research of the University of Sevilla

Good morning to all the researchers gathered here today. Good morning to the authorities and to all colleagues. First of all, on behalf of the rector of the University of Seville, I would like to extend a warm welcome to our university and to our city.

I would also like to express our sincere gratitude to the European Mathematical Society for selecting the University of Seville to host this conference. It is truly an honor to organize the second-largest mathematical event in the world, and we are extremely pleased to have had this opportunity.

This achievement is the result of the outstanding work carried out by the organizing committee, a large and dedicated team composed of many professors from the University of Seville. I see here Juan, Isabel, and many other colleagues and friends who make it possible for Seville, Andalusia, and Spain to play a leading role in mathematical research worldwide. I would like to thank them not only for organizing this

conference, but also for their daily work and tireless commitment to research, which has brought us to our current position.

I would also like to thank the scientific committee, which brings together some of the best researchers in the world. It is a great honor to have you here and to be able to attend your lectures. Scientific meetings such as this one are essential tools for discussion and collaboration, and they are of great importance for institutions like the University of Seville, as they strongly promote research. We all agree that high-quality research is essential for society, and this conference is of outstanding scientific and technological relevance.

The program covers all fields of mathematics, which gives this event particular significance. Moreover, this conference benefits from the strong mathematical structure already established in Andalusia, including the Institute of Mathematics of Seville and the Institute of Mathematics of Granada, which together form the Andalusian Institute of Mathematics, along with the universities of Almería, Cádiz, and Málaga.

This special framework, together with the support of many other institutions – the Royal Spanish Mathematical Society, the Spanish Society of Statistics and Operations Research, the Spanish Society of Applied Mathematics, and many others – has made it possible for more than 1,400 participants to gather here today, along with ten plenary speakers of the highest international standing.

By now, there is no need to emphasize the importance of mathematics. We all recognize that mathematics has accompanied humanity since its origins and has paved the way for understanding the world around us. In recent decades, mathematics has also become a key driver of technological development. We would not have mobile phones, Google Maps, or many of the technologies we use daily without mathematics and algorithms.

Mathematics therefore has an enormous impact on society. A report published a few years ago by IMUS, the Institute of Mathematics of the University of Seville, together with other institutions, showed that every euro invested in mathematical research in Spain generates approximately five euros in societal return.

Mathematics is also becoming increasingly important within our university. Allow me to mention two recent figures. Just a few days ago, the Leiden ranking was published – a ranking I personally value highly, as it is largely based on publications. According to this ranking, the University of Seville increased its share of top publications by 10% over the last year, while citations rose by 9%. More importantly, the university's strongest area continues to be mathematics and computer science, which ranks 127th worldwide.

These figures confirm what we already knew about the importance of mathematics at our university. The University of Seville is a large institution, with 65,000 students, more than 4,000 professors, seven major shared research facilities, and

fifteen research centres. Each year, we produce approximately 5,000 scientific publications and attract around 80 million euros in research funding. Over the last ten years, we have invested 50 million euros in research infrastructure, including the Institute of Mathematics, which I encourage you to visit if you have the opportunity. It offers 1,000 square meters of space to think, reflect, and work on mathematics.

I would like to conclude by encouraging you to make the most of this meeting: to engage in discussion, to exchange ideas, and to produce excellent science. And of course, do not forget to enjoy the city –especially after 9 p.m., as during the day I believe you are better off here.

Once again, congratulations to the entire mathematical community on this remarkable achievement. It is my pleasure to declare the conference officially open. Thank you very much.

Pavel Exner, chair of the EMS prize committee

Dear colleagues, distinguished guests, and dear friends,

I have no intention of competing in dance after that beautiful performance. We now move on to something entirely different. It is my task to announce and present the recipients of the EMS prizes.

In total, ten prize winners are represented here. Two are absent for a reason that may be more important than any modern academic prize: the birth of their newborn children. Before announcing the winners, let me share a few facts about the process.

First, a word about the prize committee. The committee is appointed by the EMS executive committee in accordance with the statutes, and, as you may know, its composition remains confidential until the prize ceremony. The time has now come to reveal its members.

The members of the prize committee were José Antonio Carrillo de la Plata, Heike Fassbender, Elena Fernández Areizaga, Ben Green, Kathryn Hess Bellwald, Juliette Kennedy, Tomasz Łuczak, Volodymir Mazorchuk, Pierre Pansu, Giuseppe Savaré, Teresa Martínez Seara, Kristian Seip, Ofer Zeitouni, and myself as the chair.

Let me now say a few words about the prize itself. The EMS prize was established at the founding of the European Mathematical Society and was first awarded at the inaugural congress in Paris. It quickly gained great prestige, as illustrated by the fact that in almost every group of ten prize winners, one, two, or even three later received the Fields Medal. Out of the 80 laureates to date, 15 have gone on to become Fields medalists.

For the current cycle, we received 55 eligible nominations, significantly more than four years ago. This was a source of both great joy and great frustration: joy, because it reflected the outstanding mathematical work being done by young researchers, and

frustration, because it required us to make many difficult choices. In the end, several excellent candidates could not be selected.

Without further ado, let us now turn to the prize winners, whom I will announce in alphabetical order.

Maria Colombo (EPFL, Switzerland), for her innovative results in fluid dynamics, optimal transport and kinetic theory, and for their impact on analysis in general.
(As mentioned earlier, she is one of the two recipients who could not attend, and her prize will be accepted on her behalf by her colleague, Lucio Galli.)

Cristiana de Filippis (University of Parma, Italy), for her outstanding contributions to elliptic regularity, particularly Schauder estimates for non-uniform elliptic equations and non-differentiable variational integrals, and minima of quasiconvex integrals.

Jessica Fintzen (University of Bonn, Germany), for her transformative work on the representation theory of p-adic groups, particularly for her spectacular demonstration that Yu's construction of supercuspidal representations is exhaustive.

Nina Holden (New York University, USA), for her profound contributions to probability theory and its applications to statistical physics, including results linking Liouville quantum gravity, Schramm–Loewner evolution, and random triangulations.
(She is the second recipient unable to attend, and her prize will be accepted by her uncle, Helge Holden, himself a distinguished mathematician.)

Tom Hutchcroft (California Institute of Technology, USA), for his revolutionary contributions to probability theory and geometric group theory, particularly to percolation theory on general graphs, using tools from geometry, operator theory, group theory, and functional analysis.

Jacek Jendrej (Sorbonne Paris North University, France), for his innovative proofs of the soliton resolution conjecture and the collision problem of two solitons for equivariant wave functions, developing new approaches using ideas from dynamical systems theory to describe the behavior of solutions near a multi-soliton configuration.

Adam M. Kanigowski (University of Maryland, USA), for his outstanding contributions to spectral classification and mixing properties of slowly chaotic dynamical systems.

Frederick Manners (University of California San Diego, USA), for his notable contributions to additive combinatorics and related areas, particularly to the foundations of higher-order Fourier analysis, as well as for other diverse results, such as the solution of the pajama problem.

Richard Montgomery (University of Warwick, United Kingdom), for his solution of the Ringel tree packing conjecture, development of distributive absorption techniques with applications to graph immersion problems, and solution of several classical Erdős conjectures and others on cycle lengths in sparse graphs using the novel machinery of sublinear expanders.

Danylo Radchenko (CNRS, University of Lille, France), for the construction of optimal spherical designs and his seminal contribution to the new field of Fourier interpolation, as well as his fundamental contributions to the theory of polylogarithms.

(All of them went on stage to receive their award from the president of the European Mathematical Society: an accrediting diploma and an Andalusian-themed tile.)

To conclude this part of the ceremony, let us wish these outstanding ten mathematicians every success in their future careers and reward them with a well-deserved round of applause.

Peregrina Quintela, chair of the Felix Klein prize committee

According to the statutes of the Felix Klein prize, the award seeks to honor a scientist under the age of 38 who has used sophisticated methods to provide a solution, meeting the full satisfaction of industry, to a concrete and challenging industrial problem.

At its meeting on April 1, 2024 the committee for the 2024 Felix Klein prize, composed of Professor Krzysztof Burnecki, Dr. Patrice Hauret, Professor Ralph Korn, Professor Karl Kunisch, and chaired by myself, unanimously decided to confer the 2024 Felix Klein prize to

Fabien Casenave (Safran Tech, Digital Sciences & Technologies, France), for his contributions to the integration of simulation-based design in projects related to reduced-order physical modeling in the aeronautical industry.

Thank you very much, and congratulations.

Karine Chemla, chair of the Otto Neugebauer prize committee

As in previous cases, awarding the 2024 Otto Neugebauer prize was a particularly challenging task. The nominations we received were of outstanding quality. The prize

committee, composed of Erika Luciano, Antoni Malet, Roman Murawski, and Peter Ullrich, unanimously decided – after extensive deliberation – to award the prize to

> *Reinhard Siegmund-Schultze (University of Agder, Kristiansand, Norway),* for his outstanding work in the social history of mathematics, which has helped to shape a new and much richer vision of the context of mathematics in the 20th century.

Professor Siegmund-Schultze, thank you very much, and congratulations.

Barbara Kaltenbacher, member of the EMS/ECMI Lánczos prize committee

A substantial amount of mathematical effort goes not only into proving theorems, lemmas, and propositions, but also into the development of mathematical software. Mathematical thinking is essential for the design and implementation of scientific code, and, conversely, software is an indispensable tool for disseminating mathematical ideas.

To recognize this fact, and on the initiative of CAIR, the EMS Committee for Applied and Interdisciplinary Relations, the European Mathematical Society – jointly with ECMI, the European Consortium for Mathematics in Industry – has established a prize for mathematical software. This prize is named after Cornelius Lánczos, a European mathematician and pioneer of numerical computation.

The committee received 13 strong nominations. It was composed of members nominated by ECMI and the EMS and was chaired by Patrick Farrell, whom I am honored to represent here today. It is therefore my great pleasure to announce the prize winner. Fortunately, the name is easy to pronounce:

> *MUMPS. Patrick Amestoy, Jean-Yves L'Excellent and Théo Mary,* for its important and widely used contributions to the numerical solution of sparse linear systems.

Thank you very much, and congratulations.

Konstantin Khanin, member of the Paul Lévy prize committee
Vincent Biron, grandson of Paul Lévy

It is my great pleasure and privilege to introduce a brand-new prize, integrated this year: the Paul Lévy prize in probability theory. This prize has no age or nationality restrictions. Mathematicians from all over the world are eligible, provided they have made outstanding contributions to probability theory and its applications, broadly understood.

The creation of this prize was a collective effort involving many contributors. The founding organizations include the European Mathematical Society, École Polytechnique – where Paul Lévy worked for nearly forty years – the École Polytechnique Foundation, and the Lévy family. The principal sponsor is BNP Paribas, which not only provided generous financial support but also contributed significantly to making this prize possible.

The selection committee was chaired by Wendelin Werner. While I will not list all five members here, their names will be made public shortly. We are very pleased that the committee was able, within a short time frame, to select an outstanding first laureate.

I would also like to note that the Lévy family has been involved from the very beginning. In particular, Hélène Freysz, a distinguished astrophysicist at the Observatoire de la Côte d'Azur in Nice and granddaughter of Paul Lévy, played an important role in the early stages. Unfortunately, she could not attend today, but we are delighted that Vincent Biron is here to present the award.

The first Paul Lévy prize in probability theory is awarded to

> *Jeremy Quastel (University of Toronto),* for his major advances in the fields of hydrodynamic theory, stochastic partial differential equations, and integrable probability.

Once again, congratulations to all the prize winners.

Plenary lectures

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/1

Chasing finite shadows of infinite groups through geometry

Martin R. Bridson

Abstract. There are many situations in geometry and group theory where it is natural, convenient or necessary to explore infinite groups via their actions on finite objects – that is, via the finite quotients of the group. But how much understanding can one really gain about an infinite group by examining its finite images? Which properties of the group can one recognise, and when does the set of finite images determine the group completely? How hard is it to decide what the finite images of a given infinite group are?

These notes follow my plenary lecture at the ECM in Sevilla, July 2024. The goal of the lecture was to sketch some of the rich history of the preceding problems and to present results that illustrate how the field surrounding these questions has been transformed in recent years by input from low-dimensional topology and the study of non-positively curved spaces.

1. Introduction

To what extent can one understand an infinite group by studying only its finite images? This compelling question has re-emerged with different emphases throughout the history of group theory, and in recent years it has been animated by a rich interplay with geometry and low-dimensional topology. My purpose in the lecture on which these notes are based was to report on some highlights of this interplay.

Before engaging directly with the challenge of identifying and exploring the finite images of groups, we will step back and consider the following more primitive questions: how should we describe the groups that arise in nature, and how might we go about the task of understanding a group that is presented to us? The first question leads us to a discussion of group presentations and the unlimited difficulty of extracting information from such presentations. A classical and insightful response to the second question is that since a group is, by definition, the mathematical object that captures the notion of symmetry, one ought to explore groups by seeking objects they are symmetries of, that is, one should study *actions* on various objects. But what

Mathematics Subject Classification 2020: 20E26 (primary); 20E18, 20F67, 57K32 (secondary).
Keywords: finitely presented groups, finite quotients, profinite rigidity, 3-manifolds.

objects? There are many things that are worth trying, but in the absence of further information, the most obvious objects to try first are finite sets, and that brings us back to the basic question that frames our study: to what extent can one understand a group by studying only its finite images?

Our focus will be on finitely generated and finitely presented groups. We shall discuss how hard it is to identify the finite quotients of a given group, and we shall explore which properties of a group can be detected from its set of finite quotients. We will be particularly concerned with questions of *profinite rigidity*, that is, situations where an infinite group is completely determined by its finite quotients. Groups that arise in low-dimensional geometry and topology will play a central role in our discussion. For example, one might ask to what extent the group of symmetries of the paving of the hyperbolic plane shown in Figure 1 is distinguished from all other groups by its finite quotients.

The narrative of the lecture on which these notes are based was built around the following sequence of results. The first serves to illustrate the difficulty of identifying the finite quotients of a finitely presented group.

Theorem A ([42]). *There does not exist an algorithm that, given an arbitrary finite presentation, can determine whether or not the group presented has a non-trivial finite quotient.*

Following [28], we shall discuss refinements of this result that describe the (non-)existence of algorithms that can identify quotients in particular families of finite

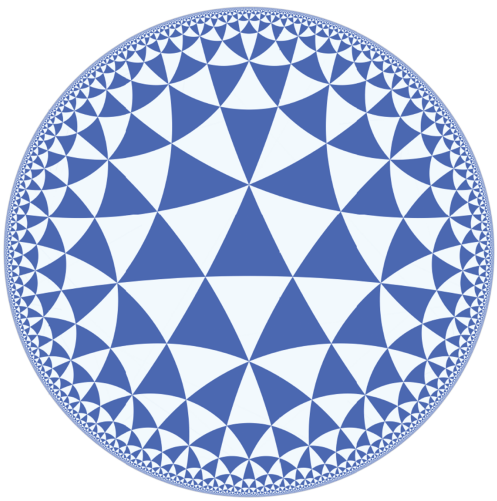


Figure 1. $\Delta(3,3,4) = \langle A, B, C \mid A^3 = B^3 = C^4 = 1 = ABC \rangle$

simple groups, for example, the groups $\mathrm{PSL}_n(q)$ with either n fixed and q varying, or vice versa.

The finite images of a group Γ form an inverse system indexed by the finite index normal subgroups $N \trianglelefteq \Gamma$: if $N < M$, then $\Gamma/N \twoheadrightarrow \Gamma/M$. The *profinite completion* $\widehat{\Gamma}$ of Γ is the inverse limit of this system; it is a compact, totally disconnected topological group. There is a natural map $\iota : \Gamma \to \widehat{\Gamma}$ with dense image. Every homomorphism from Γ to a finite group Q extends, via ι, to a map $\widehat{\Gamma} \to Q$ with the same image, and if Γ is finitely generated, every finite quotient of $\widehat{\Gamma}$ arises in this way, so Γ and $\widehat{\Gamma}$ have the same set of finite quotients. For finitely generated groups Γ and Λ, the set of finite images of Γ will be the same as the set of finite images of Λ if and only if $\widehat{\Lambda}$ and $\widehat{\Gamma}$ are isomorphic (as abstract or topological groups) [49, 87]; thus, for finitely generated groups, the reader can read the statement $\widehat{\Lambda} \cong \widehat{\Gamma}$ as "Γ and Λ have the same set of finite quotients".

When discussing profinite rigidity, it is natural to restrict attention to groups Γ that are *residually finite*, that is, every $\gamma \in \Gamma \smallsetminus \{1\}$ has non-trivial image in some finite quotient of Γ. We shall consider three versions of profinite rigidity: *absolute* profinite rigidity, in which comparisons to all finitely generated, residually finite groups are made; *relative* profinite rigidity, whereby one compares groups within a class – for example, 3-manifold groups, nilpotent groups, or lattices in Lie groups; and *Grothendieck* rigidity, where one is concerned with the existence of pairs of non-isomorphic groups $P \hookrightarrow \Gamma$ such that the inclusion induces an isomorphism of profinite completions. A central result in this last setting is the following.

Theorem B ([29]). *There exist residually finite (Gromov) hyperbolic groups H and finitely presented subgroups $P \hookrightarrow \Gamma := H \times H$ of infinite index, with $P \ncong \Gamma$, such that the inclusion $P \hookrightarrow \Gamma$ induces an isomorphism $\widehat{P} \to \widehat{\Gamma}$.*

We shall see in Section 7 that in this theorem, one can assume $H < \mathrm{SL}(n, \mathbb{Z})$.

Easy examples of residually finite groups that are profinitely rigid in the absolute sense include $\mathbb{Z}^d$, but it is much more difficult to find *full-sized* examples, that is, groups that contain a non-abelian free subgroup and hence do not satisfy any group law. This is where input from low-dimensional geometry and topology comes to the fore. The reader will recall that $\mathrm{PSL}(2, \mathbb{R})$ and $\mathrm{PSL}(2, \mathbb{C})$ are the groups of orientation-preserving isometries of the hyperbolic plane $\mathbb{H}^2$ and hyperbolic 3-space $\mathbb{H}^3$, respectively.

Theorem C ([33]). *There are arithmetic lattices in $\mathrm{PSL}(2, \mathbb{C})$ and $\mathrm{PSL}(2, \mathbb{R})$ that are profinitely rigid in the absolute sense.*

Examples of such lattices in $\mathrm{PSL}(2,\mathbb{C})$ include the Bianchi group $\mathrm{PSL}(2,\mathbb{Z}[\omega])$, where $1+\omega+\omega^2=1$, as well as certain cocompact lattices of small covolume. Examples in $\mathrm{PSL}(2,\mathbb{R})$ include the triangle group $\Delta(3,3,4)$ portrayed in Figure 1.

The remarkable nature of 3-manifold groups and the beauty and utility of hyperbolic geometry and its generalisations have been major themes in the study of profinite rigidity over recent years. Another important theme that has emerged is the importance of finiteness properties: there can be a stark difference in the behaviour of finitely generated and finitely presented groups in this context. The following recent result is perhaps the best illustration of that theme. A more refined version of this theorem, which again relies on extensive input concerning 3-dimensional manifolds, will be described in the penultimate section of these notes.

Theorem D ([37]). *There are finitely presented, residually finite groups Γ that are profinitely rigid among finitely presented groups but contain infinitely many pairwise non-isomorphic, finitely generated subgroups $H \hookrightarrow \Gamma$ such that $\widehat{H} \cong \widehat{\Gamma}$*

2. Presenting groups and understanding them

In a moment, we will decipher the caption of Figure 1 by recalling what it means to *present* a group. But first let me provoke the reader with the assertion that group theory is not really a branch of algebra – it belongs to the whole of mathematics, just as numbers do.

2.1. Groups everywhere

What kind of mathematics do you want to do? When we start doing formal mathematics, settling down to lend precision to an intuition gleaned from examples, we make precise definitions that encapsulate the essence of the objects that we want to study. As we start to manipulate and compare our objects, we have to decide what sort of maps $X \to Y$ we will allow. If we are studying sets, then any maps will do, but if we are interested in a linear problem and are studying vector spaces, we are likely to restrict to linear or affine maps; if we are doing topology, we will probably consider continuous maps; if we are modelling flexible geometric objects, we might study Lipschitz maps, etc. But no matter what the context, it is likely that sooner or later we will want to understand the *automorphisms* of our objects X – that is, the set of invertible maps (of our chosen type) that preserve the defining features of X. And, no matter what the context, these automorphisms form a *group*, $\mathrm{Aut}(X)$. According to one's nature, one might then be drawn to the study of groups themselves.

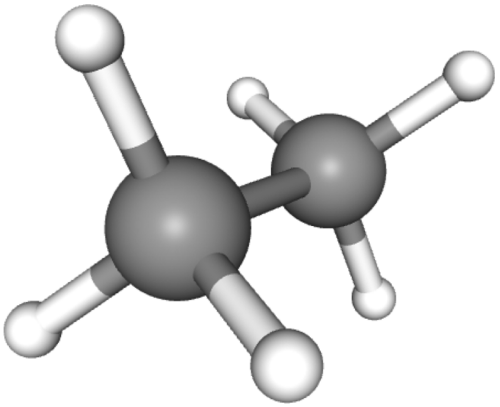


Figure 2. $\langle T, F \mid T^3 = 1 = F^2,\ FTF = T^2 \rangle$

2.2. How should we capture groups?

Maintaining the spirit of doing mathematics in the wild, let us consider how we might go about describing the symmetries of objects we encounter in nature. For an ethane molecule, we might start by noting that it has six (rigid, orientation-preserving) symmetries, but more information is contained in the observation that we need to combine two basic operations – T (twist) and F (flip) – in order to get everything. We also note that doing T three times or F twice amounts to doing nothing, and, after some more thought, we might also observe that doing F then T then F has the same effect as doing T twice. In summary, $\langle T, F \mid T^3 = 1 = F^2,\ FTF = T^2 \rangle$. We want to formalise the fact that these notes describe the group completely (see Figure 2).

2.3. Presentations of groups

A *presentation* of a group G is an expression of the form

$$G = \langle a_1, a_2, \ldots \mid r_1 = 1,\ r_2 = 1, \ldots \rangle,$$

where the r_j are finite words in the letters a_i and their inverses. The letters a_i are called *generators*, and the r_j are called *relators* (or "defining relations"). We will be mostly concerned with the case where the lists (a_i) and (r_j) are finite, in which case the presentation is said to be finite and G is said to be *finitely presented*.

If r_j is a concatenation of words $u_j v_j$, it is sometimes convenient to write $u_j = v_j^{-1}$ instead of $r_j = 1$. It is also common practice to abbreviate the list of relators by writing $r_1, r_2, \ldots$ instead of $r_1 = 1, r_2 = 1, \ldots$

When we give a presentation, we are asserting three things. First, there is a choice of elements $a_1, a_2, \ldots \in G$ (which need not be distinct) such that every element of G can be obtained by repeatedly performing the operations a_i and their inverses.

Two assertions are made concerning the relators: first, each of the equalities $r_i = 1$ is true in the group G (where 1 represents the identity element); second, *all* relations among the $a_i^{\pm 1}$ in G are consequences of the rules $r_j = 1$, where a "consequence" is a deduction made by repeatedly appealing to the following obvious facts and nothing else

$$[r = s \text{ and } t = u \implies rt = su], \quad [r = 1 \implies a^{-1}ra = 1].$$

These conditions can be formalised by saying that $G \cong F/R$ where F is the free group with basis $\{a_1, a_2, \ldots\}$ and $R \trianglelefteq F$ is the smallest normal subgroup containing $\{r_1, r_2, \ldots\}$.

With these formalities in hand, we return to the caption in Figure 1. The reader can verify (with effort) that if one takes A, B, C to be suitable rotations about the vertices of any fixed triangle in the tessellation, then the caption satisfies the three requirements of a presentation.

2.4. Groups that do not have finite presentations

A group G is termed *finitely generated* if there is a finite collection of elements $a_1, \ldots, a_n \in G$ such that every element of G is a product of the a_i and their inverses. If a group is finitely generated, then it is countable, but not all countable groups are finitely generated. For example, it is easy to verify that the rational numbers under addition $(\mathbb{Q}, +)$ cannot be finitely generated.

It is less obvious how to construct finitely generated groups that cannot be finitely presented, but I want to emphasise that one does not need obscure, monstrous constructions to find such groups: the absence of a finite presentation emerges in small, concrete examples. To illustrate this, let

$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}$$

and consider the 4-by-4 integer matrices

$$A = \begin{pmatrix} \mathbf{X} & 0 \\ 0 & \mathbf{I} \end{pmatrix}, \quad B = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{X} \end{pmatrix}, \quad C = \begin{pmatrix} \mathbf{Y} & 0 \\ 0 & \mathbf{Y} \end{pmatrix}.$$

Example 2.1. The subgroup of $\mathrm{SL}(4, \mathbb{Z})$ generated by $\{A, B, C\}$ has *no finite presentation*. To prove this, one first uses Klein's ping-pong argument to prove that the subgroup of $\mathrm{SL}(2, \mathbb{Z})$ generated by X and Y is a free group F of rank 2. Next, one observes that A, B, C lie in an obvious copy of $F \times F$. In fact, $\langle A, B, C \rangle$ is normal in $F \times F$ with quotient $\mathbb{Z}$, and such a subgroup can never be finitely presented [11, 57]. This example, which was studied by Stallings [115], is the seed for a rich vein of ideas in geometric group theory involving higher-dimensional (homological) finiteness properties [12], algebraic fibring [13, 70], and residually free groups [31].

The indirect nature of the proof sketched in the preceding example points to the fact that it can be hard to decide if an explicitly described group has a finite presentation. The fact that the following question has defied our understanding for half a century underscores this point.

Question 2.2 (Serre 1973 [110]). Can every finitely generated subgroup of $\mathrm{SL}(3,\mathbb{Z})$ be finitely presented?

The following theorem points to deeper problems and introduces a theme that we will pursue in the next section. This result relies on the existence of groups with relatively small presentations where the word problem is unsolvable (e.g., [20]). It is contained in the work of Mihailova [83] and Miller [84]; for a concise proof, see [22, Lemma 2.1].

Theorem 2.3. *Let F_2 be a free group of rank 2. There is no algorithm that, given a set of 20 elements $S \subset F_2 \times F_2$, can decide if the subgroup $\langle S \rangle$ has a finite presentation or not.*

Even in situations where one knows that finite presentations of groups exist, it can be extremely hard to construct them – for example, it is far from clear how to construct an explicit finite presentation for any uniform lattice[1] $\Gamma < \mathrm{SL}(3,\mathbb{R})$. Moving beyond lattices, if one works only with groups of integer matrices given by finite subsets $S < \mathrm{SL}(d,\mathbb{Z})$, then one can prove that there is no algorithm that allows one to construct finite presentations, even in cases where one knows that such presentations exist [40].

2.5. Problem: Hard to extract information from a presentation

Notwithstanding the preceding comments on the difficulty of finding presentations, when groups arise in nature (e.g., as discrete groups of automorphisms of a nice space), there are some general procedures that one can try to use to construct presentations, with versions of the Seifert/van Kampen theorem foremost among them. These are particularly effective in low dimensions. There are also various algebraic settings in which groups come naturally equipped with a finite presentation. But it can be ferociously hard to extract information about a group from a presentation.

[1] A *lattice* in a Lie group G is a discrete subgroup $\Gamma < G$ whose covolume $\mathrm{vol}(G/\Gamma)$, with respect to Haar measure, is finite. The lattice is *uniform* if G/Γ is compact, and otherwise it is non-uniform.

Rather than engaging in an abstract discussion of this difficulty, let us examine three similar-looking small presentations:

$$\begin{aligned} G_2 &= \langle A, B \mid BA = A^2B \rangle, \\ G_3 &= \langle a, b, c \mid ba = a^2b,\ cb = b^2c,\ ac = c^2a \rangle, \\ G_4 &= \langle \alpha, \beta, \gamma, \delta \mid \beta\alpha = \alpha^2\beta,\ \gamma\beta = \beta^2\gamma,\ \delta\gamma = \gamma^2\delta,\ \alpha\delta = \delta^2\alpha \rangle. \end{aligned} \tag{2.1}$$

The easiest of these groups to understand is G_2: for example, by adding the relation $A = 1$, we see that G_2 maps onto $\mathbb{Z}$, so in particular it is infinite. As for G_3 and G_4, one of these groups is infinite while the other is the trivial group – how are we to tell which is which? In my experience, asking an audience to vote on which of G_3 and G_4 is the trivial group will provide an even split of opinion, and I see no obvious reason why it should be otherwise. We will return to these groups shortly.

2.6. We need action!

One can quickly get frustrated by the task of trying to understand groups such as G_3 and G_4 simply by manipulating their presentations algebraically. Instead, one seeks to understand a group $G = \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$ by examining the ways in which it can be realised as a group of symmetries (automorphisms) of something. For example, one might pursue the path of classical representation theory by trying to understand linear actions of G: in finite dimensions, this amounts to searching for invertible matrices $A_1, \dots, A_n$ so that substituting each A_i for a_i in the defining relations r_j yields a product that equals the identity matrix. Most primitively of all, we can examine the ways in which G can act as a group of permutations of a finite set. Beyond this, we might look for actions on spaces that carry interesting or illuminating geometric, topological, or analytic features; this is where geometric group theory really begins.

2.7. A group springing into action: Topological models

Instead of viewing $G = \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$ as a recipe for expressing G as a quotient of a free group, one can regard it as a recipe for building a 2-dimensional complex K with fundamental group G: the complex has one vertex (0-cell), an edge (1-cell) for each a_i, orientated and labelled a_i, and for each relator r_j, there is 2-cell obtained by attaching a disc to the loop in the 1-skeleton that spells out the word r_j. If one does this with the standard presentation $\mathbb{Z}^2 = \langle a, b \mid aba^{-1}b^{-1} \rangle$, the complex K is a torus. One can then unwrap K to its universal cover to get the group $\mathbb{Z}^2 = \pi_1 K$ acting freely and cocompactly[2] (as the group of deck transformations;

[2]Thanks to Tim Riley for drawing this picture.

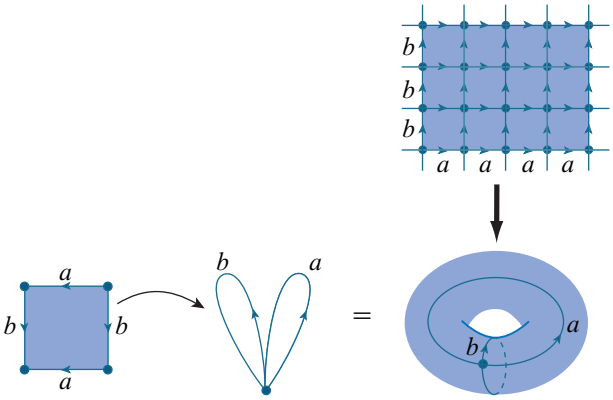


Figure 3. The 2-complex creating an action of $\mathbb{Z}^2 = \langle a, b \mid ab = ba \rangle$

see Figure 3). If we start with an arbitrary finite presentation, the 2-complex K will not be a manifold in general, but one can remedy this by embedding it in $\mathbb{R}^5$ then taking the boundary of a regular neighbourhood. This boundary can be smoothed, and with a little effort one can argue that it has the same fundamental group as the 2-complex we started with, thus proving the following proposition.

Proposition 2.4. *Every finitely presented group is the fundamental group of a smooth, compact* 4*-dimensional manifold.*

Classes of groups that are of special interest emerge when one asks how one might improve on this construction. The most obvious question, perhaps, is whether one can reduce to three dimensions; the answer is no, which provides us with the first hint that fundamental groups of compact 3-dimensional manifolds ("3-*manifold groups*" for short) have special features, a theme to which we will return.

But first let me remark that while these general constructions projecting the study of group presentations into topology are both useful and appealing, no hard work has been done, so one should expect them to preserve the hardness of most problems. Thus, for example, the task we set ourselves of deciding which of G_3 and G_4 was the trivial group is faithfully translated into the equally hard problem of deciding which of the 2-complexes built from these presentations is contractible, that is, can be continuously deformed within itself to a point. (It requires some theory to see that these problems are equivalent.)

2.8. Decision problems (Max Dehn 1911)

Combinatorial group theory is the study of groups given by generators and relators. Its origins are intertwined with the development of low-dimensional topology at the beginning of the twentieth century, particularly in the work of Max Dehn who,

in his famous paper [48], was the first to articulate clearly the algorithmic problems involved in extracting information from finite presentations $G = \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$, singling out three problems as the most fundamental: the *word problem* (deciding which words in the generators equal $1 \in G$), the *conjugacy problem* (deciding which words represent conjugate elements in G), and the *isomorphism problem* (which includes the triviality problem – deciding if the given group is trivial or not). One senses a certain level of frustration when he writes: *"One is led to such problems by necessity when working in geometry and topology."* His main interest at the time was the classification of knots and related questions in 3-manifold topology.

2.9. Unsolvable problems

In the 1950s, P. S. Novikov [89] and W. W. Boone [15–19] (followed by Adyan [1], Rabin [97], and others) proved that the sought-after algorithms do not exist in general.

Theorem 2.5 (Novikov [89], Boone [15–19]). *There exist finitely presented groups $G = \langle a_1, \dots, a_n \mid r_1, \dots, r_m \rangle$ for which there is no algorithm that, given an arbitrary word w in the generators, can decide whether or not $w = 1$ in G.*

Starting with a finitely presented group G that has an unsolvable word problem, it is not difficult to construct groups that have solvable word problem but unsolvable conjugacy problem, nor is it too difficult to construct recursive sequences of finitely presented groups so that one cannot tell which are trivial. For the triviality problem, one uses the theory of HNN extensions and amalgamated free products to make a sequence of groups, indexed by words in the generators of G, so that $w = 1$ in $G \Leftrightarrow G_w = \{1\}$.

2.10. 3-manifold groups are well behaved

We have already seen that 3-manifold groups are special, and this is particularly borne out in the context of decision problems, although it took the whole of the twentieth century to prove that the algorithms that Dehn sought in this context at the beginning of the century do exist.

Theorem 2.6. *If Γ is the fundamental group of a compact 3-manifold, then there are algorithms to solve the word and conjugacy problems in Γ. Moreover, there is an algorithm that, given two finite presentations of 3-manifold groups, will determine whether or not the groups presented are isomorphic.*

This theorem follows from Perelman's resolution of the geometrisation conjecture, but to see this one requires the work of many authors; we refer the reader to [5] for a thorough discussion.

2.11. Returning to the groups G_2, G_3, G_4

We commented earlier that the general construction illustrated in Figure 3 does not make the task of deciding which of G_3 and G_4 is $\{1\}$ any easier. Since topology has not helped us, we might fall back on representation theory, or the search for finite images. But these too fail us: we can gain further insight into $G_2 = \langle A, B \mid BA = A^2B \rangle$ by noting that the defining relation is satisfied by the matrices

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix},$$

yielding a representation that one can prove is faithful; one can also construct infinitely many finite quotients, for example, the presentation we gave of the symmetries of the ethane molecule (Figure 2) provides an action on a finite object; but a search for matrices of any size over any field satisfying the defining relations of G_3 and G_4 will prove fruitless – neither group admits any non-trivial, finite-dimensional representation over any field. Likewise, a search for non-trivial finite images of these groups will prove fruitless.

In fact, G_3 is the trivial group (this is a challenging exercise) while G_4 is an infinite group that has no non-trivial finite quotients. Since G_4 has no non-trivial matrix representations and no non-trivial actions on finite sets, where are we to look for an action that will unravel the structure of the group? The answer is that we get the group to act on a *tree*: the group decomposes as a non-trivial amalgamated free product and hence (according to Bass–Serre theory [111]) acts with compact quotient on an infinite tree. This is an important insight of Higman [62], reinterpreted by Serre [111].

2.12. Really useful actions: Trees, hyperbolicity, and cube complexes

We have just had our first encounter with actions on trees, which is the base example for two classes of actions that play a central role in what follows; both originate in the seminal work of Gromov [52]. First, there is Gromov's theory of hyperbolic groups and spaces. Whereas in a tree all triangles degenerate to tripods, hyperbolicity for geodesic metric spaces is characterised by the property that all triangles are uniformly thin: there is a constant $\delta > 0$ such that each side of each geodesic triangle is contained in the δ-neighbourhood of the union of the other two sides. A finitely generated group is *(Gromov) hyperbolic* if it acts properly and cocompactly by isometries on such a space. It will be helpful for the reader to have a nodding acquaintance with this theory (which can be gleaned from [30] or [52]), but no details will be needed to understand these notes.

From a different perspective, one can regard trees as 1-dimensional CAT(0) cube complexes. The rich and extensive theory of groups acting on CAT(0) cube complexes has played a central role in geometric group theory and low-dimensional topology over the last three decades. The reader unfamiliar with the subject can consult [30, 47, 107] for the basic theory and [106] for a concise introduction. Briefly, a *non-positively curved cube complex* is a space built by assembling Euclidean cubes, glued together by local isometries between faces, so that the resulting path metric on the space is non-positively curved in the sense of A. D. Alexandrov, that is, in the neighbourhood of each point, geodesic triangles are no fatter than triangles in the Euclidean plane with the same edge lengths (see [30]). This is equivalent to a purely combinatorial condition on links: there are no "empty simplices" in the link of any vertex of the complex; in dimension 2, this means that there are no circuits of length less than 4 in the link graphs (see Figure 4). If the complex is locally finite dimensional, all triangles in the universal cover (not just small ones) will satisfy Alexandrov's comparison condition; in other words, this universal cover will be a CAT(0) space [30].

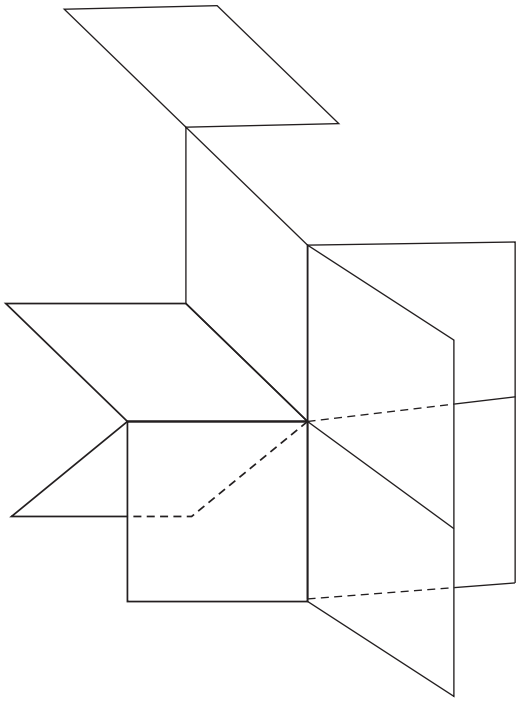


Figure 4. A 2-dimensional CAT(0) cube complex

There is a rich theory of groups that act by isometries on CAT(0) spaces and a much richer theory of groups that act properly and compactly by isometries on CAT(0) cube complexes, particularly when the complex is *special* in the sense of Haglund and Wise [59]. This last condition (which concerns the way in which cube-bisecting hyperplanes intersect) defines the class of *special* groups (sometimes called cocompact special, for clarity). Although it is defined in terms of the geometry of hyperplanes, specialness has a global characterisation that is particularly important for us: a cube complex is special if and only if there is a locally isometric embedding of it into the Salvetti complex of some right-angled Artin group [59]. For us, the important consequence of this is that any (virtually) special group can be embedded in $\mathrm{SL}(n,\mathbb{Z})$ for some n.

The powerful theory of (virtually) special groups developed by Wise and his collaborators has had a profound influence on many aspects of geometric group theory, including the study of profinite rigidity. If the reader explores the original papers on which these notes are based, they will quickly see how useful this theory is. A fundamental reason for this is the tight connection between specialness, the behaviour of subgroups in the profinite topology, and hierarchical decompositions of groups [133].

Much of the impact of the theory of specialness in the context of profinite rigidity flows through Agol's theorem and its consequences [2]. Agol proved that if a hyperbolic group acts properly and cocompactly by isometries on any CAT(0) cube complex, then it has a subgroup of finite index that is special. Agol used this theorem to prove Thurston's virtual fibring conjecture for hyperbolic 3-manifolds. This fibring theorem, as well as various other consequences of the work of Agol and Wise, are crucial ingredients in the main results of [32, 37].

3. Hard to see if finite images exist

We want to know to what extent finitely presented groups can be understood by examining only their finite quotients. In this context, we are inevitably led to ask how hard it is to discern what the finite quotients are. There are naive processes that one can run to list all finite quotients of a finitely presented group, but if such a process has not found a non-trivial finite quotient after a certain time, how is one to know whether it is worth continuing the search? One cannot know.

Theorem 3.1 (Bridson, Wilton [42]). *There does not exist an algorithm that, given an arbitrary finite presentation, can determine whether or not the group presented has a non-trivial finite quotient.*

A non-trivial finitely generated group of matrices (i.e., a linear group) always has non-trivial finite quotients, and finite groups are linear over any field, so Theorem 3.1 implies that one also cannot decide whether a finitely presented group has a non-trivial finite-dimensional, linear representation (over any field or over a fixed field).

3.1. Sketch of proof

The proof relies on:

- Logic: Slobodskoi's proof [113] of the undecidability of the universal theory of finite groups.
- Technology related to the geometry of subgroup separability in low dimensions, particularly in the context of graphs of free and finite groups and their covering spaces, as well as CAT(0) cube complexes (ideas of Stallings [116] and Wise [129, 130]).

The *universal theory of finite groups* is the set of all one-quantifier first-order sentences that are true in all finite groups. To see why it is relevant here, consider the following sentence Ψ (remembering that $\vee$ means "or"):

$$\forall a,b,c,d : (ba^2b^{-1} \neq a^3) \vee (dc^2d^{-1} \neq c^3) \vee ([a,b] \neq d) \\ \vee ([c,d] \neq b) \vee (a = b = c = d = 1).$$

The reader can convince themselves that this sentence is true in a specific group Q if and only if there is no non-trivial homomorphism $B \to Q$, where B is the group

$$B = \langle a,b,c,d \mid ba^2b^{-1}a^{-3},\ dc^2d^{-1}c^{-3},\ [a,b]d^{-1},\ [c,d]b^{-1}\rangle. \tag{3.1}$$

The reader will then see how to write a sentence Ψ_{AR} based on any finite group-presentation $\langle A \mid R\rangle$ so that Ψ_{AR} is true in a group Q if and only if there is no non-trivial homomorphism $G \to Q$, where $G = \langle A \mid R\rangle$.

In our example, the group B (which is infinite) has no finite quotients [29], so the sentence Ψ is true in every finite group, that is, it belongs to the universal theory of finite groups. In proving that this theory is undecidable, Slobodskoi [113] constructed a finitely presented group G in which there is no algorithm to determine which elements of the group survive in any finite quotient. One can phrase this as the unsolvability of a profinite version of the word problem in G. With such a group G in hand, one might hope to prove Theorem 3.1 by using amalgamated free products and HNN extensions in the manner alluded to in Section 2.9, building a sequence of finitely presented groups G_w, indexed by words in the generators of G, so that G_w has a non-trivial finite quotient if and only if the image of w is non-trivial in some finite quotient of G: in a standard notation, $w = 1$ in $\hat{G} \Leftrightarrow \hat{G}_w = 1$.

Ultimately, this outline of strategy does succeed (see [42, Theorem C]), but the details of the implication $\Rightarrow$ are much harder than in the classical situation (see Section 2.9). This is where the technology of Stallings and Wise is used. For reasons that quickly become clear when one tries to pursue the natural strategy, a key point is that one needs to be able to control the orders of the generators of G (and related groups) in finite quotients: this requires G to be crafted carefully, and it relies on refinements of Wise's work on *omnipotence* [129], which began with his proof that, given elements a, b of a non-abelian free group that are independent in homology, one can find a finite quotient in which the orders of a and b are in any given ratio. Beyond this first stage of control, one needs further technology concerning malnormal subgroups of virtually free groups, fibre products, and theorems concerning malnormal amalgams and residual finiteness.

3.2. Refinement: Finite simple images

It follows easily from Theorem 3.1 that there is no algorithm that determines the finite simple images of a finitely presented group. Famously, every non-abelian finite simple group is either an alternating group, a group from one of 16 families of Lie type, for example, $\mathrm{PSL}_d(q)$, or one of the 26 sporadic groups. Thus the question arises: for which collections of finite simple groups is there an algorithm that determines the members of the collection that are images of a given finitely presented group, and for which families do such algorithms not exist? This question is essentially settled in [28] where it is proved that if a collection of finite simple groups contains infinitely many alternating groups, or contains classical groups of unbounded ranks, then there is no algorithm, whereas for families of simple groups of Lie type with bounded rank, the desired algorithms do exist. The following special case of this result provides a concrete illustration of this dichotomy.

Theorem 3.2 (Bridson, Evans, Liebeck, Segal [28]). *Both of the following statements hold:*

(1) *There is an algorithm that, given a finitely presented group Γ, will determine whether or not Γ has infinitely many quotients $\mathrm{PSL}_n(q)$ with n fixed and q varying, and will list these quotients if there are only finitely many.*

(2) *There does not exist an algorithm that, given a finitely presented group Γ, will determine whether or not Γ has infinitely many quotients $\mathrm{PSL}_n(q)$ with q fixed and n varying, nor does there exist an algorithm that can determine whether Γ has any quotient of this form.*

The proof of (1) is rooted in representation theory and the model theory of finite fields; the algorithm is not practical.

4. Capturing groups via finite actions: Residual finiteness

We return to our main theme: *to what extent is a finitely generated (or finitely presented) group determined by its finite quotients (equivalently, its profinite completion)?*

The information contained in the kernel of the natural map $\Gamma \to \widehat{\Gamma}$ is lost when we restrict our attention to $\widehat{\Gamma}$, so it is natural to focus on groups where this kernel is trivial, that is, the residually finite groups.

Definition 4.1. A group Γ is *residually finite* if for every $\gamma \in \Gamma \smallsetminus \{1\}$ there is a homomorphism $\pi : \Gamma \to Q$ to a finite group with $\pi(\gamma) \neq 1$.

4.1. Which groups are residually finite?

There are many results in the literature establishing that groups are residually finite; the following selection is slanted towards the theme of these notes. I particularly want to emphasise the remarkable nature of items (4) and (5): they provide combinatorial criteria for verifying that a group is linear, and they have vastly extended the available reservoir of residually finite groups. The wealth of examples that these criteria provide, made available largely through the work of Dani Wise [133], underpins many advances in the field.

(1) If a finitely generated group Γ admits a faithful linear representation over any field $\Gamma \hookrightarrow \mathrm{GL}(n, K)$, then Γ is residually finite [81].

(2) If Γ is the fundamental group of a compact 3-manifold, then Γ is residually finite [60].

(3) Mapping class groups of compact surfaces and (outer) automorphism groups of finitely generated free groups are residually finite [53].

(4) Hyperbolic groups that act properly and cocompactly on CAT(0) complexes are residually finite.

(5) $C'(1/6)$ small cancellation groups are residually finite.

(6) Random finitely presented groups, in a suitable model, are residually finite.

Item (4) is a consequence of Agol's theorem, as discussed in Section 2.12. The $C'(1/6)$ small cancellation condition in (5) bounds the size of overlaps between subwords of relators in group presentations. Groups that have finite presentations satisfying this condition are hyperbolic, and Wise [132] proved that they act on CAT(0) cube complexes, so (4) applies. This way of phrasing things, though, obscures the fact that Wise's earlier work established residual finiteness in many cases, opening up the great reservoir of examples that (4) and (5) provide [130, 131, 133]. A popular model for a random group is Gromov's density model [86], whereby one fixes the number of generators for a presentation and chooses the relators uniformly at random

from cyclically reduced words of a given length, with the number of words determined by a density parameter; the focus is on what happens as the length gets large. At small densities, with overwhelming probability these groups satisfy the $C'(1/6)$ small cancellation condition [90].

The following challenge is one of the guiding problems of geometric group theory.

Question 4.2. Are all Gromov hyperbolic groups residually finite?

Residually finite groups harbour less pathology than arbitrary groups. For example, it is easy to see that a finitely presented, residually finite group Γ has a solvable word problem, although there do exist finitely generated, residually finite groups that have unsolvable word problem [82]. Note the recurrence of an important theme here: finitely generated groups can be much wilder than finitely presented ones.

We mention one other property of residually finite groups because it plays a surprisingly useful role in the results we shall look at later.

Lemma 4.3. *If a finitely generated group Γ is residually finite, then it has the Hopf property, that is, every epimorphism $\phi : \Gamma \to \Gamma$ is an isomorphism.*

Proof. If there were a non-trivial element $\gamma \in \Gamma$ in the kernel of an epimorphism $\phi : \Gamma \to \Gamma$, and if $\pi : \Gamma \to Q$ were a homomorphism to a finite group with $\pi(\gamma) \neq 1$, then the infinitely many maps $\pi \circ \phi^n : \Gamma \to Q$ would all be distinct, which is impossible because any homomorphism $\Gamma \to Q$ is determined by the image of a finite generating set. ∎

5. Profinite properties of groups

We want to understand the extent to which properties of residually finite groups are encoded in their profinite completions.

Definition 5.1. A property $\mathfrak{P}$ of abstract groups is a *profinite property* (or a profinite invariant) if, for finitely generated, residually finite groups Γ_1 and Γ_2, whenever Γ_1 has $\mathfrak{P}$ and $\widehat{\Gamma}_1 \cong \widehat{\Gamma}_2$, then Γ_2 has $\mathfrak{P}$.

5.1. Some basic facts about profinite completion

In the introduction, we defined the profinite completion $\widehat{\Gamma}$ of an abstract group Γ to be the inverse limit of the system of all finite quotients Γ/N of Γ. Equivalently, one

can define $\widehat{\Gamma}$ to be the closure of the image of Γ in the direct product of the finite quotients Γ/N (which is compact in the product topology)

$$\widehat{\Gamma} = \{(x_N) \mid x_N M = x_M \ \forall M > N\} \subset \prod_N \Gamma/N.$$

Profinite completion $\Gamma \mapsto \widehat{\Gamma}$ is a functor from the category of abstract groups to the category of profinite groups [103]: a homomorphism of abstract groups $u : \Gamma_1 \to \Gamma_2$ extends uniquely to a continuous homomorphism of profinite groups $\widehat{u} : \widehat{\Gamma}_1 \to \widehat{\Gamma}_2$.

By construction, $\widehat{\Gamma}$ is residually finite. And Γ is residually finite if and only if the natural map $\iota : \Gamma \to \widehat{\Gamma}$ is injective. Henceforth, we shall concentrate on residually finite groups, in which setting we identify Γ with its image under ι. If Γ is finitely generated, there is a 1-1 correspondence between the finite index subgroups of Γ and the finite index subgroups[3] of $\widehat{\Gamma}$: this associates to $H < \Gamma$ its (topological) closure $\overline{H} < \widehat{\Gamma}$, and it associates to $\mathcal{H} < \widehat{\Gamma}$ the intersection $\mathcal{H} \cap \Gamma$.

Remark 5.2. An important point to keep in mind is that, for finitely generated groups Γ_1 and Γ_2, the existence of an isomorphism $\widehat{\Gamma}_1 \cong \widehat{\Gamma}_2$ does not imply the existence of any non-trivial homomorphism from Γ_1 to Γ_2. On the other hand, the correspondence between finite index subgroups of Γ_i and $\widehat{\Gamma}_i$ does induce, via $\widehat{\Gamma}_1 \cong \widehat{\Gamma}_2$, an isomorphism between the lattices of finite index subgroups in Γ_1 and Γ_2; this isomorphism preserves the index of subgroups, normality, and conjugacy. It follows that any property that involves counting subgroups of a given index will be a profinite invariant, for example, subgroup growth [78]. Moreover, if $H_1 < \Gamma_1$ corresponds to $H_2 < \Gamma_2$, then $\widehat{H}_1 \cong \widehat{H}_2$ (an observation that can be extremely useful).

5.2. Profinite properties: Examples and counterexamples

The following list contains some properties that are easily seen to be profinite invariants, but it also points to many properties that are *not* profinite invariants. Rather than being disappointed by this, I hope that the reader will reflect that the difficulty of finding non-trivial profinite invariants makes the existence of profinite rigidity theorems all the more remarkable.

(1) Being *abelian* is a profinite property. *Proof:* If Γ_1 is abelian, then so are all of its finite quotients. If Γ_2 is not abelian, then for some $a, b \in \Gamma_2$, the commutator $c = [a, b]$ is non-trivial. If Γ_2 is residually finite, then $p(c) = [p(a), p(b)] \neq 1$ in some finite quotient $p : \Gamma_2 \to Q$. As this quotient is not abelian, it is not a quotient of Γ_1, and therefore $\widehat{\Gamma}_1 \ncong \widehat{\Gamma}_2$.

[3] The simplicity of this statement relies on the Nikolov–Segal theorem [87].

(2) By considering precisely which abelian p-groups Γ maps onto (varying the prime p), one can show that the isomorphism type of the *abelianisation* of a finitely generated group is a profinite invariant.

(3) A slight variation on the argument in (1) shows that satisfying any *group law* is a profinite property, for example, being nilpotent of class c, or solvable of derived length d. It is considerably harder to prove that being *polycyclic* is a profinite invariant, but Sabbagh and Wilson proved that this is true [105].

(4) The ℓ_2-Betti numbers of groups are defined as analytic invariants, but Lück's approximation theorem [79] allows one to compute the *first* ℓ_2-*Betti number* of a finitely presented, residually finite group as a limit of ordinary Betti numbers: $b_1^{(2)}(\Gamma) := \lim b_1(H_n)/[\Gamma\colon H_n]$, where (H_n) is any sequence of finite index normal subgroups intersecting in the identity. It follows from the facts about subgroup lattices in Remark 5.2 that the first ℓ_2-Betti number is a profinite invariant among finitely presented, residually finite groups [27, Corollary 3.3]. But Kammeyer and Sauer show that the higher-dimensional ℓ_2-Betti numbers are not profinite invariants [68].

(5) Lackenby [72] proved that having a subgroup of finite index that maps onto a non-abelian free group (*largeness*) is a profinite invariant among finitely presented groups.

(6) Aka [3] found lattices in semisimple Lie groups that have the same profinite completion even though the real rank of the Lie groups is different. He also proved that Kazhdan's *property (T)* is not a profinite invariant [4]. Work of Kassabov and Nikolov [69] shows that property (τ) is not a profinite invariant.

(7) Property (T) can be phrased as a fixed-point property for isometric actions on Hilbert spaces. *Fixed-point properties* for actions on other CAT(0) spaces have also been studied. Cheetham-West, Lubotzky, Reid, and Spitler [45] showed that Serre's property FA (the fixed-point property for actions on trees) is not a profinite invariant, and in [26] this was extended to fixed-point properties for actions on higher-dimensional CAT(0) spaces.

(8) Kionke and Schesler [71] proved that *amenability* is not a profinite invariant.

(9) Being finitely presented is not a profinite invariant of finitely generated groups (see Section 7), and likewise the higher *finiteness properties* of groups [77].

(10) Being *hyperbolic* (in the sense of Gromov) is not a profinite invariant among finitely generated groups, but it is unknown whether it is preserved under profinite equivalence of finitely presented groups.

(11) It is unknown whether having a faithful finite-dimensional *linear* representation is a profinite property.

(12) The theorem stated below can be used to see that the property of having a *subgroup* of a given sort is almost never a profinite invariant: for example, having a non-trivial finite subgroup, or a subgroup isomorphic to $\mathbb{Z}^3$, or a 2-generator subgroup that is neither abelian nor free. This precludes the profinite invariance of various properties that are inherited by subgroups, for example, being orderable, or locally indicable, or having cohomological dimension less than $d > 2$.

Theorem 5.3 ([25]). *Given a finitely generated, recursively presented group Γ that is residually finite, one can construct a finitely generated, residually finite free-by-free group $M_\Gamma = F_\infty \rtimes F_4$ and an embedding $M_\Gamma \hookrightarrow (F_4 * \Gamma) \times F_4$ that induces an isomorphism of profinite completions.*

This theorem is proved by combining ideas described in Section 7 with the universal embedding theorem in [23], which elaborates on ideas of Higman [63] and Baumslag, Dyer, and Heller [10].

6. Profinite rigidity

When studying profinite rigidity, one is interested in theorems whose conclusion is

$$\widehat{\Lambda} \cong \widehat{\Gamma} \implies \Lambda \cong \Gamma.$$

6.1. Three forms of rigidity

The following definition of what it means for a group to be profinitely rigid has become standard, although it is still common to add "in the absolute sense" for clarity (as we shall do several times in this article).

Definition 6.1 (*Absolute* profinite rigidity). A finitely generated, residually finite group Γ is *profinitely rigid* if, whenever Λ is finitely generated and residually finite, $\widehat{\Lambda} \cong \widehat{\Gamma} \Rightarrow \Lambda \cong \Gamma$.

In the previous section, we saw that when Γ satisfies a group law, one only needs to consider groups Λ satisfying the same group law. In such cases, Definition 6.1 reduces to an instance of the following problem.

Definition 6.2 (*Relative* profinite rigidity in a class $\mathcal{G}$). If a finitely generated, residually finite group Γ belongs to a class of groups $\mathcal{G}$, then one says that Γ is *profinitely rigid among groups in* $\mathcal{G}$ if whenever $\Lambda \in \mathcal{G}$ is finitely generated and residually finite, $\widehat{\Lambda} \cong \widehat{\Gamma} \Rightarrow \Lambda \cong \Gamma$.

Remark 6.3 (The absolute case is much harder). The literature contains many more theorems about relative profinite rigidity than it does about absolute profinite rigidity, for the obvious reason that even if one starts with a group Γ that has rich structure, when one finds a group Λ "lying in the gutter" and knows only that it is finitely generated, residually finite and has the same finite quotients as Γ, how is one to know that Λ shares any of the fine features of Γ? For example, if Γ is a group of matrices, why should Λ have any interesting linear representations? If Γ is a 3-manifold group, how would one begin to argue that Λ is a 3-manifold group?

In contrast, if one is trying to argue that a group is profinitely rigid in a class $\mathcal{G}$, then one can study special features of the groups in $\mathcal{G}$ and ask if these are visible in $\widehat{\Lambda}$ when $\Lambda \in \mathcal{G}$. For example, we shall see in Section 8 that many features of compact 3-manifolds M are invariants of $\widehat{\pi_1 M}$.

The third form of profinite rigidity that is widely studied is *Grothendieck* rigidity, where one is interested in whether it is possible to obtain isomorphisms $\hat{u} : \widehat{\Lambda} \to \widehat{\Gamma}$ induced by homomorphisms of discrete groups $u : \Lambda \to \Gamma$. This is the subject of Section 7.

6.2. Early work, sobering examples, and a guiding question

(1) We have seen that finitely generated abelian groups are profinitely rigid. When one moves from abelian to *nilpotent* groups, one finds that some are profinitely rigid (e.g., free nilpotent groups) while others are not [101]. But in all cases, the ambiguity is limited: only finitely many finitely generated nilpotent groups can have the same profinite completion [91]. Following [56], an appealing way to express this finite ambiguity is to use the language of *profinite genus* (within a class of groups). Thus Pickel's theorem is that the profinite genus of any finitely generated nilpotent group is finite. Grunewald, Pickel, and Segal [58] proved that *virtually polycyclic* groups also have *finite genus*.

(2) In 1974, Baumslag [8] discovered the following striking example which underscores how quickly one can lose profinite rigidity when one moves away from the abelian setting. Let $G_1 = (\mathbb{Z}/25) \rtimes_\alpha \mathbb{Z}$ and $G_2 = (\mathbb{Z}/25) \rtimes_\beta \mathbb{Z}$, where α is the automorphism $x \mapsto x^6$ and $\beta = \alpha^2$ is the automorphism $x \mapsto x^{11}$. The reader can verify directly that $G_1 \ncong G_2$, but G_1 and G_2 have the same finite quotients.

(3) Baumslag drew inspiration from an important example of Serre [109]. In 1964, Serre constructed a smooth projective variety V defined over a number field K such that for the varieties V_{ϕ_1} and V_{ϕ_2} obtained by extension of scalars along different embeddings $\phi_1, \phi_2 : K \hookrightarrow \mathbb{C}$, the fundamental groups $\Gamma_i = \pi_1 V_{\phi_i} \cong \mathbb{Z}^{p-1} \rtimes (\mathbb{Z}/p)$ are not isomorphic, but $\widehat{\Gamma}_1 \cong \widehat{\Gamma}_2$ because each $\widehat{\Gamma}_i$ is the algebraic (étale) fundamental group of the underlying scheme. Pairs of *Galois conjugate* varieties exhibiting this behaviour can also be found among complex surfaces [7]. Galois conjugation will play an important role in Section 9.2.

(4) The examples of Stebe [117] are based on a different phenomenon. He constructed pairs of matrices in $\mathrm{SL}(2,\mathbb{Z})$ that are not conjugate in $\mathrm{GL}(2,\mathbb{Z})$ but are conjugate when one projects them to $\mathrm{GL}(2,\mathbb{Z}/d)$ for every integer $d > 2$. One such pair is

$$\phi(1) = \begin{pmatrix} 188 & 275 \\ 121 & 177 \end{pmatrix} \quad \text{and} \quad \phi(2) = \begin{pmatrix} 188 & 11 \\ 3025 & 177 \end{pmatrix}.$$

Writing $\Gamma_i = \mathbb{Z}^2 \rtimes_{\phi(i)} \mathbb{Z}$, we have $\Gamma_1 \ncong \Gamma_2$ but $\widehat{\Gamma}_1 \cong \widehat{\Gamma}_2$. Every element of $\mathrm{SL}(2,\mathbb{Z})$ can be realised as an orientation-preserving homeomorphism of the torus, so Γ_i is the fundamental group of a torus bundle over the circle. In fact, it is a lattice in the Lie group Sol, which is one of the eight geometries in dimension 3 (see [108]). Thus we have examples of closed, geometric 3-manifolds that are not profinitely rigid. These types of torus-bundle examples were explored further by Funar [50]. The corresponding punctured-torus bundles are more rigid [38].

(5) Hempel [61] produced further pairs of geometric 3-manifolds with $\pi_1 M_1 \ncong \pi_1 M_2$ but $\widehat{\pi_1 M}_1 \cong \widehat{\pi_1 M}_2$. These examples, which have the local geometry of $\mathbb{H}^2 \times \mathbb{R}$, are higher-genus surface bundles over the circle, where the holonomy has finite order. They are very close in spirit to Baumslag's finite-by-cyclic examples in (2).

(6) Following the work of Platonov and Tavgen' [93] discussed in Section 7, one knows that the direct product of two or more non-abelian free groups is not profinitely rigid.

A noticeable absentee from the list of groups considered above is the free group F_n. The following question [88, Question 15] is a guiding challenge in the field.

Question 6.4 (Remeslennikov). Are finitely generated free groups profinitely rigid?

6.3. Relative profinite rigidity and hyperbolic geometry

There is now a large literature on relative profinite rigidity with many positive and negative results, some of which are referenced in Sections 5.2, 6.2, and 8. It is a very active field. I highlight the following result because it marked the beginning of a flowering of interest in the profinite rigidity of low-dimensional orbifold groups, and also because it underscores the point I was making in Remark 6.3: this theorem shows that free groups are profinitely rigid within lattices, but Question 6.4 is a vastly more difficult question because it is hard to extract any geometry from the hypothesis $\widehat{\Lambda} \cong \widehat{F}_n$.

Theorem 6.5 (Bridson, Conder, Reid [27]). *If* Γ *is a lattice in* $\mathrm{PSL}(2,\mathbb{R})$ *and* Λ *is a lattice in any connected Lie group, then* $\widehat{\Gamma} \cong \widehat{\Lambda} \Rightarrow \Gamma \cong \Lambda$.

The profinite invariance of the first ℓ_2-Betti number plays an important role in the proof of this theorem. In particular, the work of Lott and Lück [76] allows one to reduce fairly quickly to the case where Λ is a lattice in $\mathrm{PSL}(2,\mathbb{R})$, where one understands the lattices very explicitly.

We shall discuss the profinite rigidity of lattices in $\mathrm{PSL}(2,\mathbb{C})$ in a moment. Stover [118, 119] showed that there are lattices in other rank-1 Lie groups (isometries of complex hyperbolic space) that are not profinitely rigid.

7. Grothendieck pairs

The study of Grothendieck pairs was initiated by Alexander Grothendieck in 1970 [54]. He was investigating the extent to which the topological fundamental group of a smooth complex projective variety is determined by the étale fundamental group of the underlying scheme, and he was interested in the extent to which a finitely generated, residually finite group can be recovered from a complete knowledge of its finite-dimensional representation theory.

Definition 7.1 (Grothendieck pairs). A monomorphism of residually finite groups $u : P \hookrightarrow \Gamma$ is called a *Grothendieck pair* if $\widehat{u} : \widehat{P} \to \widehat{\Gamma}$ is an isomorphism but u is not. One says that a finitely generated, residually finite group Γ is *Grothendieck rigid*[4] if there are no Grothendieck pairs $P \hookrightarrow \Gamma$ with $P \neq \Gamma$ finitely generated.

[4]There is also interest in fixing Γ and considering Grothendieck pairs $\Gamma \hookrightarrow \Lambda$. When pursuing this, it is natural to distinguish between the two settings with the terminology left/right (or up/down) Grothendieck rigidity [46, 67].

Grothendieck discovered a remarkably close connection between the representation theory of a finitely generated group and its profinite completion: if $A \neq 0$ is a commutative ring and $u : \Gamma_1 \to \Gamma_2$ is a homomorphism of finitely generated groups, then $\widehat{u} : \widehat{\Gamma}_1 \to \widehat{\Gamma}_2$ is an isomorphism if and only if the restriction functor $u_A^* : \mathrm{Rep}_A(\Gamma_2) \to \mathrm{Rep}_A(\Gamma_1)$ is an equivalence of categories, where $\mathrm{Rep}_A(\Gamma)$ is the category of finitely presented A-modules with a Γ-action. This led him to pose his famous question about the existence of finitely presented Grothendieck pairs.

Question 7.2 (Grothendieck 1970 [54]). If Γ_1 and Γ_2 are finitely presented and residually finite, must $u : \Gamma_1 \to \Gamma_2$ be an isomorphism if $\widehat{u} : \widehat{\Gamma}_1 \to \widehat{\Gamma}_2$ is an isomorphism?

Platonov and Tavgen' [93] answered the corresponding question for finitely generated groups, and various other constructions followed [6, 85, 96]. Grothendieck's question was eventually settled by myself and Fritz Grunewald [29]. The following is an enhancement of what is stated in [29] (see Remark 7.9).

Theorem 7.3 (Bridson, Grunewald [29]). *There exist (Gromov) hyperbolic groups $\Gamma < \mathrm{SL}(n, \mathbb{Z})$ and finitely presented subgroups $P \hookrightarrow G := \Gamma \times \Gamma$ of infinite index, with $P \ncong G$, such that the inclusion $P \hookrightarrow G$ induces an isomorphism $\widehat{P} \to \widehat{G}$.*

7.1. The ideas in the proof

A natural place to begin a search for Grothendieck pairs is with short exact sequences

$$1 \to N \to \Gamma \to Q \to 1 \tag{7.1}$$

with N and Γ finitely generated and residually finite, and Q an infinite group with no non-trivial finite quotients, for example, G_4 from (2.1) or B from (3.1). If profinite completion were an exact functor, we could simply put hats on the groups in (7.1) to conclude that $\widehat{N} \to \widehat{\Gamma}$ was an isomorphism. But it is not an exact functor, and while $\widehat{N} \to \widehat{\Gamma}$ will be surjective if $\widehat{Q} = 1$, it need not be injective. To understand why, the reader will need some familiarity with the (co)homology of groups [43]. Consider the 5-term exact sequence in homology associated to (7.1)

$$H_2(Q, \mathbb{Z}) \to H_0(Q,\, H_1(N, \mathbb{Z})) \to H_1(\Gamma, \mathbb{Z}) \to H_1(Q,\, \mathbb{Z}) \to 0. \tag{7.2}$$

As N is finitely generated, the automorphism group of its abelianisation $H_1(N, \mathbb{Z})$ is residually finite, so the action $Q \to \mathrm{Aut}(H_1(N, \mathbb{Z}))$ must be trivial, since $\widehat{Q} = 1$, and the group of coinvariants $H_0(Q,\, H_1(N, \mathbb{Z}))$ is simply $H_1(N, \mathbb{Z})$. Moreover, $\widehat{Q} = 1$ implies $H_1(Q,\, \mathbb{Z}) = 0$, so (7.2) simplifies to an exact sequence

$$H_2(Q, \mathbb{Z}) \to H_1(N, \mathbb{Z}) \to H_1(\Gamma, \mathbb{Z}) \to 0,$$

from which we deduce that if $H_2(Q,\mathbb{Z}) \to H_1(N,\mathbb{Z})$ is not the zero map, then N will have finite abelian quotients that Γ does not have, whence $\widehat{N} \ncong \widehat{\Gamma}$. To avoid this difficulty, we assume $H_2(Q,\mathbb{Z}) = 0$, a condition satisfied by our examples G_4 and B.

Lemma 7.4. *With $1 \to N \to \Gamma \to Q \to 1$ as above, if $H_2(Q,\mathbb{Z}) = 0$, then $N \hookrightarrow \Gamma$ is a Grothendieck pair.*

With this lemma in hand, we will have finitely generated Grothendieck pairs if we can construct short exact sequences of the desired type. To do this, we appeal to the Rips construction [104] which, given a finite presentation of a group Q, will produce a short exact sequence $1 \to N \to \Gamma \to Q \to 1$, as desired, with N finitely generated and Γ a 2-dimensional hyperbolic group that satisfies a small cancellation condition; we saw in Section 4.1 that Γ is residually finite. To obtain further examples, one can verify that in this situation, if $H < \Gamma$ is any subgroup that contains N and if $\widehat{H/N} = 1$, then $N \hookrightarrow H$ and $H \hookrightarrow \Gamma$ are both Grothendieck pairs. By taking Q to be one of the universal groups constructed in [23, Theorem A], we deduce the following result.

Theorem 7.5. *There exist residually finite, (Gromov) hyperbolic groups Γ of dimension 2 with uncountably many pairwise non-isomorphic subgroups $\iota_H : H \hookrightarrow \Gamma$ such that $\widehat{\iota}_H : \widehat{H} \to \widehat{\Gamma}$ is an isomorphism. Moreover, one can arrange for infinitely many of these subgroups to be finitely generated.*

This is not how Platonov and Tavgen' [93] produced the first examples of finitely generated Grothendieck pairs. (A version of the Rips construction that guaranteed residual finiteness was not available at that time.) Instead, they considered *fibre products*, modifying the proof of Lemma 7.4 to establish the following criterion, which they applied to $F_4 \to G_4$ to construct a Grothendieck pair $P \hookrightarrow F_4 \times F_4$ with P finitely generated.

Proposition 7.6 (Platonov, Tavgen' [93]). *Let $f : \Gamma \to Q$ be an epimorphism of groups, with Γ finitely generated and Q finitely presented. Suppose that $\widehat{Q} = 1$ and $H_2(Q,\mathbb{Z}) = 0$. Then the fibre product $P = \{(g,h) \mid f(g) = f(h)\} < \Gamma \times \Gamma$ is finitely generated and $P \hookrightarrow \Gamma \times \Gamma$ induces an isomorphism of the profinite completions $\widehat{P} \xrightarrow{\cong} \widehat{\Gamma \times \Gamma}$.*

It is easy to see that P is finitely generated: if Γ is generated by X and $Q = \langle X \mid R \rangle$, then P is generated by $\{(x,x) \mid x \in X\} \cup \{(r,1) \mid r \in R\}$. This fact is often called the 0-1-2 lemma. But if Γ is free and Q is infinite, the fibre product P will never have a finite presentation. Indeed, while the Platonov–Tavgen' construction produces finitely generated Grothendieck pairs $P \hookrightarrow F_4 \times F_4$ (and one can replace F_4

by F_2 using [135]), the following theorem implies that there are no such finitely presented Grothendieck pairs, even if one allows the product of more free groups, or other residually free groups, such as surface groups.

Theorem 7.7 (Bridson, Wilton [39]). *If Π is finitely presented and residually free, then every finitely presented subgroup $P < \Pi$ is closed in the profinite topology; in particular, $P \hookrightarrow \Pi$ is not a Grothendieck pair.*

Remark 7.8. Once more, I want to emphasise the contrast in behaviour between finitely presented subgroups and those that are merely finitely generated.

Theorem 7.3 is proved by combining the ideas described above and appealing to results on the finite presentability of fibre products. The 1-2-3 *theorem* of Baumslag, Bridson, Miller, and Short [9] states that if Q has a classifying space with a finite 3-skeleton (as our examples G_4 and B do), while Γ is finitely presented and $N = \ker(\Gamma \to Q)$ is finitely generated, then the fibre product $P < \Gamma \times \Gamma$ of $\Gamma \to Q$ is finitely presented. To prove Theorem 7.3, one can take $Q = B$, use the Rips construction to generate $1 \to N \to \Gamma \to B \to 1$, and then appeal to Wise's work [132] to see that Γ is a cubulated hyperbolic group, which by Agol's theorem [2] can be embedded in $\mathrm{SL}(d, \mathbb{Z})$ for some integer d. The 1-2-3 theorem assures us that the associated fibre product P is finitely presented, and Proposition 7.6 assures us that $P \hookrightarrow \Gamma \times \Gamma$ is a Grothendieck pair.

Remark 7.9. When [29] was written, neither Agol's theorem nor the virtually special Rips construction in [59] was available, so the penultimate sentence of the proof sketched above was replaced by an appeal to Wise's residually finite version of the Rips construction [131]. This accounts for the difference between Theorems B and 7.3.

7.2. Elaborations, refinements, and Grothendieck rigidity

In the last twenty years, many papers have been written constructing Grothendieck pairs $P \hookrightarrow \Gamma$ where P and Γ have different properties. Indeed, this has been perhaps the main mechanism by which various properties have been shown not to be profinite. Typically, this is achieved by encoding additional properties into a group Q that has a finite classifying space with $\widehat{Q} = 1$ and $H_2(Q, \mathbb{Z}) = 0$. One then follows the above template of proof, perhaps adjusting Γ in some way, by refining the Rips construction, or by passing to a central extension [23,41], or replacing the hyperbolic group Γ by something entirely different [25]. Often, a more refined version of the 1-2-3 theorem is needed, as in [21] where it is shown that there exist finitely presented, residually finite groups Γ that contain infinitely many pairwise non-isomorphic, finitely presented subgroups P_n such that $P_n \hookrightarrow \Gamma$ is a Grothendieck pair.

In the opposite direction, there are notable results proving that certain groups are Grothendieck rigid. Grothendieck himself [54] proved this for groups of the form $\mathrm{G}(A)$ where A is a commutative ring and G is an affine group scheme of finite type over A; see Jaikin-Zapirain and Lubotzky [67] for a concise account of this and further results. In the setting of 3-manifolds, Long and Reid [75] established Grothendieck rigidity for the fundamental groups of all closed geometric 3-manifolds and all finite-volume hyperbolic 3-manifolds. Following the work of Agol and Wise, Sun was able to extend this to cover the fundamental group Γ of every compact 3-manifold [120]. In the hyperbolic case, [36, Theorem A] proves something stronger, ruling out abstract isomorphisms $\widehat{H} \cong \widehat{\Gamma}$ for subgroups $H < \Gamma$, not just those induced by inclusions $H \hookrightarrow \Gamma$. This stronger statement is false in the non-hyperbolic setting [50, 61].

8. Profinite rigidity and 3-manifolds

Thurston's geometrisation conjecture, proved by Perelman, tells us that every closed orientable 3-manifold can be cut into pieces by a system of embedded spheres and tori so that the interior of each piece is homeomorphic to a finite-volume quotient of one of the eight 3-dimensional geometries [108]. Hyperbolic geometry $\mathbb{H}^3$ is the richest and most ubiquitous of these geometries, and we encountered Sol in Section 6.2 (4). The pieces modelled on the remaining six geometries are Seifert fibre spaces, which we shall discuss in a moment.

Many beautiful results have been proved in recent years concerning the great extent to which properties of 3-manifolds are captured by the profinite completions of their fundamental groups. *As a result, we now know that specifying* $\widehat{\pi_1 M}$ *determines a compact* 3*-manifold* M *up to finite ambiguity, and we know that many features of the* 3*-manifold are captured by* $\widehat{\pi_1 M}$.

A 3-manifold is *irreducible* if every 2-sphere in it bounds a ball. Every closed, orientable 3-manifold can be written as a connected sum of irreducible manifolds and copies of $S^2 \times S^1$. In the spirit of the geometrisation conjecture, the first thing that one wants to know is whether the prime decomposition of M is visible in $\widehat{\pi_1 M}$, and Wilton and Zalesskii [128] proved that it is: if M and N are closed, orientable 3-manifolds with prime decompositions $M = M_1 \# \cdots \# M_m \# r(S^1 \times S^2)$ and $N = N_1 \# \cdots \# N_n \# s(S^1 \times S^2)$, and $\widehat{\pi_1 M} \cong \widehat{\pi_1 N}$, then $m = n$, $r = s$, and after permuting the indices, $\widehat{\pi_1 M_i} \cong \widehat{\pi_1 N_i}$. Wilkes [124] showed that this remains valid if M has an incompressible boundary. These results allow us to concentrate on irreducible manifolds, where attention falls on the process of cutting the manifold along tori. In this context, Wilton and Zalesskii [128] showed that $\widehat{\pi_1 M}$ encodes the canonical JSJ decomposition of the manifold, including the profinite completions of the

pieces into which the tori cut the manifold; again, Wilkes shows that one can allow boundary [122, 123]. Wilton and Zalesskii also proved that the profinite completion distinguishes the hyperbolic pieces in the geometric decomposition from the Seifert fibred pieces [127].

We saw in Section 6.2 (4) that there are distinct quotients of Sol whose fundamental groups have the same profinite completion, but up to homeomorphism there can be only finitely many with a given completion [58]. The same is true of $\mathbb{H}^2 \times \mathbb{R}$ manifolds. Wilkes [121] extended this finite genus result to manifolds that consist entirely of Seifert pieces, giving a complete account of the ambiguities.

The first example of a hyperbolic 3-manifold that is distinguished from all other 3-manifolds by $\widehat{\pi_1 M}$ was the figure-8 knot complement (Boileau–Friedl [14] and Bridson–Reid [35]). Bridson, Reid, and Wilton later proved that every punctured-torus bundle over the circle is distinguished among 3-manifold groups by $\widehat{\pi_1 M}$. (Punctured-torus bundles account for all groups of the form $F_2 \rtimes \mathbb{Z}$. It remains unknown whether groups of the form $F_n \rtimes \mathbb{Z}$ are distinguished from each other by their profinite completions – cf. [34, 64].) These early results relied on recovering fibre-bundle structure from $\widehat{\pi_1 M}$. Subsequently, Jaikin-Zapirain [65] proved that $\widehat{\pi_1 M}$ always detects whether a 3-manifold fibres over the circle with compact fibre. Other notable results include the fact that 3-manifold groups are all good in the sense of Serre [99, 112, 125], that is, if Γ is a 3-manifold group, then for all $n > 0$ and all finite Γ-modules A, the natural map $H^n(\widehat{\Gamma}, A) \to H^n(\Gamma, A)$ from the continuous cohomology of $\widehat{\Gamma}$ to the group cohomology of Γ is an isomorphism. We refer to Reid's ICM talk [100] for an overview of related results up to 2018 and [98, 102] for updates.

A breakthrough that deserves particular attention comes from the work of Liu [73]. He proved that only finitely many finite-volume hyperbolic 3-manifolds can have the same profinite completion. Building on much of the work described above, Xiaoyu Xu recently extended this finite genus theorem to all compact, orientable, 3-manifolds with toral or empty boundary [134].

The following discussion of Seifert fibre spaces is needed for Theorem D.

8.1. Seifert fibre spaces

A *Seifert fibre space* is a 3-manifold M that is foliated by circles. The circles of the foliation are divided into regular fibres and exceptional fibres: a circle is a regular fibre if it has a neighbourhood that admits a fibre-preserving homeomorphism to the solid torus $\mathbb{D}^2 \times S^1$ foliated by the circles $\{x\} \times S^1$; in a neighbourhood of an exceptional fibre, the foliation is modelled on the foliation of $\mathbb{D}^2 \times S^1$ that is obtained by cutting the product fibration open along a disc $\mathbb{D}^2 \times \{*\}$ and then regluing after twisting through an angle $2q\pi/p$, where $p, q \in \mathbb{Z}$ are coprime integers with $0 < q < p$.

By collapsing each circle fibre in M to a point, we obtain a 2-manifold with an orbifold structure given by marking each exceptional fibre with a cyclic group recording the order p of the twisting; this is called the base orbifold of the Seifert fibration.

We will be concerned only with Seifert fibred manifolds where $\pi_1 M$ is infinite, in which case there is a central extension

$$1 \to \mathbb{Z} \to \pi_1 M \to \Delta \to 1,$$

where the central $\mathbb{Z}$ is the fundamental group of a regular fibre and Δ is the fundamental group of the base orbifold. For example, if M has 3 exceptional fibres, with twisting data (p_1, q_1), (p_2, q_2), (p_3, q_3) and the underlying surface of the base orbifold is a sphere, then $\Delta = \Delta(p_1, p_2, p_3)$ is a triangle group and

$$\pi_1(M) = \langle c_1, c_2, c_3, z \mid z \text{ is central, } c_1^{p_1} z^{-q_1} = c_2^{p_2} z^{-q_2} = c_3^{p_3} z^{-q_3} = 1, \\ c_1 c_2 c_3 = z^d \rangle$$

for some $d \in \mathbb{Z}$.

The examples of Hempel (Section 6.2 (5)) are surface bundles over the circle with finite holonomy. Wilkes [121] proved that all Seifert manifolds apart from these are profinitely rigid in the class of 3-manifold groups.

9. Absolute profinite rigidity

We seek finitely generated, residually finite groups Γ that are profinitely rigid in the absolute sense. We observed in Section 6.2 (3) that if Γ satisfies a group law, then this immediately reduces to a problem within the class of groups that satisfy this law, that is, it reduces to a challenge concerning relative profinite rigidity. But we also alluded (Remark 6.3) to the difficulty of knowing where to begin if there is no obvious reduction to a relative problem. That is the fundamental challenge that we want to address in this section. To avoid simple reductions, we consider only groups that are *full sized*, meaning that they contain a non-abelian free group. *Do there exist finitely generated, residually finite, full-sized groups that are profinitely rigid?* Might 3-manifolds and hyperbolic geometry[5] help us?

9.1. Profinitely rigid Kleinian groups

We begin by describing some hyperbolic 3-orbifolds of small volume. The reader will likely be familiar with the tiling of the hyperbolic plane by ideal triangles. In the same

[5] Always remember, $\mathbb{H}$yperbolic geometry is good for you!

way, one can tile hyperbolic 3-space $\mathbb{H}^3$ with copies of the regular ideal tetrahedron, that is, the tetrahedron with vertices at infinity where the dihedral angle between each pair of faces is $\pi/3$. Barycentric subdivision divides each tetrahedron in the tiling into 24 congruent tetrahedra, each with one vertex at infinity. The reflections in the faces of any one of these smaller tetrahedra will generate the full isometry group of the tiling. This group of isometries is

$$\Lambda_0 = \langle A, B, C, D \mid A^2, B^2, C^2, D^2, (AB)^3, (AC)^3, (AD)^2, (BC)^3, (BD)^2, (CD)^6 \rangle.$$

This is the unique non-uniform lattice of smallest covolume in $\mathrm{Isom}(\mathbb{H}^3)$. It contains the Bianchi group $\mathrm{PSL}(2, \mathbb{Z}[\omega])$ as a subgroup of index 4, where $\omega^2 + \omega + 1 = 0$.

Among torsion-free uniform lattices in $\mathrm{PSL}(2, \mathbb{C})$, there is a unique one of smallest covolume, namely Γ_W, the fundamental group of the Weeks manifold [51], which can be obtained by Dehn surgery on the Whitehead link.

Theorem 9.1 (Bridson, McReynolds, Reid, Spitler [32]). *Each of the groups* Γ_W, Λ_0, $\mathrm{PSL}(2, \mathbb{Z}[\omega])$, *and* $\mathrm{PGL}(2, \mathbb{Z}[\omega])$ *is profinitely rigid among all finitely generated, residually finite groups.*

The proofs in [32] cover a number of related lattices, but only finitely many. The list of lattices in $\mathrm{PSL}(2, \mathbb{C})$ that are known to be profinitely rigid (in the absolute sense) was extended in [36, 44] but remains finite for the moment.

The following are some of the key elements of the proof of Theorem 9.1 for the given examples Γ:

- Arithmetic: The trace field K_Γ of Γ (defined below) has small degree, and Γ is commensurable with the group of units of a maximal order in the associated quaternion algebra.
- Character varieties: Γ has very few Zariski-dense representations into $\mathrm{PSL}(2, \mathbb{C})$, up to conjugacy, and these are all accounted for by the arithmetic of K_Γ (*"Galois rigidity"*).
- Mostow rigidity and volume calculations are useful, as are various elements of 3-manifold topology, including the existence of compact cores, some knot theory, and the existence of finite covers that fibre. The consequences of cubulation, following Agol [2] and Wise [133], play a crucial role [94, 95], not least in ensuring cohomological goodness in the sense of Serre [55, 112].
- Intricate arguments concerning the subgroups of small finite index in each example were needed in [32], but these were superseded by the results in [36].

9.2. Trace fields and Galois rigidity

A remarkable feature of hyperbolic 3-manifolds is that their fundamental groups have *arithmetic structure* intrinsically associated to them (see Maclachlan and Reid [80]). Much of the proof of Theorem 9.1 revolves around this arithmetic structure and the way it controls representations into $\mathrm{PSL}(2,\mathbb{C})$ in our examples. A key notion in this regard is *Galois rigidity*. I will explain very briefly what this is and how it is used.

Let $\phi\colon \mathrm{SL}(2,\mathbb{C}) \to \mathrm{PSL}(2,\mathbb{C})$ be the quotient homomorphism, and if H is a finitely generated subgroup of $\mathrm{PSL}(2,\mathbb{C})$, set $H_1 = \phi^{-1}(H)$. It will be convenient to say that H is *Zariski dense* in $\mathrm{PSL}(2,\mathbb{C})$ when what we actually mean is that H_1 is a Zariski-dense subgroup of $\mathrm{SL}(2,\mathbb{C})$. The *trace field* of H is defined to be the field

$$K_H = \mathbb{Q}(\mathrm{tr}(\gamma)\colon \gamma \in H_1).$$

If K_H is a number field with ring of integers R_{K_H}, we say that H has *integral traces* if $\mathrm{tr}(\gamma) \in R_{K_H}$ for all $\gamma \in H_1$.

An important object associated to H is the subalgebra of $\mathrm{Mat}(2,\mathbb{C})$ generated by H_1; this is a *quaternion algebra* over K_H.

Suppose now that Λ is an abstract finitely generated group and $\rho\colon \Lambda \to \mathrm{PSL}(2,\mathbb{C})$ a Zariski-dense representation with $K = K_{\rho(\Lambda)}$ a number field of degree n_K. If $K = \mathbb{Q}(\theta)$ for some algebraic number θ, then the Galois conjugates of θ, say $\theta = \theta_1, \ldots, \theta_{n_K}$, provide embeddings $\sigma_i\colon K \to \mathbb{C}$ defined by $\theta \mapsto \theta_i$. These in turn can be used to build n_K Zariski-dense non-conjugate representations $\rho_{\sigma_i}\colon \Lambda \to \mathrm{PSL}(2,\mathbb{C})$ with the property that $\mathrm{tr}(\rho_{\sigma_i}(\gamma)) = \sigma_i(\mathrm{tr}(\rho(\gamma)))$ for all $\gamma \in \Lambda$. One refers to these as *Galois conjugate representations*. The existence of these Galois conjugates shows that $|\mathrm{X}_{\mathrm{zar}}(\Gamma,\mathbb{C})| \geq n_{K_{\rho(\Lambda)}}$, where $\mathrm{X}_{\mathrm{zar}}(\Lambda,\mathbb{C})$ denotes the set of Zariski-dense representations $\Gamma \to \mathrm{PSL}(2,\mathbb{C})$ up to conjugacy.

Definition 9.2. Let Λ be a finitely generated group and $\rho\colon \Lambda \to \mathrm{PSL}(2,\mathbb{C})$ a Zariski-dense representation whose trace field $K_{\rho(\Lambda)}$ is a number field. If $|\mathrm{X}_{\mathrm{zar}}(\Lambda,\mathbb{C})| = n_{K_{\rho(\Gamma)}}$, we say that Γ is *Galois rigid* (with associated field $K_{\rho(\Lambda)}$).

9.3. An outline of the proof of Theorem 9.1

An explicit understanding of the character variety of each of our examples enables us to prove that each is Galois rigid. For example, we prove that the only Zariski-dense representations of $\mathrm{PSL}(2,\mathbb{Z}[\omega])$, whose trace field is $\mathbb{Q}(\sqrt{-3})$, are the inclusion map and its complex conjugate. In each case, one also has integrality of traces. The integrality of traces is used to ensure the boundedness of the non-archimedean representations $\Gamma \to \mathrm{SL}(2,\overline{\mathbb{Q}_p})$ obtained by transporting Zariski-dense representations $\Gamma \to \mathrm{SL}(2,\mathbb{C})$ via (non-continuous) field isomorphisms $\mathbb{C} \cong \overline{\mathbb{Q}_p}$, for all primes p. This boundedness enables one to extend the representations to $\widehat{\Gamma}$, then restrict to any

finitely generated Λ with $\widehat{\Lambda} \cong \widehat{\Gamma}$ to obtain bijections between the sets of bounded, Zariski-dense characters of Γ and Λ at every finite place [32, Lemma 4.6].

With this control established, one can argue that the abstract group Λ is Galois rigid with a Zariski-dense representation $\rho\colon \Lambda \to \mathrm{PSL}(2,\mathbb{C})$ whose arithmetic data match those of Γ; see [32, Theorem 4.8] for a precise statement. In good situations (which include the low-degree number fields of our examples), this matching of data forces the number fields $K_{\rho(\Lambda)}$ and K_Γ to be equal, and likewise the quaternion algebras of Γ and $\rho(\Lambda)$. This in turn allows one to conclude that (up to Galois conjugation) the image of $\rho : \Lambda \to \mathrm{PSL}(2,\mathbb{C})$ is contained in Γ or a small extension of it, and in all of the examples one can force the image to lie in Γ.

At this point of the argument in [32], a host of results about subgroups of 3-manifolds in general, and the finite index subgroups of our examples in particular, were used to argue that if the image of ρ was not equal to Γ, then Λ would have a finite quotient that Γ did not have. These arguments can now be replaced by [36, Theorem A]: *if* $\Pi < \mathrm{PSL}(2,\mathbb{C})$ *is a lattice and* $H < \Pi$ *is a finitely generated, proper subgroup, then* $\widehat{H} \ncong \widehat{\Pi}$*; if* H *has finite index in* Π*, then there is a finite group onto which* H *maps but* Π *does not.*

At this stage, we have proved that there is a surjection $\rho : \Lambda \twoheadrightarrow \Gamma$, hence $\widehat{\rho} : \widehat{\Lambda} \twoheadrightarrow \widehat{\Gamma}$, and the Hopf property for finitely generated profinite groups (a variant of Lemma 4.3) enables us to conclude that $\widehat{\rho}$ is injective and hence so is ρ. ∎

9.4. Profinitely rigid Fuchsian groups

Given integers (p,q,r) with $1/p + 1/q + 1/r < 1$, one can tile the hyperbolic plane with triangles whose interior angles are $2\pi/p$, $2\pi/q$, $2\pi/r$. The group generated by reflections in the sides of a fixed triangle acts transitively on the set of triangles in the tiling; we denote this group $\Delta^{\pm}(p,q,r)$. The index-2 subgroup consisting of orientation-preserving isometries is $\Delta(p,q,r)$. We restrict our attention to the following triples (p,q,r):

$$(3,3,4),\quad (3,3,5),\quad (3,3,6),\quad (2,5,5),\quad (4,4,4). \tag{9.1}$$

$$(2,3,8),\quad (2,3,10),\quad (2,3,12),\quad (2,4,5),\quad (2,4,8). \tag{9.2}$$

Theorem 9.3 (Bridson, McReynolds, Reid, Spitler [33]). *Both* $\Delta(p,q,r)$ *and* $\Delta^{\pm}(p,q,r)$ *are profinitely rigid if* (p,q,r) *belongs to one of these lists.*

A key fact about these groups is that they are Galois rigid (as subgroups of $\mathrm{PSL}(2,\mathbb{C})$), and in each case the trace field is a real quadratic number field where the ramification of primes is simple enough to ensure that the argument from [32] sketched above applies.

We have discovered some lattices in $\mathrm{PSL}(2,\mathbb{R})$ *and* $\mathrm{PSL}(2,\mathbb{C})$ *that are profinitely rigid in the absolute sense, but conjecturally there should be many more.*

10. Seifert fibre spaces and the importance of finiteness properties

The following theorem provides the first examples of finitely presented groups that are profinitely rigid among finitely presented groups but not among finitely generated groups. Like our earlier examples, the groups Γ in this theorem are 3-manifold groups with particular arithmetic properties. Here, $S^2(p,q,r)$ denotes the quotient $\mathbb{H}^2/\Delta(p,q,r)$ of the hyperbolic plane by the triangle group $\Delta(p,q,r)$.

Theorem 10.1 (Bridson, Reid, Spitler [37]). *There exist finitely presented, residually finite groups* Γ *with the following properties:*

(1) $\Gamma \times \Gamma$ *is profinitely rigid among all finitely presented, residually finite groups.*

(2) *There are infinitely many pairwise distinct, finitely generated groups* Λ *such that* $\widehat{\Lambda} \cong \widehat{\Gamma} \times \widehat{\Gamma}$.

(3) *If* Λ *is as in* (2)*, then there is an embedding* $\Lambda \hookrightarrow \Gamma \times \Gamma$ *that induces the isomorphism* $\widehat{\Lambda} \cong \widehat{\Gamma \times \Gamma}$ *(in other words,* $\Lambda \hookrightarrow \Gamma \times \Gamma$ *is a Grothendieck pair).*

If M *is any Seifert fibred space with base orbifold* $S^2(3,3,4)$ *or* $S^2(3,3,6)$ *or* $S^2(2,5,5)$*, then* $\Gamma = \pi_1 M$ *has these properties.*

10.1. An outline of the proof

Let M be a Seifert fibre space with base orbifold $S^2(p,q,r) = \mathbb{H}^2/\Delta(p,q,r)$, where (p,q,r) is one of triples from list (9.1) or (9.2). We know from the previous section that $\Delta = \Delta(p,q,r)$ is profinitely rigid. There is a central extension

$$1 \to \mathbb{Z} \to \pi_1 M \to \Delta \to 1.$$

Theorem 10.2. *Let* $\Delta(p,q,r)$ *be a triangle group from list* (9.1) *or* (9.2)*, and let* M *be a Seifert fibred space with base* $S^2(p,q,r)$*. Then* $\pi_1 M$ *is profinitely rigid (in the absolute sense).*

In order to prove this theorem, one has to extend the arguments about profinite rigidity in the previous sections to cover central extensions. It is not the case that an arbitrary central extension of Δ will be profinitely rigid [92], but special arguments apply in the case where the centre is cyclic. Wilkes's work on the profinite rigidity of Seifert fibred spaces within the class of 3-manifold groups [121] is used here.

The key step in the process of reducing abstract profinite isomorphism to the study of Grothendieck pairs also requires an extension of the scope of our arguments about Galois rigidity, this time to cover direct products, fibre products, as well as central extensions.

Theorem 10.3. *Let $\Delta(p,q,r)$ be a triangle group from list (9.1) or (9.2), let M be a Seifert fibred space with base orbifold $S^2(p,q,r)$, and let $\Gamma = \pi_1 M$. Then, for every finitely generated, residually finite group Λ with $\widehat{\Lambda} \cong \widehat{\Gamma \times \Gamma}$, there is an embedding $\Lambda \hookrightarrow \Gamma \times \Gamma$ that induces the isomorphism.*

With this reduction to Grothendieck pairs in hand, the following straightforward extension of Theorem 7.7 establishes the desired profinite rigidity among finitely presented groups.

Theorem 10.4. *For every Fuchsian group F and every Seifert fibred space M with base orbifold $\mathbb{H}^2/F$, there are no Grothendieck pairs $\Lambda \hookrightarrow \pi_1 M \times \pi_1 M$ with $\Lambda \neq \pi_1 M \times \pi_1 M$ finitely presented.*

To complete the proof of Theorem 10.1, we need finitely generated Grothendieck pairs.

Theorem 10.5. *If $\Delta(p,q,r)$ is a triangle group from list (9.1), then there are infinitely many Seifert fibred spaces with base orbifold $S^2(p,q,r)$ whose fundamental group Γ has the property that there are infinitely many pairwise non-isomorphic, finitely generated groups Λ and inclusions $\Lambda \hookrightarrow \Gamma \times \Gamma$ inducing isomorphisms $\widehat{\Lambda} \cong \widehat{\Gamma \times \Gamma}$.*

Note that at this stage, not all of the groups from Theorem 10.2 are used: we need constraints on the Seifert invariants that ensure $\Gamma = \pi_1 M$ has finite index in $[G,G]$ where G is a group that maps onto a non-elementary hyperbolic group and has $H_1(G,\mathbb{Z})$ finite and $H_2(G,\mathbb{Z}) = 0$. These conditions facilitate a homological argument, adapted from an idea of Bass and Lubotzky [6], that allows us to modify the template for constructing Grothendieck pairs that was described in Section 7.

11. Open questions

The most celebrated problem concerning profinite rigidity is Remeslennikov's question (6.4), where the answer is widely believed to be positive: see [24] for a discussion of related results and recent progress, most notably [66].

Conjecture 11.1. *Finitely generated free groups are profinitely rigid.*

The consensus on whether the following conjecture is true is less uniform: it is widely believed that lattices in $\mathrm{PSL}(2,\mathbb{R})$ are profinitely rigid, but many would favour a more cautious conjecture in the Kleinian case, simply asking that a finitely generated, residually finite group Λ with the same finite quotients as a lattice in $\mathrm{PSL}(2,\mathbb{C})$ should itself be a lattice. If we knew that Λ was a lattice in $\mathrm{PSL}(2,\mathbb{C})$, Liu's finite genus theorem [73] would tell us that there are only finitely many possibilities for Λ; it is unclear if this finite ambiguity can be removed, but recent results of [73, 74, 126] constrain it.

Conjecture 11.2. *All lattices in* $\mathrm{PSL}(2,\mathbb{R})$ *and* $\mathrm{PSL}(2,\mathbb{C})$ *are profinitely rigid (in the absolute sense).*

With Theorem 10.1 in mind, one might weaken the above conjectures and ask only for rigidity among finitely presented groups. We saw in Section 7 that the difference between finite generation and finite presentation definitely matters in the context of direct products of free groups.

Conjecture 11.3. *Let* F *and* F' *be finitely generated free groups. If a finitely presented, residually finite group* Γ *has the same finite quotients as* $F \times F'$*, then* $\Gamma \cong F \times F'$.

For groups of higher-dimensional origin, the most compelling question in the field is the following.

Question 11.4. Is $\mathrm{SL}(n,\mathbb{Z})$ profinitely rigid when $n \geq 3$?

Using results from Spitler's thesis [114] that were inspired by ideas of Lubotzky (and are close in spirit to our discussion of Galois rigidity), one can reduce this question to the challenge of deciding whether $\mathrm{SL}(n,\mathbb{Z})$ is Grothendieck rigid. But since the subgroups of $\mathrm{SL}(n,\mathbb{Z})$ are so varied and mysterious, this remains a very difficult challenge.

Acknowledgements. I thank my longstanding co-author Alan Reid for the many insights that he has shared as we have explored profinite rigidity together. I acknowledge with gratitude my debt to the late Fritz Grunewald, who enticed me into this field. I also offer warm thanks to my other co-authors: Marston Conder, David Evans, Martin Liebeck, Ben McReynolds, Dan Segal, Ryan Spitler, and Henry Wilton. In addition, I thank Henry for his helpful comments on my first draft of these notes.

I was honoured by the invitation to speak at the 9th European Congress of Mathematics in Sevilla, and I thank the organisers for making it such a stimulating and enjoyable event. I am also grateful to Stanford University and to my hosts,

Tadashi Tokieda and Gregory Lieb, for the hospitality that enabled me to complete this manuscript.

References

[1] S. I. Adyan, Unsolvability of some algorithmic problems in the theory of groups (in Russian). *Trudy Moskov. Mat. Obšč.* **6** (1957), 231–298 Zbl 0080.24101 MR 0095872

[2] I. Agol, The virtual Haken conjecture. With an appendix by Ian Agol, Daniel Groves and Jason Manning. *Doc. Math.* **18** (2013), 1045–1087 Zbl 1286.57019 MR 3104553

[3] M. Aka, Arithmetic groups with isomorphic finite quotients. *J. Algebra* **352** (2012), no. 1, 322–340 Zbl 1254.20026 MR 2862189

[4] M. Aka, Profinite completions and Kazhdan's property (T). *Groups Geom. Dyn.* **6** (2012), no. 2, 221–229 Zbl 1244.20025 MR 2914858

[5] M. Aschenbrenner, S. Friedl, and H. Wilton, *3-manifold groups*. EMS Ser. Lect. Math., European Mathematical Society, Zürich, 2015, 215 pp. Zbl 1326.57001 MR 3444187

[6] H. Bass and A. Lubotzky, Nonarithmetic superrigid groups: counterexamples to Platonov's conjecture. *Ann. of Math. (2)* **151** (2000), no. 3, 1151–1173 Zbl 0963.22005 MR 1779566

[7] I. Bauer, F. Catanese, and F. Grunewald, Faithful actions of the absolute Galois group on connected components of moduli spaces. *Invent. Math.* **199** (2015), no. 3, 859–888 Zbl 1318.14034 MR 3314516

[8] G. Baumslag, Residually finite groups with the same finite images. *Compos. Math.* **29** (1974), no. 3, 249–252 Zbl 0309.20007 MR 0357615

[9] G. Baumslag, M. R. Bridson, C. F. Miller III, and H. Short, Fibre products, non-positive curvature, and decision problems. *Comment. Math. Helv.* **75** (2000), no. 3, 457–477 Zbl 0973.20034 MR 1793798

[10] G. Baumslag, E. Dyer, and A. Heller, The topology of discrete groups. *J. Pure Appl. Algebra* **16** (1980), no. 1, 1–47 Zbl 0419.20026 MR 0549702

[11] G. Baumslag and J. E. Roseblade, Subgroups of direct products of free groups. *J. Lond. Math. Soc. (2)* **30** (1984), no. 1, 44–52 Zbl 0559.20018 MR 0760871

[12] M. Bestvina and N. Brady, Morse theory and finiteness properties of groups. *Invent. Math.* **129** (1997), no. 3, 445–470 Zbl 0888.20021 MR 1465330

[13] R. Bieri, W. D. Neumann, and R. Strebel, A geometric invariant of discrete groups. *Invent. Math.* **90** (1987), no. 3, 451–477 Zbl 0642.57002 MR 0914846

[14] M. Boileau and S. Friedl, The profinite completion of 3-manifold groups, fiberedness and the Thurston norm. In D. P. Thurston (ed.), *What's next? The mathematical legacy of William P. Thurston*, pp. 21–44, Ann. of Math. Stud. 205, Princeton University Press, Princeton, NJ, 2020 Zbl 1452.57012 MR 4205634

[15] W. W. Boone, Certain simple, unsolvable problems of group theory. I. *Nederl. Akad. Wetensch. Proc. Ser. A* **57** (1954), 231–237 = *Indag. Math.* **16**, 231–237 Zbl 0055.00602 MR 0066372

[16] W. W. Boone, Certain simple, unsolvable problems of group theory. II. *Nederl. Akad. Wetensch. Proc. Ser. A* **57** (1954), 492–497 = *Indag. Math.* **16**, 492–497 Zbl 0057.01701 MR 0066373

[17] W. W. Boone, Certain simple, unsolvable problems of group theory. III. *Nederl. Akad. Wetensch. Proc. Ser. A* **58** (1955), 252–256 = *Indag. Math.* **17**, 252–256 Zbl 0067.25706 MR 0066374

[18] W. W. Boone, Certain simple, unsolvable problems of group theory. IV. *Nederl. Akad. Wetensch. Proc. Ser. A* **58** (1955), 571–577 Zbl 0067.25706 MR 0098775

[19] W. W. Boone, Certain simple unsolvable problems of group theory. V, VI. *Nederl. Akad. Wetensch. Proc. Ser. A* **60** = *Indag. Math.* **19** (1957), 22–27, 227–232 Zbl 0067.25706 MR 0098776

[20] V. V. Borisov, Simple examples of groups with unsolvable word problem (in Russian). *Mat. Zametki* **6** (1969), no. 5, 521–532 Zbl 0211.34102. English translation: *Math. Notes* **6** (1969), no. 5, 768–775 Zbl 0215.10803 MR 0260851

[21] M. R. Bridson, The strong profinite genus of a finitely presented group can be infinite. *J. Eur. Math. Soc. (JEMS)* **18** (2016), no. 9, 1909–1918 Zbl 1350.20023 MR 3531665

[22] M. R. Bridson, On the recognition of right-angled Artin groups. *Glasg. Math. J.* **62** (2020), no. 2, 473–475 Zbl 1481.20116 MR 4085052

[23] M. R. Bridson, The homology of groups, profinite completions, and echoes of Gilbert Baumslag. In P. Baginski, B. Fine, A. Moldenhauer, G. Rosenberger, and V. Shpilrain (eds.), *Elementary theory of groups and group rings, and related topics*, pp. 11–28, De Gruyter Proc. Math., De Gruyter, Berlin, 2020 Zbl 1528.20094 MR 4204977

[24] M. R. Bridson, Profinite rigidity and free groups. In J.-M. Morel and B. Teissier (eds.), *Mathematics going forward. Collected mathematical brushstrokes*, pp. 233–240, Lecture Notes in Math. 2313, Springer, Cham, 2023 Zbl 1536.20039 MR 4627962

[25] M. R. Bridson, Profinite completions of free-by-free groups contain everything. *Q. J. Math.* **75** (2024), no. 1, 139–142 Zbl 1552.20139 MR 4732948

[26] M. R. Bridson, Profinite isomorphisms and fixed-point properties. *Algebr. Geom. Topol.* **24** (2024), no. 7, 4103–4114 Zbl 1555.20043 MR 4840391

[27] M. R. Bridson, M. D. E. Conder, and A. W. Reid, Determining Fuchsian groups by their finite quotients. *Israel J. Math.* **214** (2016), no. 1, 1–41 Zbl 1361.20037 MR 3540604

[28] M. R. Bridson, D. M. Evans, M. W. Liebeck, and D. Segal, Algorithms determining finite simple images of finitely presented groups. *Invent. Math.* **218** (2019), no. 2, 623–648 Zbl 1446.20051 MR 4011707

[29] M. R. Bridson and F. J. Grunewald, Grothendieck's problems concerning profinite completions and representations of groups. *Ann. of Math. (2)* **160** (2004), no. 1, 359–373 Zbl 1083.20023 MR 2119723

[30] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*. Grundlehren Math. Wiss. 319, Springer, Berlin, 1999, 643 pp. Zbl 0988.53001 MR 1744486

[31] M. R. Bridson, J. Howie, C. F. Miller III, and H. Short, Subgroups of direct products of limit groups. *Ann. of Math. (2)* **170** (2009), no. 3, 1447–1467 Zbl 1196.20047 MR 2600879

[32] M. R. Bridson, D. B. McReynolds, A. W. Reid, and R. Spitler, Absolute profinite rigidity and hyperbolic geometry. *Ann. of Math. (2)* **192** (2020), no. 3, 679–719 Zbl 1455.57028 MR 4172619

[33] M. R. Bridson, D. B. McReynolds, A. W. Reid, and R. Spitler, On the profinite rigidity of triangle groups. *Bull. Lond. Math. Soc.* **53** (2021), no. 6, 1849–1862 Zbl 1523.20089 MR 4386043

[34] M. R. Bridson and P. Piwek, Profinite rigidity for free-by-cyclic groups with centre. *J. Lond. Math. Soc. (2)* **111** (2025), no. 6, article no. e70181, 28 pp. Zbl 1567.20076 MR 4920376

[35] M. R. Bridson and A. W. Reid, Profinite rigidity, fibering, and the figure-eight knot. In D. P. Thurston (ed.), *What's next? The mathematical legacy of William P. Thurston*, pp. 45–64, Ann. of Math. Stud. 205, Princeton University Press, Princeton, NJ, 2020 Zbl 1452.57013 MR 4205635

[36] M. R. Bridson and A. W. Reid, Profinite rigidity, Kleinian groups, and the cofinite Hopf property. *Michigan Math. J.* **72** (2022), 25–49 [Special issue in honor of Gopal Prasad] Zbl 1517.20077 MR 4460248

[37] M. R. Bridson, A. W. Reid, and R. Spitler, Absolute profinite rigidity, direct products, and finite presentability. *Duke Math. J.* **175** (2026), no. 3, 511–554 Zbl 08169266 MR 5036002

[38] M. R. Bridson, A. W. Reid, and H. Wilton, Profinite rigidity and surface bundles over the circle. *Bull. Lond. Math. Soc.* **49** (2017), no. 5, 831–841 Zbl 1439.57006 MR 3742450

[39] M. R. Bridson and H. Wilton, Subgroup separability in residually free groups. *Math. Z.* **260** (2008), no. 1, 25–30 Zbl 1158.20013 MR 2413340

[40] M. R. Bridson and H. Wilton, On the difficulty of presenting finitely presentable groups. *Groups Geom. Dyn.* **5** (2011), no. 2, 301–325 Zbl 1244.20033 MR 2782175

[41] M. R. Bridson and H. Wilton, The isomorphism problem for profinite completions of finitely presented, residually finite groups. *Groups Geom. Dyn.* **8** (2014), no. 3, 733–745 Zbl 1309.20025 MR 3267521

[42] M. R. Bridson and H. Wilton, The triviality problem for profinite completions. *Invent. Math.* **202** (2015), no. 2, 839–874 Zbl 1360.20020 MR 3418245

[43] K. S. Brown, *Cohomology of groups*. Grad. Texts in Math. 87, Springer, New York, 1982, 309 pp. Zbl 0584.20036 MR 0672956

[44] T. Cheetham-West, Absolute profinite rigidity of some closed fibered hyperbolic 3-manifolds. *Math. Res. Lett.* **31** (2024), no. 3, 615–638 Zbl 1554.57027 MR 4831034

[45] T. Cheetham-West, A. Lubotzky, A. W. Reid, and R. Spitler, Property FA is not a profinite property. *Groups Geom. Dyn.* **19** (2025), no. 3, 1081–1087 Zbl 08098353 MR 4945540

[46] O. Cotton-Barratt, Detecting ends of residually finite groups in profinite completions. *Math. Proc. Cambridge Philos. Soc.* **155** (2013), no. 3, 379–389 Zbl 1303.20032 MR 3118408

[47] M. W. Davis, *Infinite group actions on polyhedra*. Ergeb. Math. Grenzgeb. (3) 77, Springer, Cham, 2024, 271 pp. Zbl 1542.20002 MR 4769472

[48] M. Dehn, Über unendliche diskontinuierliche Gruppen. *Math. Ann.* **71** (1911), no. 1, 116–144 JFM 42.0508.03 MR 1511645

[49] J. D. Dixon, E. W. Formanek, J. C. Poland, and L. Ribes, Profinite completions and isomorphic finite quotients. *J. Pure Appl. Algebra* **23** (1982), no. 3, 227–231 Zbl 0477.20014 MR 0644274

[50] L. Funar, Torus bundles not distinguished by TQFT invariants. Appendix by Louis Funar and Andrei Rapinchuk. *Geom. Topol.* **17** (2013), no. 4, 2289–2344 Zbl 1278.57020 MR 3109869

[51] D. Gabai, R. Meyerhoff, and P. Milley, Minimum volume cusped hyperbolic three-manifolds. *J. Amer. Math. Soc.* **22** (2009), no. 4, 1157–1215 Zbl 1204.57013 MR 2525782

[52] M. Gromov, Hyperbolic groups. In S. M. Gersten (ed.), *Essays in group theory*, pp. 75–263, Math. Sci. Res. Inst. Publ. 8, Springer, New York, 1987 Zbl 0634.20015 MR 0919829

[53] E. K. Grossman, On the residual finiteness of certain mapping class groups. *J. Lond. Math. Soc. (2)* **9** (1974), no. 1, 160–164 Zbl 0292.20032 MR 0405423

[54] A. Grothendieck, Representations lineaires et compactification profinie des groupes discrets. *Manuscripta Math.* **2** (1970), 375–396 Zbl 0239.20065 MR 0262386

[55] F. Grunewald, A. Jaikin-Zapirain, and P. A. Zalesskii, Cohomological goodness and the profinite completion of Bianchi groups. *Duke Math. J.* **144** (2008), no. 1, 53–72 Zbl 1194.20029 MR 2429321

[56] F. Grunewald and P. Zalesskii, Genus for groups. *J. Algebra* **326** (2011), no. 1, 130–168 Zbl 1222.20019 MR 2746056

[57] F. J. Grunewald, On some groups which cannot be finitely presented. *J. Lond. Math. Soc. (2)* **17** (1978), no. 3, 427–436 Zbl 0385.20020 MR 0500627

[58] F. J. Grunewald, P. F. Pickel, and D. Segal, Polycyclic groups with isomorphic finite quotients. *Ann. of Math.* **111** (1980), no. 1, 155–195 Zbl 0431.20033 MR 0558400

[59] F. Haglund and D. T. Wise, Special cube complexes. *Geom. Funct. Anal.* **17** (2008), no. 5, 1551–1620 Zbl 1155.53025 MR 2377497

[60] J. Hempel, Residual finiteness for 3-manifolds. In S. M. Gersten and J. R. Stallings (eds.), *Combinatorial group theory and topology (Alta, Utah, 1984)*, pp. 379–396, Ann. of Math. Stud. 111, Princeton University Press, Princeton, NJ, 1987 Zbl 0772.57002 MR 0895623

[61] J. Hempel, Some 3-manifold groups with the same finite quotients. 2014, arXiv: 1409.3509v2

[62] G. Higman, A finitely generated infinite simple group. *J. Lond. Math. Soc.* **26** (1951), no. 1, 61–64 Zbl 0042.02201 MR 0038348

[63] G. Higman, Subgroups of finitely presented groups. *Proc. Roy. Soc. London Ser. A* **262** (1961), no. 1311, 455–475 Zbl 0104.02101 MR 0130286

[64] S. Hughes and M. Kudlinska, On profinite rigidity amongst free-by-cyclic groups I: the generic case. *Proc. Lond. Math. Soc. (3)* **130** (2025), no. 6, article no. e70059, 43 pp. Zbl 1569.20060 MR 4920397

[65] A. Jaikin-Zapirain, Recognition of being fibered for compact 3-manifolds. *Geom. Topol.* **24** (2020), no. 1, 409–420 Zbl 1455.57024 MR 4080486

[66] A. Jaikin-Zapirain, The finite and solvable genus of finitely generated free and surface groups. *Res. Math. Sci.* **10** (2023), no. 4, article no. 44, 24 pp. Zbl 1523.20050 MR 4651132

[67] A. Jaikin-Zapirain and A. Lubotzky, Some remarks on Grothendieck pairs. *Groups Geom. Dyn.* **19** (2025), no. 2, 597–616 Zbl 08098336 MR 4940654

[68] H. Kammeyer and R. Sauer, S-arithmetic spinor groups with the same finite quotients and distinct ℓ^2-cohomology. *Groups Geom. Dyn.* **14** (2020), no. 3, 857–869 Zbl 1511.20194 MR 4167024

[69] M. Kassabov and N. Nikolov, *Profinite properties and property Tau*, pp. 1550–1551, Oberwolfach Report No. 28/2008

[70] D. Kielak, Residually finite rationally solvable groups and virtual fibring. *J. Amer. Math. Soc.* **33** (2020), no. 2, 451–486 Zbl 1480.20102 MR 4073866

[71] S. Kionke and E. Schesler, Amenability and profinite completions of finitely generated groups. *Groups Geom. Dyn.* **17** (2023), no. 4, 1235–1258 Zbl 1527.20038 MR 4641368

[72] M. Lackenby, Detecting large groups. *J. Algebra* **324** (2010), no. 10, 2636–2657 Zbl 1231.20026 MR 2725193

[73] Y. Liu, Finite-volume hyperbolic 3-manifolds are almost determined by their finite quotient groups. *Invent. Math.* **231** (2023), no. 2, 741–804 Zbl 1511.57022 MR 4542705

[74] Y. Liu, Finite quotients, arithmetic invariants, and hyperbolic volume. *Peking Math. J.* **8** (2025), no. 1, 143–189 Zbl 07990815 MR 4870949

[75] D. D. Long and A. W. Reid, Grothendieck's problem for 3-manifold groups. *Groups Geom. Dyn.* **5** (2011), no. 2, 479–499 Zbl 1250.57027 MR 2782181

[76] J. Lott and W. Lück, L^2-topological invariants of 3-manifolds. *Invent. Math.* **120** (1995), no. 1, 15–60 Zbl 0876.57050 MR 1323981

[77] A. Lubotzky, Finiteness properties and profinite completions. *Bull. Lond. Math. Soc.* **46** (2014), no. 1, 103–110 Zbl 1295.20029 MR 3161766

[78] A. Lubotzky and D. Segal, *Subgroup growth*. Progr. Math. 212, Birkhäuser, Basel, 2003, 453 pp. Zbl 1071.20033 MR 1978431

[79] W. Lück, Approximating L^2-invariants by their finite-dimensional analogues. *Geom. Funct. Anal.* **4** (1994), no. 4, 455–481 Zbl 0853.57021 MR 1280122

[80] C. Maclachlan and A. W. Reid, *The arithmetic of hyperbolic 3-manifolds*. Grad. Texts in Math. 219, Springer, New York, 2003, 463 pp. Zbl 1025.57001 MR 1937957

[81] A. Malcev, On isomorphic matrix representations of infinite groups (in Russian). *Rec. Math. [Mat. Sb.] N.S.* **8(50)** (1940), no. 3, 405–422 Zbl 0025.00804 MR 0003420

[82] S. Meskin, A finitely generated residually finite group with an unsolvable word problem. *Proc. Amer. Math. Soc.* **43** (1974), no. 1, 8–10 Zbl 0261.20020 MR 0335645

[83] K. A. Mihailova, The occurrence problem for direct products of groups (in Russian). *Dokl. Akad. Nauk SSSR* **119** (1958), 1103–1105 Zbl 0084.25302 MR 0100018. English translation: *Mat. Sb. (N.S.)* **70(112)** (1966), no. 2, 241–251 Zbl 0161.02102 MR 0194497

[84] C. F. Miller III, *On group-theoretic decision problems and their classification*. Ann. of Math. Stud. 68, Princeton University Press, Princeton, NJ; University of Tokyo Press, Tokyo, 1971, 106 pp. Zbl 0277.20054 MR 0310044

[85] V. Nekrashevych, An uncountable family of 3-generated groups with isomorphic profinite completions. *Internat. J. Algebra Comput.* **24** (2014), no. 1, 33–46 Zbl 1297.20029 MR 3189664

[86] G. A. Niblo and M. A. Roller (eds.), *Geometric group theory. Volume 2: Asymptotic invariants of infinite groups, M. Gromov (Sussex, 1991)*, pp. 1–295, London Math. Soc. Lecture Note Ser. 182, Cambridge University Press, Cambridge, 1993 Zbl 0841.20039 MR 1253544

[87] N. Nikolov and D. Segal, On finitely generated profinite groups. I: strong completeness and uniform bounds. *Ann. of Math. (2)* **165** (2007), no. 1, 171–238 Zbl 1126.20018 MR 2276769

[88] G. A. Noskov, V. N. Remeslennikov, and V. A. Roman'kov, Infinite groups (in Russian). *Itogi Nauki i Tekhniki. Ser. Algebra. Topol. Geom.* **17** (1979), 65–157. English translation: *J. Soviet Math.* **85** (1982), no. 5, 669–735 Zbl 0429.20001 MR 0584569

[89] P. S. Novikov, On the algorithmic unsolvability of the word problem in group theory (in Russian). *Trudy Mat. Inst. Steklov.* **44** (1955), 3–143 Zbl 0068.01301 MR 0075197

[90] Y. Ollivier, Some small cancellation properties of random groups. *Internat. J. Algebra Comput.* **17** (2007), no. 1, 37–51 Zbl 1173.20025 MR 2300404

[91] P. F. Pickel, Finitely generated nilpotent groups with isomorphic finite quotients. *Trans. Amer. Math. Soc.* **160** (1971), 327–341 Zbl 0235.20027 MR 0291287

[92] P. Piwek, Profinite rigidity properties of central extensions of 2-orbifold groups. *Algebr. Geom. Topol.* **26** (2026), no. 2, 565–597 Zbl 08167296 MR 5034308

[93] V. P. Platonov and O. I. Tavgen', Grothendieck's problem on profinite completions and representations of groups. *K-Theory* **4** (1990), no. 1, 89–101 Zbl 0722.20020 MR 1076526

[94] P. Przytycki and D. T. Wise, Separability of embedded surfaces in 3-manifolds. *Compos. Math.* **150** (2014), no. 9, 1623–1630 Zbl 1305.57039 MR 3260144

[95] P. Przytycki and D. T. Wise, Mixed 3-manifolds are virtually special. *J. Amer. Math. Soc.* **31** (2018), no. 2, 319–347 Zbl 1511.57025 MR 3758147

[96] L. Pyber, Groups of intermediate subgroup growth and a problem of Grothendieck. *Duke Math. J.* **121** (2004), no. 1, 169–188 Zbl 1057.20019 MR 2031168

[97] M. O. Rabin, Recursive unsolvability of group theoretic problems. *Ann. of Math. (2)* **67** (1958), 172–194 Zbl 0079.24802 MR 0110743

[98] A. W. Reid, *Profinite rigidity, Grothendieck pairs and low-dimensional topology*, to appear in KIAS Lecture Note Series, https://math.rice.edu∼ar99/KIAS_Reid2.pdf visited on 8 January 2026

[99] A. W. Reid, Profinite properties of discrete groups. In C. M. Campbell, M. R. Quick, E. F. Robertson, and C. M. Roney-Dougal (eds.), *Groups St Andrews 2013*, pp. 73–104, London Math. Soc. Lecture Note Ser. 422, Cambridge University Press, Cambridge, 2015 Zbl 1354.20018 MR 3445488

[100] A. W. Reid, Profinite rigidity. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. II. Invited lectures*, pp. 1193–1216, World Scientific, Hackensack, NJ, 2018 Zbl 1445.20026 MR 3966805

[101] V. N. Remeslennikov, Conjugacy of subgroups in nilpotent groups (in Russian). *Algebra Logika* **6** (1967), no. 2, 61–76 Zbl 0189.30902 MR 0218459

[102] B. Rémy, Exposé Bourbaki 1221: complétion profinie et géométrie des groupes (d'après Martin R. Bridson, Alan Reid,...). *Astérisque* **454** (2024), 429–465 Zbl 08132298 MR 4860594

[103] L. Ribes and P. Zalesskii, *Profinite groups*. Ergeb. Math. Grenzgeb. (3) 40, Springer, Berlin, 2000, 435 pp. Zbl 0949.20017 MR 1775104

[104] E. Rips, Subgroups of small cancellation groups. *Bull. Lond. Math. Soc.* **14** (1982), no. 1, 45–47 Zbl 0481.20020 MR 0642423

[105] G. Sabbagh and J. S. Wilson, Polycyclic groups, finite images, and elementary equivalence. *Arch. Math. (Basel)* **57** (1991), no. 3, 221–227 Zbl 0702.20001 MR 1119894

[106] M. Sageev, $CAT(0)$ cube complexes and groups. In *Geometric group theory*, pp. 7–54, IAS/Park City Math. Ser. 21, American Mathematical Society, Providence, RI, 2014 Zbl 1440.20015 MR 3329724

[107] P. Schwer, CAT(0) *cube complexes. An introduction*. Lecture Notes in Math. 2324, Springer, Cham, 2023, 188 pp. Zbl 1537.57001 MR 4738388

[108] P. Scott, The geometries of 3-manifolds. *Bull. Lond. Math. Soc.* **15** (1983), no. 5, 401–487 Zbl 0561.57001 MR 0705527

[109] J.-P. Serre, Exemples de variétés projectives conjuguées non homéomorphes. *C. R. Acad. Sci. Paris* **258** (1964), 4194–4196 Zbl 0117.38003 MR 0166197

[110] J.-P. Serre, Problems. In *Proceedings of the second international conference on the theory of groups, Australian National University, August 13–24, 1973*, pp. 734–735, Lecture Notes in Math. 372, Springer, Berlin, 1974

[111] J.-P. Serre, *Trees*. Springer, Berlin, 1980, 142 pp. Zbl 0548.20018 MR 0607504

[112] J.-P. Serre, *Cohomologie galoisienne*. 5th edn., Lecture Notes in Math. 5, Springer, Berlin, 1994, 181 pp. Zbl 0812.12002 MR 1324577

[113] A. M. Slobodskoi, Undecidability of the universal theory of finite groups (in Russian). *Algebra Logika* **20** (1981), no. 2, 207–230 Zbl 0519.03006 MR 0641581

[114] R. F. Spitler, *Profinite completions and representations of finitely generated groups*. Ph.D. thesis, Purdue University, 2019 MR 4625560

[115] J. Stallings, A finitely presented group whose 3-dimensional integral homology is not finitely generated. *Amer. J. Math.* **85** (1963), no. 4, 541–543 Zbl 0122.27301 MR 0158917

[116] J. R. Stallings, Topology of finite graphs. *Invent. Math.* **71** (1983), no. 3, 551–565 Zbl 0521.20013 MR 0695906

[117] P. F. Stebe, Conjugacy separability of groups of integer matrices. *Proc. Amer. Math. Soc.* **32** (1972), 1–7 Zbl 0242.20048 MR 0289666

[118] M. Stover, Lattices in PU(n, 1) that are not profinitely rigid. *Proc. Amer. Math. Soc.* **147** (2019), no. 12, 5055–5062 Zbl 1453.20040 MR 4021068

[119] M. Stover, Algebraic fundamental groups of fake projective planes. *J. Eur. Math. Soc. (JEMS)* (2024), DOI 10.4171/JEMS/1536

[120] H. Sun, All finitely generated 3-manifold groups are Grothendieck rigid. *Groups Geom. Dyn.* **17** (2023), no. 2, 385–402 Zbl 1514.20098 MR 4584669

[121] G. Wilkes, Profinite rigidity for Seifert fibre spaces. *Geom. Dedicata* **188** (2017), 141–163 Zbl 1369.57003 MR 3639628

[122] G. Wilkes, Profinite completions, cohomology and JSJ decompositions of compact 3-manifolds. *New Zealand J. Math.* **48** (2018), 101–113 Zbl 1411.57003 MR 3884906

[123] G. Wilkes, Profinite rigidity of graph manifolds and JSJ decompositions of 3-manifolds. *J. Algebra* **502** (2018), 538–587 Zbl 1404.57039 MR 3774902

[124] G. Wilkes, Relative cohomology theory for profinite groups. *J. Pure Appl. Algebra* **223** (2019), no. 4, 1617–1688 Zbl 1443.20048 MR 3906519

[125] G. Wilkes, *Profinite groups and residual finiteness*. EMS Textbk. Math., EMS Press, Berlin, 2024, pp. 434 Zbl 1557.20001 MR 4809945

[126] H. Wilton and A. Sisto, The congruence subgroup property for mapping class groups and the residual finiteness of hyperbolic groups (with an appendix by A. Sisto). 2024, arXiv:2410.00556v2

[127] H. Wilton and P. Zalesskii, Distinguishing geometries using finite quotients. *Geom. Topol.* **21** (2017), no. 1, 345–384 Zbl 1361.57023 MR 3608716

[128] H. Wilton and P. Zalesskii, Profinite detection of 3-manifold decompositions. *Compos. Math.* **155** (2019), no. 2, 246–259 Zbl 1436.57020 MR 3896566

[129] D. T. Wise, Subgroup separability of graphs of free groups with cyclic edge groups. *Q. J. Math.* **51** (2000), no. 1, 107–129 Zbl 0991.05056 MR 1760573

[130] D. T. Wise, The residual finiteness of negatively curved polygons of finite groups. *Invent. Math.* **149** (2002), no. 3, 579–617 Zbl 1040.20024 MR 1923477

[131] D. T. Wise, A residually finite version of Rips's construction. *Bull. Lond. Math. Soc.* **35** (2003), no. 1, 23–29 Zbl 1027.20014 MR 1934427

[132] D. T. Wise, Cubulating small cancellation groups. *Geom. Funct. Anal.* **14** (2004), no. 1, 150–214 Zbl 1071.20038 MR 2053602

[133] D. T. Wise, *The structure of groups with a quasiconvex hierarchy*. Ann. of Math. Stud. 209, Princeton University Press, Princeton, NJ, 2021, 376 pp. Zbl 1511.20002 MR 4298722

[134] X. Xu, Profinite almost rigidity in 3-manifolds. *Adv. Math.* **480** (2025), no. part B, article no. 110505, 79 pp. Zbl 08105107 MR 4953004

[135] A. Yu. Ol'shanskii and M. V. Sapir, A 2-generated 2-related group with no non-trivial finite quotients. *Algebra Discrete Math.* (2007), no. 2, 111–114 Zbl 1164.20343 MR 2364067

Martin R. Bridson
Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford OX2 6GG, UK; bridson@maths.ox.ac.uk

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/2

Generalized Riemann–Hilbert correspondence, wall-crossing and resurgence

Maxim Kontsevich

Abstract. Together with Yan Soibelman, we propose a new version of homological mirror symmetry, a hypothetical equivalence between two categories associated with a complex algebraic symplectic variety. In the case of cotangent bundles, it gives the classical Riemann–Hilbert correspondence, which is the equivalence between the category of holonomic D-modules with regular singularities on a complex algebraic variety and the category of perverse constructible sheaves. In general, the de Rham side is replaced by the algebraic deformation quantization, whereas the Betti side is replaced by the Fukaya category. New theory is applicable not only to D-modules, but also to q-difference and elliptic difference equations. The key role is played by the theory of wall-crossing, hypothetically leading to resurgence properties of WKB expansions.

This note is an extended version of my talk at ECM 2024, Seville, where I tried to give an introduction to the general program we (Yan Soibelman and I) have been working on for the last 10 years, which we have dubbed "Holomorphic Floer Theory". Most of the text focuses on the foundational issues; the wall-crossing and resurgence only appear at the end. I apologize in advance to the reader for not always being precise, and often quite vague. One can also consult our recent preprint [21] where we considered a more manageable part of the general program related to closed 1-forms and also made several speculative predictions concerning resurgence and path integrals. Also, in the case of complex symplectic surfaces, most of the hypothetical constructions can be made precise; a detailed exposition will be given in [22] which we are preparing for publication.

The emerging picture is magnificent: it unifies many branches of mathematics into a new "algebraic analysis".

Mathematics Subject Classification 2020: 53D55 (primary); 53D37 (secondary).
Keywords: symplectic geometry, deformation quantization, Fukaya category.

1. Main conjecture

Let (M, ω) be a smooth algebraic *symplectic* variety over $\mathbb{C}$. Assume that it is *A-good* and *B-good* at infinity (or (A, B)-good in short), meaning

A-good: (a complex version of *geometrically bounded real symplectic manifolds*, cf. [8]) $\exists$ *complete* Hermitian metric g on $M(\mathbb{C})$ such that the curvature tensor of g, (injectivity radius of $g)^{-1}$, 2-form $\omega \in \Gamma(M, \wedge^2 T^*_M)$ and Poisson bivector $\omega^{-1} \in \Gamma(M, \wedge^2 T_M)$ are all *uniformly* bounded.

B-good: $\exists$ a smooth compact algebraic Poisson variety P over $\mathbb{C}$ such that (M, ω) is its open Zariski-dense symplectic leaf (see Section 2.5 for a criticism).

Examples. For *any* smooth algebraic variety X over $\mathbb{C}$ (not necessarily compact), its cotangent bundle $M = T^*X$ endowed with the standard symplectic form is (A, B)-good. Also, the complex algebraic torus $(\mathbb{C}^\times)^{2n}$ endowed with the symplectic form

$$\omega = \sum_{i=1}^{n} \frac{dx_i}{x_i} \wedge \frac{dy_i}{y_i} \tag{1.1}$$

is (A, B)-good. Any compact algebraic symplectic manifold (M, ω) (e.g., a K3 surface) is automatically (A, B)-good at infinity. An example of a symplectic variety which is neither A- nor B-good is $M = \mathbb{C}^2 - 0, \omega = dx \wedge dy$. It looks plausible that the condition of being B-good implies that of being A-good.

In what follows, when I speak about categories, I mean $(\infty, 1)$-categories associated with A_∞- or dg-categories, linear over some base ring R.

Informal conjecture (not yet properly formulated...).

- *(A-side $\Leftrightarrow$ Betti side) we have an analytic family of Fukaya categories (twisted by the so-called B-field $\beta_\hbar$) over $\mathbb{C}^\times_\hbar = \mathbb{C}_\hbar - 0$:*

$$C_{A,\hbar}(M, \omega) = \mathcal{F}\left(M, \Re\left(\frac{\omega}{\hbar}\right), \beta_\hbar\right), \quad \beta_\hbar := \mathrm{i} \cdot \left[\Im\left(\frac{\omega}{\hbar}\right)\right] \in H^2(M, \mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z}).$$

- *(B-side $\Leftrightarrow$ de Rham side) we have an analytic family of categories of deformation quantization modules over $\mathbb{C}_\hbar$:*

$$C_{B,\hbar}(M, \omega), \quad \textit{fiber at } \hbar = 0 \textit{ is } C_{B,0}(M, \omega) \overset{\text{full}}{\hookrightarrow} \mathrm{Perf}_c(M).$$

- *Generalized Riemann–Hilbert correspondence (GRH in short):*

$$C_{A,\hbar}(M, \omega) \simeq C_{B,\hbar}(M, \omega) \quad \textit{for } \hbar \neq 0,$$

with the equivalence holomorphically dependent on $\hbar$. Moreover, both categories are endowed with bounded t-structures preserved by GRH, $\Rightarrow$ equivalence of abelian hearts $C^{\mathrm{heart}}_{A,\hbar}(M,\omega) \simeq C^{\mathrm{heart}}_{B,\hbar}(M,\omega)$.

The notations will be explained later, in both cases we deal implicitly with objects whose "support" is compact Lagrangian *(whatever it means...). The important point is that for any given complex parameter $\hbar \neq 0$, the notion of Fukaya category belongs to the realm of symplectic topology for real C^∞-manifolds. The B-side categories are (hypothetically) defined in terms of algebraic/analytic deformation quantization. The main challenge is to find a satisfactory general definition for both the A and B cases.*

I did not explain here what I mean speaking above about an "analytic family" of categories. The full definition is based on a quite technical notion of higher derived analytic stacks. The basic example which one should have in mind is the following. Consider an analytic family $(Y_s)_{s\in S}$ of compact complex *not necessarily algebraic* manifolds parametrized by a complex-analytic space S. Then the categories of perfect complexes $\mathrm{Perf}(Y_s)$ should form an analytic family over S. As a part of the structure, for any analytic map $\widetilde{S} \to S$, one gets a *sheaf* of A_∞-categories on $\widetilde{S}$ enriched over the sheaf $\mathrm{Perf}_{\widetilde{S}}$ of symmetric monoidal $(\infty,1)$-categories of perfect complexes.

Finally, I have to explain the origin of the names A and B for the sides of GRH correspondence. One can view the generalized Riemann–Hilbert correspondence as a cousin of *homological mirror symmetry* (see [14]). The difference is that in HMS the A-model on one variety is equivalent to the B-model on *another* variety, the mirror dual to the original one, whereas in the GRH correspondence we deal on both sides with the *same* variety. Also, the B-side in GRH correspondence is not the derived category of coherent sheaves (or rather, a certain full subcategory of it), but a deformation (non-perturbative quantization).

1.1. Relation to the classical RH correspondence

Let $M = T^*X$ where X is a smooth *compact* algebraic variety over $\mathbb{C}$, with the standard algebraic symplectic form ω. Assume that $c_1(X) \in H^2(X,\mathbb{Z})$ is divisible by 2, and a square root line bundle $\sqrt{K_X}$ is chosen.

- $C_{A,\hbar}(M,\omega)$ does not depend on $\hbar \neq 0$, and its abelian heart $C^{\mathrm{heart}}_{A,\hbar}(M,\omega)$ coincides with category of $\mathbb{C}$-local systems of finite rank on X.
- For $\hbar \neq 0$ category, $C^{\mathrm{heart}}_{B,\hbar}(M,\omega)$ consists of holonomic D_X-modules which are algebraic vector bundles with algebraic flat connections on X (i.e., in local coordinates, modules over the algebra generated by operators x_i and $\hbar\partial_{x_i}$). If X is projective, the fiber at $\hbar = 0$ is the category of semistable Higgs bundles $(\mathcal{E},\theta)$ where $c_1(\mathcal{E}) = 0$ and $\theta \in \Gamma(X, \mathrm{End}\,\mathcal{E} \otimes \Omega^1_X)$, $\theta \wedge \theta = 0$. It is the same as the category of coherent sheaves with compact support on M which are semistable of

slope 0 for some (asymptotic) Bridgeland stability condition on the triangulated category $\mathrm{Perf}_c(M)$ of compactly supported perfect complexes on M.

One can omit the somewhat annoying condition that $c_1(T_X)$ is divisible by 2. In general, for both categories $C_{A,\hbar}(M,\omega)$ and $C_{B,\hbar}(M,\omega)$ (and their hearts), one can define *twists* by any class $\beta \in H^2(M, \mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z})$. The convenient twist for $M = T^*X$ is 2-torsion and is given by the pullback to M of the image of $c_1(T_X) \in H^2(X,\mathbb{Z})$ via the map on coefficients $1 \in \mathbb{Z} \mapsto \pi\mathrm{i} \pmod{2\pi\mathrm{i}\mathbb{Z}}$.

1.2. Hodge-theoretic invariant

With each of the categories $C_{A,\hbar}(M,\omega)$, $C_{B,\hbar}(M,\omega)$, we expect to be able to associate a $\mathbb{Z}/2$-graded finitely generated abelian group, its *topological K-theory*, which in both A and B cases is canonically identified (by its own reasons in A and B cases) with the compactly supported topological K-theory $K_c^{\mathrm{top}}(M) = K_{c,\mathrm{even}}^{\mathrm{top}} \oplus K_{c,\mathrm{odd}}^{\mathrm{top}}(M)$. The equivalence is compatible with this identification. Moreover, we should have a cycle map

$$\mathrm{cl} : K_0(C_{A \text{ or } B,\hbar}(M,\omega)) \to K_{c,\mathrm{even}}^{\mathrm{top}}(M).$$

Also, on $K_{c,\mathrm{even}}^{\mathrm{top}}(M) \otimes \mathbb{C} \simeq H_c^{\mathrm{even}}(M,\mathbb{C})$ (the isomorphism is given by $[\mathcal{E}] \mapsto \sqrt{\mathrm{Todd}(T_M)} \cdot \mathrm{ch}(\mathcal{E})$), we have *Hodge filtration* (cf. [17, point 1.36]):

$$F^i := e^{\omega/\hbar}(H_c^0(M,\mathbb{C}) \oplus H_c^2(M,\mathbb{C}) \oplus \cdots \oplus H_c^{2n-2i}(M,\mathbb{C})).$$

The image of the cycle map cl is expected to lie in F^0.

The Hodge filtration for the category $C_{B,\hbar}(M,\omega)$ at $\hbar = 0$ is defined as the limit of the above filtration at $\hbar \to 0$. We expect that there exists a limiting Bridgeland stability condition such that $C_{B,\hbar}^{\mathrm{heart}}(M,\omega)$ is an Artinian abelian full subcategory characterized as the category of semistable objects with slope 0. In particular, classes in $K_{c,\mathrm{even}}^{\mathrm{top}}(M)$ should belong to a certain subgroup (lying in F^0) cut out by the central charge Z.

1.3. Example related to the irregular RH correspondence

Let (M_0,ω_0) be $(\mathbb{C}^2, dx \wedge dy)$, the cotangent bundle to the affine line $\mathbb{A}^1_x$. It is an open symplectic leaf in the compact Poisson surface $P_0 = \mathbb{C}P^2$. After making several iterated blowups at points of the Poisson boundary $P_0 - M_0$, we embed M_0 as the open symplectic leaf into another Poisson surface P_1 which contains two boundary divisors D_0, D_1, each endowed with a coordinate $D_i \simeq \mathbb{C}^1$, $i = 0, 1$, such that

$$\lim_{x\to\infty} \underbrace{(x, -2x + t/x)}_{\in \mathbb{C}^2 = M_0} = \underbrace{t}_{\in \mathbb{C} = D_0}, \quad \lim_{x\to\infty} \underbrace{(x, +2x + t/x)}_{\in \mathbb{C}^2 = M_0} = \underbrace{t}_{\in \mathbb{C} = D_1}.$$

Symplectic form ω_0 has poles of order 1 at divisors D_0, D_1.

Let us make a *choice* of two complex numbers $\lambda_0, \lambda_1 \in \mathbb{C}$, and define Poisson surface $P = P_{(\lambda_0,\lambda_1)}$ to be the blowup of P_1 at two points in D_0, D_1 with coordinates λ_0, λ_1, respectively.

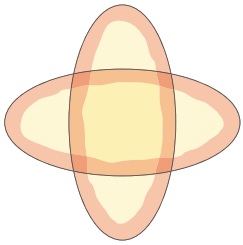
The open symplectic leaf (M, ω) of P is $(M_0 = \mathbb{C}^2) \sqcup \mathbb{C}^1_{(0)} \sqcup \mathbb{C}^1_{(1)}$ as a set; we add two open exceptional divisors. Here M is neither affine nor projective.

1.3.1. Category $C^{\text{heart}}_{A,\hbar}(M, \omega)$ in the example. Abelian category $C^{\text{heart}}_{A,\hbar}(M, \omega)$ is the category of tuples consisting of a vector space V and 4 subspaces $V_0, V_1, V_2, V_3 \subset V$ such that V_{i+1} is complementary to V_i (here put $V_4 := V_0$), and such that two compositions of isomorphisms

$$V_0 \xrightarrow{\sim} V/V_1 \xrightarrow{\sim} V_2 \xrightarrow{\sim} V/V_3 \xrightarrow{\sim} V_0, \quad V_1 \xrightarrow{\sim} V/V_2 \xrightarrow{\sim} V_3 \xrightarrow{\sim} V/V_0 \xrightarrow{\sim} V_1$$

are equal to $\exp(2\pi \mathrm{i}\lambda_0/\hbar) \cdot \mathrm{id}_{V_0}$ and $\exp(2\pi \mathrm{i}\lambda_1/\hbar) \cdot \mathrm{id}_{V_1}$, respectively. More precisely, enumeration of V_i is determined up to the involution $0 \leftrightarrow 2, 1 \leftrightarrow 3$, depending on the choice $\theta/2 \in \mathbb{R}/2\pi\mathbb{Z}$ where $\theta = \arg \hbar$. Speaking in a more invariant way, subspaces should be labeled by $\theta/2, \theta/2 + \pi/2, \theta/2 + \pi, \theta/2 + 3\pi/2$ instead of $0, 1, 2, 3$.

This category coincides with the category of constructible sheaves on the plane $\mathbb{R}^2 \simeq \mathbb{C}$ with the *microlocal support* (see [12]) contained in a closed bounded domain in zero section $\mathbb{R}^2 \subset T^*\mathbb{R}^2$ together with the positive conormal bundles to two co-oriented closed curves, depicted as



1.3.2. Category $C^{\text{heart}}_{B,\hbar}(M, \omega)$ in the example. Abelian category $C^{\text{heart}}_{B,\hbar}(M, \omega)$ for $\hbar \neq 0$ is the category of holonomic D-modules on $\mathbb{A}^1_x$ which are algebraic vector bundles with connections $(\mathcal{E}, \nabla)$ such that over $\mathbb{C}(\!(x^{-1})\!)$ the bundle splits into the sum of copies of two types of the trivial rank 1 bundles with connections

$$d - \hbar^{-1}\Big(-2x + \frac{\lambda_0}{x}\Big)dx, \quad d - \hbar^{-1}\Big(+2x + \frac{\lambda_1}{x}\Big)dx.$$

These two special equations have the following solutions near ∞:

$$x^{\lambda_0/\hbar} e^{-x^2/\hbar}, \quad x^{\lambda_1/\hbar} e^{+x^2/\hbar}.$$

For $\hbar = 0$, the abelian category $C^{\text{heart}}_{B,0}(M,\omega)$ should be defined as a full subcategory of $\text{Coh}_c(M)$ consisting of coherent sheaves with compact support which are semistable with slope 0 for certain asymptotic Bridgeland stability condition. This stability condition guarantees that the support of any non-zero object is purely 1 dimensional, so that the sheaves with 0-dimensional support are excluded.

1.3.3. Equivalence of categories. The equivalence $C_{A,\hbar}(M,\omega) \simeq C_{B,\hbar}(M,\omega)$ (at least on hearts) is (after a little work) a special case of 1-dimensional *irregular Riemann–Hilbert correspondence* by P. Deligne and B. Malgrange; see [4] (the higher-dimensional generalization was worked out in [3]). For a connection $(\mathcal{E},\nabla)$, the corresponding space V is the space of analytic flat sections of $\mathcal{E}$ over $\mathbb{C}$. The subspaces V_j (or, more precisely, $V_{\theta/2+j\cdot\pi/2}$) are subspaces of sections with the exponentially fast decay along rays with arguments $\theta/2 + j\cdot\pi/2$, $j = 0,\dots,3$.

The dependence of (any of two equivalent) categories $C_{A,\hbar}(M,\omega) \simeq C_{B,\hbar}(M,\omega)$ on parameters $\lambda_0, \lambda_1 \in \mathbb{C}$ is quite drastic – the category vanishes unless $a\lambda_0 + b\lambda_1 = 0$ for some pair of non-negative integers $(a,b) \in \mathbb{Z}^2_{\geq 0} - \{(0,0)\}$. In fact, one has $a/b = \dim V_0/\dim V_1$ for any non-zero object in the $C^{\text{heart}}_{A,\hbar}(M,\omega)$. The dependence on parameters can be promoted to an analytic family of categories over $\mathbb{C}^2_{\lambda_0,\lambda_1} \times \mathbb{C}^\times_\hbar$.

1.4. Non-compact support: Log-Poisson extensions

In the previous example, we made a certain non-trivial effort in the definition of symplectic variety (M,ω), by enlarging the more familiar cotangent bundle. One can follow another route, which is more directly related to the classical (irregular) RH correspondence.

Let us assume that M is Zariski open symplectic leaf in a compact algebraic Poisson variety (P,γ), $\gamma \in \Gamma(P, \wedge^2 T_P)$ such that the complement $P - M$ is a divisor with *normal crossings*. Let $P_{\log} \subset P$ be the set of points such that there exist local analytic coordinates $(x_1,\dots,x_n;\ y_1,\dots,y_n)$ in which the Poisson structure is given by

$$\gamma = \sum_{i=1}^{k} x_i \partial_{x_i} \wedge \partial_{y_i} + \sum_{i=k+1}^{n} \partial_{x_i} \wedge \partial_{y_i}$$

for some integer k, $0 \leq k \leq n$. It follows that $P_{\log}$ is Zariski open subset in P containing M. The (possibly non-compact) Poisson variety $(P, \gamma_{|P})$ is called the *log-Poisson extension* of (M,ω).

In the case $M = T^*X$ where X is a smooth (possibly non-compact) curve, log-Poisson extensions of M correspond to arbitrary finite collections of irregular terms

for holonomic D_X-modules, that is, expressions of the type

$$\exp\left(\sum_{j=1}^{N'} c_j x^{-j/N} + O(\log|x|)\right) \quad \text{up to } c_j \rightsquigarrow e^{\frac{2\pi \mathrm{i}}{N} j} c_j, \quad N, N' \in \mathbb{N},\ c_j \in \mathbb{C},$$

where x is a local coordinate at a point of $\overline{X}$.

Different Poisson compactifications $P, P' \supset M$ can give the *same* log-Poisson extensions $P_{\log} \simeq P'_{\log}$. It looks plausible that one can define non-compactly supported versions of categories $C_{A,\hbar}(M,\omega)$ and $C_{B,\hbar}(M,\omega)$ which depend *only* on the log-Poisson extension, without the choice of a full Poisson compactification.

The non-compactly supported version of $C_{A,\hbar}(M,\omega)$ for $\hbar \neq 0$ should be an appropriately defined *partially wrapped* Fukaya category, where objects are roughly real Lagrangian submanifolds in $\left(M, \Re\left(\frac{\omega}{\hbar}\right)\right)$ whose closure in $P_{\log}$ is compact and behaves well in an appropriate sense at infinity $P_{\log} - M$.

We expect that in the case $M = T^*X$ and a given log-Poisson extension $P \supset M$, the non-compactly supported version of $C^{\mathrm{heart}}_{B,\hbar}(M,\omega)$ for $\hbar \neq 0$ can be directly identified with the full category of holonomic D_X-modules consisting of objects for which the possible irregular terms of restrictions to all analytic punctured discs in X are specified, in a certain way associated with P.

An important open question here is "symplectic resolution of singularities".

Conjecture. *For any B-good symplectic variety (M,ω) and possibly singular algebraic closed Lagrangian subset $L \subset M$, there exists a log-Poisson extension $P_{\log} \supset M$ such that the closure of L in $P_{\log}$ is compact.*

This conjecture holds by elementary means in the case $\dim M = 2$ (see, e.g., the discussion of blowups in [19, Section 8.3] where irregular terms were called "singular terms") and, more surprisingly, for the case of an algebraic torus of *any* dimension, with the "trigonometric" symplectic structure as in (1.1).

1.5. Trigonometric case: Quantum tori

I will present a version of GRH correspondence for the algebraic torus (1.1), implicitly using some log-Poisson extensions.

Let us start with the two-dimensional case, $(M_0,\omega) = \left((\mathbb{C}^\times)^2, \frac{dx}{x} \wedge \frac{dy}{y}\right)$. Its natural algebraic quantization is algebra

$$A_\hbar = \mathbb{C}\langle X^{\pm 1}, Y^{\pm 1}\rangle / \text{relation } YX = qXY, \quad q := e^\hbar. \tag{1.2}$$

For $\hbar \notin 2\pi\mathrm{i}\mathbb{Q}$, a finitely generated module E over $A_\hbar$ is called *holonomic* iff it admits an exhaustive increasing filtration $E_0 \subset E_1 \subset \cdots$ such that

$$X^{\pm 1} E_i, Y^{\pm 1} E_i \subset E_{i+1}, \quad \dim E_i = O(i). \tag{1.3}$$

Theorem (Can be extracted from [26]). *In the case $\hbar \notin \mathrm{i}\mathbb{R}$, the category of holonomic $A_\hbar$ modules is equivalent to the category of triples consisting of a coherent sheaf $\mathcal{F}$ on elliptic curve $\mathbb{C}^\times/q^{\mathbb{Z}}$ and two exhaustive ascending filtrations of $\mathcal{F}$ by the ordered set $\mathbb{Q} \cup \{+\infty\}$ such that the associated graded sheaves are torsion sheaves for the grading $+\infty$, and semistable vector bundles with the slope* $\mathrm{rk}/\deg = a/b$ *for gradings $a/b \in \mathbb{Q}$.*

Notice that these filtrations are labeled in the order *opposite* to the standard Harder–Narasimhan filtration. Indices of two filtrations can be encoded geometrically as a finite collection of rays with rational slopes in $\mathbb{R}^2$.

Fukaya category perspective leads to a similar description on the *same* elliptic curve but now written as $\mathbb{C}^\times/\widetilde{q}^{\mathbb{Z}}$ where $\widetilde{q} = e^{(2\pi\mathrm{i})^2/\hbar}$.

One can generalize (hypothetically) this GRH correspondence to arbitrary dimension as in (1.1). The quantum algebra here is $A_\hbar^{\otimes n}$. Condition (1.3) in the definition of a holonomic module is replaced by the condition

$$\dim E_i = O(i^n).$$

On the A-side, one replaces the notion of two anti-Harder–Narasimhan filtrations by a certain mix of the local Fukaya category associated with a closed rational polyhedral Lagrangian cone in $\mathbb{R}^{2n}$, and the bounded derived category of coherent sheaves on the abelian variety $(\mathbb{C}^\times/\widetilde{q}^{\mathbb{Z}})^n$; see Section 3.3.3 for more details.

2. *B*-side: Deformation quantization

2.1. Formal deformation quantization

One part of the main conjecture implies that for a B-good algebraic symplectic variety (M, ω), we have a canonical A_∞-category enriched over the $\mathrm{Perf}(\mathbb{C}\{\hbar\})$. Therefore, by base change, we get a category enriched over $\mathrm{Perf}(\mathbb{C}[\![\hbar]\!])$.

The usual *formal deformation quantization* offers a natural guess. Namely, it follows from formality theorem (see [16]) that for any complex Poisson variety (P, γ) (here $\gamma \in \Gamma(P, \wedge^2 T_P)$, $[\gamma, \gamma] = 0 \in \Gamma(P, \wedge^3 T_P)$), without any assumption on algebraicity, one get a canonical sheaf of $\mathbb{C}[\![\hbar]\!]$-linear algebroids $\mathcal{O}_{q,P}$ on P (algebroids are roughly algebras up to inner automorphisms; see [15]) and the corresponding canonical sheaf of flat $\mathbb{C}[\![\hbar]\!]$-linear triangulated A_∞-categories on P. In the case when γ is non-degenerate (i.e., we have a symplectic manifold (M, ω) and $\gamma = \omega^{-1}$), locally this sheaf is equivalent to the sheaf of perfect complexes over the sheaf of flat $\mathbb{C}[\![\hbar]\!]$-linear algebras given in local Darboux coordinates by the standard Moyal

product on $\mathcal{O}_{\mathbb{C}^{2n}}[\![\hbar]\!]$:

$$f \star g = \sum_{i_1,\dots,i_n \geq 0} \frac{\hbar^{i_1+\cdots+i_n}}{i_1!\cdots i_n!} \partial_{y_1}^{i_1} \cdots \partial_{y_n}^{i_n} f \cdot \partial_{x_1}^{i_1} \cdots \partial_{x_n}^{i_n} g. \tag{2.1}$$

In the polynomial case, one can associate with the function f a differential operator $f(x_1, \dots, x_n, \hbar\partial_{x_1}, \dots, \hbar\partial_{x_n})$, where one puts all derivatives on the right. Then the Moyal product gives the usual composition of operators.

We expect that the $\mathbb{C}[\![\hbar]\!]$-linear version of the B-model category $C_{B,\hbar}(M, \omega)$ is naturally equivalent to the category of perfect complexes of $\mathcal{O}_{q,M}$-modules which are *holonomic* and with *compact support*. Recall that the support of any perfect complex is a coisotropic analytic subset, and the object is called holonomic iff the support is Lagrangian (possibly singular); see [13]. There is a natural bounded t-structure on this category with the heart consisting of perfect complexes concentrated in degree 0.

Alternatively, one can use arbitrary Poisson compactification $P \supset M$ such that M is an open dense symplectic leaf of P, and consider the category of perfect complexes of $\mathcal{O}_{q,P}$-modules whose supports are compact Lagrangian subsets contained in M.

Finally, one can twist the sheaf of categories of perfect $\mathcal{O}_{q,P}$-modules by any 2-cocycle with coefficients in $\mathbb{C}[\![\hbar]\!]^\times$.

2.1.1. Deformation quantization of Poisson divisors. Here we present a useful result concerning formal deformation quantization.

An effective divisor $D \subset (P, \gamma)$ is called *Poisson* iff the ideal sheaf $\mathcal{I}_D = \mathcal{O}_P(-D) \subset \mathcal{O}_P$ is a sheaf of Poisson ideals: $\{\mathcal{I}_D, \mathcal{O}_P\} \subset \mathcal{I}_D$. An equivalent property is that the locus of smooth points of the associated reduced divisor D^{red} is a locally closed Poisson hypersurface.

If $M \subset P$ is an open, everywhere dense symplectic leaf, then the zero locus of $\gamma^{\dim P/2}$ (a section of the anti-canonical bundle $-K_P$) is a Poisson divisor, the associated reduced divisor is the complement $P - M$.

We claim that the canonical sheaf of algebroids $\mathcal{O}_{q,P}$ contains a canonical two-sided ideal $\mathcal{I}_{q,D}$, which is a flat deformation of $\mathcal{I}_D$. Moreover, if $f \in \mathcal{O}_{P,\mathrm{pt}}$ is a germ at a point $\mathrm{pt} \in P$ of an analytic function with zero locus D, then it can be lifted to a generator of $\widetilde{f} \in \mathcal{I}_{q,D,\mathrm{pt}}$ of $\mathcal{I}_{q,D,\mathrm{pt}}$ as a free rank 1 left $\mathcal{O}_{q,P,\mathrm{pt}}$-module (the same is true for $\mathcal{I}_{D_{\mathrm{pt}}}$ considered as a right module). As a corollary, we have a canonical sheaf of quotient algebroids $\mathcal{O}_{q,P}/\mathcal{I}_{q,D}$ supported on D and deforming $\mathcal{O}_D = \mathcal{O}_X/\mathcal{I}_D$.

The definition is easy at smooth points of D^{red}. Indeed, one can choose local coordinates $x_1, \dots, x_n$ near $\mathrm{pt} \in P$ (here $n = \dim P$) such that $\mathcal{I}_{D,\mathrm{pt}}$ is generated by x_1^k for some $k \geq 1$ (the multiplicity of D at pt). The condition of D being Poisson means that the bivector field γ is $\mathcal{O}_{P,\mathrm{pt}}$-linear combination of

$$x_1\partial_{x_1} \wedge \partial_{x_i}, \quad \partial_{x_i} \wedge \partial_{x_j}, \quad i, j \geq 2.$$

Explicit formulas for the $\star$-product in deformation quantization (see [15]) imply that $\mathcal{I}_{D,\mathrm{pt}} := x_1^k \mathcal{O}_{P,\mathrm{pt}}[\![\hbar]\!]$ is a two-sided $\star$-ideal, and x_1^k is its free generator as a left and as a right $\mathcal{O}_{q,P,\mathrm{pt}}$-module. Finally, one can extend the sheaf of ideals to the whole D by the Hartogs principle.

2.2. Hypothetical non-formal quantization

In [15], Questions 1 at the end of Section 2.3, I discussed a possibility to pass from $\mathbb{C}[\![\hbar]\!]$ to germs of analytic functions $\mathbb{C}\{\hbar\}$ for a class of *algebraic* affine symplectic manifolds admitting an algebraic Poisson compactification $P \supset M$ such that $P - M$ is an ample divisor and $H^1(P, \mathcal{O}) = H^2(P, \mathcal{O}) = 0$. Examples suggest that the dependence on $\hbar$ is analytic not only in a neighborhood of 0, but is defined *everywhere* in $\mathbb{C}_\hbar$-plane. Now I am ready to propose a more ambitious conjecture, going beyond the algebraic case.

Very optimistic conjecture. *For any compact complex Poisson manifold (P, γ) of Kähler type, and such that $H^2(P, \mathcal{O}) = 0$, one has a* canonical *category* $\mathrm{Perf}_{P,\gamma}$ *enriched over* $\mathrm{Perf}(\mathbb{C})$*, with a bounded t-structure. Moreover, the dependence on (P, γ) is analytic, and the corresponding family in the formal neighborhood of the locus of (P, γ) with $\gamma = 0$ coincides with the one given by the formal deformation quantization.*

The assumption of being of Kähler type (i.e., to admit a Kähler metric) is not really justified. It guarantees Hodge decomposition and is stable under small deformations. I included it just to be on the safe side.

One can also try to remove the constraint that the Hodge number $h^{0,2}(P)$ vanishes. In the general case of a compact Poisson manifold of Kähler type, we should also choose a 2-cocycle with values in $U(1) = \mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z} \subset \mathbb{C}^\times$. The dependence on additional parameters should be real analytic (i.e., extends to a neighborhood of compact torus $H^2(P, U(1))$ in complex torus $H^2(P, \mathbb{C}^\times)$).

By adding gerbes, one gets certain redundancy of parameters. For example, at the commutative point $\gamma = 0$, the tangent space to the moduli stack of categories is (by the formality theorem)

$$H^2(P, \mathcal{O}_P \to T_P \to \wedge^2 T_P \to \cdots),$$

whereas the tangent space to Poisson structures together with $U(1)$ gerbe is

$$H^2(P, \mathrm{i}\mathbb{R} \to \mathcal{T}_P \to \wedge^2 T_P \to \cdots).$$

There is a natural epimorphism from the second to the first.

The general conjecture gives an attractive way to "name" quantum categories. In concrete cases, when P is obtained by blowups from compactified cotangent bundles,

or compactified tori (trigonometric case), or, for example, Odesskii–Feigin projective spaces with elliptic brackets, one can typically make a reasonable guess of what the category $\mathrm{Perf}_{P,\gamma}$ is. For example, when P contains an open dense symplectic leaf (M,ω), $\gamma_{|M} = \omega^{-1}$, the cohomology class $[\omega] \in H^2(M,\mathbb{C})$ corresponds to the so-called Deligne class of the localization $\mathrm{Perf}_{M,\gamma_{|M}}$ of $\mathrm{Perf}_{P,\gamma}$; see Section 1.2 and [17, points 1.30–1.37].

2.3. Objects with compact support

Now we are almost ready to propose a definition of $C_{B,\hbar}(M,\omega)$. If (M,ω^{-1}) is the open dense symplectic leaf in (P,γ), we can try to define abelian B-model category $C^{\mathrm{heart}}_{B,\hbar}(M,\omega)$ as the full subcategory of the heart of $\mathrm{Perf}_{P,\gamma}$ consisting of objects which are "vanishing after the restriction to $P - M$".

The above proposal makes perfect sense for the formal deformation quantization (see Section 2.1.1). The category over $\mathbb{C}[\![\hbar]\!]$ can be alternatively defined as the abelian category of perfect $\mathcal{O}_{q,M}$-modules with compact support.

One can see clearly some potential problems with such a definition. First, the limiting category at $\hbar = 0$ is $\mathrm{Coh}_c(M)$, which is "too large", as it contains coherent sheaves with 0-dimensional support, or, in the case when M is compact, with the full $2n$-dimensional support. One would like to exclude such objects and keep only those with n-dimensional Lagrangian support. Therefore, we like to specify a subgroup of $K^{\mathrm{top}}_{c,\mathrm{even}}(M)$ of "admissible" classes of objects. We expect that this constraint comes from a limiting stability condition, as mentioned in Section 1.2.

In the quantized case for generic $\hbar \neq 0$, by general philosophy, objects should have support of dimension $\geq \dim M/2$ (see, e.g., [11]) and can be thought of as *coisotropic* compact branes. Again, as in the case $\hbar = 0$, we would like to put a constraint via a limiting stability condition, forcing the classes to stay in a smaller subgroup of $K^{\mathrm{top}}_{c,\mathrm{even}}(M)$ corresponding roughly to compact Lagrangian cycles.

2.4. Remarks about non-compact support

It is interesting to get some understanding of quantum modules whose classical limit is a coherent sheaf on M with a *non-compact* coisotropic support.

The first example is the so-called big brane (also called *canonical coisotropic brane*), a deformation of the structure sheaf $\mathcal{O}_M$ as was suggested by physicists; see [9, 11]. In the case, for example, $M = \mathbb{A}^{2n}_{\mathbb{C}}$ with the standard symplectic form $\sum_i dx_i \wedge dy_i$, the endomorphism algebra of the big brane should be the n-th Weyl algebra, that is, the algebra of polynomial differential operators generated by $x_i, \hbar\partial_{x_i}$, $i = 1,\dots,n$. If one pursues the idea that the quantized algebra *should not depend* on the choice of Poisson compactification, one inevitably comes to the conclusion that the group of polynomial symplectomorphisms should act by automorphisms

of Weyl algebra (see [15, Question 4 in Section 4.1], main conjecture in [1] and its announced proof in [5]).

Another reason to be interested in non-compactly supported objects is GRH correspondence, where the supports are Lagrangian subsets in the classical limit $\hbar \to 0$, like, for example, in Section 1.4. This is natural if one tries to keep the contact with the traditional RH correspondence for holonomic D-modules, or its version for quantum tori (see Sections 1.5 and 3.3.3). Here one should analyze the behavior of Lagrangian subsets at infinity.

In this lecture, I will try to focus more on the approach when one deals only with compactly supported objects, which is achieved by enlarging the symplectic leaf like as, for example, in Section 1.3. The formulation of the main conjecture is a bit cleaner in the compactly supported case.

2.5. Beyond smooth Poisson compactifications

In general, the assumption that M admits a *smooth* Poisson compactification seems to be a bit unnatural. The reason is that Poisson structures behave badly under blowups, and one can envision a situation when the complement $P - M$ has singularities of high codimension which do not affect the general construction, but which cannot be resolved in the class of Poisson manifolds (see, e.g., [15, Questions 3 in Section 4.1]).

In differential geometry, it is natural to consider hyperkähler manifolds with *complete* metric, which seems to be the framework of [9]. Presumably, the "optimal" condition at infinity for B-model should go in this direction.

A more drastic change would be to allow some mild singularities even for symplectic manifolds. Both hypothetical approaches to A and B models, presented in this text, fail for non-smooth varieties.

3. A-side: Fukaya categories, real case

3.1. Almost-complex structures, conditions at infinity

Recall that for a *real* symplectic C^∞-manifold (M, ω), an almost-complex structure $J \in \Gamma(M, \mathrm{Aut}(T_M))$, $J^2 = -1$ is called *compatible* with ω iff

$$\langle \omega_x, v \wedge Jv \rangle > 0 \quad \forall x \in M,\ \forall v \in T_x M - 0.$$

In the definition of Fukaya A_∞-category of a *compact* symplectic manifold, one usually starts with a *choice* of compatible almost-complex structure J and then considers J-holomorphic discs with boundaries on unions of Lagrangian submanifolds. The resulting category does not depend (up to a contractible space of equivalences)

on the choice of J, and in fact can be *canonically* defined, as follows from the contractibility of the space of almost-complex structures compatible with ω.

Structure constants of higher multiplications of Fukaya categories are given by virtual fundamental classes of spaces of J-holomorphic discs of bounded area, and one needs to ensure a compactness property in order to get a well-defined count. For compact M, this follows from Gromov compactness, whereas in the non-compact case one has to impose an appropriate behavior at infinity.

A sufficient condition on (M, ω) is the following: M is separable (i.e., contains a countable dense subset), and there exists a compatible almost-complex structure J such that for any compact subset $K \Subset M$ and any real $A > 0$, there exists a larger compact $K \subset K' \Subset M$ such that all J-pseudo-holomorphic discs $D \subset M$ with $\partial D \subset K$ and $\int_D \omega < A$ are contained in K'.

If this condition is satisfied, then one can define the compactly supported version of the Fukaya category (and genus zero Gromov–Witten invariants). It follows from the stronger condition of geometrical boundedness (see [8]) – the property of being A-good is its complex version – which is easier to check in practice.

3.2. Traditional approach to Fukaya categories

Let now (M, ω) be a C^∞ symplectic manifold of real dimension $2n$. Assume that (M, ω) is A-good and is Calabi–Yau in the symplectic sense, that is, $c_1(T_M) = 0 \in H^2(M, \mathbb{Z})$. Moreover, we need a trivialization of $c_1(T_X)$ on cochain level, for example, a choice of almost-complex structure J and a non-vanishing complex-valued n-form $\mathrm{vol} \in \Gamma(M, \Omega^{n,0})$ (not necessarily closed). More generally, one can choose any non-vanishing complex-valued n-form on Lagrangian subspaces; see [10].

After making such a choice, we will obtain the Fukaya category $\mathcal{F}(X, \omega)$ defined over the Novikov field

$$\mathrm{Nov} := \mathbb{Q}((e^{-\mathbb{R}/\hbar})) := \left\{ \sum_i c_i e^{-A_i/\hbar} \mid c_i \in \mathbb{Q},\ A_i \in \mathbb{R},\ \lim_i A_i = +\infty \right\}. \quad (3.1)$$

Usual definition (see, e.g., [6]) involves compact oriented Lagrangian submanifolds with spin structure and with the trivialized vanishing of Maslov index $\mu_L \in H^1(L, \mathbb{Z})$ associated with vol. In the lucky case when there is no J-holomorphic discs with boundary on L, such L gives a *canonical* object of $\mathcal{F}(X, \omega)$.

The complexes of morphisms are defined via (transversal) intersection points, and the differential is the usual Floer differential, counting J-holomorphic 2-gons with signs and weights $\exp(-\int_D \omega/\hbar)$.

Fukaya category can be twisted by a B-field with class $\in H^2(M, \mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z})$ (or, more generally, in $H^2(M, \mathbb{C}^\times)$). In this case, the field of definition is the complex analog of the Novikov field; coefficients c_i in (3.1) are allowed to lie in $\mathbb{C}$. Also, in

the case when $[\omega]$ belongs to the image of $H^2(X,\mathbb{Z})$ in $H^2(X,\mathbb{R})$, one can replace the Novikov field by the usual Laurent series field $\mathbb{Q}((q))$ or $\mathbb{C}((q))$, where q corresponds to $e^{-1/\hbar}$ (equivalently, all areas A_i in (3.1) are integers).

In what follows, I will describe *another* potential approach to Fukaya categories, first introduced in [18]; see also [25].

3.3. Local Fukaya categories

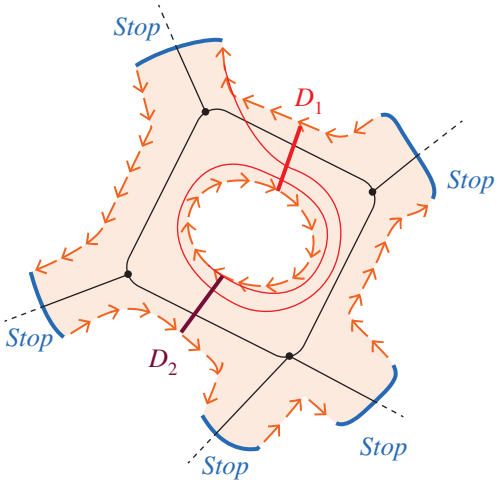
Let us assume that (M,ω) is real analytic, no condition at infinity, with the trivialization of $c_1(T_X)$ on the cochain level, as above. The following is essentially known, but not yet properly articulated.

Claim. With any subanalytic locally closed subset $L \subset M$ which is Lagrangian at smooth points, one can associate a *canonical* small A_∞-category $\mathcal{WF}^{\mathrm{local}}(L)$ defined over $\mathbb{Z}$ and depending *only* on the germ of (M,ω) at L such that

- set $\mathrm{Ob}(\mathcal{WF}^{\mathrm{local}}(L))$ consists of smooth points $x \in L$ together with a local orientation, local spin structure near X and a local Maslov lift of $\arg(\mathrm{vol}_{T_x L})$ from $\mathbb{R}/2\pi\mathbb{Z}$ to $\mathbb{R}$,
- if L is an embedded open d-dimensional disc, then all objects are isomorphic up to cohomological shift, and the endomorphism A_∞-algebra of each object is quasi-isomorphic to $\mathbb{Z}$,
- every open embedding $L_1 \hookrightarrow L_2$ induces a functor (tautological on objects) $\mathcal{WF}^{\mathrm{local}}(L_1) \to \mathcal{WF}^{\mathrm{local}}(L_2)$, so we get a co-presheaf of categories on every locally closed L,
- this co-presheaf is a co-sheaf for every locally closed L,
- every closed embedding $L_1 \hookrightarrow L_2$ induces a functor (obvious on objects) $\mathcal{WF}^{\mathrm{local}}(L_2) \to \mathcal{WF}^{\mathrm{local}}(L_1) \sqcup \{0\}$ which identifies $\mathcal{WF}^{\mathrm{local}}(L_1)$ with the localization of $\mathcal{WF}^{\mathrm{local}}(L_1)$ by objects corresponding to $x \in L_2^{\mathrm{smooth}} - L_1$.

The co-sheaf of small categories on L gives a sheaf $\mathcal{WF}_L^{\mathrm{local},\vee}$ of large categories by associating with each open $L' \subset L$ the category of triangulated functors from $\mathcal{WF}^{\mathrm{local}}(L')$ to complexes of $\mathbb{Z}$-modules.

3.3.1. Heuristic picture: Wrapping with stops. *Assume* that L has a Liouville tubular neighborhood U such that one has a complete Liouville flow generated by vector field η, $\mathrm{Lie}_\eta\, \omega = -\omega$, contracting U to L. One should interpret an object of $\mathcal{WF}^{\mathrm{local}}(L)$ as a small Lagrangian disc D closed in U, which intersects *transversally* L at a smooth point $\mathrm{pt} \in L$. Local orientation, local spin structure and local Maslov lift of L near x induce similar structures on D. The morphisms in $\mathcal{WF}^{\mathrm{local}}(L)$ are defined via infinite wrapping by a Hamiltonian vector field on U with stops at "infinity" of L; see [7] for the general theory of wrapping with stops.



The main technical trouble with this heuristic picture is that the existence of a Liouville neighborhood is not known.

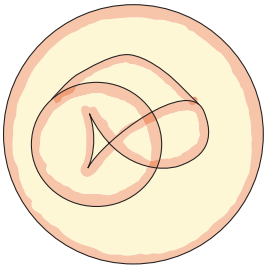
3.3.2. One can calculate $\mathcal{WF}^{\mathrm{local}}(L)$ explicitly. Here we sketch a rigorous alternative definition based on the theory of microsupport. First, we describe sheaf $\mathcal{WF}^{\mathrm{local}}$ locally *up to an equivalence* in the following way. For a given point $\mathrm{pt} \in L$, first choose local Darboux coordinates $(x_i, y_i)_{1\le i\le n}$, $n = \dim M/2$ such that $\omega = \sum_i dx_i \wedge dy_i$. There exists a unique germ of C^0-function z at $\mathrm{pt} \in L$ which vanishes as pt and whose restriction to any locally closed smooth submanifold $Y \subset L$ near pt is also smooth, and the differential of $z_{|Y}$ is equal to the restriction of 1-form $\sum_i y_i dx_i$ to Y. Functions $(x_i, y_i)_{1\le i\le n}$ together with function z give an embedding of the germ of L at pt to $\mathbb{R}^{2n+1}$, as a singular Legendrian subset. Theory of microlocal support (see [12]) gives a canonical germ of a sheaf of categories, which in our picture is the sheaf of representations $\mathcal{WF}^{\mathrm{local},\vee}_{L'}$ for sufficiently small neighborhood L' of $\mathrm{pt} \in L$. Finally, the global choice of the trivialization of $c_1(T_X)$ gives rise to the global co-sheaf of categories on L (see [2] for the discussion of ambiguities of the gluing).

The co-sheaf $\mathcal{WF}^{\mathrm{local}}$ on L is locally constant along some Whitney stratification. The co-stalks are certain nice A_∞-categories over $\mathbb{Z}$, of finite type in the sense of B. Toën and M. Vaquié; see [28]. Up to derived Morita equivalence, they are localizations of smooth proper categories (in fact, just finite posets) by finite collections of compact objects.

In the case $\dim M = 2$, a singular Lagrangian is just a ribbon graph. The co-stalk at vertex of degree $d \ge 1$ is Morita equivalent to Dynkin quiver A_{d-1}. In general, the co-stalk can be identified with the category of constructible sheaves on $\mathbb{R}^n$ whose

microlocal support is contained in a given closed subanalytic *conical* Lagrangian $L' \subset T^*B^n$ where $B^n \subset \mathbb{R}^n$ is the closed n-dimensional ball.

A typical picture of such a Lagrangian is given below.



3.3.3. Application: RH correspondence for quantum tori. Here we give a sketch of the description of the A-model side of the hypothetical RH correspondence for higher-dimensional quantum algebraic tori, promised in Section 1.5. Let $L \subset \mathbb{R}^{2n}$ be a closed conical subset which is a finite union of rational Lagrangian simplicial cones, for the standard symplectic form $\sum_{i=1}^{n} dx_i \wedge dy_i$.

Symplectic group $\mathrm{Sp}(2n, \mathbb{Z})$ acts transitively on the set of rational Lagrangian subspaces in $\mathbb{Q}^{2n}$. Similarly, a central extension of $\mathrm{Sp}(2n, \mathbb{Z})$ acts on $\mathrm{Perf}(E^n)$ where E is an elliptic curve, by generalized Fourier–Mukai transforms. With each rational Lagrangian subspace $\Lambda \subset \mathbb{Q}^{2n}$, we associate a full triangulated subcategory $C_\Lambda \subset \mathrm{Perf}(E^n)$. This association is characterized by two properties: (1) it is $\mathrm{Sp}(2n, \mathbb{Z})$-equivariant, and (2) the category C_{Λ_0} corresponding to the standard Lagrangian subspace Λ_0 spanned by the first n base vectors has the heart given by coherent sheaves with 0-dimensional support on E^n. The latter is invariant under the linear action of $\mathrm{GL}(n, \mathbb{Z})$ on E^n, and the tensor product with line bundles, matching the stabilizer $\mathrm{GL}(n, \mathbb{Z}) \ltimes (\mathbb{Z})^{n(n+1)/2} \subset \mathrm{Sp}(2n, \mathbb{Z})$ of Λ_0.

Let us consider the category of functors from $\mathcal{WF}^{\mathrm{local}}(L)$ to $\mathrm{Perf}(E^n)$. This category contains a full subcategory consisting of functors F such that when F applied to an object corresponding to a smooth point x with an extra data (local orientation, local spin structure, local Maslov lift) as above, the result lies in the subcategory $C_{T_x L}$ and moreover belongs to a cohomological shift $[k]$ of the heart, where $k \in \mathbb{Z}$ is determined (in certain natural way) by the local Maslov lift of $\arg(\mathrm{vol}_{T_x L})$. The resulting category is an Artinian abelian A_∞-category. Inclusions of rational Lagrangian conical subset give rise to inclusions of full subcategories. The inductive limit over all $L \subset \mathbb{R}^{2n}$ is (hypothetically) equivalent to the category of holonomic modules over the $2n$-dimensional quantum torus.

3.4. Hypothetical "correct" definition of Fukaya category

For a symplectic C^ω-manifold (M,ω) geometrically bounded at infinity, with trivialized $c_1(T_M)$, with a tame choice of almost-complex structure J and a compact singular subanalytic Lagrangian $L\subset M$, we expect that J-holomorphic discs with boundary on L give rise to a *canonical* deformation of $\mathcal{WF}^{\mathrm{local}}(L)$ over Novikov ring $\mathrm{Nov}_+ := \mathbb{Q}[\![e^{-\mathbb{R}_{\geq 0}/\hbar}]\!]$. To my knowledge, the general problem of the definition of this canonical deformation was not addressed, except the cases when L is smooth or $n=2$. In the case when an appropriate Liouville neighborhood exists, it is possible to define the expected deformation using traditional tools of symplectic topology.

Let us define the global Fukaya category *supported* on L (and denoted by $\mathcal{F}^{\mathrm{global}}_L(M,\omega)$), as the category of finite-dimensional over Nov representations of the *deformed* algebra $\mathcal{WF}^{\mathrm{local}}(L)$, admitting a model over Nov_+ (subscript "+" indicates that the areas A_i in (3.1) are ≥ 0). The localization property implies that for $L_1\subset L_2$ we have a fully faithful embedding $\mathcal{F}^{\mathrm{global}}_{L_1}(M,\omega)\hookrightarrow \mathcal{F}^{\mathrm{global}}_{L_2}(M,\omega)$. The universal Fukaya category $\mathcal{F}^{\mathrm{global}}(M,\omega)$ can be defined as the inductive limit of $\mathcal{F}^{\mathrm{global}}_L(M,\omega)$ over the poset of all subanalytic compact singular Lagrangians $L\subset M$.

Categories $\mathcal{F}^{\mathrm{global}}_L(M,\omega)$ and $\mathcal{F}^{\mathrm{global}}(M,\omega)$ can be modified by a B-field, a closed 2-cochain with coefficients in $U(1)=\mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z}$. The global category does not change (up to homotopy) under deformations of geometrically bounded almost-complex structure J.

It is expected that Bridgeland stability structures on $\mathcal{F}^{\mathrm{global}}_L(M,\omega)$ appear when one chooses a compatible almost-complex structure J and *closed* complex volume form $\mathrm{vol}\in\Gamma(X,\Omega^{n,0})$. More generally, the constraint on vol to be of type $(n,0)$ can be weakened; see [10].

3.5. Moduli stack of non-formal Fukaya categories

The Fukaya categories, defined either in the traditional way or via deformations of local Fukaya categories as above, still depend on the small parameter $\hbar$ in a formal way, as a series in the Novikov field, without any control of convergence. It is still not proven (although expected) that the counting of pseudo-holomorphic discs exhibits an exponential growth.

It is well known that local deformations of *compact* symplectic manifolds, up to symplectomorphisms close to identity, are governed by deformations of the cohomology class of symplectic form $[\omega]\in H^2(M,\mathbb{R})$. Adding the cohomology class of B-field $[\beta]\in H^2(M,U(1))$, one arrives to a domain in complex torus $H^2(M,\mathbb{C}^\times)$ (so-called *complexified Kähler cone*). Moreover, one can expect the same picture for *non-compact* symplectic manifolds (M,ω) when the symplectic structure is geometrically bounded at infinity and $\mathrm{rk}\,H^2(M,\mathbb{Z})<\infty$ (e.g., M is topologically cylindrical at infinity, as typically happens).

Homological mirror symmetry gives a very strong indication that there should be a "non-formal" Fukaya category associated with (M, ω) and B-field β, at least when the class $[\omega] \in H^2(M, \mathbb{R})$ is "sufficiently large". We denote this putative category as $\mathcal{F}(M, \omega, \beta)$. It is a category enriched over $\mathrm{Perf}(\mathbb{C})$, and not over the Novikov field.

Let us now restrict ourselves to the case of *compact* X, class $[\omega]$ is integral, and we are in the situation of mirror symmetry, which means roughly that X is fibered by Lagrangian tori, that is, endowed with the structure of a real integrable system. In this case, homological mirror symmetry identifies $\mathcal{F}^{\mathrm{global}}(M, \omega)$ with the category of perfect complexes on the mirror family of complex Calabi–Yau varieties defined over the formal disc $\mathrm{Spf}\, \mathbb{Z}[\![q]\!]$ where q corresponds to $e^{-1/\hbar}$. Moreover, the formal disc with parameter q embedded at the cusp of the moduli space of mirror varieties is in fact defined analytically by the so-called *inverse mirror map*. For example, in the case of quintic 3-folds and $[\omega] = c_1(\mathcal{O}(1))$, the moduli space of mirrors is $\mathbb{P}^1_z - \{0, 5^{-5}, \infty\}$. The correspondence between the A-model coordinate q and the B-model coordinate z is

$$q = \exp\left(2\pi \mathrm{i} \frac{\int_{\gamma_1} \mathrm{vol}}{\int_{\gamma_0} \mathrm{vol}}\right) = z + 770z^2 + 1014275z^3 + \cdots,$$

where vol is the holomorphic volume form on the dual variety $M^\vee$, and $\gamma_0, \gamma_1 \in H_3(M^\vee, \mathbb{Z})$ are certain cycles. The inverse mirror map $z(q) = q - 770q^2 + \cdots$ becomes multivalued when $|q|$ is too large, that is, when the rescaled symplectic form $\hbar^{-1}\omega$ is too small. Nevertheless, all this suggests that the Fukaya category is *well defined* as an analytic family of categories over certain domain in the complexified Kähler cone.

In the case of *complex* symplectic manifolds, even non-compact ones, we expect more: the global Fukaya category is well defined even for *arbitrary small* non-zero values of $\hbar$. Usual mirror symmetry predictions do not produce singularities for the inverse mirror map, as all non-trivial Gromov–Witten invariants of complex symplectic manifolds vanish identically. The resulting moduli stack of categories is rather mysterious, as there are, for example, automorphisms which act trivially on cohomology, like squares of spherical twists in the case $\dim_{\mathbb{C}} M = 2$ (case of K3 surfaces).

It was proposed (see [11]) that the "true" A-model category should include objects with "coisotropic support", of dimension $\in \{n, n+2, n+4, \ldots\}$. One possible mathematical model of such an object would be a smooth coisotropic submanifold endowed with a Lagrangian foliation and a transversal complex structure. In the framework of the generalized RH correspondence discussed here, the heart of "true" B-model category, one should be the full subcategory of $\mathrm{Perf}^{\mathrm{heart}}_{P,\gamma}$ consisting of objects which are "vanishing after the restriction to $P - M$", and *omitting* the constraint (see the end of Section 2.3) on classes in topological K-theory.

I have tried to make a guess (based on Hodge-theoretic invariant from Section 1.2) of what parametrizes "true" A-model categories of compact tori $(S^1)^{2n}$ corresponding to translationally invariant symplectic structure and the B-field. The expected answer is

$$\mathcal{U}_n/O(2n,2n;\mathbb{Z}),$$

where $\mathcal{U}_n$ is the open subset of the isotropic Grassmannian of $\mathbb{C}\otimes(\mathbb{Z})^{2n,2n}$ consisting of $2n$-dimensional complex isotropic subspaces of $\mathbb{C}^{4n}$ intersecting trivially $\mathbb{R}^{4n}$. Remarkably, the *same* stack appears in the B-model, corresponding to translationally invariant holomorphic Poisson structures on compact tori, together with $U(1)$ gerbes, up to (potential) derived equivalences of Fourier–Mukai type.

4. A-side: Fukaya categories, complex case

4.1. Holomorphic symplectic manifolds, Lagrangian subsets

Let $M = M_{\mathbb{C}}$ be a complex-analytic manifold with a holomorphic closed non-degenerate 2-form $\omega = \omega_{\mathbb{C}}$. Then for any $\theta \in \mathbb{R}/2\pi\mathbb{Z}$, we have a *real-analytic* symplectic manifold

$$(M,\omega_\theta), \quad \omega_\theta := \Re(e^{-\mathrm{i}\theta}\omega_{\mathbb{C}}) = \cos\theta\cdot\Re\,\omega_{\mathbb{C}} + \sin\theta\cdot\Im\,\omega_{\mathbb{C}}.$$

Let us assume that $(M,\omega_{\mathbb{C}})$ is geometrically bounded at infinity, and let us choose a geometrically bounded almost-hyperkähler structure on M such that $\Re\,\omega_{\mathbb{C}}$ and $\Im\,\omega_{\mathbb{C}}$ are the first two standard symplectic forms among three $(\omega_I,\omega_J,\omega_K)$ (the space of such choices is contractible). Then for any θ, we get an ω_θ-compatible almost-complex structure J_θ.

Let $L\subset M$ be a compact closed $\mathbb{C}$-analytic singular Lagrangian subset. Then L is a singular ω_θ-Lagrangian for any $\theta\in\mathbb{R}/2\pi\mathbb{Z}$, with the canonical orientation and Maslov lift. Small categories $\mathcal{WF}_\theta^{\mathrm{local}}(L)$ form a *local system* over $\mathbb{R}/2\pi\mathbb{Z}$, with the monodromy functor identical on objects. Let us pick a point on each connected component of L^{smooth} and consider the endomorphism algebra $\mathcal{A}_{L,\theta}$ of the formal sum of corresponding objects of $\mathcal{WF}_\theta^{\mathrm{local}}(L)$. The monodromy along varying $\theta\in\mathbb{R}/2\pi\mathbb{Z}$ gives an automorphism ψ_L of $\mathcal{A}_L := \mathcal{A}_{L,0}$. Presumably, $\mathcal{A}_L$ is the path algebra of finite dg quiver defined over $\mathbb{Z}$ such that all the arrows have *non-positive* degrees. The category of finite-dimensional complex representations of $\mathcal{A}_L$ carries a canonical bounded t-structure, with the heart consisting of representations in degree 0. This is the perverse t-structure described in [2]. In the case when L is smooth, connected and has a trivial spin structure, in the sense $c_1(T_L)\in H^2(L,\mathbb{Z})$ is divisible by 2, algebra $\mathcal{A}_L$ is isomorphic to the algebra of $\mathbb{Z}$-chains of the based loop monoid of P, and the automorphism ψ_L is trivial.

Algebras $\mathcal{A}_{L,\theta}$ are naturally graded by the abelian group $H_1(L,\mathbb{Z})$. The reason is that the algebraic group $H^1(L,\mathbb{G}_m)$ acts on the category of all (not necessarily finite-dimensional) representations of $\mathcal{A}_{L,\theta}$, by twisting in the sheaf-theoretic description (this remark is also applicable to local Fukaya categories in the general real case).

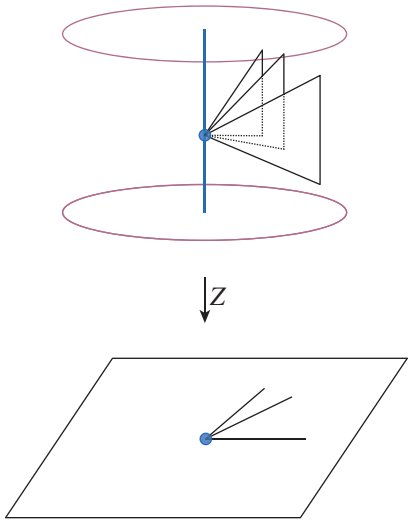
Define a finite rank lattice $\Gamma := H_2(M,L;\mathbb{Z})/\text{torsion}$, and denote by $Z : \Gamma \to \mathbb{C}$ the linear map (called central charge) given by the integration of $\omega_{\mathbb{C}}$.

Key observation. *For any J_θ-holomorphic disc D with boundary on L, one has $Z([D]) \in e^{i\theta}\mathbb{R}_{\geq 0}$.*

Indeed, the restriction of $\omega_{\mathbb{C}}$ to any non-trivial J_θ-holomorphic curve has argument θ at every point. Hence for $\theta \notin \{\arg Z(\gamma) \mid \gamma \in \Gamma,\ Z(\gamma) \neq 0\}$ (i.e., $e^{i\theta}\mathbb{R}_{\geq 0}$ is *not a Stokes* ray), *there are no* J_θ-holomorphic discs with boundary on L, and hence $\mathcal{WF}_\theta^{\text{local}}(L)$ is *not deformed*.

4.2. Complex Lagrangians and Stokes automorphisms

Staying (infinitesimally) a little bit on the left, or on the right of a given Stokes ray, we obtain *two* fully faithful embeddings of the local Fukaya category into the global one. These two embeddings coincide at the leading order (ignoring non-trivial disc corrections), hence should differ by certain *autoequivalence* $A_{\theta,\text{cat}}$ of $\mathcal{WF}_\theta^{\text{local}}(L)$ which is equal to identity in the leading order. Naturally, the coefficients for the deviation from identity should correspond to non-trivial J_θ-holomorphic discs. It follows from the uniform bounds on almost-hyperkähler structure that the possible homology classes $\gamma \in \Gamma$ of these discs belong to some strict closed convex cones $\text{Cone}_\theta \subset Z_{\mathbb{R}}^{-1} e^{i\theta}\mathbb{R}_{\geq 0} \subset \Gamma \otimes \mathbb{R}$ which are *uniformly* separated from $\ker Z_{\mathbb{R}}$.



Categories $\mathcal{WF}_\theta^{\mathrm{local}}(L)$ and categorical Stokes autoequivalences $A_{\theta,\mathrm{cat}}$ *do not* depend on the auxiliary choice of almost-complex structures J_θ as above.

The Stokes autoequivalence comes from an automorphism A_θ of the dg path algebra $\mathcal{A}_{L,\theta}$ close to the identity, and hence of the form $\exp(\delta)$ where δ is a derivation of $\mathcal{A}_{L,\theta}$ with coefficients in the non-unital semigroup algebra $\mathbb{Q}[\![\mathrm{Cone}_\theta \cap \Gamma - 0]\!]$. For each monomial $\gamma \in \mathrm{Cone}_\theta \cap \Gamma - 0$, the γ-coefficient of δ is expected to be an element of $\mathrm{Der}(\mathcal{A}_{L,\theta})$ which is homogeneous of degree $\partial\gamma \in H_1(L,\mathbb{Z})/\mathrm{torsion}$ where $\partial : H_2(M,L;\mathbb{Z}) \to H_1(L,\mathbb{Z})$ is the boundary map, and $H_1(L,\mathbb{Z})$-grading on $\mathrm{Der}(\mathcal{A}_{L,\theta})$ is induced from the $H_1(L,\mathbb{Z})$-grading on $\mathcal{A}_{L,\theta}$ described above in Section 4.1.

4.2.1. Analytic wall-crossing structures.

- The collection of automorphisms (A_θ) associated with Stokes rays can be interpreted as a *stability structure* (see [20]) for certain local system on S^1 of countable-dimensional Γ-graded Lie algebras:

$$\mathfrak{g}_\theta = \bigoplus_{\gamma\in\Gamma} \mathfrak{g}_{\theta,\gamma}, \quad g_{\theta,\gamma} := \mathrm{Der}_{\partial\gamma}(\mathcal{A}_{L,\theta}).$$

- If we vary parameters (i.e., $M_{\mathbb{C}}, \omega_{\mathbb{C}}$ and L as well) preserving the equivalence class of $\mathcal{WF}_{\theta=0}^{\mathrm{local}}(L)$ (it is a structure of a "combinatorial" nature, no continuous parameters), then the stability structure varies in a universal way (*wall-crossing* equations) controlled *only* by the varying homomorphism $Z \in \mathrm{Hom}(\Gamma,\mathbb{C})$.
- It looks that one can "understand" all (A_θ) at once, for example, in some examples all information is encoded by finitely many non-commutative cluster transformations in free variables.

4.3. Generalization: Compatible pairs of symplectic structures

It is not even necessary to work only with complex symplectic manifolds; the whole theory makes sense in a much larger generality of *pairs* of real symplectic structures such that any non-trivial linear combination of corresponding 2-forms is non-degenerate; see [21, Section 4.6]. Compact singular subset L should be Lagrangian with respect to both structures. The importance of this generalization is that the parameter space (possible pairs of cohomology classes of symplectic forms) is locally an open domain in $H^2(M,L;\mathbb{C})$, and it is in general *strictly larger* than the image in $H^2(M,L;\mathbb{C})$ of the parameter space of complex symplectic manifolds with analytic Lagrangian subsets.

A particularly interesting class is of triples (M,ω,L) such that the cohomology class of ω belongs to $H^2(M,L;\mathbb{Q}\oplus \mathrm{i}\mathbb{Q})$. There are only *countably many* isotopy classes of such triples, producing therefore countably many automorphisms

A_θ, for example, only for $\theta = 0$, which control *all* wall-crossing structures in GRH correspondence.

All the above deals exclusively with the A-model.

5. Local RH correspondence, and a generalization to fractional powers of $\hbar$

One can associate with the pair $(\mathcal{A}_L, \psi_L)$ a new $\mathbb{C}(\!(\hbar)\!)$-linear category (see [21, Section 5.5]). Namely, introduce the universal ring with an automorphism

$$\mathbf{B} := \lim_{N\to+\infty} \mathbb{C}(\!(\hbar^{1/N})\!)[\log\hbar], \quad \psi_{\mathbf{B}}(\hbar^{1/N}) = e^{2\pi\mathrm{i}/N}\hbar^{1/N},$$
$$\psi_{\mathbf{B}}(\log\hbar) = \log\hbar + 2\pi\mathrm{i}$$

and consider representations of $\mathcal{A}$ in $\mathrm{Perf}(\mathbf{B})$ that admit a model over $\mathbf{B}_+$ and are equivariant with respect to $(\psi_L, \psi_{\mathbf{B}})$. The resulting category is linear over $\mathbf{B}^{\psi_{\mathbf{B}}} = \mathbb{C}(\!(\hbar)\!)$. In the case of smooth L and hence trivial ψ_L, we essentially obtain the category of families over $\mathbb{C}[\![\hbar]\!]$ of finite-dimensional representations of $\mathcal{A}_L$, with inverted $\hbar$.

We have made a conjecture (*local RH correspondence*) in [21] that this category is equivalent to certain (not specified) full subcategory of the category of holonomic modules over the *formal* deformation quantization of (M, ω), after inversion of $\hbar$. For example, in the case when L is smooth, the relevant category consists of modules which in local coordinates are isomorphic to finite sums of copies of

$$i_*\mathcal{O}_{\mathbb{C}^n}[\![\hbar]\!], \quad \text{where } i : \mathbb{C}^n \to \mathbb{C}^{2n} \text{ is } (x_1, \dots, x_n) \mapsto (x_1, \dots, x_n, 0, \dots, 0).$$

Here the standard coordinates $(x_1, \dots, x_n, y_1, \dots, y_n)$ in $\mathbb{C}^{2n}$ endowed with the Moyal product (2.1) act as actions by operators x_i and $\hbar\partial_{x_i}$ (see [13]).

5.1. Fractional powers of $\hbar$ and enhanced Lagrangian subsets

The above discussed hypothetical local RH correspondence *does not* give the whole category of holonomic modules supported on L (up to $\hbar$-torsion modules), even in the case of smooth L. The basic example is (in the above notations)

$$e^{\sum_{k=1}^{N-1}\hbar^{-k/N} f_k(x_1,\dots,x_n)}\, j_*\mathcal{O}_{\mathbb{C}^n}[\![\hbar^{1/N}]\!],$$

where $N \geq 1$ is an integer and $f_1, \dots, f_N$ are germs at $0 \in \mathbb{C}^n$ of analytic functions (here $j : \mathbb{C}^n \to \mathbb{C}^{2n}$ is the coordinate embedding). It is a holonomic $\mathcal{O}_{q,M}$-module, with the Lagrangian support given by $y_i = \partial_{x_i} f_N$, $i = 1, \dots, n$.

The appearance of fractional powers of $\hbar$ opens a new can of worms! We (Y. Soibelman and I) are trying now to figure out a precise definition of *enhanced singular Lagrangian subsets* controlling geometrically WKB-like modules with fractional powers of $\hbar$ as above. Heuristically, one should think of a family of singular Lagrangian subsets $L_\hbar$ depending on formal or small analytic parameter $\hbar$, converging at $\hbar \to 0$ to the support in the usual sheaf-theoretic sense. Hypothetically, by adding all possible enhanced singular Lagrangian subsets, both in A-model and B-model versions, we expect to get a full local RH correspondence between *all* holonomic formal $\mathcal{O}_{q,M}$-modules after inversion of $\hbar$, and $\mathbb{C}(\!(\hbar)\!)$-linear categories associated with dg algebras endowed with automorphism, as in the beginning of the current section.

Finally, we observe that the picture with wall-crossing structures should generalize to the enhanced case. The lattice $\Gamma = H_1(M, L, \mathbb{Z})/\text{torsion}$ should be replaced by the local system of lattices $\Gamma_\theta = H_1(M, L, \mathbb{Z})/\text{torsion}$ on S^1_θ, and the central charge $Z : \Gamma \to \mathbb{C}$ should be replaced by the local system of central charges with values in $\mathbb{C}[\hbar^{1/N}]/\hbar\,\mathbb{C}[\hbar^{1/N}]$.

6. Reconstruction of global categories

- We have a local system on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$ of categories of finite-dimensional modules over $\mathcal{WF}^{\text{local}}_\theta(L)$.
- Using homotopy equivalence $S^1 \sim \mathbb{C} - 0$, we obtain a local system of categories on $\mathbb{C} - 0$.
- The putative local RH correspondence should give an extension of this family to an analytic family over $\mathbb{C}$ such that over $\mathbb{C}[\![\hbar]\!]$ we get the usual formal deformation quantization.
- Wall-crossing formalism, together with certain assumption of *analyticity* (purely A-model statement), allows us to modify the obtained family over $\mathbb{C}$ (without changing anything over $\mathbb{C}[\![\hbar]\!]$) by making finitely many cuts in $\mathbb{C}_\hbar$ (see [20]) and modifying the bundle along cuts by analytic transformations in sectors which have trivial asymptotic expansion at $\hbar = 0$ (like function $e^{-1/\hbar}$ in sector $|\arg(\hbar)| \le \theta_0$ for any $\theta_0 \in (0, \pi)$).
- The resulting family should coincide with an open part of family $\mathcal{C}_{B,\hbar}(M, \omega)$.

Alternatively, one can formulate a version of the hypothetical global RH correspondence without mentioning local RH correspondence at all. Namely, consider the inductive limit over all enhanced compact singular Lagrangians in M of the categories enriched over Perf R where $R = \mathbb{C}\{\hbar\}[\hbar^{-1}]$ is the field of germs of

meromorphic functions. The categories under consideration are the ones obtained by modifications along cuts as above. Then the resulting category should be canonically equivalent to the localization (by inverting $\hbar$) of $\mathbb{C}\{\hbar\}$-linear category associated with the germ at $\hbar = 0$ of the analytic family $C_{B,\hbar}(M, \omega)$.

6.1. Application to resurgence

In [19], we made a Conjecture 4.2.1, which was formulated for analytic stability structure in the case of a finite rank lattice Γ with trivial automorphism, and a central charge $Z : \Gamma \to \mathbb{C}$. In the framework of GRH correspondence, it is the case when ψ_L is trivial (holds for smooth L). Here I present a slight generalization of this conjecture.

Input.

- a complex-analytic space/stack Y,
- integer $N \geq 0$,
- a finite collection of rays $r_j = e^{\mathrm{i}\theta_j} \cdot \mathbb{R}_{\geq 0}$ in $\mathbb{C}_\hbar$,
- for each ray r_j, a germ f_j at $Y \times \{0\} \subset Y \times \mathbb{C}^N$ of a map to Y, which coincides with id_Y on $Y \times 0$,
- a finite collection of numbers $z_{i,j} \in \mathbb{C} - 0$ for $1 \leq i \leq N$ such that $|\arg z_{i,j} - \theta_j| < \pi/2$.

Then one modifies the trivial bundle $Y \times \mathbb{C}_\hbar \to \mathbb{C}_\hbar$ by cutting at rays r_j and re-gluing for small $\hbar \in r_j$ by the map

$$(y, \hbar) \mapsto (f_j(y, e^{-z_{1,j}/\hbar}, \ldots, e^{-z_{N,j}/\hbar}), \hbar).$$

The resulting new analytic germ of bundle is trivialized over $\mathbb{C}[\![\hbar]\!] \Rightarrow$ for a holomorphic germ s of section of the modified bundle, and for a germ of analytic function F on Y at $s(0)$, one get a formal series $g = s^*F \in \mathbb{C}[\![\hbar]\!]$.

Conjecture. Series $g = \sum_{n\geq 0} g_n \hbar^n$ is *resurgent*, that is, $\sum_{n\geq 0} g_n/n! \cdot \zeta^n$ admits endless analytic continuation.

This conjecture will imply resurgence of WKB series for solutions of linear differential equations depending on small parameter $\hbar$.

In the more general case of a local system of lattices and central charges with values in fractional powers of $\hbar$, one expect to get certain (not yet developed) generalization of the notion of resurgence.

6.2. Painlevé non-linear flat connection

Categories $C_{A,\hbar}(M, \omega)$ for $\hbar \neq 0$ should (locally) depend *only* on the cohomology class $\hbar^{-1}[\omega] \in H^2(M, \mathbb{C})$, as it is a general feature of Fukaya categories. Therefore,

the same should be true for categories $C_{B,\hbar}(M,\omega)$. The latter is defined via a Poisson compactification.

Assume that we have a holomorphic family of compact Poisson manifolds $(P_s,\gamma_s)_{s\in S}$ where S is an analytic space such that each P_s has an open dense symplectic leaf M_S such that topologically M_s form a locally trivial family and the cohomology class of the symplectic form $\omega_s := (\gamma_s)^{-1}_{|M_s}$ stays *constant* in s. Then the family of categories $C_{B,\hbar}(M_s,\omega_s)$ should also be analytically locally constant in s, by the GRH correspondence.

As a concrete example, consider a family of compact curves C_s, with P_s being the standard compactification of $M_s := T^*C_s$ obtained by adding a copy of C_s at infinity. Then we get a family over s of moduli stacks of vector bundles with (automatically flat) connections on C_s. Classical Riemann–Hilbert correspondence identifies these stacks analytically with the moduli stacks of $\mathbb{C}$-local systems, which stay locally constant as we deform $s\in S$. Algebraically, we get an algebraic non-linear flat connection on the bundle of algebraic de Rham moduli stacks over S. In the case of SL_2 bundles over $\mathbb{C}P^1$ without 4 points, we get classical Painlevé equations.

6.3. Twistor family, harmonic objects

Let us recapitulate: generalized RH correspondence deals with the family over $\mathbb{C}^\times_\hbar - 0$ of Fukaya categories where [the class of real symplectic form]+ [the class of B-field] gives an element of $H^2(M,\mathbb{C}/2\pi\mathrm{i}\mathbb{Z})$ depending holomorphically on $\hbar$ in the following way:

$$\hbar\in\mathbb{C}^\times\mapsto\hbar^{-1}[\omega^{2,0}]+\beta_0,\quad \text{for a given } \beta_0\in H^2(M,\mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z}).$$

The extension of this family to $\hbar=0$ uses local RH correspondence and formal deformation quantization. The special fiber should coincide with the category of semistable objects in $\mathrm{Perf}_c(M)$ of slope 0 for certain stability condition ("Higgs bundles").

It seems natural to consider another family of categories, where the dependence on $\hbar$ of the total class in $H^2(M,\mathbb{C}/2\pi\mathrm{i}\mathbb{Z})$ is slightly different:

$$\hbar\in\mathbb{C}^\times\mapsto\hbar^{-1}[\omega^{2,0}]+\hbar[\overline{\omega^{2,0}}]+\beta_0,\quad \beta_0\in H^2(M,\mathrm{i}\mathbb{R}/2\pi\mathrm{i}\mathbb{Z}).$$

In this way we get another holomorphic in $\hbar\in\mathbb{C}P^1$ family of categories $(C_\hbar)$ such that $C_{-1/\bar\hbar}\simeq\overline{C_\hbar}^{\mathrm{opp}}$, a *twistor* family. The categories for $\hbar\neq 0$ should be defined in general as Fukaya categories (or B-model categories for appropriately chosen complex symplectic manifolds with B-fields, up to the ambiguity of Painlevé type from the previous section).

One can try to generalize Deligne–Simpson picture of harmonic representatives to such families. Hypothetically, *all* categories from the twistor family are equivalent to each other as abstract A_∞-categories, but the equivalence is not holomorphic in

the moduli stacks of objects. T. Mochizuki established it for holonomic D-modules in arbitrary dimensions (see the review [27]). In [22], in the case $n = \dim_{\mathbb{C}} M/2 = 1$, we worked out the relation between harmonic objects and periodic monopoles, based on Fukaya-theoretic considerations. Recently, T. Mochizuki proved encouraging results, see [23, 24], confirming our predictions.

Acknowledgments. We thank Yan Soibelman for comments.

Funding. This work was partially supported by the ERC grant "ReNewQuantum".

References

[1] A. Belov-Kanel and M. Kontsevich, Automorphisms of the Weyl algebra. *Lett. Math. Phys.* **74** (2005), no. 2, 181–199 Zbl 1081.16031 MR 2191954

[2] L. Côté, C. Kuo, D. Nadler, and V. Shende, Perverse microsheaves. [v1] 2022, [v3] 2025, arXiv:2209.12998v3

[3] A. D'Agnolo and M. Kashiwara, Riemann–Hilbert correspondence for holonomic D-modules. *Publ. Math. Inst. Hautes Études Sci.* **123** (2016), no. 1, 69–197 Zbl 1351.32017 MR 3502097

[4] P. Deligne, B. Malgrange, and J.-P. Ramis, *Singularités irrégulières: correspondance et documents*. Doc. Math. (Paris) 5, Société Mathématique de France, Paris, 2007, 188 pp. Zbl 1130.14001 MR 2387754

[5] C. Dodd, The p-cycle of holonomic D-modules and quantization of exact algebraic lagrangians. [v1] 2015, [v4] 2024, arXiv:1510.05734v4

[6] K. Fukaya, Y.-G. Oh, H. Ohta, and K. Ono, *Lagrangian intersection Floer theory, part 2: anomaly and obstruction, part II*, pp. 397–805, AMS/IP Stud. Adv. Math. 46.2, American Mathematical Society, Providence, RI; International Press, Somerville, MA, 2009, Zbl 1181.53003 MR 2548482

[7] S. Ganatra, J. Pardon, and V. Shende, Covariantly functorial wrapped Floer theory on Liouville sectors. *Publ. Math. Inst. Hautes Études Sci.* **131** (2020), no. 1, 73–200 Zbl 1508.53091 MR 4106794

[8] Y. Groman, Floer theory and reduced cohomology on open manifolds. *Geom. Topol.* **27** (2023), no. 4, 1273–1390 Zbl 1535.53084 MR 4602417

[9] S. Gukov and E. Witten, Branes and quantization. *Adv. Theor. Math. Phys.* **13** (2009), no. 5, 1445–1518 Zbl 1247.81378 MR 2672467

[10] F. Haiden, An extension of the Siegel space of complex abelian varieties and conjectures on stability structures. *Manuscripta Math.* **163** (2020), no. 1–2, 87–111 Zbl 1445.53059 MR 4131992

[11] A. Kapustin and D. Orlov, Remarks on A-branes, mirror symmetry, and the Fukaya category. *J. Geom. Phys.* **48** (2003), no. 1, 84–99 Zbl 1029.81058 MR 2006226

[12] M. Kashiwara and P. Schapira, *Sheaves on manifolds. With a short history. «Les débuts de la théorie des faisceaux». By Christian Houzel*. Grundlehren Math. Wiss. 292, Springer, Berlin, 1990, 512 pp. Zbl 0709.18001 MR 1074006

[13] M. Kashiwara and P. Schapira, Deformation quantization modules. [v1] 2010, [v3] 2011, arXiv:1003.3304v3

[14] M. Kontsevich, Homological algebra of mirror symmetry. In S. D. Chatterji (ed.), *Proceedings of the international congress of mathematicians: August 3–11, 1994, Zürich, Switzerland*, pp. 120–139, Birkhäuser, Basel, 1995 Zbl 0846.53021 MR 1403918

[15] M. Kontsevich, Deformation quantization of algebraic varieties. *Lett. Math. Phys.* **56** (2001), no. 3, 271–294 Zbl 1081.14500 MR 1855264

[16] M. Kontsevich, Deformation quantization of Poisson manifolds. *Lett. Math. Phys.* **66** (2003), no. 3, 157–216 Zbl 1058.53065 MR 2062626

[17] M. Kontsevich, XI Solomon Lefschetz Memorial Lecture series: Hodge structures in non-commutative geometry. In G. Dito, H. García-Compeán, E. Lupercio, and F. J. Turrubiates (eds.), *Non-commutative geometry in mathematics and physics*, pp. 1–21, Contemp. Math. 462, American Mathematical Society, Providence, RI, 2008 Zbl 1155.53060 MR 2444365

[18] M. Kontsevich, Symplectic geometry of homological algebra, talk on Arbeitstaging 2009, https://www.mpim-bonn.mpg.de/preblob/4024 visited on 8 January 2026

[19] M. Kontsevich and Y. Soibelman, Wall-crossing structures in Donaldson–Thomas invariants, integrable systems and mirror symmetry. In R. Castano-Bernard, F. Catanese, M. Kontsevich, T. Pantev, Y. Soibelman, and I. Zharkov (eds.), *Homological mirror symmetry and tropical geometry*, pp. 197–308, Lect. Notes Unione Mat. Ital. 15, Springer, Cham, 2014 Zbl 1326.14042 MR 3330788

[20] M. Kontsevich and Y. Soibelman, Analyticity and resurgence in wall-crossing formulas. *Lett. Math. Phys.* **112** (2022), no. 2, article no. 32, 56 pp. Zbl 1487.32048 MR 4403710

[21] M. Kontsevich and Y. Soibelman, Holomorphic Floer theory I: exponential integrals in finite and infinite dimensions. 2024, arXiv:2402.07343v2

[22] M. Kontsevich and Y. Soibelman, Riemann–Hilbert correspondence in dimension one, Fukaya categories and periodic monopoles, in preparation

[23] T. Mochizuki, Doubly periodic monopoles and q-difference modules. 2019, arXiv: 1902.03551v1

[24] T. Mochizuki, Triply periodic monopoles and difference modules on elliptic curves. *SIGMA Symmetry Integrability Geom. Methods Appl.* **16** (2020), article no. 048, 23 pp. Zbl 1441.53016 MR 4106616

[25] D. Nadler and E. Zaslow, Constructible sheaves and the Fukaya category. *J. Amer. Math. Soc.* **22** (2009), no. 1, 233–286 Zbl 1227.32019 MR 2449059

[26] J.-P. Ramis, J. Sauloy, and C. Zhang, Local analytic classification of q-difference equations. *Astérisque* **355** (2013), 151 pp. Zbl 1320.39008 MR 3185985

[27] C. Sabbah, Théorie de Hodge et correspondance de Hitchin-Kobayashi sauvages [d'après T. Mochizuki]. *Astérisque* **352** (2013), Exp. No. 1050, 205–241 Zbl 1312.14032 MR 3087348

[28] B. Toën and M. Vaquié, Moduli of objects in dg-categories. *Ann. Sci. Éc. Norm. Super. (4)* **40** (2007), no. 3, 387–444 Zbl 1140.18005 MR 2493386

Maxim Kontsevich
Institut des Hautes Etudes Scientifiques, 35 route de Chartres, 91440 Bures-sur-Yvette, France; maxim@ihes.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/3

Restricted subgraphs of edge-colored graphs and applications

Benny Sudakov

Abstract. A properly edge-colored graph is a graph with a coloring of its edges such that no vertex is incident to two or more edges of the same color. A subgraph is called rainbow if all its edges have different colors. The problem of finding rainbow subgraphs or other restricted structures in edge-colored graphs has a long history, dating back to Euler's work on Latin squares. It has also proven to be a powerful method for studying several well-known questions in other areas.

In this survey, we will provide a brief introduction to this topic, discuss several results in this area and demonstrate their applications to problems in graph decomposition, additive combinatorics, theoretical computer science and coding theory.

1. Introduction

An edge-colored graph is a graph G with assignment of colors to its edges. A subgraph H of an edge-colored graph G is called rainbow if all the edges of H have different colors. The study of rainbow subgraphs and other restricted structures in edge-colored graphs goes back more than 200 years to the work of Euler [33] on Latin squares. A Latin square of order n is an $n \times n$ array filled with symbols from $\{1, \dots, n\}$ such that each symbol appears once in every row and column. Well-known examples of Latin squares are multiplication tables of finite groups. Latin squares have also connections to 2-dimensional permutations, design theory, finite projective planes and error correcting codes (see, e.g., [62, 102]). A pair of Latin squares $P = (p_{i,j})$ and $Q = (q_{i,j})$ are called mutually orthogonal if all the n^2 ordered pairs $(p_{i,j}, q_{i,j})$ are distinct. According to folklore, Catherine the Great asked Euler, who was residing at her court at St. Petersburg at the end of 18th century, the following question. This question is known as the 36 officers problem and is equivalent to the existence of mutually orthogonal Latin squares of order 6. Euler introduced it as follows:

Mathematics Subject Classification 2020: 05C35 (primary); 05C70, 05B15, 68R05 (secondary).
Keywords: extremal problems, graph decompositions, rainbow subgraphs, Latin squares.

> *"The question revolves around arranging* 36 *officers to be drawn from* 6 *different regiments so that they are ranged in a square so that in each line (both horizontal and vertical) there are* 6 *officers of different ranks and different regiments."*

Euler could not solve the problem, but was able to demonstrate the existence of the mutually orthogonal Latin squares when n is odd or a multiple of 4. Observing that no Latin square of order 2 exists and being unable to find a construction for $n = 6$, he conjectured that no mutually orthogonal Latin squares of order n exist for any even $n = 2 \pmod 4$.

Euler's question is equivalent to asking for which values of n there is a Latin square which can be decomposed into n disjoint transversals, where a transversal in a Latin square is a collection of cells which do not share the same row, column or symbol. Indeed, given a Latin square $P = (p_{i,j})$ which is a disjoint union of transversals $T_1, \ldots, T_n$, one can construct another $n \times n$ array $Q = (q_{i,j})$ by changing every entry in the transversal T_i to be the symbol i. It is easy to check from the definition of transversal that the resulting Q is a Latin square which is orthogonal to P. The non-existence of orthogonal squares of order 6 was confirmed by Gaston Tarry [100] only in 1900. However, Euler's conjecture resisted solution for another 60 years until Parker, Bose and Shrikhande [16] in 1960 constructed mutually orthogonal Latin squares for all orders $n \geq 7$, showing that Euler's conjecture is false.

Note that every Latin square $L = (\ell_{i,j})$ of order n corresponds to an edge coloring of the complete bipartite graph $K_{n,n}$ by coloring the edge ij by the symbol $\ell_{i,j}$ from the corresponding cell (i, j). It is easy to check that this gives a proper edge coloring of $K_{n,n}$, using n colors. An edge coloring is proper if no two edges sharing a vertex receive the same color. Identifying the cell (i, j) with the edge ij, a transversal in L corresponds to a rainbow perfect matching (a collection of disjoint edges covering all the vertices) (see Figure 1). Thus, finding transversals in Latin squares is a special case of finding rainbow subgraphs in properly edge-colored graphs. Another reason

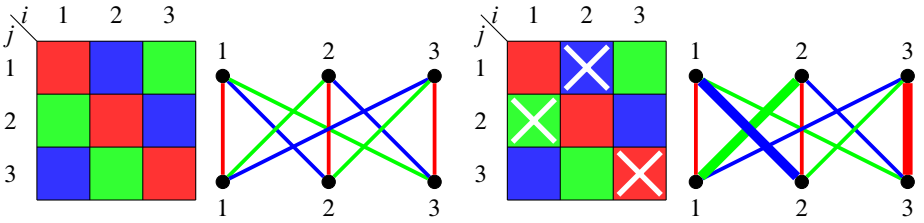


Figure 1. Correspondence between a Latin square and a properly edge-colored complete bipartite graph, on the left, and between a transversal and rainbow perfect matching, on the right.

to study rainbow subgraphs arises in Ramsey theory, more precisely in the canonical version of Ramsey's theorem proved by Erdös and Rado [32]. This theorem shows that locally bounded edge colorings of the complete graph K_n contain rainbow copies of large cliques. An edge coloring is locally k-bounded if each vertex is in at most k edges of any one color.

In this short survey, we discuss the following general question: Given a particular class of edge-colored graphs, what rainbow or other edge-constrained subgraphs do they contain? We present several results in this area and demonstrate their applications to problems in graph decomposition, additive combinatorics, theoretical computer science and coding theory. It is of course impossible to cover everything in such a short article, and therefore the choice of results we present is inevitably biased. Yet, we hope to describe enough examples and problems from this fascinating subject to appeal to researchers not only in discrete mathematics but in other areas as well.

2. Cycles with edge-coloring constraints

The study of cycles in graphs has long been one of the central topics in graph theory. A basic result in this area says that a maximal acyclic graph (with respect to the number of edges) is a tree, that is, an n-vertex graph with n edges contains a cycle. Finding cycles in an appropriately defined auxiliary graph often corresponds to finding a certain dependency between the objects studied. In many situations, this is too simplistic since one cannot capture all relevant properties of the underlying objects by a simple graph. As we will see, in some cases, encoding the relations among studied objects via a properly edge-colored graph gives one significantly more power. In this situation, however, one usually needs to find a cycle with some special properties. This leads us to the following very natural question. How many edges in a properly edge-colored graph will guarantee that there exists a cycle satisfying certain edge-coloring constraints?

One of the most basic constraints is to ask that a cycle have a color which appears uniquely, namely a color that is distinct from the colors of all the remaining edges of this cycle. This is clearly a very special case of the rainbow cycle problem in which all edges must have a unique color. This problem was first proposed in 2007 by Keevash, Mubayi, Sudakov and Verstraëte [61], who observed that, somewhat surprisingly, a linear number of edges is no longer enough to guarantee such a cycle. The construction which shows this uses Cayley sum graphs together with their canonical edge coloring (see Figure 2).

Definition 2.1. Given an abelian group Γ and a subset $S \subset \Gamma$, let G be the graph whose vertices are the elements of Γ. Two vertices x, y are adjacent if and only if

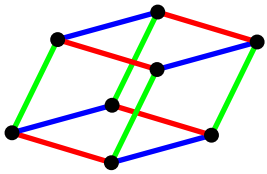


Figure 2. Cayley sum graph of the group $\mathbb{F}_2^3$ with respect to the standard basis and its canonical edge coloring.

$x + y = s$ for some $s \in S$. Moreover, G has a canonical proper edge coloring in which we color edge (x, y) by color $s = x + y$.

To get an example of a properly edge-colored graph without a cycle with a unique color, consider the Cayley sum graph G of the abelian group $\mathbb{F}_2^k$ with respect to the standard basis $S = \{e_1, e_2, \ldots, e_k\}$, where e_i is a vector whose i-th coordinate is one and the rest are zero. This graph has $n = 2^k$ vertices; it is k-regular and properly edge colored with the color of the edge (x, y) being i if $x + y = e_i$. Hence, G has $nk/2 = (n/2)\log_2 n$ edges. Moreover, the sum of the basis vectors corresponding to colors of the edges of a cycle in G equals twice the sum of elements corresponding to the vertices of this cycle. In other words, since we are working over $\mathbb{F}_2$, the colors of any cycle in G correspond to the collection of basis vectors whose sum is zero. Since these vectors are linearly independent, we have that each basis vector (or each color) should appear in every cycle an even number of times. Thus, no cycle has an edge of unique color, and we have a construction which shows that more than $(n/2)\log_2 n$ edges are needed in order to guarantee a cycle with a unique color in a graph.

The above bound has the right order of magnitude. This was shown in [61], using an argument which counts the number of rainbow paths (a path is a walk in which no vertex is repeated) in the edge-colored graph. Moreover, the proof produces a cycle with an edge of unique color that has at most logarithmic length.

Lemma 2.2. *Any properly edge-colored n-vertex graph with at least $2n\log_2 n$ edges contains a cycle of length at most $2\log_2 n$ with an edge of unique color.*

Proof. Let G be a properly edge-colored n-vertex graph with at least $2n\log_2 n$ edges. By deleting vertices of low degree, we can assume that the minimum degree of G is at least $r = 2\log_2 n$. Let $\ell = \log_2 n$, and for the sake of contradiction, suppose G contains no cycle of length at most 2ℓ with an edge of unique color. We will double-count the number of rainbow paths in G of length ℓ. By our assumption, note that counting rainbow paths is the same as counting rainbow walks, since if the rainbow walk intersects itself, we immediately get a desired rainbow cycle. In order to give a lower bound on the number of rainbow paths, we can then take a vertex $v_1 \in V(G)$

and greedily count the number of rainbow paths in G of the form $v_1v_2v_3\cdots v_{\ell+1}$. By the minimum degree assumption and since G is properly edge colored, we have at least r options for v_2, then at least $r-1$ options for v_3, as the second edge of the path should have a different color from the first edge. With the same reasoning, we have at least $r-(i-2)$ options for v_i, concluding that the number of such rainbow paths is at least $n\prod_{j=0}^{\ell-1}(r-j)\geq n\ell^{\ell}$.

On the other hand, since there is no cycle with an edge of unique color, we can upper bound the number of rainbow paths as follows. For each pair of vertices x, y (there are at most n^2 of them), observe that if there is a rainbow path of length ℓ between them using a set of colors $\mathcal{C}$, then any other rainbow path between x and y must use the same set of colors $\mathcal{C}$. Therefore, there are at most $\ell!$ rainbow paths of length ℓ between any pair of vertices. Since $n^2\ell!<n\ell^{\ell}$ for $\ell=\log_2 n$ and large enough n, this contradiction completes the proof. ∎

2.1. Even covers and locally decodable codes

The above simple lemma has interesting applications to problems in theoretical computer science and coding theory. To describe these applications, we need the following definition.

Definition 2.3. Let X be a finite ground set, and let $\mathcal{H}$ be a collection of subsets of X. A non-empty subfamily $\mathcal{F}\subset\mathcal{H}$ is called an even cover if it covers every vertex an even number of times. Equivalently, $\mathcal{F}$ is an even cover if $\sum_{E\in\mathcal{F}}v_E=0$ over $\mathbb{F}_2$, where v_E denotes the characteristic vector of E. The characteristic vector of a subset E is a 0/1-vector whose coordinates corresponding to the elements of E are 1 and the rest of whose coordinates are 0.

A hypergraph is just a collection of subsets called edges, and a hypergraph is k-uniform if every subset has size k. When $\mathcal{H}$ is a 2-uniform hypergraph, that is, a usual graph, an even cover is simply an even subgraph, that is, a subgraph whose vertices all have even degree. Such a subgraph is a union of edge-disjoint cycles in $\mathcal{H}$, and so an even cover with the smallest number of edges must in fact be a simple cycle. Hence, in the case of graphs, the length of the smallest even cover equals the length of the shortest cycle of $\mathcal{H}$, which we call girth. The extremal trade-off between the size of $\mathcal{H}$ (i.e., the number of edges in the graph) and its girth was conjectured by Bollobás and confirmed by Alon, Hoory and Linial [7] who proved the *irregular Moore bound*, which says that every graph on n vertices with average degree d has girth at most $2\lceil\log_{d-1}(n)\rceil$. It is an outstanding open problem whether the constant 2 in this bound can be further improved (for the best known constant, see [71]).

Even covers in k-uniform hypergraphs also correspond to linearly dependent subsets of a system of homogeneous linear equations over $\mathbb{F}_2$ which have exactly k

non-zero coefficients (such equations are called k-sparse). To see why, let us associate a variable x_v with each vertex $v \in V(\mathcal{H})$ and the k-sparse equation $\sum_{v \in E} x_v = 0$ with each edge E in $\mathcal{H}$. Then, observe that for any even cover $\mathcal{F}$, the left-hand sides of the equations corresponding to $E \in \mathcal{F}$ add up to 0 and are thus linearly dependent. Thus, size vs. girth trade-offs for k-uniform hypergraphs correspond to the largest possible size ℓ of the minimum linear dependency in a system of k-sparse linear equations with m equations in n variables. Such k-sparse linear equations arise naturally as parity check equations for *low-density parity check* (LDPC) error correcting codes, and the size vs. girth trade-offs for k-uniform hypergraphs thus correspond to rate vs. distance trade-offs for LDPC codes.

By the equivalence between even covers and linear dependencies, it is clear that every hypergraph with $m \geq n + 1$ hyperedges must have an even cover of size at most $n + 1$. This is tight as can be seen, for example, by considering a 1-uniform hypergraph consisting of n singletons. This suggests the following natural question.

Question 2.4. How does the trade-off between the number of edges m of the k-uniform hypergraph $\mathcal{H}$ and the maximum possible girth of $\mathcal{H}$ look as m increases beyond $n + 1$?

For k-uniform hypergraphs with $k > 2$, the size vs. girth trade-offs were first studied by Naor and Verstraëte [83] through applications to rate vs. distance trade-offs for LDPC codes mentioned above. They showed that every $\mathcal{H}$ with $m \geq n^{k/2} \log^{O(1)}(n)$ hyperedges on n vertices must contain an even cover of length $O(\log n)$. The $\log^{O(1)}(n)$ factor was further improved to an $O(\log \log n)$-factor in a subsequent work of Feige [34]. For $k = 2$, this recovers a coarse version of the irregular Moore bound. For $k > 2$, however, there is an interesting regime between the two extreme thresholds of $m = n + 1$ (with maximum possible girth of $n + 1$) and $m \sim n^{k/2} \log^{O(1)}(n)$ (with maximum possible girth of $O(\log n)$).

In 2008, Feige [34] formulated a conjecture about this in-between regime that suggests a smooth interpolation between the two extremes noted above.

Conjecture 2.5. *Fix any $k \geq 2$. Then, there exists a sufficiently large $C > 0$ such that for sufficiently large n and every ℓ, every k-uniform hypergraph H with $m \geq C n (n/\ell)^{k/2-1}$ hyperedges has an even cover of length at most $O(\ell \log_2 n)$.*

Random hypergraphs witness the quantitative behavior in Feige's conjecture, up to a multiplicative factor of $\log(n)$ in m. Indeed, Feige's conjecture was based on the hypothesis that random hypergraphs are approximately extremal for the purpose of avoiding short even covers. The motivation for this conjecture comes from the question of showing the existence of (and/or efficiently finding) polynomial size *refutation witnesses*. These are easily verifiable certificates of unsatisfiability of randomly

chosen k-SAT formulas parameterized by the number of clauses. Feige's conjecture implies that the result of Feige, Kim and Ofek [35], that random 3-SAT formulas with $m \geq \Omega(n^{1.4})$ clauses admit a polynomial size refutation witness with high probability, will extend to the significantly more general setting of *smoothed* 3-SAT formulas, in addition to simplifying the construction and arguments based on the second moment method in [35].

Until recently, not much was known about Feige's conjecture except for the work of Alon and Feige [6] that showed a suboptimal version for the case of $k = 3$, and that of Feige and Wagner [36] that built an approach to the problem of even covers by viewing them as an instance of generalized girth problems about hypergraphs. In 2022, Guruswami, Kothari and Manohar [47] proved Feige's conjecture up to an additional loss of a $\log^{2k}(n)$ multiplicative factor in m via a rather involved spectral argument applied to the so-called Kikuchi graph, which we discuss in more detail below. Their argument was tightened to reduce the loss down to an $O(\log n)$ multiplicative factor in m by Hsieh, Kothari and Mohanty in [54]. Using Lemma 2.2, together with Hsieh, Kothari, Mohanty and Munhá Correia [55], we can give a substantially simpler, purely combinatorial argument that recovers their result and improves the logarithmic factors for hypergraphs of odd uniformity. This proof is particularly short in the case of even uniformity, so we present it here.

Theorem 2.6. *For all even k and sufficiently large constant C, every k-uniform n-vertex hypergraph with at least $Cn(n/\ell)^{k/2-1} \cdot \log_2 n$ hyperedges contains an even cover of size $2\ell \log_2 n$.*

Proof. Let $\mathcal{H}$ be a hypergraph satisfying the assertion of the theorem. We define an edge-colored graph G, called the Kikuchi graph, as follows (see Figure 3). The vertex set of G consists of all ℓ-element subsets of the vertex set $V(\mathcal{H})$. For two sets S, T of size ℓ, define an edge $S \leftrightarrow T$ if there exists an edge E of the hypergraph $\mathcal{H}$ such that characteristic vectors v_S, v_T satisfy $v_S + v_T = v_E$, that is, $|S \cap E| = |T \cap E| = k/2$ and $|S \cap T| = \ell - k/2$. Note that G is a subgraph of the corresponding Cayley sum graph. Color the edge $S \leftrightarrow T$ in G with color E. An important observation is that

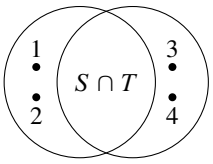


Figure 3. An illustration of the Kikuchi graph defined above with an edge $E = \{1, 2, 3, 4\}$ of $\mathcal{H}$ and an edge $S \leftrightarrow T$ of G colored with E.

this edge coloring is proper. Indeed, given an ℓ-set S and a color $E \in E(\mathcal{H})$, there exists at most one ℓ-set T satisfying $v_S + v_T = v_E$, since this implies $v_T = v_S + v_E$.

Next we apply Lemma 2.2 to show that G contains a cycle of length $2\ell \log_2 n$ with a unique color. Note that G has $N = \binom{n}{\ell}$ vertices and each edge $E \in E(\mathcal{H})$ creates at least $\binom{n-k}{\ell-k/2}$ edges of G, since we need to choose the intersection $S \cap T$ which is disjoint from E and combine it with an arbitrary partition of E into two equal parts. Therefore, the number of edges in G is

$$e(G) \geq e(\mathcal{H})\binom{n-k}{\ell-k/2} \geq Cn\left(\frac{n}{\ell}\right)^{k/2-1} \log_2 n \cdot \binom{n-k}{\ell-k/2} \geq 2N \log_2 N.$$

The last inequality follows from a simple but tedious computation by choosing the constant C sufficiently large and using that $N = \binom{n}{\ell}$ and $\log_2 N \leq \log_2 n^\ell = \ell \log_2 n$. Since G is properly colored, Lemma 2.2 implies that it contains the desired cycle of length at most $r \leq 2 \log_2 N \leq 2\ell \log_2 n$.

To finish the proof, take the cycle $S_1, S_2, \ldots, S_r$ in G given above and consider the collection of edges $\{E_i\}$ in $\mathcal{H}$ which appear as colors in this cycle. By the definition of G, we have $v_{S_i} + v_{S_{i+1}} = v_{E_i}$. Thus, $\sum_i v_{E_i} = \sum_i (v_{S_i} + v_{S_{i+1}}) = 0$. Since some edge E_i appears in this sum uniquely (i.e., does not cancel out), we got a nontrivial even cover of size at most $r = 2\ell \log_2 n$, as desired. ∎

The bound in Lemma 2.2 on the number of edges which guarantees a cycle with a unique color also has an application to another well-known problem related to locally decodable codes. A binary error-correcting code is a map $\mathcal{C} : \{0,1\}^m \to \{0,1\}^n$, where we view the input as an m-bit "message" and the output as an n-bit codeword. By a slight abuse of notation, we use $\mathcal{C}$ to also denote the set of all codewords: $\{y : \exists x \in \{0,1\}^m, \mathcal{C}(x) = y\}$. We say that $\mathcal{C}$ is linear if, when viewing the input and output as $\mathbb{F}_2^m$ and $\mathbb{F}_2^n$, respectively, $\mathcal{C}$ is an $\mathbb{F}_2$-linear map. A code $\mathcal{C}$ is called (q, δ)-locally decodable (LDC) if, in addition, it admits a randomized *local decoding* algorithm. Such a local decoding algorithm takes as input any target message bit i for $1 \leq i \leq m$ and a corrupted codeword y such that $\mathsf{dist}(y, \mathcal{C}(x)) \leq \delta n$ for some $x \in \{0,1\}^m$, where dist counts the number of coordinates in which y and $C(x)$ differ. The goal of the algorithm is to access at most q locations in y and output x_i correctly with high probability over the choice of the q locations. In other words, the local decoder can decode any bit of the message by reading at most q locations of the received corrupted codeword.

Locally decodable codes are intensely investigated in computer science (see the survey [110] for background and applications) with applications to probabilistically checkable proofs, private information retrieval [109] and worst-case to average-case reductions in computational complexity theory. They also have deep connections with additive combinatorics and incidence geometry [28]. We are typically concerned with

codes that are locally decodable with very few queries, such as $q = 2$ or 3, and the fundamental question is the smallest possible $n = n(m)$ such that there is a (q, δ)-binary LDC $\mathcal{C} : \{0,1\}^m \to \{0,1\}^n$. Classical results have essentially completely resolved the case of $q = 2$, and we know that codewords must have blocklength of at least $n \geq 2^{\Omega(m)}$ [43, 64], and a matching upper bound is provided by the Hadamard codes (when δ is constant). The case of $q = 3$ already presents wide gaps, where, until recently, the best known lower bound [43, 64] was $n \geq \widetilde{O}(m^2)$, while the best known construction [30, 108] gives a 3-query binary linear code with $n \leq \exp(\exp(O(\sqrt{\log m \log\log m})))$. Recently, Alrabiah et al. [10] improved the quadratic bound above to obtain a lower bound of $n \geq m^3/\operatorname{polylog} m$ for 3-query binary locally decodable codes. As in the case of the results on even covers (mentioned above), their proof involves spectral arguments on signed adjacency matrices of Kikuchi graphs based on matrix concentration inequalities. In [55], we use the variant of Lemma 2.2 to give a simple and purely combinatorial proof of the following result, which recovers their cubic lower bound and has slightly better logarithmic dependence.

Theorem 2.7. *Let $C : \mathbb{F}_2^m \to \mathbb{F}_2^n$ be a linear map that gives a 3-query locally decodable code with distance $\delta > 0$. Then, $m \leq Kn^{1/3}\log n$ for $K = 10^7/\delta^2$.*

2.2. Rainbow cycles

As we already mentioned in the introduction, the most natural edge-coloring constraint is to ask for a subgraph to be rainbow. In this subsection, we discuss a problem of finding rainbow cycles in edge-colored graphs. The study of this question was initiated by Keevash, Mubayi, Sudakov and Verstraëte [61], who asked how many edges one can have in a properly edge-colored n-vertex graph without containing a rainbow cycle. The example of the Cayley sum graph of $\mathbb{F}_2^k$ with respect to the standard basis, which we described earlier, gives a lower bound for this problem of $\Omega(n \log n)$, and it was conjectured in [61] that this is best possible.

Conjecture 2.8. *There exists a constant C such that every properly colored n-vertex graph with at least $Cn\log n$ edges contains a rainbow cycle.*

The first nontrivial upper bound for the rainbow cycle problem was obtained by Das, Lee and Sudakov [25], who showed that, for sufficiently large n, having $n^{1+o(1)}$ edges is enough to guarantee such a cycle. This paper also introduced one of the two main approaches to the rainbow cycle problem, in which one passes to an expander subgraph to find in it a rainbow cycle. Expanders are graphs with strong connectivity properties, and they have many applications across mathematics and computer science (see [53]). Another important approach, based on the homomorphism counts,

was pioneered by Janzer in [57], who proved the first polylogarithmic bound of $O(n(\log n)^4)$ for this problem. Keevash, Mubayi, Sudakov and Verstraëte [61] also proved that if G is a properly edge-colored n-vertex graph with at least $n\log_2(n+3)$ edges, then for some k it contains a cycle of length k which has more than $k/2$ different colors. Because of the Cayley sum construction, described at the beginning of this section, this is tight up to a constant factor. Using a novel variant of the homomorphism count technique, Janzer and Sudakov [58] strengthened this substantially by finding a cycle which is almost rainbow.

Theorem 2.9. *If n is sufficiently large, $0 < \varepsilon < 1/2$ and G is a properly edge-colored n-vertex graph with at least $(4n/\varepsilon)\log n$ edges, then for some k it contains a cycle of length k with more than $(1-\varepsilon)k$ different colors.*

Conjecture 2.8 is still open. After several further improvements in [58, 66, 101], the current best bound was obtained in an impressive paper of Alon, Bucić, Sauermann, Zakharov and Zamir [5], who came very close to proving the conjecture.

Theorem 2.10. *There exists a constant $C > 0$ such that every properly edge-colored n-vertex graph with average degree at least $C\log n\log\log n$ contains a rainbow cycle.*

The rainbow cycle problem has interesting applications to questions in other areas. Next we discuss one such application to a problem in additive combinatorics.

Definition 2.11. Let $(\Gamma,\cdot)$ be a group with identity element e. A subset $S\subseteq\Gamma$ is called dissociated if there is no solution to the equation $g_1^{\varepsilon_1}\cdots g_m^{\varepsilon_m} = e$ for all choices $\varepsilon_i = \pm 1$, $m \geq 1$ and distinct $g_1,\dots,g_m \in S$.

This definition means that there is no nontrivial relation between elements of S with exponents ± 1. Therefore, the maximal dissociated sets play to some extent a similar role in arbitrary groups as linearly independent sets play in vector spaces. From this definition, it is easy to see that a maximal dissociated subset of a set $A\subseteq\Gamma$ generates the set A, using only exponents ± 1. In addition, one can show [73] that for abelian groups all maximal dissociated subsets of any set $A\subseteq\Gamma$ have almost the same size. It is therefore not surprising that controlling the size of a maximal dissociated subset is relevant for various problems in additive number theory; see [14, 18, 22, 69, 94]. Dissociated sets also play an important role in harmonic analysis; see, for example, [68, 91]. The above-mentioned properties lead naturally to a notion of dimension of a finite subset A of a group Γ which was defined by Schoen and Shkredov [96].

Definition 2.12. The additive dimension $\dim A$ of subset A of group Γ is the size of the largest dissociated subset of A.

A subset $A \subset \Gamma$ has a small doubling if the cardinality of $A \cdot A = \{aa' : a, a' \in A\}$ is of order $O(|A|)$. The study of the structure of sets with small doubling in the abelian setting traces back to a celebrated result of Freiman [39] from 1964 and has been the subject of extensive study ever since; see, for example, [45, 95] and references therein. There has been a lot of work extending results about sets with small doubling to nonabelian groups, with a foundational paper of Tao [99] fueling the work in this direction. One can naturally expect that a set with small doubling should be rather structured and could be "well approximated" by a "subspace/subgroup" generated by its maximal dissociated subset. The following theorem, proved in [93] (part 1) and [5] (part 2), supports this intuition.

Theorem 2.13. *Let Γ be a group, and let $A \subseteq \Gamma$ be a finite subset with $|A \cdot A| \leq K|A|$ for some positive integer K.*

(1) *If Γ is abelian, then* $\dim A \leq O_K(\log|A|)$.

(2) *For general group Γ,* $\dim A \leq O_K(\log^{1+o(1)}|A|)$.

The proof of the first statement can be deduced from Lemma 2.2, and we omit it here. We illustrate the application of Theorem 2.10 by proving the second statement.

Proof. Let $A \subseteq \Gamma$ with $|A \cdot A| \leq K|A|$, and let S be a dissociated subset of A of maximum size. Then $|S| = \dim A$. Define an auxiliary bipartite graph H with disjoint parts A and $B = A \cdot S = \{as : a \in A,\ s \in S\}$ such that (a, b) is an edge of H if $b = as$ for some $s \in S$. Also, color such an edge of H by color s.

By definition, this edge coloring is proper, and the number of vertices of H satisfies $|V(H)| \leq |A| + |A \cdot A| \leq O_K(|A|)$. Moreover, the number of edges in H is $|A||S|$. Therefore, if $|S| > C_K \log|A| \log\log|A|$, for a sufficiently large constant C_K depending on K, we can use Theorem 2.10 to find a rainbow cycle in H. This cycle has an even length, since H is bipartite. A rainbow cycle with distinct colors $s_1, s_2, \ldots, s_{2\ell}$ appearing in that order implies

$$s_1 s_2^{-1} s_3 s_4^{-1} \cdots s_{2\ell-1} s_{2\ell}^{-1} = e.$$

This contradicts that S is dissociated and shows that $|S| \leq C_K \log|A| \log\log|A| = O(\log^{1+o(1)}|A|)$. ∎

The rainbow cycle problem has several additional interesting applications. For example, Theorem 2.10 was recently used by Alrabiah and Guruswami [9] to obtain nearly tight bounds on the dimension of 3-query locally correctable binary linear

codes. These codes are closely related to locally decodable codes, which we already mentioned above. For more definitions and history of the problem, we refer the interested reader to [9].

3. Matchings and transversals in Latin squares

Recall that a Latin square of order n is an $n \times n$ array filled with n symbols so that every symbol appears only once in each row and in each column. A partial transversal of size t is a collection of t cells of the Latin square which do not share the same row, column or symbol. A transversal (also known as a full transversal) is a partial transversal of order n. As we already discussed in the introduction, the study of transversals in Latin squares goes back to the work of Euler on orthogonal Latin squares.

It is easy to see that for even n, there are many Latin squares without full transversals. For example, consider a Latin square $L = (\ell_{i,j})$ of order $2k$ which is an addition table of the cyclic group $\mathbb{Z}_{2k}$. Suppose it has a transversal $T = \ell_{1,\sigma(1)}, \dots, \ell_{2k,\sigma(2k)}$, where σ is some permutation of $1, \dots, 2k$. By definition, $\ell_{i,\sigma(i)}$ is a permutation of the numbers $1, \dots, 2k$ and also $\ell_{i,\sigma(i)} = i + \sigma(i) \bmod 2k$. This implies that the sum of the symbols in the transversal T is

$$\sum_{i=1}^{2k} i = \sum_{i=1}^{2k} \ell_{i,\sigma(i)} = \sum_{i=1}^{2k} (i + \sigma(i)) = 2 \sum_{i=1}^{2k} i \bmod 2k.$$

As $\sum_{i=1}^{2k} i = k \bmod 2k$, this is a contradiction and shows that the addition table of $\mathbb{Z}_{2k}$ has no transversal (see Figure 4).

One central problem concerning Latin squares is to determine which of them have transversals. This question is very difficult even in the case of multiplication tables of finite groups. In 1955, Hall and Paige [51] conjectured that the multiplication table of a group G has a full transversal exactly if the 2-Sylow subgroups of G are trivial

	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

Figure 4. The addition table of the cyclic group $\mathbb{Z}_4$, for which the corresponding Latin square does not contain a full transversal.

or non-cyclic. It took 50 years to establish this conjecture. Its proof, which is a combination of works by Wilcox, Evans and Bray, is based on the classification of finite simple groups (see [103] and the references therein). Recently, two alternative proofs of this conjecture, using very different methods, were found for large groups.

The first proof was obtained by Eberhard, Manners and Mrazović [29] using tools from analytic number theory. Remarkably they were also able to obtain a very accurate quantitative bound for the number of transversals in the multiplication tables of groups which satisfy Hall–Paige conjecture. A different proof of this conjecture, this time by combinatorial techniques, was given by Müyesser and Pokrovskiy [82]. They prove a very general theorem about multiplication tables of large groups which gives a solution to several old combinatorial problems in group theory, including the proof of Hall–Paige conjecture for large groups.

Another central problem in this area is to determine the size of the largest partial transversal one can find in every Latin square. As we already explained in the introduction, this is equivalent to finding a maximum rainbow matching in a proper edge coloring of the complete bipartite graph $K_{n,n}$ with n colors. The answer to this problem was famously conjectured more than 50 years ago by Ryser, Brualdi and Stein [20, 92, 98].

Conjecture 3.1. *Every proper edge coloring of the complete graph $K_{n,n}$ with n colors contains a rainbow matching of size $n-1$. Moreover, if n is odd, it contains a perfect rainbow matching.*

Most research towards the Ryser–Brualdi–Stein conjecture has focused on proving better and better lower bounds on the size of a maximum rainbow matching, trying to get as close to $n-1$ as possible. Here Koksma [67] found rainbow matchings of size $2n/3 + O(1)$, and Drake [27] improved this to $3n/4 + O(1)$. The first asymptotic proof of the conjecture was obtained by Brouwer, de Vries and Wieringa [19], and independently by Woolbright [105], who found rainbow matchings of size $n - \sqrt{n}$. This was improved in 1982 by Shor [97] to $n - O(\log^2 n)$. His paper had a mistake which was later rectified, using the original approach, by Hatami and Shor [52]. For nearly 40 years, this was the best known bound for the Ryser–Brualdi–Stein conjecture. Recently, Keevash, Pokrovskiy, Sudakov and Yepremyan [62] improved this old bound and showed that one can find a rainbow matching of size $n - O(\log n/\log\log n)$ in every proper edge coloring of $K_{n,n}$ which uses n colors. Finally, last year, in a groundbreaking development, Montgomery [77] proved the Ryser–Brualdi–Stein conjecture for large even n.

Theorem 3.2. *For every sufficiently large n, every proper edge coloring of the complete graph $K_{n,n}$ with n colors contains a rainbow matching of size $n-1$.*

While this confirms the conjecture for even n, it falls short of the conjecture by an additive constant 1 when n is odd. This slack is important in many places in the proof, and it seems that further ideas will be needed to find a perfect rainbow matching for odd n.

The proofs of the recent bounds on the Ryser–Brualdi–Stein conjecture are rather involved, so we only briefly discuss some highlights here. One of the key ideas needed to find transversals of size $n - O(\log n/\log\log n)$ in [62] is a claim that a union of a random nearly perfect matching and an arbitrary regular bipartite graph produces with high probability a good expander, that is, a graph with strong connectivity properties.

Given a copy of $K_{n,n}$ which is properly colored with n colors, the authors of [62] begin by finding a large random nearly perfect rainbow matching, containing $n - n^{1-\varepsilon}$ edges for some fixed small $\varepsilon > 0$. This initial matching is found using the so-called Rödl nibble [88], which is a celebrated probabilistic technique with many applications. This matching is then successively modified in several steps, each time getting a rainbow matching that is one edge larger. At each step, given a current rainbow matching M, one takes two vertices $x, y \in V(K_{n,n}) \setminus V(M)$ which are in different parts, and an x, y-alternating path P, that is, a path such that even edges in this path are in M and odd edges (starting with the edge containing x) all have different colors not used on M. Then, removing the even edges of P from M and adding the odd edges gives a rainbow matching with one more edge. Let us call G the graph consisting of the edges with the colors not used on M and note that G is regular (since every vertex touches edges of all the colors). Since M contains an initial random matching, its union with G is an expander. This essentially implies that for every pair of vertices not in the matching and every set of $C\log n/\log\log n$ colors (for some fixed constant C) not on M, such a path P can be found. Moreover, this property is sufficiently robust that it can be used to make iterative adjustments (with some additional care) until the rainbow matching uses all but $C\log n/\log\log n$ of the original n colors.

An error term of order $\Omega(\log n/\log\log n)$ is a natural barrier for this method, since the diameter of the expander graph in this proof is at least $\Omega(\log n/\log\log n)$. This means that if M contains more than $n - \Omega(\log n/\log\log n)$, it will not be possible to find a path P described above, since such paths must contain at least $\Omega(\log n/\log\log n)$ colors which do not appear in M.

A significant part of the work of Montgomery [77] is to exploit the structure of the proper edge coloring of $K_{n,n}$ and to find some way to identify approximate algebraic properties of this edge coloring, as well as the introduction of an "addition structure" to use these properties. Recall that if the edge coloring was coming from the addition table of a finite abelian group and we had three edges (x, y) of color c_1 and (x', y') of color c_2 and (x, x') of color c_3, then $x + y = c_1, x' + y' = c_2$ and $x + x' = c_3$. Therefore, the color of the edge (y, y') is determined and is equal to $c_1 + c_2 - c_3$.

Having many such three-edge paths with colors c_1, c_3, c_2 for which the color of the fourth edge is the same can be used to find an absorbing structure which, by switching the edges in the rainbow matching, can free pairs of vertices (that are no longer covered by rainbow matching). Such an absorbing structure can be found and set aside, before a large rainbow matching is found disjointly from it, using the method of distributive absorption.

Absorption is a versatile technique for extending approximate results into exact ones, which was first systematically used by Rödl, Ruciński and Szemerédi [89]. Such ideas also appeared earlier in work of Erdős, Gyárfás and Pyber [31], and of Krivelevich [70]. It has since been implemented in many different settings using a variety of creative constructions. Distributive absorption is an efficient and powerful variant of this method, introduced by Montgomery [76], which can be used to build a global absorbing structure from small local absorbers. For more details about this very impressive proof, we refer the interested reader to the excellent survey by Montgomery [78].

A notion closely related to that of a Latin square is the Latin array. A Latin array is an $n \times n$ square array filled with an arbitrary number of symbols such that no symbol appears twice in the same row or column. Latin arrays are natural extensions of Latin squares and have also been extensively studied. A familiar example of such an array is a multiplication table for the elements of two subsets of equal size in some group. Having a Latin array is equivalent to having a proper edge coloring of the complete bipartite graph $K_{n,n}$ without any restriction on the number of colors. It is generally believed that extra colors in a proper edge coloring should help to find large rainbow matchings. Although there are proper edge colorings of $K_{n,n}$ with strictly more than n colors and not even a single rainbow perfect matching, there are several results which confirm this intuition. For example, Montgomery, Pokrovskiy and Sudakov [79] showed that with some very weak additional constraints, one can find many disjoint rainbow perfect matchings.

Theorem 3.3. *For every $\varepsilon > 0$ and sufficiently large n, the following holds. Every proper edge coloring of $K_{n,n}$ with at least εn colors appearing at most $(1-\varepsilon)n$ times contains at least $(1-\varepsilon)n$ edge-disjoint perfect rainbow matchings.*

Note that in a proper edge coloring, each color can appear on at most n edges. Hence, this result suggests that the only Latin arrays without transversals are small perturbations of Latin squares.

The methods used to study the Ryser–Brualdi–Stein conjecture usually also apply to give results about general proper edge colorings of $K_{n,n}$. For example, the authors of [62] showed that any such coloring contains a rainbow matching of size

$n - O(\log n / \log\log n)$. As a consequence of Theorem 3.3, we only need to consider colorings which are close to the colorings that use exactly n colors. Similarly, the techniques developed by Montgomery [77] apply to show that every proper edge coloring of the complete graph $K_{n,n}$ contains a rainbow matching of size $n - 1$.

We will finish this section by discussing two extensions of the problem of finding large rainbow matchings in properly edge-colored bipartite graphs $K_{n,n}$.

The first extension considers host graphs that are not complete bipartite graphs. This problem was proposed by Aharoni and Berger [2], who conjectured that if G is a bipartite multigraph (i.e., more than 1 edge is allowed between each pair of vertices) which is properly colored such that at least $n - 1$ colors have at least n edges, then it has a rainbow matching of size $n - 1$. In fact, this is a generalization of Conjecture 3.1, since a proper n edge coloring of $K_{n,n}$ clearly contains $n - 1$ colors of size n. For the problem of Aharoni and Berger, even finding an approximate solution appeared to be very difficult. It took a sequence of improvements (see the recent survey by Pokrovskiy [85] for this history, and further related problems), before Pokrovskiy [84] gave such a solution. He showed that, for each $\varepsilon > 0$ and sufficiently large n, having $n - 1$ colors with at least $(1 + \varepsilon)n$ edges is sufficient for a rainbow matching with $n - 1$ edges. His original proof was rather involved and a much simpler proof, using a probabilistic sampling trick, was found in [24].

For our second extension, we consider what happens if we no longer require that the coloring of $K_{n,n}$ to be proper, asking only that each of n colors appears n times (this is equivalent to what is known as an equi-n-square). In 1975, Stein [98] conjectured that every such coloring has a rainbow matching with $n - 1$ edges. This ambitious conjecture was open for 40 years but was eventually disproved by Pokrovskiy and Sudakov [86], showing that one cannot hope for rainbow matching of size larger than $n - c \log n$. However, there was a belief that every such coloring at least contains a rainbow matching of asymptotically correct size. After a sequence of improving bounds [3, 11, 98], Chakraborti, Christoph, Hunter, Montgomery, and Petrov [21] confirmed this by showing that every coloring of $K_{n,n}$ in which each of n colors appears n times contains a rainbow matching with $(1 - o(1))n$ edges (and moreover can even be almost decomposed into such matchings).

The result in [21] is very recent and shows that the study of rainbow subgraphs remains an active and exciting area. One very interesting open question is an outstanding conjecture of Stein [98] from 1975, giving a midway point between his later disproved conjecture and the Ryser–Brualdi–Stein conjecture. It says that, given a coloring of $K_{n,n}$ in which each of the n colors appears n times such that in one of the vertex classes no two edges of the same color meet (i.e., the coloring is proper on one side), is there always a rainbow matching with $n - 1$ edges?

4. Rainbow trees and decomposition problems

One of the most natural problems in the study of rainbow structures is to determine which graphs appear as rainbow subgraphs in every properly edge-colored complete graph K_n. Since for even n one can properly color K_n using only $n-1$ colors, we may ask in general only for rainbow subgraphs with at most $n-1$ edges. This leads to a natural question: which n-vertex trees have a rainbow copy in any proper coloring of K_n?

The special case of this question was considered in 1980 by Hahn [50] who conjectured that there is always a rainbow n-vertex path. This was disproved in [74] by considering a complete graph on the vertex set $\mathbb{F}_2^k$, where the edge (x, y) is colored using the color $x + y \in \mathbb{F}_2^k \setminus \{0\}$. Recall that this graph is precisely the Cayley sum graph of $\mathbb{F}_2^k$ with respect to all non-zero elements of this group, along with its canonical edge coloring. For any path $P = x_1x_2 \cdots x_n$ in this coloring, it is easy to see that the sum of the colors of the edges of this path is $x_1 + x_n \neq 0$. But the sum of all the colors (i.e., all non-zero elements of the group) is zero, showing that there is no rainbow path covering all the vertices. Nevertheless, it was widely believed that any properly colored K_n contains a rainbow path covering all but exceptionally few vertices. In particular, Andersen [12] in 1989 conjectured that one can always find a rainbow path covering $n-1$ vertices. The progress on this conjecture was initially slow, despite the efforts of various researchers; see, for example, [4, 23, 41, 48, 49]. Several years ago, the problem was resolved asymptotically by Alon, Pokrovskiy and Sudakov [8], who showed that any properly colored K_n contains both a rainbow path and a rainbow cycle with $n - O(n^{3/4})$ vertices (this error term was slightly improved in [13]).

For a general tree, until recently, there were no results that showed the existence of rainbow copies of large trees in properly edge-colored complete graphs. The above example for the n-vertex path can be extended (see [15]) to show that there are proper edge colorings of K_n which do not contain rainbow copies of certain spanning trees. However, it is still possible that nearly spanning trees exist in all proper colorings of complete graphs. This follows from the following general result, proved by Montgomery, Pokrovskiy and Sudakov [80]. Recall that an edge coloring of K_n is locally k-bounded if each vertex is in at most k edges of any one color. Therefore, a proper coloring is exactly a locally 1-bounded coloring.

Theorem 4.1. *For $\varepsilon > 0, k \in \mathbb{N}$ and sufficiently large n, any locally k-bounded edge coloring of K_n contains a rainbow copy of every tree with at most $(1-\varepsilon)n/k$ vertices.*

Since a locally k-bounded edge coloring of K_n may have only $(n-1)/k$ distinct colors, Theorem 4.1 is tight up to the constant ε for each k. One of the distinguishing

features of this result is that it places no conditions on the trees other than the number of vertices they may have. In comparison, all of the previous graph packing and labeling results required a maximum degree assumption, which was conjectured to be unnecessary. An important tool in the proof of this theorem is to demonstrate that, in properly colored complete graphs, a large rainbow matching can typically be found in a random vertex set using a random set of colors chosen with the same density. In fact, such a matching can cover almost all the vertices in the random vertex set. This allows most of a large tree to be embedded if it can be decomposed into certain large matchings. Contrastingly, one needs deterministic methods to embed vertices in the tree with high degree. These techniques for embedding trees with high degree vertices may offer new approaches to related problems.

4.1. Graph decompositions, labelings and orthogonal double covers

Theorem 4.1 has applications in several different areas of graph theory. We now discuss three such applications to graph decompositions, graph labelings and orthogonal double covers. With each application, this theorem proves an asymptotic form of a well-known conjecture. In each case, we consider some special edge coloring of K_n, coming from a graph-theoretic problem, to which we apply Theorem 4.1.

Graph decompositions. We say that a graph G has a decomposition into copies of a graph H if the edges of G can be partitioned into edge-disjoint subgraphs isomorphic to H. In 1847, Kirkman studied decompositions of complete graphs K_n and showed that they can be decomposed into copies of a triangle if, and only if, $n \equiv 1$ or $3 \pmod 6$. Wilson [104] generalized this result by proving necessary and sufficient conditions for a complete graph K_n to be decomposed into copies of *any* graph, for large n. A very old problem in this area, posed in 1853 by Steiner, states that, for every k, modulo an obvious divisibility condition, every sufficiently large complete r-uniform hypergraph can be decomposed into edge-disjoint copies of a complete r-uniform hypergraph on k vertices. This problem was the so-called "existence of designs" question and has practical relevance to experimental designs. It was resolved only very recently in spectacular work by Keevash [60] (see the subsequent work of [42] for an alternative proof of this result). Over the years, graph and hypergraph decomposition problems have been extensively studied and by now this has become a vast topic with many exciting results and conjectures (see, for example, [40, 106, 107]).

Decompositions of complete graphs into large trees, whose size is comparable with the size of the complete graph (in contrast with the existence of designs mentioned above), also have a long history. The earliest such result was obtained more than a century ago by Walecki. In 1882, he proved that a complete graph K_n on an even number of vertices can be partitioned into edge-disjoint Hamilton paths (this is

a path which visits every vertex of the graph exactly once). Since paths are a very special kind of tree, it is natural to ask which other large trees can be used to decompose a complete graph. One of the oldest and best-known conjectures in this area, posed by Ringel in 1963 [87], deals with the decomposition of complete graphs into edge-disjoint copies of a tree.

Conjecture 4.2. *Any tree with $n+1$ vertices packs $2n+1$ times into the complete graph K_{2n+1}.*

Note that a complete graph K_{2n+1} has precisely $(2n+1)\cdot n$ edges. Therefore, it is conceivable to partition it into $2n+1$ copies of the same tree which has n edges.

This conjecture is known for many very special classes of trees such as caterpillars, trees with ≤ 4 leaves, diameter ≤ 5 trees, trees with ≤ 35 vertices and several other classes of trees (see [40, Chapter 2] and the references therein). There are also some partial general results in the direction of Conjecture 4.2. Typically, for these results, an extensive technical method is developed which is capable of almost-packing any appropriately sized collection of certain sparse graphs; see, for example, [17, 37, 65, 75]. In particular, Joos, Kim, Kühn and Osthus [59] proved the above conjecture for very large bounded-degree trees. Ferber and Samotij [38] obtained an almost-perfect packing of almost-spanning trees with maximum degree $O(n/\log n)$ (giving what is known as an approximate version of the conjecture). A different proof of the approximate version of Ringel's conjecture for trees with maximum degree $O(n/\log n)$ was obtained by Adamaszek, Allen, Grosu, and Hladký [1], using graph labelings. Theorem 4.3 can be used to give the first asymptotic solution for Ringel's conjecture applicable with no degree restriction.

Theorem 4.3. *For $\varepsilon > 0$ and sufficiently large n, any $(n+1)$-vertex tree packs at least $2n+1$ times into the complete graph $K_{(2+\varepsilon)n}$.*

To see the connection between Theorem 4.1 and Conjecture 4.2, consider the following edge coloring of the complete graph with vertex set $\{0, 1, \ldots, 2n\}$. This coloring, which we call the near distance (ND-)coloring, was introduced by Kotzig to study Ringel's conjecture; see [90]. Color the edge ij by color k, where $k \in [n]$, if either $i = j + k$ or $j = i + k$ with addition modulo $2n+1$. As we explain below, it is easy to show (see also Figure 5) that if an $(n+1)$-vertex tree has a rainbow embedding into the ND-coloring of K_{2n+1}, then Conjecture 4.2 holds for that tree.

Proof of Theorem 4.3. Let $\ell = (1+\varepsilon/3)n$. Consider the ND-coloring of the complete graph $K_{2\ell+1}$, defined above. This is a locally 2-bounded coloring of $K_{2\ell+1}$, which thus, by Theorem 4.1, contains a rainbow copy of T, S_0 say. Now, for

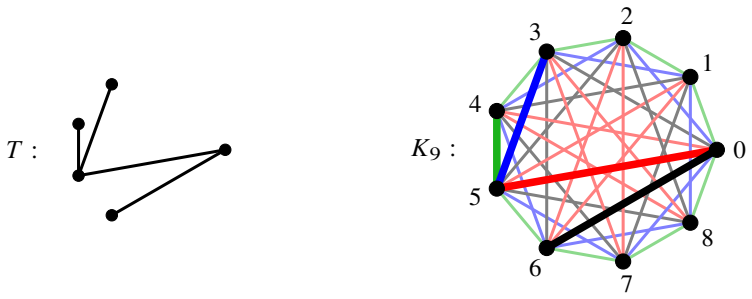


Figure 5. The ND-coloring of K_9 and a rainbow copy of a tree T with four edges. The color of each edge corresponds to its Euclidean length. By taking cyclic shifts of this tree around the center of the picture, we obtain 9 disjoint copies of the tree decomposing K_9 (and thus a proof of Ringel's conjecture for this particular tree). To see that this gives 9 disjoint trees, notice that each edge must be shifted to the other edges of the same color (since shifts are isometries).

each $i \in [\ell]$, let S_i be the tree with vertex set $\{v + i : v \in V(S_0)\}$ and edge set $\{\{v + i, w + i\} : vw \in E(S_0)\}$.

Note that the color of $vw \in E(K_{2\ell+1})$ is the same as the color of the edge $\{v + i, w + i\}$ for each $i \in [2\ell]$, and under the translation $x \mapsto x + 1$ the edge vw moves around all $2\ell + 1$ edges with the same color. Thus, each tree S_i is rainbow, and all the trees S_i are disjoint. ∎

Graph labeling. Graph labeling originated in methods introduced by Rosa [90] in 1967 as a potential path towards proving Ringel's conjecture. In the intervening decades, a large body of work has steadily developed concerning different labelings. Labeled graphs serve as useful models for a broad range of applications in coding theory, circuit design, communication networks, database management, secret sharing schemes, cryptology and models for constraint programming over finite domains (see the extensive survey [40]). One old, well-known, conjecture in this area concerns harmonious labelings. This type of labeling was introduced by Graham and Sloane [44] and arose naturally out of the study of additive bases. Given an abelian group Γ and a graph G, we say that a labeling $f : V(G) \to \Gamma$ is Γ-harmonious if the map $f' : E(G) \to \Gamma$ defined by $f'(xy) = f(x) + f(y)$ is injective. In the case when Γ is a group of integers modulo n, we omit it from our notation and simply call such a labeling harmonious. In the particular case of an n-vertex tree, Graham and Sloane asked for a harmonious labeling using $\mathbb{Z}_{n-1}$ in which each label is used on some vertex, so that a single label is used on two vertices – where this exists we call the tree harmonious. More generally, we also call a graph with m edges and at most m vertices harmonious if it has an injective harmonious labeling with $\mathbb{Z}_m$. According to an unpublished result attributed by Graham and Sloane [44] to Erdős,

almost all graphs are not harmonious. On the other hand, Graham and Sloane [44] in 1980 made the following conjecture for trees.

Conjecture 4.4. *All trees are harmonious.*

This conjecture is known for many special classes of trees such as caterpillars, trees with at most 31 vertices and some other classes of trees (see [40, Chapter 2] and the references therein). Moreover, Żak [111] proposed an asymptotic weakening of this conjecture. He asked to prove that every tree has an injective $\mathbb{Z}_{n+o(n)}$-harmonious labeling.

Note that, for any injective labeling of the vertices of the complete graph by elements of an abelian group, the edge coloring which is obtained by taking sums of labels of vertices is proper. Therefore, we can use Theorem 4.1 to study such colorings. In particular, we can obtain the following general result which shows that every tree is almost harmonious. If Γ is taken to be the cyclic group, this theorem proves Żak's conjecture from [111].

Theorem 4.5. *Every n-vertex tree T has an injective Γ-harmonious labeling for any abelian group Γ of order $n + o(n)$.*

Proof. Suppose $|\Gamma| = \ell = (1 + \varepsilon)n$ for some $\varepsilon > 0$ and sufficiently large n. Let T be a tree on n vertices. Identify the vertices of K_ℓ with the elements of Γ and consider an edge coloring that colors the edge ij by $i + j$. This is a proper coloring, so, when n is large, by Theorem 4.1 it contains a rainbow copy of T, which corresponds to a harmonious labeling. ∎

Orthogonal double covers. Theorem 4.1 can also be used to obtain an asymptotic solution for another old graph decomposition problem. An orthogonal double cover of the complete graph K_n by a graph G is a collection $G_1, \ldots, G_n$ of subgraphs of K_n such that each G_i is a copy of G, every edge of K_n belongs to exactly two of the copies and any two copies have exactly one edge in common. The study of orthogonal double covers was originally motivated by problems in statistical design theory (see [26, Chapter 2]). Since the number of copies in the double cover is n, it follows that G must have $n - 1$ edges. The central problem here is to determine for which graphs G there is an orthogonal double cover of K_n by G. In full generality, this extends several well-known problems, such as the existence question for biplanes or symmetric 2-designs (see, e.g., [56]). About 20 years ago, Gronau, Mullin, and Rosa [46] made the following conjecture about trees.

Conjecture 4.6. *Let T be an n-vertex tree which is not a path on 3 edges. Then, K_n has an orthogonal double cover by copies of T.*

This conjecture is known to hold for certain classes of trees including trees with diameter ≤ 3 and trees with ≤ 13 vertices (see [46, 72] for a full list). To see the connection between Conjecture 4.6 and Theorem 4.1, we will consider a Cayley sum coloring of a complete graph on 2^k vertices, where edges are colored by the sum of their endpoints in the abelian group $\mathbb{Z}_2^k$. By considering such a coloring, we can show that Conjecture 4.6 is asymptotically true whenever n is a power of 2.

Theorem 4.7. *Let $n = 2^k$, and let T be a tree on $n - o(n)$ vertices. Then, K_n contains n copies of T such that every edge of K_n belongs to at most two copies and any two copies have at most one edge in common.*

Proof. Identify $V(K_n)$ with the group $\mathbb{Z}_2^k$. Color each edge ij using the color $i + j \in \mathbb{Z}_2^k$. By Theorem 4.1, K_n has a rainbow copy S of T. For all $x \in \mathbb{Z}_2^k$, define a permutation $\phi_x : V(K_n) \to V(K_n)$ by $\phi_x(v) = x + v$ (with addition in $\mathbb{Z}_2^k$). Use $\phi_x(S)$ to denote the subgraph of K_n with edges $\{\phi_x(a)\phi_x(b) : ab \in E(S)\}$. Notice that, since ϕ_x is a permutation of $V(K_n)$, $\phi_x(S)$ is a tree isomorphic to T. We claim that the family of n trees $\{\phi_x(S) : x \in \mathbb{Z}_2^k\}$ satisfies the theorem.

Notice that since $\phi_x(a) + \phi_x(b) = a + x + b + x = a + b$, the permutations ϕ_x preserve the colors of edges. This implies that the trees $\phi_x(S)$ are all rainbow. Notice that the only edges fixed by the permutations ϕ_x are those colored by x. Finally, notice that $\phi_{x-y} \circ \phi_y = \phi_x$. Combining these, we have that if a color c edge is in two trees $\phi_x(S)$ and $\phi_y(S)$, then $x - y = c$ in $\mathbb{Z}_2^k$. This implies that the trees $\phi_x(S)$ cover any edge at most twice, and any pair of them have at most one edge in common. ∎

Ringel's conjecture revisited. Recall the near distance (ND-)coloring, which was used to prove the asymptotic version of Ringel's conjecture. It colors the edge ij of the complete graph K_{2n+1} by color k, where $k \in [n]$, if either $i = j + k$ or $j = i + k$ with addition modulo $2n + 1$. In the application of Theorem 4.1, we only used the fact that this coloring is 2-bounded to find a rainbow copy of an arbitrary tree of size $(1 - \varepsilon)n$. But the (ND-)coloring has a lot of structural properties and is very symmetric. So it is natural to ask whether one can exploit the structure of the (ND-)coloring to embed in it rainbow trees with exactly n edges. Using several critical new ideas combined with the approach from Theorem 4.1, this was done by Montgomery, Pokrovskiy and Sudakov [81].

Theorem 4.8. *For every sufficiently large n, the complete graph K_{2n+1} can be decomposed into copies of any tree with n edges.*

This theorem constructs a decomposition of the complete graph by considering cyclic shifts of one copy of a given tree (see Figure 5). The existence of such a cyclic decomposition was separately conjectured by Kotzig [90]. Therefore, this also gives

a proof of the conjecture by Kotzig for large trees. A different proof of Ringel's conjecture for large n was obtained in [63].

The proof of Theorem 4.8 in [81] is rather involved for discussing in detail. It introduces three key new methods. Firstly, when dealing with very high degree vertices, it uses a completely deterministic approach for finding a rainbow copy of the tree. This approach heavily relies on the features of the ND-coloring. Secondly, it uses techniques to embed 99% of any tree T randomly into the ND-coloring in a rainbow fashion. The techniques in [81] are based on the probabilistic methods in [80], except we need the random embedding to maintain some degree of independence between the vertices and the colors used for this embedding. High-degree vertices in T force dependencies between the vertices and the colors used on any randomized embedding of large subgraphs of T. These dependencies were a major barrier to generalizing previous random approaches to Ringel's conjecture to trees with high-degree vertices. Maintaining some independence between vertices and colors here is a subtle task and a critical part of the proof.

Finally, having embedded 99% of the tree T, the embedding in [81] is completed using absorption. This is a particularly challenging task, and so we prepare for it during our initial embedding by carefully constructing an absorbing structure which assists with the precise task of completing the embedding while using each unused color exactly once. The novelty in [81], compared to the classical applications of the absorption method, is in the interplay of our constructions with the randomized parts of our embedding, the constructions themselves and the use of absorption to embed graphs with high-degree vertices.

Acknowledgments. B. Sudakov would like to express his gratitude to Aleksa Milojević for many thoughtful discussions, carefully reading the manuscript and drawing several illustrations. He also wishes to thank Richard Montgomery and Matija Bucić for their helpful remarks and suggestions.

Funding. This research was supported in part by SNSF grant 200021-228014.

References

[1] A. Adamaszek, P. Allen, C. Grosu, and J. Hladký, Almost all trees are almost graceful. *Random Structures Algorithms* **56** (2020), no. 4, 948–987 Zbl 1456.05134 MR 4101349

[2] R. Aharoni and E. Berger, Rainbow matchings in r-partite r-graphs. *Electron. J. Combin.* **16** (2009), no. 1, article no. R119, 9 pp. Zbl 1186.05118 MR 2546322

[3] R. Aharoni, E. Berger, D. Kotlar, and R. Ziv, On a conjecture of Stein. *Abh. Math. Semin. Univ. Hambg.* **87** (2017), no. 2, 203–211 Zbl 1379.05115 MR 3696146

[4] S. Akbari, O. Etesami, H. Mahini, and M. Mahmoody, On rainbow cycles in edge colored complete graphs. *Australas. J. Combin.* **37** (2007), 33–42 Zbl 1130.05024 MR 2284366

[5] N. Alon, M. Bucić, L. Sauermann, D. Zakharov, and O. Zamir, Essentially tight bounds for rainbow cycles in proper edge-colourings. *Proc. Lond. Math. Soc. (3)* **130** (2025), no. 4, article no. e70044, 37 pp. Zbl 08031028 MR 4894351

[6] N. Alon and U. Feige, On the power of two, three and four probes. In *Proceedings of the 20th annual ACM-SIAM symposium on discrete algorithms (SODA 2009), January 4–6, 2009, New York, NY, USA*, pp. 346–354, SIAM, Philadelphia, PA, 2009 Zbl 1422.68040 MR 2809335

[7] N. Alon, S. Hoory, and N. Linial, The Moore bound for irregular graphs. *Graphs Combin.* **18** (2002), no. 1, 53–57 Zbl 0990.05075 MR 1892433

[8] N. Alon, A. Pokrovskiy, and B. Sudakov, Random subgraphs of properly edge-coloured complete graphs and long rainbow cycles. *Israel J. Math.* **222** (2017), no. 1, 317–331 Zbl 1378.05180 MR 3736509

[9] O. Alrabiah and V. Guruswami, Near-tight bounds for 3-query locally correctable binary linear codes via rainbow cycles. In *Proceedings of the 2024 IEEE 65th annual symposium on foundations of computer science (FOCS 2024), October 27–30, 2024, Chicago, IL, USA*, pp. 1874–1882, IEEE, Los Alamitos, CA, 2024 MR 4849308

[10] O. Alrabiah, V. Guruswami, P. K. Kothari, and P. Manohar, A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation. In *Proceedings of the 55th annual ACM symposium on theory of computing (STOC 2023), June 20–23, 2023, Orlando, FL, USA*, pp. 1438–1448, ACM, New York, 2023 Zbl 07844683 MR 4617479

[11] M. Anastos and P. Morris, A note on finding large transversals efficiently. *J. Combin. Des.* **33** (2025), no. 9, 338–342 Zbl 1570.05142 MR 4933410

[12] L. D. Andersen, Hamilton circuits with many colours in properly edge-coloured complete graphs. *Math. Scand.* **64** (1989), no. 1, 5–14 Zbl 0665.05028 MR 1036426

[13] J. Balogh and T. Molla, Long rainbow cycles and Hamiltonian cycles using many colors in properly edge-colored complete graphs. *European J. Combin.* **79** (2019), 140–151 Zbl 1414.05174 MR 3922201

[14] M. Bateman and N. H. Katz, New bounds on cap sets. *J. Amer. Math. Soc.* **25** (2012), no. 2, 585–613 Zbl 1262.11010 MR 2869028

[15] F. Benzing, A. Pokrovskiy, and B. Sudakov, Long directed rainbow cycles and rainbow spanning trees. *European J. Combin.* **88** (2020), article no. 103102, 21 pp. Zbl 1443.05061 MR 4111714

[16] R. C. Bose, S. S. Shrikhande, and E. T. Parker, Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture. *Canad. J. Math.* **12** (1960), 189–203 Zbl 0093.31905 MR 0122729

[17] J. Böttcher, J. Hladký, D. Piguet, and A. Taraz, An approximate version of the tree packing conjecture. *Israel J. Math.* **211** (2016), no. 1, 391–446 Zbl 1334.05111 MR 3474969

[18] J. Bourgain, On triples in arithmetic progression. *Geom. Funct. Anal.* **9** (1999), no. 5, 968–984 Zbl 0959.11004 MR 1726234

[19] A. E. Brouwer, A. J. de Vries, and R. M. A. Wieringa, A lower bound for the length of partial transversals in a Latin square. *Nieuw Arch. Wiskd. (3)* **26** (1978), no. 2, 330–332 Zbl 0395.05012 MR 0480083

[20] R. A. Brualdi and H. J. Ryser, *Combinatorial matrix theory*. Encyclopedia Math. Appl. 39, Cambridge University Press, Cambridge, 1991, 367 pp. Zbl 0746.05002 MR 1130611

[21] D. Chakraborti, M. Christoph, Z. Hunter, R. Montgomery, and T. Petrov, Almost-full transversals in equi-n-squares. 2024, arXiv:2412.07733v1

[22] M.-C. Chang, A polynomial bound in Freiman's theorem. *Duke Math. J.* **113** (2002), no. 3, 399–419 Zbl 1035.11048 MR 1909605

[23] H. Chen and X. Li, Long rainbow path in properly edge-colored complete graphs. 2015, arXiv:1503.04516v1

[24] D. M. Correia, A. Pokrovskiy, and B. Sudakov, Short proofs of rainbow matchings results. *Int. Math. Res. Not. IMRN* **2023** (2023), no. 14, 12441–12476 Zbl 1519.05071 MR 4615235

[25] S. Das, C. Lee, and B. Sudakov, Rainbow Turán problem for even cycles. *European J. Combin.* **34** (2013), no. 5, 905–915 Zbl 1260.05079 MR 3021520

[26] J. H. Dinitz and D. R. Stinson (eds.), *Contemporary design theory. A collection of surveys*. Wiley-Intersci. Ser. Discrete Math. Optim., John Wiley & Sons, New York, 1992, 639 pp. Zbl 0746.00028 MR 1178497

[27] D. A. Drake, Maximal sets of Latin squares and partial transversals. *J. Statist. Plann. Inference* **1** (1977), no. 2, 143–149 Zbl 0392.05014 MR 0485434

[28] Z. Dvir, Incidence theorems and their applications. *Found. Trends Theor. Comput. Sci.* **6** (2012), no. 4, 257–393 Zbl 1278.68012 MR 3004132

[29] S. Eberhard, F. Manners, and R. Mrazović, An asymptotic for the Hall–Paige conjecture. *Adv. Math.* **404** (2022), no. part A, article no. 108423, 73 pp. Zbl 1494.05018 MR 4416136

[30] K. Efremenko, 3-query locally decodable codes of subexponential length. In *Proceedings of the 41st annual ACM symposium on theory of computing (STOC 2009), May 31–June 2, 2009, Bethesda, MD, USA*, pp. 39–44, ACM, New York, 2009 Zbl 1304.94124 MR 2780048

[31] P. Erdős, A. Gyárfás, and L. Pyber, Vertex coverings by monochromatic cycles and trees. *J. Combin. Theory Ser. B* **51** (1991), no. 1, 90–95 Zbl 0766.05062 MR 1088628

[32] P. Erdös and R. Rado, A combinatorial theorem. *J. Lond. Math. Soc. (2)* **25** (1950), no. 4, 249–255 Zbl 0038.15301 MR 0037886

[33] L. Euler, *Recherches sur un nouvelle espéce de quarrés magiques*, pp. 85–239, Verhandelingen uitgegeven door het zeeuwsch Genootschap der Wetenschappen te Vlissingen 9, Middelburg, 1782

[34] U. Feige, Small linear dependencies for binary vectors of low weight. In *Building bridges. Between mathematics and computer science*, pp. 283–307, Bolyai Soc. Math. Stud. 19, Springer, Berlin, 2008 Zbl 1157.05317 MR 2484644

[35] U. Feige, J. H. Kim, and E. Ofek, Witnesses for non-satisfiability of dense random 3CNF formulas. In *Proceedings of the 2006 47th annual IEEE symposium on foundations of computer science (FOCS 2006), October 21–24, 2006, Berkeley, California, USA*, pp. 497–508, IEEE, Los Alamitos, CA, 2006

[36] U. Feige and T. Wagner, Generalized girth problems in graphs and hypergraphs. 2016, preprint

[37] A. Ferber, C. Lee, and F. Mousset, Packing spanning graphs from separable families. *Israel J. Math.* **219** (2017), no. 2, 959–982 Zbl 1364.05057 MR 3649613

[38] A. Ferber and W. Samotij, Packing trees of unbounded degrees in random graphs. *J. Lond. Math. Soc. (2)* **99** (2019), no. 3, 653–677 Zbl 1415.05147 MR 3977885

[39] G. A. Freiman, On the addition of finite sets (in Russian). *Dokl. Akad. Nauk SSSR* **158** (1964), no. 5, 1038–1041. English translation: *Sov. Math. Dokl.* **5** (1964), 1366–1370 Zbl 0163.29501 MR 0168529

[40] J. A. Gallian, A dynamic survey of graph labeling. *Electron. J. Combin.* **18** (2011), no. DS6, 256 pp.

[41] H. Gebauer and F. Mousset, On rainbow cycles and paths. 2012, arXiv:1207.0840v1

[42] 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.* **284** (2023), no. 1406, 131 pp. Zbl 1515.05005 MR 4572074

[43] O. Goldreich, H. Karloff, L. J. Schulman, and L. Trevisan, Lower bounds for linear locally decodable codes and private information retrieval. *Comput. Complexity* **15** (2006), no. 3, 263–296 Zbl 1113.68049 MR 2268406

[44] R. L. Graham and N. J. A. Sloane, On additive bases and harmonious graphs. *SIAM J. Algebr. Discrete Methods* **1** (1980), no. 4, 382–404 Zbl 0499.05049 MR 0593849

[45] B. Green and I. Z. Ruzsa, Freiman's theorem in an arbitrary abelian group. *J. Lond. Math. Soc. (2)* **75** (2007), no. 1, 163–175 Zbl 1133.11058 MR 2302736

[46] H.-D. O. F. Gronau, R. C. Mullin, and A. Rosa, Orthogonal double covers of complete graphs by trees. *Graphs Combin.* **13** (1997), no. 3, 251–262 Zbl 0885.05093 MR 1469825

[47] V. Guruswami, P. K. Kothari, and P. Manohar, Algorithms and certificates for Boolean CSP refutation: smoothed is no harder than random. In *Proceedings of the 54th annual ACM SIGACT symposium on theory of computing (STOC 2022), June 20–24, 2022, Rome, Italy*, pp. 678–689, ACM, New York, 2022 Zbl 07774370 MR 4490032

[48] A. Gyárfás and M. Mhalla, Rainbow and orthogonal paths in factorizations of K_n. *J. Combin. Des.* **18** (2010), no. 3, 167–176 Zbl 1216.05112 MR 2656391

[49] A. Gyárfás, M. Ruszinkó, G. N. Sárközy, and R. H. Schelp, Long rainbow cycles in proper edge-colorings of complete graphs. *Australas. J. Combin.* **50** (2011), 45–53 Zbl 1236.05082 MR 2829278

[50] G. Hahn, Un jeu de coloration. In *Regards sur la theorie des graphes*. Actes Colloq. Cerisy 12, Presses Polytechniques Romandes, Lausanne, 1980, 18 pp.

[51] M. Hall and L. J. Paige, Complete mappings of finite groups. *Pacific J. Math.* **5** (1955), no. 4, 541–549 Zbl 0066.27703 MR 0079589

[52] P. Hatami and P. W. Shor, A lower bound for the length of a partial transversal in a Latin square. *J. Combin. Theory Ser. A* **115** (2008), no. 7, 1103–1113 Zbl 1159.05303 MR 2450332

[53] S. Hoory, N. Linial, and A. Wigderson, Expander graphs and their applications. *Bull. Amer. Math. Soc. (N.S.)* **43** (2006), no. 4, 439–561 Zbl 1147.68608 MR 2247919

[54] J.-T. Hsieh, P. K. Kothari, and S. Mohanty, A simple and sharper proof of the hypergraph Moore bound. In *Proceedings of the 2023 annual ACM-SIAM symposium on discrete algorithms (SODA 2023), January 22–25, 2023, Florence, Italy*, pp. 2324–2344, SIAM, Philadelphia, PA, 2023 Zbl 07848049 MR 4538227

[55] J.-T. Hsieh, P. K. Kothari, S. Mohanty, D. Munhá Correia, and B. Sudakov, Small even covers, locally decodable codes and restricted subgraphs of edge-colored Kikuchi graphs. *Int. Math. Res. Not. IMRN* **2025** (2025), no. 5, article no. rnaf045, 18 pp. Zbl 08103140 MR 4875645

[56] D. Hughes, Biplanes and semi-biplanes. In D. A. Holton and J. Seberry (eds.), *Combinatorial mathematics (proc. internat. conf. combinatorial theory, Australian Nat. Univ., Canberra, 1977)*, pp. 55–58, Lecture Notes in Math. 686, Springer, Berlin, 1978 Zbl 0418.05010 MR 0526729

[57] O. Janzer, Rainbow Turán number of even cycles, repeated patterns and blow-ups of cycles. *Israel J. Math.* **253** (2023), no. 2, 813–840 Zbl 1512.05196 MR 4575428

[58] O. Janzer and B. Sudakov, On the Turán number of the hypercube. *Forum Math. Sigma* **12** (2024), article no. e38, 19 pp. Zbl 1535.05143 MR 4717647

[59] F. Joos, J. Kim, D. Kühn, and D. Osthus, Optimal packings of bounded degree trees. *J. Eur. Math. Soc. (JEMS)* **21** (2019), no. 12, 3573–3647 Zbl 1429.05164 MR 4022711

[60] P. Keevash, The existence of designs. [v1] 2014, [v4] 2024, arXiv:1401.3665v4

[61] P. Keevash, D. Mubayi, B. Sudakov, and J. Verstraëte, Rainbow Turán problems. *Combin. Probab. Comput.* **16** (2007), no. 1, 109–126 Zbl 1119.05058 MR 2286514

[62] 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), no. 8, 288–321 Zbl 1490.05269 MR 4408405

[63] P. Keevash and K. Staden, Ringel's tree packing conjecture in quasirandom graphs. *J. Eur. Math. Soc. (JEMS)* **27** (2025), no. 5, 1769–1826 Zbl 1562.05308 MR 4889236

[64] I. Kerenidis and R. de Wolf, Exponential lower bound for 2-query locally decodable codes via a quantum argument. *J. Comput. System Sci.* **69** (2004), no. 3, 395–420 Zbl 1084.68044 MR 2087942

[65] J. Kim, D. Kühn, D. Osthus, and M. Tyomkyn, A blow-up lemma for approximate decompositions. *Trans. Amer. Math. Soc.* **371** (2019), no. 7, 4655–4742 Zbl 1409.05114 MR 3934464

[66] J. Kim, J. Lee, H. Liu, and T. Tran, Rainbow cycles in properly edge-colored graphs. *Combinatorica* **44** (2024), no. 4, 909–919 Zbl 07915723 MR 4777873

[67] K. K. Koksma, A lower bound for the order of a partial transversal in a Latin square. *J. Combin. Theory* **7** (1969), no. 1, 94–95 Zbl 0172.01504 MR 0239988

[68] S. V. Konyagin and I. D. Shkredov, A quantitative version of the Beurling–Helson theorem (in Russian). *Funktsional. Anal. i Prilozhen.* **49** (2015), no. 2, 39–53. English translation: *Funct. Anal. Appl.* **49** (2015), no. 2, 110–121 Zbl 1326.42005 MR 3374902

[69] S. V. Konyagin and I. D. Shkredov, On subgraphs of random Cayley sum graphs. *European J. Combin.* **70** (2018), 61–74 Zbl 1384.05141 MR 3779604

[70] M. Krivelevich, Triangle factors in random graphs. *Combin. Probab. Comput.* **6** (1997), no. 3, 337–347 Zbl 0886.05101 MR 1464569

[71] F. Lazebnik, V. A. Ustimenko, and A. J. Woldar, A new series of dense graphs of high girth. *Bull. Amer. Math. Soc. (N.S.)* **32** (1995), no. 1, 73–79 Zbl 0822.05039 MR 1284775

[72] U. Leck and V. Leck, On orthogonal double covers by trees. *J. Combin. Des.* **5** (1997), no. 6, 433–441 Zbl 0914.05014 MR 1472880

[73] V. F. Lev and R. Yuster, On the size of dissociated bases. *Electron. J. Combin.* **18** (2011), no. 1, article no. P117, 5 pp. Zbl 1217.05046 MR 2811086

[74] M. Maamoun and H. Meyniel, On a problem of G. Hahn about coloured Hamiltonian paths in K_2t. *Discrete Math.* **51** (1984), no. 2, 213–214 Zbl 0549.05044 MR 0758880

[75] S. Messuti, V. Rödl, and M. Schacht, Packing minor-closed families of graphs into complete graphs. *J. Combin. Theory Ser. B* **119** (2016), 245–265 Zbl 1334.05121 MR 3486345

[76] R. Montgomery, Spanning trees in random graphs. *Adv. Math.* **356** (2019), article no. 106793, 92 pp. Zbl 1421.05080 MR 3998769

[77] R. Montgomery, A proof of the Ryser–Brualdi–Stein conjecture for large even n. 2023, arXiv:2310.19779v1

[78] R. Montgomery, Transversals in Latin squares. In F. Fischer and R. Johnson (eds.), *Surveys in combinatorics 2024*, pp. 131–158, London Math. Soc. Lecture Note Ser. 493, Cambridge University Press, Cambridge, 2024 Zbl 08100150 MR 4749040

[79] R. Montgomery, A. Pokrovskiy, and B. Sudakov, Decompositions into spanning rainbow structures. *Proc. Lond. Math. Soc. (3)* **119** (2019), no. 4, 899–959 Zbl 1429.05024 MR 3964824

[80] R. Montgomery, A. Pokrovskiy, and B. Sudakov, Embedding rainbow trees with applications to graph labelling and decomposition. *J. Eur. Math. Soc. (JEMS)* **22** (2020), no. 10, 3101–3132 Zbl 1508.05024 MR 4153104

[81] R. Montgomery, A. Pokrovskiy, and B. Sudakov, A proof of Ringel's conjecture. *Geom. Funct. Anal.* **31** (2021), no. 3, 663–720 Zbl 1473.05249 MR 4311581

[82] A. Müyesser and A. Pokrovskiy, A random Hall–Paige conjecture. *Invent. Math.* **240** (2025), no. 3, 779–867 Zbl 1565.20065 MR 4902157

[83] A. Naor and J. Verstraëte, Parity check matrices and product representations of squares. *Combinatorica* **28** (2008), no. 2, 163–185 Zbl 1164.05005 MR 2399017

[84] A. Pokrovskiy, An approximate version of a conjecture of Aharoni and Berger. *Adv. Math.* **333** (2018), 1197–1241 Zbl 1390.05236 MR 3818098

[85] A. Pokrovskiy, Rainbow subgraphs and their applications. In A. Nixon and S. Prendiville (eds.), *Surveys in combinatorics 2022*, pp. 191–214, London Math. Soc. Lecture Note Ser. 481, Cambridge University Press, Cambridge, 2022 Zbl 1529.05072 MR 4421403

[86] A. Pokrovskiy and B. Sudakov, A counterexample to Stein's equi-n-square conjecture. *Proc. Amer. Math. Soc.* **147** (2019), no. 6, 2281–2287 Zbl 1411.05041 MR 3951411

[87] G. Ringel, *Theory of graphs and its applications (Proc. sympos. Smolenice, 1963)*. Publishing House of the Czechoslovak Academy of Sciences, Prague, Czechoslovakia, 1963

[88] V. Rödl, On a packing and covering problem. *European J. Combin.* **6** (1985), no. 1, 69–78 Zbl 0565.05016 MR 0793489

[89] V. Rödl, A. Ruciński, and E. Szemerédi, A Dirac-type theorem for 3-uniform hypergraphs. *Combin. Probab. Comput.* **15** (2006), no. 1–2, 229–251 Zbl 1082.05057 MR 2195584

[90] A. Rosa, On certain valuations of the vertices of a graph. In *Theory of graphs (international symposium, Rome, 1966)*, pp. 349–355, Gordon & Breach, New York, 1967 Zbl 0193.53204 MR 0223271

[91] W. Rudin, *Fourier analysis on groups*. Wiley Classics Lib., John Wiley & Sons, New York, 1990, 285 pp. Zbl 0698.43001 MR 1038803

[92] H. J. Ryser, Neuere Probleme der Kombinatorik. In *Vorträge über Kombinatorik*, pp. 69–91, Mathematisches Forschungsinstitut Oberwolfach, Germany, 1967

[93] T. Sanders, On a theorem of Shkredov. *Online J. Anal. Comb.* (2010), no. 5, article no. 4, 4 pp. Zbl 1261.11014 MR 2789255

[94] T. Sanders, On Roth's theorem on progressions. *Ann. of Math. (2)* **174** (2011), no. 1, 619–636 Zbl 1264.11004 MR 2811612

[95] T. Sanders, The structure theory of set addition revisited. *Bull. Amer. Math. Soc. (N.S.)* **50** (2013), no. 1, 93–127 Zbl 1337.11014 MR 2994996

[96] T. Schoen and I. D. Shkredov, Additive dimension and a theorem of Sanders. *J. Aust. Math. Soc.* **100** (2016), no. 1, 124–144 Zbl 1364.11022 MR 3436775

[97] P. W. Shor, A lower bound for the length of a partial transversal in a Latin square. *J. Combin. Theory Ser. A* **33** (1982), no. 1, 1–8 Zbl 0489.05012 MR 0665651

[98] S. K. Stein, Transversals of Latin squares and their generalizations. *Pacific J. Math.* **59** (1975), no. 2, 567–575 Zbl 0302.05015 MR 0387083

[99] T. Tao, Product set estimates for non-commutative groups. *Combinatorica* **28** (2008), no. 5, 547–594 Zbl 1254.11017 MR 2501249

[100] G. Tarry, Le problème des 36 officiers. *C. R. Assoc. Française Avancement Sci.* **29** (1900), no. 2, 170–203 JFM 32.0219.04

[101] I. Tomon, Robust (rainbow) subdivisions and simplicial cycles. *Adv. Comb.* **2024** (2024), article no. 1, 37 pp. Zbl 1562.05404 MR 4695962

[102] I. M. Wanless, Transversals in Latin squares: a survey. In R. Chapman (ed.), *Surveys in combinatorics 2011*, pp. 403–437, London Math. Soc. Lecture Note Ser. 392, Cambridge University Press, Cambridge, 2011 Zbl 1226.05067 MR 2866738

[103] S. Wilcox, Reduction of the Hall–Paige conjecture to sporadic simple groups. *J. Algebra* **321** (2009), no. 5, 1407–1428 Zbl 1172.20024 MR 2494397

[104] R. M. Wilson, Decompositions of complete graphs into subgraphs isomorphic to a given graph. In *Proceedings of the 5th British combinatorial conference (Univ. Aberdeen, 1975)*, pp. 647–659, Congr. Numer., No. XV, Util. Math., Winnipeg, MB, 1976 Zbl 0322.05116 MR 0396347

[105] D. E. Woolbright, An $n \times n$ Latin square has a transversal with at least $n - \sqrt{n}$ distinct symbols. *J. Combin. Theory Ser. A* **24** (1978), no. 2, 235–237 Zbl 0368.05012 MR 0472562

[106] M. Woźniak, Packing of graphs and permutations—a survey. *Discrete Math.* **276** (2004), no. 1–3, 379–391 Zbl 1031.05041 MR 2046650

[107] H. P. Yap, Packing of graphs-a survey. *Ann. Discrete Math.* **38** (1988), 395–404 Zbl 0685.05036 MR 0975562

[108] S. Yekhanin, Towards 3-query locally decodable codes of subexponential length. *J. ACM* **55** (2008), no. 1, article no. 1, 16 pp. Zbl 1311.94125 MR 2388223

[109] S. Yekhanin, Locally decodable codes and private information retrieval schemes. Inf. Secur. Cryptography, Springer, Berlin, 2010, 82 pp. Zbl 1211.68181

[110] S. Yekhanin, Locally decodable codes. *Found. Trends Theor. Comput. Sci.* **6** (2012), no. 3, 139–255 Zbl 1278.94002 MR 2996521

[111] A. Żak, Harmonious order of graphs. *Discrete Math.* **309** (2009), no. 20, 6055–6064 Zbl 1188.05138 MR 2552639

Benny Sudakov
Department of Mathematics, ETH Zürich, Zürich, Switzerland;
benjamin.sudakov@math.ethz.ch

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/4

Large-scale behavior of out-of-equilibrium systems at and above the critical dimension

Giuseppe Cannizzaro and Fabio Toninelli

Abstract. The mathematical understanding of out-of-equilibrium, statistical physics systems at and above their critical space dimension has witnessed spectacular progress in the last few years. These systems include, for instance, driven diffusive systems (such as stochastic interacting particle systems with asymmetric jump rates) in dimension $d \geq 2$, stochastic $(d+1)$-dimensional growth models with $d \geq 2$, non-linear singular (super-)critical stochastic partial differential equations (SPDEs) and (self-)interacting diffusions. We give here a review of recent results on their large-scale (i.e., infrared) behavior, focusing on singular SPDEs at and above the so-called critical dimension (in particular, the stochastic Burgers equation (SBE) in space dimension $d \geq 2$ and the anisotropic KPZ (AKPZ) equation in $d = 2$) as well as on critical diffusions in random divergence-free vector field and self-repelling diffusions. A highlight of the main results includes sharp bounds on the large-time logarithmic divergence of the diffusion coefficient at the critical dimension, and Gaussian infrared scaling limits, at both the critical and super-critical dimensions.

Most of the results presented here have been obtained in collaboration with Dirk Erhard, Harry Giles, Massimiliano Gubinelli, Levi Haunschmid-Sibitz and Quentin Moulard, in different combinations.

1. Introduction

Out-of-equilibrium physical systems (driven diffusive systems, stochastic growth, etc.) are usually mathematically described by Markov processes, either of discrete type (interacting particle systems on a lattice, discrete deposition processes, ...) or of continuum nature, such as stochastic partial differential equations (SPDEs) or stochastic differential equations (SDEs). In particular, the relevant SPDEs (and SDEs) in this context turn out to be *non-linear and singular*, in that the expression derived in the physics literature is ill posed and some form of renormalization is needed in

Mathematics Subject Classification 2020: 60H15 (primary); 60K50 (secondary).
Keywords: stochastic partial differential equations, critical dimension, stochastic Burgers equation, anisotropic KPZ equation, superdiffusion, scaling limits.

order to say something meaningful about them. Thanks to the groundbreaking theory of regularity structures [19], paracontrolled calculus [18] or renormalization group techniques [21], the problems of *renormalization* as well as *small-scale existence, uniqueness and regularity* have been settled for virtually any SPDE in the so-called subcritical (in the sense of scaling, as explained below) dimensions (for KPZ, such dimension is $d = 1$). In contrast, in the critical and super-critical dimensions, a local solution theory is not even expected, and the most compelling question concerning their large-scale behavior is, from a mathematical viewpoint, much less explored and still poorly understood.

This is precisely what we have been studying in a series of works in collaboration with Dirk Erhard, Massimiliano Gubinelli and Quentin Moulard (in various combinations). For a class of singular (super-)critical SPDEs, we were able to (1) obtain the large-time asymptotics for the diffusion coefficient, (2) precisely identify the logarithmic corrections to scaling at the critical dimension which were conjectured in the physics literature and (3) derive their large-scale (infrared) behavior under the correct scaling. As detailed below, the class of equations to which our approach applies includes the two-dimensional anisotropic KPZ equation [5,34], the stochastic Burgers equation in dimension $d \geq 2$ [32] and the incompressible stochastic Navier–Stokes equation with divergence-free noise in dimension $d = 2$. Moreover, the methods we established have proven versatile enough to apply to critical, singular SDEs, such as the diffusion in the curl of the $2d$-Gaussian free field (a toy model for a particle in a turbulent fluid), as well as the self-repelling Brownian polymer [31]. In collaboration with Levi Haunschmid-Sibitz [11] and Harry Giles [8], respectively, substantial progress has been made in the analysis of these models, and our results have later inspired several groups of other researchers (see below for references).

The aim of this short note is to give an account of these recent developments. We keep definitions and formulations of theorems somewhat informal, and refer to the original works for precise statements and details. For the methodology and the proofs, we refer the reader to the lecture notes [13], where the mathematical ideas and key points of the arguments are explained.

Organization. In Section 2, we introduce the main mathematical models and the questions we are interested in, while in the later sections, we give a panorama of the recent mathematical results. We start the presentation of new results with the most interesting case of the critical dimension $d = 2$: large-time behavior of the diffusion matrix (Section 3), scaling limit in the weak coupling regime (Section 3.2), full scaling limit in the strong coupling regime for the stochastic Burgers equation (Section 3.3). Section 4 is instead dedicated to scaling limits of the stochastic Burgers equation in the super-critical dimensions $d \geq 3$. Finally, Section 5 discusses perspectives and open problems.

2. The models

In this section, we present the main mathematical models. Each of them corresponds to a different out-of-equilibrium irreversible physical phenomenon.

2.1. Diffusive systems with one conserved quantity: Asymmetric simple exclusion process and stochastic Burgers equation

A typical physical example of driven diffusive system is a gas or fluid of interacting particles, subject to an external field (e.g., an electric field) that induces a stationary current (e.g., an electric current).

There are several ways of modeling such system mathematically. A classical one is in terms of stochastically interacting particles on a lattice. Each vertex of $\mathbb{Z}^d$ can be either empty or occupied by a single particle. Each particle, with rate 1, attempts a jump: if the present particle position is x, it jumps to y with probability $p(y-x)$, where $z \mapsto p(z)$ is a probability measure on $\mathbb{Z}^d$; if the vertex y is already occupied, then the jump is suppressed. The stochastic process just described is a continuous-time Markov process that goes under the name of exclusion process [25], which is further dubbed "Simple" if $p(x-y)$ is non-zero only when $x-y$ has length 1 and "Symmetric" if p is even symmetric ($p(x) = p(-x)$) so that, in particular, $\mathfrak{v} := \sum_{x\in\mathbb{Z}^d} xp(x) = 0$. Here, we are interested instead in the case in which $\mathfrak{v} \neq 0$, that is, in the *asymmetric simple* exclusion process (ASEP). The vector $\mathfrak{v}$, which we write as $\mathfrak{v} = \lambda\mathfrak{w}$ with $\mathfrak{w}$ a vector of norm 1 and $\lambda = |\mathfrak{v}| > 0$, quantifies the asymmetry of p and mimics the external driving field. We note that the particle number is a locally conserved quantity (particles jump but are neither created nor annihilated) and, in general, it is the only one.

A more mesoscopic, continuum modelization of driven diffusive systems with a single conserved quantity was introduced in [32]. If for the exclusion process a state of the system consists in a function $\eta : \mathbb{Z}^d \to \{0,1\}$ with $\eta(x)$ assessing whether vertex x is occupied by a particle or not, in the continuum model the particle density profile is described by a real-valued function u defined on $\mathbb{R}^d$ instead of $\mathbb{Z}^d$. More precisely, $u(t,x)$ represents the fluctuation at point x and time t of the locally conserved quantity (say, the mass) with respect to its average, and therefore it can be either positive or negative. The function u solves the so-called stochastic Burgers equation (SBE), that is, the stochastic partial differential equation (SPDE)

$$\partial_t u = \frac{1}{2}\Delta u + \lambda(\mathfrak{w}\cdot\nabla)u^2 + \nabla\cdot\xi, \tag{2.1}$$

where

- $u = u(t,x)$ for $t \in \mathbb{R}^+$ and $x \in \mathbb{R}^d$ (alternatively, one can define the equation on the d-dimensional torus),

- $\xi = (\xi_1, \dots, \xi_d)$ is a collection of d independent space-time white noises, that is, ξ is a centered vector-valued Gaussian field indexed by space and time, with covariance formally given by $\mathbb{E}(\xi_i(t,x)\xi_j(s,y)) = \mathbf{1}_{i=j}\delta(t-s)\delta(x-y)$, $s,t \in \mathbb{R}$, $x, y \in \mathbb{R}^d$, $1 \le i, j \le d$,
- $\mathfrak{w}$ is a unit vector in $\mathbb{R}^d$ that represents the direction of the external driving field and $\lambda > 0$ is its intensity. By rotation invariance, the specific choice of $\mathfrak{w}$ is irrelevant.

Note that u is locally conserved, as is the case for the particle number of ASEP. Local conservation means that the time derivative of the integral of u on a domain $V \subset \mathbb{R}^d$ equals the flux of a (stochastic) current j across ∂V. In formulas, this is equivalent to saying that all terms in the right-hand side of (2.1) are in divergence form: in fact, the equation can be equivalently written as

$$\partial_t u = \nabla \cdot j, \quad j = \frac{1}{2}\nabla u + \lambda \mathfrak{w} u^2 + \xi, \tag{2.2}$$

and j plays the role of the current.

Apart from local conservation, the SBE retains also other characterizing mathematical features of the interacting particle system: the Laplacian is a diffusive term that corresponds to the symmetric part of the jump kernel p, the non-linear term proportional to λ mimics the asymmetric part of the jump kernel and the noise is taken to be white both in space, because randomness in a realistic particle system is expected to have short-range correlations, and in time, as we expect fluctuations to be Markovian. The constant λ plays the role of a "coupling constant". When $\lambda = 0$, the equation is linear and, assuming that the initial condition $u(t = 0)$ is Gaussian, the whole space-time process is Gaussian and therefore uniquely determined by its covariance (as it is easy to see that u has expectation zero).

Let us point out two other important properties of (2.1):

- The law $\mathbb{P}$ of the spatial white noise is formally stationary. That is, if the initial condition $u(t = 0)$ is the centered Gaussian field η with covariance $\mathbb{E}(\eta(x)\eta(y)) = \delta(x-y)$, then $u(t)$ has the same law $\mathbb{P}$ at every given time $t > 0$. This is the continuum analog of the well-known fact that, for ASEP, the product iid Bernoulli measure $\bigotimes_{x\in\mathbb{Z}^d} \mathrm{Ber}(\rho)$ is stationary for any choice of $\rho \in [0,1]$. In fact, the continuum limit (in an appropriate sense, and after centering) of $\bigotimes_{x\in\mathbb{Z}^d} \mathrm{Ber}(\rho)$ is given by η.
- Under diffusive scaling, the equation maintains the same form, but for the coefficient of the non-linearity. That is, let $\varepsilon > 0$ and define the diffusively rescaled solution

$$u^\varepsilon(t,x) := \varepsilon^{-d/2} u\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right). \tag{2.3}$$

Then, u^ε formally satisfies

$$\partial_t u^\varepsilon = \frac{1}{2}\Delta u^\varepsilon + \lambda \varepsilon^{(d-2)/2}(\mathfrak{w}\cdot\nabla)(u^\varepsilon)^2 + \nabla\cdot\xi^\varepsilon, \tag{2.4}$$

where the noise ξ^ε has the same law as ξ.

We will see later that this scaling argument is misleading in dimension $d = 2$, since it suggests that the $2d$-SBE is diffusively scale invariant, which turns out to be false. However, an important piece of information that we gain from (2.4) is that dimension $d = 2$ should play a special (critical) role: in fact, looking at large scales, which corresponds to $\varepsilon \to 0$, the prefactor of the non-linearity diverges for $d < 2$ (subcritical) and vanishes for $d > 2$ (super-critical).

The SPDE (2.1) cannot be taken literally, since it is analytically ill posed. Indeed, regularity-wise, the noise ξ is extremely rough and thus u is at best a generalized function, in which case it is unclear what meaning should be attributed to the non-linear term u^2. As we are interested in large scales, it is reasonable to regularize the equation thus obtaining a well-defined microscopic object. A classical way to proceed is to smooth the noise in space, for instance, by convolving it (with respect to the space variable) with a C^∞, compactly supported function. Another option is to leave the noise unchanged and regularize the non-linearity. It turns out to be convenient to do both at the same time: that is, one replaces

$$u^2 \to \rho^{*2} * u^2 = \rho * \rho * u^2 \tag{2.5}$$

and

$$\xi \to \rho * \xi =: \xi^1, \tag{2.6}$$

where $*$ denotes convolution in space and ρ is a rotationally invariant, compactly supported C^∞ function with total mass equal to 1. The equation then becomes

$$\partial_t u = \frac{1}{2}\Delta u + \lambda \mathcal{N}^1(u) + \nabla\cdot\xi^1 \tag{2.7}$$

with

$$\mathcal{N}^1(u) := (\mathfrak{w}\cdot\nabla)(\rho^{*2} * u^2). \tag{2.8}$$

Since the noise is now smooth in space, the equation admits a unique classical global-in-time solution [12], which, as we left the noise white in time, is still Markovian. The reason for this specific choice of regularization is that, as one can explicitly check, the law of the space white noise convolved with ρ is stationary. That is, if $\mathbb{P}^\rho$ denotes the law of $\rho * \eta$ with η a space white noise, then $\mathbb{P}^\rho$ is a stationary probability measure for the Markov process u.

Both for ASEP and for the SBE in (2.7), our goal is to identify their behavior on large space-time scales and, more precisely, to see if the process has a non-trivial scaling limit when space and time are suitably rescaled so that one looks at $t, |x| \to \infty$, which, as we will see, also corresponds to removing the regularization ρ.

To do so, we first need to understand what is the correct rate at which we should "zoom out", that is, what is the right scaling under which the universal space-time fluctuations become visible. A convenient observable in this respect is the so-called diffusion matrix. Let S be the two-point correlation function, that is, the map on $\mathbb{R}^+ \times \mathbb{R}^2$ given by

$$S(t,x) := \boldsymbol{E}\left(u(t,x)u(0,0)\right), \tag{2.9}$$

where $\boldsymbol{E}$ denotes the expectation with respect to the noise ξ^1 and to the initial condition $u(0)$, sampled from the stationary measure $\mathbb{P}^\rho$. Then, the so-called "diffusion matrix" $D(t)$ is defined as the $d \times d$ symmetric matrix

$$D_{ij}(t) = \frac{1}{t}\int_{\mathbb{R}^d} x_i x_j \left(S(t,x) - S(0,x)\right) \mathrm{d}x, \tag{2.10}$$

where the integral at the right-hand side converges since, as proven in [12] for $d = 2$ (but the proof works in any dimension), for every fixed time t, the function $x \mapsto S(t,x)$ decays faster than any inverse power for $|x| \to \infty$.

Intuitively, if $e \in \mathbb{R}^d$ is a unit vector, one can interpret $\langle e, D(t)e\rangle$ as the ratio $(L_e(t))^2/t$, with $L_e(t)$ the time-dependent correlation length in the direction e at time t, defined as the distance beyond which the correlations in direction e are negligible.

For the linear equation with $\lambda = 0$, it is immediately checked that

$$S(t,x)|_{\lambda=0} = (\rho^{*2} * \mathcal{P}_t)(x), \tag{2.11}$$

with $\mathcal{P}_t$ the heat semigroup, and therefore $D(t)|_{\lambda=0}$ is the $d \times d$ identity matrix. As a consequence, if for large t the diffusion matrix of the SBE (with $\lambda \neq 0$) has $O(1)$ entries, then its large-scale behavior is said to be diffusive; if it diverges with t, it is superdiffusive; and if it tends to zero, subdiffusive. The formal scaling argument (2.4) seems to suggest that the SBE is superdiffusive in dimension $d = 1$, diffusive in dimension $d > 2$ and possibly diffusive also in dimension $d = 2$. We will see that, while the first two predictions are correct, the third is not.

2.2. Stochastic (2 + 1)-dimensional interface growth: The KPZ and AKPZ equations

Another intrinsically out-of-equilibrium physical phenomenon is *stochastic interface growth* [2], that occurs in many situations in nature, from crystal growth to the evolution of fire fronts. Note that, if the ambient physical space has dimension D, the

interface between two regions is a $d = (D-1)$-dimensional surface. We usually speak of $(d+1)$-dimensional growth models, where d is the geometric dimension of the interface and "+1" refers to the time variable. From the physical point of view, the most interesting cases are $d = 1$ (one-dimensional interfaces in two-dimensional physical space) or $d = 2$, but from the mathematical perspective, any dimension $d \geq 1$ has its own relevance.

As it is the case for driven diffusive systems, stochastic interface growth is usually modeled via Markov processes, either discrete (e.g., the corner growth model, ballistic deposition, polynuclear growth, ...) or continuous. Here, we focus only on the latter, whose canonical example is an SPDE, the celebrated Kardar–Parisi–Zhang (KPZ) equation [20]. In its general form, it can be (formally) written as

$$\partial_t h = \frac{1}{2}\Delta h + \lambda \langle \nabla h, Q \nabla h \rangle + \xi \tag{2.12}$$

with $h = h(t,x), t \geq 0, x \in \mathbb{R}^d$, ξ is a space-time white noise, λ is a real parameter (coupling constant) and Q a $d \times d$ symmetric matrix. The Laplacian physically corresponds to a "surface tension" (diffusive) term, that tends to flatten out deviations of the height function from a locally linear profile. The noise is the driver of the dynamics and has to be taken (locally) uncorrelated in space and time. The quadratic non-linearity corresponds to the so-called "lateral growth effect": in any non-trivial model, the speed of growth is a non-linear function of the local slope. Phenomenologically, the matrix λQ provides the connection between the KPZ equation and any microscopic stochastic growth model: indeed, to establish the link, one should choose Q to be proportional to $D^2 v$, the Hessian of the interface growth velocity v of the microscopic model, seen as a function of the interface slope [30]. In view of this link, while in the original KPZ equation introduced in [20] Q is the identity matrix, Q has in general no reason to be diagonal or positive definite. In fact, there are discrete stochastic $(2+1)$-dimensional growth models both with $\det(D^2 v) > 0$ and with $\det(D^2 v) < 0$, such as the Gates–Westcott model [28] and certain integrable, interlaced particle systems [3] (see [30] for more references).

It is well known that in dimension $d = 1$ the KPZ equation is equivalent to the $1d$ SBE via the relation $\partial_x h = u$, but such correspondence fails as soon as $d \geq 2$. However, a formal scaling argument very similar to that given above for the SBE suggests that, for the KPZ equation too, at large scales (i.e., under diffusive scaling) the non-linearity grows for $d < 2$, while it vanishes for $d > 2$. On the other hand, for $\lambda = 0$, the equation is exactly scale invariant in any dimension, and it is known as the "stochastic heat equation with additive noise" or "Edwards–Wilkinson" (EW) equation in the physics literature.

Specializing to the critical dimension $d = 2$, physicists soon realized that, as the scaling argument fails, finer features of the equation, and in particular, the signature

of Q, play an important role in the large-scale behavior of the solution to (2.12). Indeed, heuristic renormalization group arguments [34] suggest that, if $\det(Q) \leq 0$, then the non-linearity should not modify the critical exponents of the equation with respect to EW, while if $\det(Q) > 0$, the large-scale fixed point should be qualitatively different from the latter. In this sense, in dimension $d = 2$ the KPZ equation(s) (2.12) can be divided into two universality classes, according to the sign of the determinant of Q.

An especially interesting case of equation (2.12) is the so-called anisotropic KPZ (or AKPZ) equation, corresponding to the case of Q diagonal with diagonal elements ± 1 (i.e., $\det(Q) = -1 < 0$), which is

$$\partial_t h = \frac{1}{2}\Delta h + \lambda((\partial_1 h)^2 - (\partial_2 h)^2) + \xi. \tag{2.13}$$

A special feature of the AKPZ equation is that, formally, the law of the two-dimensional Gaussian free field ($2d$-GFF) is a stationary measure for its evolution [5]. We recall here that the $2d$-GFF is a centered Gaussian distribution on $\mathbb{R}^2$, with covariance function $\mathbb{E}(\eta(x)\eta(y)) \propto -\log(|x-y|)$. Equivalently, defining $U(t,x) = (-\Delta)^{1/2}h(t,x)$, then the stationary measure for U is a space white noise. In fact, the AKPZ equation is, among the equations of the form (2.12) in dimension $d = 2$, the only one (up to rotations of the space coordinates) that admits a Gaussian stationary measure [16].

As is the case for the SBE (and for the same reasons), also the (A)KPZ equation is highly singular, and a regularization is needed. The choice adopted in [5–7], is to leave the noise ξ unchanged and regularize the non-linearity with a Fourier cutoff. That is, one replaces

$$((\partial_1 h)^2 - (\partial_2 h)^2) \to \mathcal{M}^1(h) := \Pi_1((\partial_1 \Pi_1 h)^2 - (\partial_2 \Pi_1 h)^2), \tag{2.14}$$

where Π_a is the orthogonal projection on the Fourier components with wave vector k satisfying $|k| \leq a$. The regularized AKPZ equation is then

$$\partial_t h = \frac{1}{2}\Delta h + \lambda \mathcal{M}^1(h) + \xi. \tag{2.15}$$

Looking at the equation on large diffusive scales is equivalent to removing the cutoff. In fact, $h^\varepsilon(t,x) := h(t/\varepsilon^2, x/\varepsilon)$ solves in law the SPDE

$$\partial_t h^\varepsilon = \frac{1}{2}\Delta h^\varepsilon + \lambda \mathcal{M}^\varepsilon(h^\varepsilon) + \xi, \tag{2.16}$$

with $\mathcal{M}^\varepsilon$ defined as in (2.14), but with Π_1 replaced by $\Pi_{1/\varepsilon}$, that formally tends to the identity operator as $\varepsilon \to 0$. It is a simple but important remark made in [5] that, despite the regularization, the law of the GFF is stationary for (2.15).

As we are interested in the large-scale properties of the regularized AKPZ equation (2.15), we will use the "bulk diffusion coefficient" which is the counterpart of the diffusion matrix (2.10) for SBE. (Since AKPZ has a rotation invariance property, a single coefficient captures as much information as the full matrix.) The relevant definition adopted in [6] is analogous to it,

$$D_{\mathrm{AKPZ}}(t) = \frac{1}{t}\int_{\mathbb{R}^2} |x|^2 S(t,x)\,\mathrm{dx}, \tag{2.17}$$

with $S(t,x)$ given by

$$S(t,x) = \boldsymbol{E}\left((-\Delta)^{1/2}h(t,x)(-\Delta)^{1/2}h(0,0)\right), \tag{2.18}$$

where $\boldsymbol{E}$ denotes expectation with respect to the law of the stationary process. The reason why, in contrast to (2.10), we do not subtract $S(0,x)$ is that this equals $\delta_0(x)$, that would give a vanishing contribution.

For the linear equation where $\lambda = 0$, it is immediately checked that $D_{\mathrm{AKPZ}}(t)$ does not depend on t.

2.3. (Self-)interacting critical diffusions: The Brownian particle in the curl of the GFF and the self-repelling Brownian polymer

The so-called "Brownian particle (or diffusion) in the curl of the Gaussian free field", in short DCGFF, was introduced in the theoretical physics literature as a toy model for the motion of a tracer particle in a turbulent two-dimensional fluid. The evolution of the particle position X_t at time t is given by the stochastic differential equation

$$\mathrm{dX}_t = \omega(X_t)\,\mathrm{dt} + \mathrm{dB}_t, \quad X_0 = x, \tag{2.19}$$

where B_t is a standard two-dimensional Brownian motion, while the drift $\omega : \mathbb{R}^2 \to \mathbb{R}^2$ is a position-dependent *random* vector field. Specifically, ω is taken to be the curl of a two-dimensional GFF η independent of B, that is, $\omega(x_1,x_2) = (\partial_{x_2}\eta(x), -\partial_{x_1}\eta(x))$. It was conjectured that the mean square displacement $\mathbb{E}(|X_t|^2)$ grows like $t\sqrt{\log t}$, both in an annealed sense [31] (where the average is with respect to both the randomness of B and of ω) and in a quenched sense [4], that is, almost surely with respect to ω.

The self-repelling Brownian polymer (SRBP), on the other hand, is a model for a Brownian particle that tries to avoid its past, or for a polymer that is repelled by itself. Mathematically, it is (formally) defined by the solution of the stochastic differential equation

$$\mathrm{dX}_t = -\nabla L_t(X_t)\,\mathrm{dt} + \mathrm{dB}_t, \tag{2.20}$$

where X_t denotes the location of the particle at time t (or the tip of the polymer of length t), and $L_t(x)$ is its local time at x. Because of the minus sign, the drift

tends to push X_t away from locations it has already visited. Again, it was conjectured [31] (with some important caveat, see below, about the initial condition) that as $t \to \infty$, $\mathbb{E}(|X_t|^2) \asymp t\sqrt{\log t}$, that is, the self-repulsion effect makes the process logarithmically superdiffusive.

Both equations (2.19) and (2.20) are very singular and analytically ill posed, since η is a distribution and $L_t(\cdot)$ can be expected to be very rough too. To make sense of them, one usually replaces the GFF η and the local time L_t by $\eta * \rho$ and $L_t * \rho$, respectively, with ρ a smooth mollifier and $*$ denoting the spatial convolution. With such regularization in place, global existence of the solution X_t is classical for both processes, but the quantitative study of their large-t behavior remains highly non-trivial.

A priori, these two processes seem very different, not only phenomenologically but also mathematically as, for example, the DCGFF is a Markov process (at least in a quenched sense), while the SRBP is not, and also look very different from the SPDEs discussed above. However, if one takes a different point of view and lifts them to the *environment level*, then some similarities can be drawn. To witness, let us first look at the DCGFF and define the *environment seen by the particle* as the space-time random vector field

$$\zeta(t,\cdot) := \omega(X_t + \cdot),$$

that is nothing but the (time-independent) drift field ω, recentered at the instantaneous position of the particle. For the SRBP, one proceeds analogously and defines η as above but replacing ω with $-\nabla(L_t * \rho)(\cdot)$.

A classical and crucial fact for the DCGFF is that, since ω is ergodic and divergence-free ($\nabla \cdot \omega = 0$ pointwise), then the law of ζ is independent of time. That is, at all fixed times t, the environment seen by the particle has the law of the curl of the (smoothed) GFF. This point of view turns the DCGFF into a stationary Markov process with Gaussian stationary (non-reversible) measure. Something similar happens for the SRBP: while the process X itself is not Markovian, the environment process is, and in addition, the latter admits the gradient of the (smoothed) GFF as stationary measure. This means that if at time zero the field $-\nabla(L_0 * \rho)(\cdot)$ has the law of the gradient of the (smoothed) GFF, the same remains true at all times. We can then make the conjecture [31] about logarithmic superdiffusivity of the SRBP more precise: if the local time field is initialized at time zero so that $-\nabla(L_0 * \rho)$ is distributed according to the stationary measure, then $\mathbb{E}(|X_t|^2) \asymp t\sqrt{\log t}$, where the expectation is with respect to both the Brownian noise B and to the initial condition for the local time. In terms of particle process, starting from a drift field $-\nabla(L_0 * \rho)$ that is not zero but stationary corresponds to imagining that the particle has evolved from time $-\infty$ to time 0, so that the drift field has reached its stationary distribution. We emphasize, however, that at the moment there is no mathematical result

(or even physics conjecture we are aware of) about how stationarity is approached if the process starts from $L_0 \equiv 0$.

In other words, upon suitably initializing it, the environment seen by the particle process ζ is a stationary Markov process with stationary Gaussian measure, pretty much in the same vein as the solution to AKPZ and SBE. More than that, upon invoking Itô's formula, it can be shown that ζ solves an SPDE with quadratic non-linearity and that the diffusivity can be expressed solely as a function of it.

3. Critical dimension

In this section, we focus on the critical dimension $d = 2$ and begin by presenting the results concerning the large-time asymptotics of the diffusion matrix.

3.1. Superdiffusivity bounds

We start with the stochastic Burgers equation, for which from now on we assume, without loss of generality, that the vector $\mathfrak{w}$ in the non-linearity is $(1, 0, \dots, 0)$, that is, the driving field is horizontal. Then, it is not hard to see that, for symmetry reasons, $D_{ij}(t) = 0$ for $i \neq j$ and $D_{ii}(t) = 1$, $i \neq 1$. That is, the only non-trivial component of the diffusion matrix is $D_{11}(t)$. In [32], it was conjectured (on the basis of mode-coupling theory) that $D_{11}(t) \asymp t^{1/3}$ as $t \to \infty$ in dimension $d = 1$, $D_{11}(t) \asymp 1$ in dimension $d \geq 3$ while, in the critical dimension $d = 2$, $D_{11}(t) \asymp (\log t)^{2/3}$. In the rest of this section, we focus exclusively on dimension $d = 2$.

Remark 3.1. The exponent $2/3$ of the logarithm for $2d$-SBE is quite subtle and due to the (a)symmetry of the equation. Indeed, the same heuristic argument as in [32] (or an alternative one given in [31]) leads to the prediction of a $(\log t)^{1/2}$ divergence of the diffusion coefficient for the AKPZ equation and (as already mentioned above) for the DCGFF and for the self-repelling Brownian polymer. In fact, the $(\log t)^{2/3}$ divergence is typical of critical models which are *anisotropic*, in that the non-linearity acts only in one direction, while the $(\log t)^{1/2}$ divergence occurs for isotropic critical models. In this sense, *$2d$-SBE belongs to a universality class different from that of AKPZ, SRBP or DCGFF.*

The first rigorous result that was proven in support of the conjectured logarithmic divergence of the diffusion coefficient is due to H.-T. Yau, who in the landmark paper [36] proved that, for $2d$-ASEP, $D_{11}(t)$ diverges indeed like $(\log t)^{2/3}$, though in a Tauberian sense (i.e., for the Laplace transform of $t \mapsto tD_{11}(t)$), and with diverging, subdominant, multiplicative corrections (see discussion after Theorem 3.3).

His work has remained an isolated result in this area for almost 20 years. In particular, progress for SPDEs (the AKPZ and SBE) is extremely recent, and the first significant step was made in [6], for the AKPZ equation.

Theorem 3.2 ([6]). *Consider the AKPZ equation, with coupling constant $\lambda > 0$. The Laplace transform $\mathcal{D}_{\mathrm{AKPZ}}(\mu)$ of $t \mapsto tD_{\mathrm{AKPZ}}(t)$ satisfies*

$$\frac{|\log \mu|^{1/2}}{\mu^2}(\log|\log \mu|)^{-C} \lesssim \mathcal{D}_{\mathrm{AKPZ}}(\mu) \lesssim \frac{|\log \mu|^{1/2}}{\mu^2}(\log|\log \mu|)^{C} \tag{3.1}$$

as $\mu \to 0^+$ for some $C > 0$.

While these bounds do not immediately translate into upper and lower bounds on $D_{\mathrm{AKPZ}}(t)$, morally they say that $D_{\mathrm{AKPZ}}(t)$ diverges like $\sqrt{\log t}$, up to subdominant multiplicative corrections polynomial in $\log\log t$.

The analogous result (with a slightly different form of the subdominant corrections) has been proven around the same time in [11] for the annealed mean square displacement of the DCGFF. For the latter, [15] has subsequently obtained the sharper bound $c_-\sqrt{\log t} \le (1/t)\mathbb{E}(|X_t|^2) \le c_+\sqrt{\log t}$, without the subdominant diverging corrections and without going through the Laplace transform.

As for the $2d$-stochastic Burgers equation, the following theorem was, until very recently, the best available result on the large-time behavior of the diffusion coefficient.

Theorem 3.3 ([17]). *For the SBE in dimension $d = 2$, the Laplace transform $\mathcal{D}_{11}(\mu)$ of $t \mapsto tD_{11}(t)$, with $D_{11}(t)$ the first entry of the diffusion matrix $D(t)$ in (2.10), satisfies*

$$\frac{|\log \mu|^{2/3}}{\mu^2}(\log\log|\log \mu|)^{-C} \lesssim \mathcal{D}_{11}(\mu) \lesssim \frac{|\log \mu|^{2/3}}{\mu^2}(\mathrm{loglog}|\log \mu|)^{C} \tag{3.2}$$

as $\mu \to 0^+$ for some $C > 0$.

This theorem is actually the analog of the above-mentioned result [36], for $2d$-ASEP (the theorem in [36] has worse bounds on the subdominant corrections, of order $\exp((\mathrm{loglog}|\log \mu|)^2)$ instead of polynomial in $\mathrm{loglog}|\log \mu|$).

Without entering technical details, let us mention that Theorems 3.2 and 3.3 do not actually work directly with $D_{\mathrm{AKPZ}}(t)$ and $D_{11}(t)$, but with functions that are (at least heuristically) equivalent to them. It is only in (see [12, Proposition 2.5]) that the actual equivalence of the two definitions has been shown.

The culmination of this sequence of works is the following theorem, that represents the only instance where the conjectured logarithmic divergence of the diffusion matrix of a singular critical SPDE has been fully and sharply identified.

Theorem 3.4 ([12]). *Consider the $2d$-SBE with $\lambda > 0$. Then,*

$$D_{11}(t) = C_{\text{eff}}(\lambda)(\log t)^{2/3}(1 + o(1)) \tag{3.3}$$

as $t \to \infty$, with $C_{\text{eff}}(\lambda) = (3/2\pi)^{2/3}\lambda^{4/3}$.

Remark 3.5. It is quite remarkable that Theorem 3.4 identifies the large-time behavior to the precision of the constant prefactor. This is even more true considering that the proportionality $C_{\text{eff}}(\lambda) \propto \lambda^{4/3}$ reveals even deeper universal properties of driven diffusive systems. Indeed, exactly the same holds for the *one-dimensional* stochastic Burgers equation, despite the fact that the t-dependence of the diffusion coefficients is completely different in the two cases (of order $t^{1/3}$ in one dimension and of order $(\log t)^{2/3}$ in two dimensions). As discussed in [12, Remark 1.2], the equality of the exponents of λ in $d = 1, 2$ is not a coincidence, but rather a reflection of the fact that the diffusion coefficient is asymptotically (for large t) independent of the microscopic diffusion coefficient in the direction $\mathfrak{w}$ where the non-linearity acts.

3.2. Scaling limits: The weak coupling regime

In view of Theorem 3.3, in dimension $d = 2$, if we scale the solution of the stochastic Burgers equation diffusively, that is, according to (2.3), there is no hope for it to converge in the limit for $\varepsilon \to 0$. In fact, a simple change of variables argument shows that the $(\log t)^{2/3}$ divergence in Theorem 3.3 is equivalent to a $|\log \varepsilon|^{2/3}$ divergence of the diffusion coefficient for u^ε. In order to tame this divergence, two ways are possible. The most natural one is to change the diffusive space-time rescaling. This will be discussed in Section 3.3. The second one (that is referred to as the "weak coupling limit" or "intermediate disorder regime") is to insist on diffusive scaling and tame the growth of the non-linearity by modifying the coefficients of the SBE in a scale-dependent way. More precisely, one replaces the coupling constant λ with an ε-dependent coupling constant λ_ε, tending to zero with ε in a precisely chosen way, such that the effect of the non-linearity remains of order 1 as $\varepsilon \to 0$. As it turns out, the correct choice is to take the coupling constant λ_ε inversely proportional to $\sqrt{|\log \varepsilon|}$.

Theorem 3.6 ([10]). *Let $d = 2$, $\lambda > 0$ and $T > 0$. Let u be the solution to (2.7) started from the stationary measure and with λ replaced by*

$$\lambda_\varepsilon := \frac{\lambda}{\sqrt{|\log \varepsilon|}}$$

for $\varepsilon > 0$, and u^ε be defined according to (2.3). Then, as $\varepsilon \to 0$, u^ε converges in distribution in $C([0, T], \mathcal{S}')$ to the stationary solution of

$$\partial_t \eta = \frac{1}{2}(D_{\text{SBE}}(\lambda)(\mathfrak{w} \cdot \nabla)^2 + \Delta)\eta + \big(\sqrt{D_{\text{SBE}}(\lambda)}\mathfrak{w} \cdot \nabla + \nabla\cdot\big)\xi, \tag{3.4}$$

where

$$D_{\mathrm{SBE}}(\lambda) = \left(\frac{3\lambda^2}{2\pi} + 1\right)^{2/3} - 1. \tag{3.5}$$

Here, $C([0,T], \mathcal{S}')$ denotes the space of continuous (in time) functions with values in the space of distributions (in space).

Remark 3.7. To be precise, in [10] the result is proven for the equation defined on the two-dimensional torus of size ε^{-1} rather than on the infinite Euclidean space, so that u^ε is defined on the unit two-dimensional torus $\mathbb{T}^2$ and $\mathcal{S}'$ is the space of distributions on $\mathbb{T}^2$. Also, the regularization used in [10] is slightly different from the one presented in Section 2.1: only the non-linearity is regularized, while the noise is kept white in space and time, so that for every ε the stationary measure is the law of the (non-regularized) white noise on the torus. This being said, the proof of Theorem 3.6 would work with minor modifications for the regularization of Section 2.1.

Remark 3.8. An analogous result was proven earlier for the AKPZ equation [7], later in [8] for the SRBP and in [35] for the DCGFF, all under weak coupling scaling. In the former, it was shown that the stationary solution of

$$\partial_t h^\varepsilon = \frac{1}{2}\Delta h^\varepsilon + \frac{\lambda}{\sqrt{|\log \varepsilon|}}\mathcal{M}^\varepsilon(h^\varepsilon) + \xi^\varepsilon \tag{3.6}$$

(to be compared with (2.16)) converges in distribution, as $\varepsilon \to 0$, to the stationary solution of the linear heat equation

$$\partial_t \eta = \frac{1}{2}(D_{\mathrm{AKPZ}}(\lambda) + 1)\Delta\eta + \sqrt{D_{\mathrm{AKPZ}}(\lambda) + 1}\,\xi, \tag{3.7}$$

with ξ a space-time white noise and

$$D_{\mathrm{AKPZ}}(\lambda) = \sqrt{\frac{2\lambda^2}{\pi} + 1} - 1. \tag{3.8}$$

For the SRBP and the DCGFF, an invariance principle for the particle position process $(X_t)_{t\geq 0}$ holds (see [8] for the SRBP and [35] for the DCGFF). In both cases, the diffusion constant of the limiting Brownian motion is strictly larger than 1, that is, strictly larger than the diffusivity of the Brownian noise B_t that appears in the equations (2.19) and (2.20) defining the two models, implying that non-linearity has a non-trivial effect in the weak coupling limit. While at a very high level the proofs of these results have many common features (in particular, coming from the fact that they all have a Gaussian stationary measure), each has its specific technical difficulties and challenges.

What is remarkable is that, while the $2d$-SBE and AKPZ/DCGFF/SRBP have different exponents of the log-divergence of the diffusion coefficient and therefore belong to two different universality classes (see Remark 3.1), yet the weak scaling regime in all of these cases amounts to choose the coupling constant of the same order, that is, proportional to $|\log \varepsilon|^{-1/2}$.

3.3. Scaling limit: Superdiffusive central limit theorem

We now explore the second way in which one can account for the logarithmic divergence of the diffusion matrix, namely by suitably changing the diffusive scaling to a *superdiffusive scaling*. This is the *physically relevant* approach as we are not modifying the coefficients of the equation but only zooming out at *correct scale*. Our work [12] is the first result that fully determines the large-scale behavior of a *critical, singular SPDE* (the $2d$-SBE) beyond weak coupling and identifies its infrared fixed point under superdiffusive scaling.

Before stating the full result, let us see what the correct superdiffusive scaling should be. In view of Theorem 3.4, we should expect normal diffusion in the direction orthogonal to $\mathfrak{w} = (1, 0)$ and logarithmic superdiffusion in the direction $\mathfrak{w}$. That is, the natural scaling to consider is

$$u^\varepsilon(t,x) := \varepsilon^{-1}|\log\varepsilon|^{1/6} u\left(\frac{t}{\varepsilon^2}, R_\varepsilon^{-1/2}\frac{x}{\varepsilon}\right), \quad R_\varepsilon = \begin{pmatrix} |\log\varepsilon|^{-2/3} & 0 \\ 0 & 1 \end{pmatrix} \tag{3.9}$$

and then to take the limit $\varepsilon \to 0$. This formula requires some explanation. First of all, the matrix R_ε allows us to rescale the horizontal and the vertical space direction differently and thus take into account the anisotropy of the equation (and the shape of the diffusion matrix as stated in Theorem 3.4). Second, the prefactor $\varepsilon^{-1}|\log\varepsilon|^{1/6}$ is easily explained: it is a normalization chosen so that the stationary measure of the process u^ε converges to the spatial white noise as $\varepsilon \to 0$. At last, the scaling in (3.9) is indeed logarithmically superdiffusive in the horizontal direction as in fact, $\varepsilon^{-1}|\log\varepsilon|^{1/3} \gg \varepsilon^{-1}$.

We are now ready to state the superdiffusive central limit theorem for the $2d$-SBE.

Theorem 3.9 ([12]). *Let $d = 2$, and $\lambda > 0$. Define u^ε as in (3.9), with u the solution of the SBE (2.7), with initial condition that is absolutely continuous with respect to the stationary measure. Then, as $\varepsilon \to 0$, u^ε converges in the sense of finite-dimensional distributions to the solution of the linear SPDE*

$$\partial_t u^{\mathrm{eff}} = \frac{1}{2}\nabla \cdot D_{\mathrm{eff}} \nabla u^{\mathrm{eff}} + \nabla \cdot \sqrt{D_{\mathrm{eff}}}\,\xi, \quad \textit{for } D_{\mathrm{eff}} = \begin{pmatrix} C_{\mathrm{eff}}(\lambda) & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.10}$$

where $\xi = (\xi_1, \xi_2)$ is a vector-valued space-time white noise and

$$C_{\mathrm{eff}}(\lambda) = \left(\frac{3}{2\pi}\right)^{2/3} \lambda^{4/3}$$

is the same constant as in Theorem 3.4.

We refer the reader to [12, Theorem 1.4] for a more precise and actually stronger formulation, that (roughly speaking) states among other things that $(P_t^\varepsilon F)(u_0)$ (with P_t^ε the semigroup associated to the equation for u^ε and a large enough class of functions F) converges to $(P_t^{\mathrm{eff}} F)(u_0)$ (with obvious notation for P_t^{eff}), in L^p with respect to stationary measure for any $p \geq 1$.

Remark 3.10. It is interesting to compare Theorems 3.6 and 3.9. In some sense, the limit equation (3.10) is the $\lambda \to \infty$ limit of (3.4). For instance, one sees immediately that $D_{\mathrm{SBE}}(\lambda) = C_{\mathrm{eff}}(\lambda)(1 + o(1))$. Also, it is possible to use the method of [12] to control the $2d$-SBE in the whole crossover regime between weak and strong coupling, where the coupling constant $\lambda = \lambda_\varepsilon$ is $|\log \varepsilon|^{-1/2} \ll \lambda_\varepsilon \ll 1$ and space-time is rescaled suitably. In this whole regime, the limit equation is (3.10).

Remark 3.11. There is a fundamental reason why the study of the SBE in the strong coupling regime $\lambda = O(1)$ is much harder than in the weak coupling regime, where the non-linearity is proportional to $|\log \varepsilon|^{-1/2}$. In fact, while at the right-hand side of the SBE equation in the weak coupling scaling there is a full Laplacian, that is crucial in obtaining apriori estimates and in proving tightness of the law of the process, in the equation solved by u^ε (see [12, equation (1.9)]) the Laplacian in direction $\mathfrak{w}$ is multiplied by the asymptotically vanishing prefactor $|\log \varepsilon|^{-2/3}$. This is highly problematic, since the regularizing effect of the Laplacian is lost. That is, the equation for u^ε is much more singular than the one corresponding to the weak coupling scaling. The fact that the Laplacian in direction $\mathfrak{w}$ formally vanishes means that the effective diffusivity in direction $\mathfrak{w}$ is fully produced (in the limit $\varepsilon \to 0$) by the non-linearity.

Remark 3.12. While Theorem 3.9 is the first superdiffusive CLT for a critical, singular SPDE, an analogous result was proven in [1] for the Brownian particle in the curl of the GFF (DCGFF) introduced in Section 2.3. That is, [1] proves a quenched central limit theorem (under logarithmically superdiffusive space-time rescaling) for the trajectory of the particle.

It should be emphasized that, since the DCGFF and the $2d$-SBE belong to two different universality classes (recall Remark 3.1), the space-time rescaling in [1] for DCGFF is different from the one in (3.9) for $2d$-SBE. More precisely, the rescaling for the DCGFF is $(t', x') = (t/\varepsilon^2, x/(\varepsilon|\log \varepsilon|^{-1/4}))$: in particular, it is isotropic in space, since the DCGFF is rotationally invariant in distribution.

The tools exploited in [1] are completely different from those of [12], and it is unclear whether they could be adapted to the SPDE setting (and vice versa).

4. The super-critical case

The (formal) scaling argument outlined in Section 2.1 suggests that, for $d \geq 3$, driven diffusive systems behave diffusively at large scales, since the factor $\varepsilon^{(d-2)/2}$ multiplying the non-linearity tends to zero as $\varepsilon \to 0$. A series of works [14, 22, 24] has confirmed this conjecture for ASEP and, specifically, has proved the convergence of the particle density fluctuation process to a linear stochastic heat equation at large scales (under diffusive scaling). A counterpart on the SPDE side is the recent [10], where the analogous result is obtained for the SBE in dimension $d \geq 3$.

Theorem 4.1 ([10]). *Let $d \geq 3$, $\lambda > 0$ and $T > 0$. Let u be the solution of (2.7), started from the stationary measure and define u^ε according to (2.3).*

Then, as $\varepsilon \to 0$, u^ε converges in distribution in $C([0,T], \mathcal{S}')$ to the stationary solution of (3.4), for some strictly positive constant $D_{\mathrm{SBE}}(\lambda)$.

In contrast with Theorem 3.6, the limiting diffusion constant $D_{\mathrm{SBE}}(\lambda)$ in dimension $d \geq 3$ is not explicit, and in fact we do not expect it to have an elementary form. That said, it is strictly positive. This means that, even though the non-linearity in the equation for u^ε is multiplied by a vanishing factor (see (2.4)), it is *not irrelevant*, strictly speaking, as it has a non-trivial effect on the limit (and thus it would be dangerous to take the scaling argument of Section 2.1 too literally). Morally, it produces a new noise and a new Laplacian which though do not affect the overall Gaussian nature of the fixed point.

5. Perspectives and open problems

As briefly summarized in this note, these last few years have witnessed substantial progress on critical singular SPDEs, and more broadly on out-of-equilibrium irreversible systems at the critical dimension. In the case of the $2d$-SBE and of the DCGFF, the conjectures of, respectively, $(\log t)^{2/3}$ [32] and $(\log t)^{1/2}$ [4, 31] divergence of the diffusion coefficient have been fully proven [12, 15], and the infrared fixed point of the processes has been identified, in the form of a superdiffusive central limit theorem [1, 12]. In a broader perspective, these works can be seen as a first step towards a more systematic understanding of out-of-equilibrium irreversible systems at the critical dimension.

This progress also raises a number of natural and intriguing new questions, such as

- How universal is the $\lambda^{4/3}$ scaling of the diffusion coefficient with respect to the coupling constant λ, for models with anisotropic driving field? (see Remark 3.5).

For instance, does the scaling hold for $2d$-ASEP? Does it hold for the SBE in dimension $d > 2$, in the limit of large λ?

- It is very tempting to conjecture that a superdiffusive central limit theorem of the type of Theorem 3.9 holds also for the particle fluctuation field of two-dimensional ASEP and other asymmetric, two-dimensional interacting particle systems. Can the techniques of [12] be adapted to the discrete setting?
- All the models mentioned in these notes have an explicit stationary but non-reversible measure (either for the process itself or for the associated environment process). How to go beyond cases where the stationary measure is known?
- Is it possible to combine the functional analytic methods of [12] with the stochastic homogenization methods of [1, 15]? For the moment, the two lines of research are complementary.
- There is a number of natural and mathematically non-trivial ways in which the SBE can be generalized. First of all, one would like to go beyond the case where the non-linearity is exactly quadratic. The scaling argument and the heuristics work the same for polynomial (or even more general) non-linearities, but from the technical point of view, the generalization is not at all immediate. Secondly, the SBE models driven diffusive systems with a single conserved quantity. However, systems like driven fluids have $N > 1$ conserved quantities (for a fluid they would be mass, energy and d components of momentum, for a total of $N = 2 + d$). The natural modelization suggested by the theory of fluctuating hydrodynamics [29] is then a system of coupled stochastic Burgers equations. It is known in one dimension [29] that the large-scale behavior can depend qualitatively on N (see [9] for an example). In dimension $d = 2$, the same is conjectured: in fact, while the single SBE exhibits $(\log t)^{2/3}$ superdiffusivity, two-dimensional fluids (corresponding to $N = 4$) are expected [33] to exhibit $(\log t)^{1/2}$ superdiffusivity. This problem is wide open, the only available result being a suboptimal $\log\log t$ lower bound on superdiffusivity for a discrete lattice fluid model [23]. Unfortunately, such result is too weak to distinguish between the two different powers of $\log t$.

Lastly, a very fascinating line of works [26, 27] has displayed compelling features of the DCGFF such as intermittency properties and an unexpected connection with a diffusion on the Lie group SL(2), and it would be interesting to see if analogous relations can be found for the SBE, AKPZ equation or even more broadly for other out-of-equilibrium systems at the critical dimension.

Acknowledgments. F.T. wishes to thank the scientific committee of the European Congress of Mathematics for the honor of delivering a plenary lecture and the organizers of the congress for their great logistic job.

Both authors are very grateful to Dirk Erhard, Harry Giles, Massimiliano Gubinelli, Levi Haunschmid-Sibitz and Quentin Moulard; the works described in this review are the fruit of collaborations with them (in various combinations).

Funding. G.C. gratefully acknowledges financial support via the UKRI FL fellowship "Large-scale universal behaviour of Random Interfaces and Stochastic Operators" MR/W008246/1. F.T. was supported by the Austrian Science Fund (FWF): P35428-N.

References

[1] S. Armstrong, A. Bou-Rabee, and T. Kuusi, Superdiffusive central limit theorem for a Brownian particle in a critically-correlated incompressible random drift. 2024, arXiv: 2404.01115v2

[2] A.-L. Barabási and H. E. Stanley, *Fractal concepts in surface growth*. Cambridge University Press, Cambridge, 1995, 366 pp. Zbl 0838.58023 MR 1600794

[3] A. Borodin and P. L. Ferrari, Anisotropic growth of random surfaces in $2 + 1$ dimensions. *Comm. Math. Phys.* **325** (2014), no. 2, 603–684 Zbl 1303.82015 MR 3148098

[4] J.-P. Bouchaud, A. Comtet, A. Georges, and P. Le Doussal, Anomalous diffusion in random media of any dimensionality. *J. Phys. France* **48** (1987), no. 9, 1445–1450

[5] G. Cannizzaro, D. Erhard, and P. Schönbauer, 2D anisotropic KPZ at stationarity: scaling, tightness and nontriviality. *Ann. Probab.* **49** (2021), no. 1, 122–156 Zbl 1457.60112 MR 4203334

[6] G. Cannizzaro, D. Erhard, and F. Toninelli, The stationary AKPZ equation: logarithmic superdiffusivity. *Comm. Pure Appl. Math.* **76** (2023), no. 11, 3044–3103 Zbl 1525.35263 MR 4642815

[7] G. Cannizzaro, D. Erhard, and F. Toninelli, Weak coupling limit of the anisotropic KPZ equation. *Duke Math. J.* **172** (2023), no. 16, 3013–3104 Zbl 1540.60150 MR 4679957

[8] G. Cannizzaro and H. Giles, An invariance principle for the $2d$ weakly self-repelling Brownian polymer. *Probab. Theory Related Fields* (2025), DOI 10.1007/s00440-025-01363-y

[9] G. Cannizzaro, P. Gonçalves, R. Misturini, and A. Occelli, From ABC to KPZ. *Probab. Theory Related Fields* **191** (2025), no. 1–2, 361–420 Zbl 1558.60149 MR 4869258

[10] G. Cannizzaro, M. Gubinelli, and F. Toninelli, Gaussian fluctuations for the stochastic Burgers equation in dimension $d \geq 2$. *Comm. Math. Phys.* **405** (2024), no. 4, article no. 89, 60 pp. Zbl 1536.60052 MR 4719971

[11] G. Cannizzaro, L. Haunschmid-Sibitz, and F. Toninelli, $\sqrt{\log t}$-Superdiffusivity for a Brownian particle in the curl of the 2D GFF. *Ann. Probab.* **50** (2022), no. 6, 2475–2498 Zbl 1502.82015 MR 4499841

[12] G. Cannizzaro, Q. Moulard, and F. Toninelli, Superdiffusive Central Limit Theorem for the Stochastic Burgers Equation at the critical dimension. [v1] 2024, [v2] 2025, arXiv: 2501.00344v2

[13] G. Cannizzaro and F. Toninelli, Lecture notes on stationary critical and super-critical SPDEs. 2024, arXiv:2403.15006v1, to appear in Lect. Notes Math., CIME Found. Subser., Springer

[14] C.-C. Chang, C. Landim, and S. Olla, Equilibrium fluctuations of asymmetric simple exclusion processes in dimension $d \geq 3$. *Probab. Theory Related Fields* **119** (2001), no. 3, 381–409 Zbl 1006.60096 MR 1821140

[15] G. Chatzigeorgiou, P. Morfe, F. Otto, and L. Wang, The Gaussian free-field as a stream function: asymptotics of effective diffusivity in infra-red cut-off. *Ann. Probab.* **53** (2025), no. 4, 1510–1536 Zbl 1570.35025 MR 4930533

[16] A. da Silveira and M. Kardar, Nonlinear stochastic equations with calculable steady states. *Phys. Rev. E* **68** (2003) no. 4, article no. 046108, 9 pp.

[17] D. De Gaspari and L. Haunschmid-Sibitz, $(\log t)^{\frac{2}{3}}$-superdiffusivity for the 2d stochastic Burgers equation. *Electron. J. Probab.* **29** (2024), 1–34 Zbl 1564.60063 MR 4838438

[18] M. Gubinelli, P. Imkeller, and N. Perkowski, Paracontrolled distributions and singular PDEs. *Forum Math. Pi* **3** (2015), article no. e6, 75 pp. Zbl 1333.60149 MR 3406823

[19] M. Hairer, A theory of regularity structures. *Invent. Math.* **198** (2014), no. 2, 269–504 Zbl 1332.60093 MR 3274562

[20] M. Kardar, G. Parisi, and Y.-C. Zhang, Dynamic scaling of growing interfaces. *Phys. Rev. Lett.* **56** (1986), no. 9, 889–892 Zbl 1101.82329

[21] A. Kupiainen, Renormalization group and stochastic PDEs. *Ann. Henri Poincaré* **17** (2016), no. 3, 497–535 Zbl 1347.81063 MR 3459120

[22] C. Landim, S. Olla, and S. R. S. Varadhan, Diffusive behaviour of the equilibrium fluctuations in the asymmetric exclusion processes. In T. Funaki and H. Osada (eds.), *Stochastic analysis on large scale interacting systems*, pp. 307–324, Adv. Stud. Pure Math. 39, Math. Soc. Japan, Tokyo, 2004 Zbl 1080.60094 MR 2073338

[23] C. Landim, J. A. Ramírez, and H.-T. Yau, Superdiffusivity of two dimensional lattice gas models. *J. Stat. Phys.* **119** (2005), no. 5–6, 963–995 Zbl 1088.82016 MR 2157854

[24] C. Landim and H. T. Yau, Fluctuation-dissipation equation of asymmetric simple exclusion processes. *Probab. Theory Related Fields* **108** (1997), no. 3, 321–356 Zbl 0884.60092 MR 1465163

[25] T. M. Liggett, *Interacting particle systems*. Grundlehren Math. Wiss. 276, Springer, New York, 1985, 488 pp. Zbl 0559.60078 MR 0776231

[26] P. Morfe, F. Otto, and C. Wagner, A critical drift-diffusion equation: connections to the diffusion on **SL**(2). [v1] 2024, [v2] 2025, arXiv:2410.15983v2

[27] F. Otto and C. Wagner, A critical drift-diffusion equation: intermittent behavior. [v1] 2024, [v3] 2025, arXiv:2404.13641v3

[28] M. Prähofer and H. Spohn, An exactly solved model of three-dimensional surface growth in the anisotropic KPZ regime. *J. Stat. Phys.* **88** (1997), no. 5–6, 999–1012 Zbl 0945.82543 MR 1478060

[29] H. Spohn, Nonlinear fluctuating hydrodynamics for anharmonic chains. *J. Stat. Phys.* **154** (2014), no. 5, 1191–1227 Zbl 1291.82119 MR 3176405

[30] F. Toninelli, $(2 + 1)$-Dimensional interface dynamics: mixing time, hydrodynamic limit and anisotropic KPZ growth. In *Proceedings of the international congress of mathematicians—Rio de Janeiro 2018. Vol. III. Invited lectures*, pp. 2733–2758, World Scientific, Hackensack, NJ, 2018 Zbl 1453.82046 MR 3966870

[31] B. Tóth and B. Valkó, Superdiffusive bounds on self-repellent Brownian polymers and diffusion in the curl of the Gaussian free field in $d = 2$. *J. Stat. Phys.* **147** (2012), no. 1, 113–131 Zbl 1245.82092 MR 2922762

[32] H. van Beijeren, R. Kutner, and H. Spohn, Excess noise for driven diffusive systems. *Phys. Rev. Lett.* **54** (1985), no. 18, 2026–2029 MR 0789756

[33] T. E. Wainwright, B. J. Alder, and D. M. Gass, Decay of time correlations in two dimensions. *Phys. Rev. A* **4** (1971), 233–237

[34] D. E. Wolf, Kinetic roughening of vicinal surfaces. *Phys. Rev. Lett.* **67** (1991), no. 13, 1783–1786

[35] H. Yang and Z. Yang, Weak coupling limit of a Brownian particle in the curl of the 2D GFF. 2024, arXiv:2405.05778v1

[36] H.-T. Yau, $(\log t)^{2/3}$ law of the two dimensional asymmetric simple exclusion process. *Ann. of Math. (2)* **159** (2004), no. 1, 377–405 Zbl 1060.60099 MR 2052358

Giuseppe Cannizzaro
Department of Statistics, University of Warwick, Coventry CV4 74L, UK;
giuseppe.cannizzaro@warwick.ac.uk

Fabio Toninelli
Institut für Stochastik und Wirtschaftsmathematik, TU Wien, Wiedner Hauptstraße 8-10/105-7, 1040 Vienna, Austria; fabio.toninelli@tuwien.ac.at

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/5

Anomalous diffusion via iterative quantitative homogenization: An overview of the main ideas

Scott Armstrong and Vlad Vicol

Abstract. Anomalous diffusion is the fundamental ansatz of phenomenological theories of passive scalar turbulence and has been confirmed numerically and experimentally to an extraordinary extent. The purpose of this survey is to discuss our recent result, in which we construct a class of incompressible vector fields that have many of the properties observed in a fully turbulent velocity field, and for which the associated scalar advection-diffusion equation generically displays anomalous diffusion. Our main contribution is to propose an analytical framework in which to study anomalous diffusion via a backward cascade of renormalized eddy viscosities.

1. Introduction

Let $u = u(t,x)\colon [0,\infty)\times \mathbb{T}^d \to \mathbb{R}^d$ be a given divergence-free velocity field, which is $\mathbb{T}^d$-periodic in space. Here, $d \geq 2$ is the space dimension. We are interested in the setting where u is continuous (in space and time), but has *low regularity*; e.g., $u \in C^\alpha_{x,t} := C^0_t C^\alpha_x \cap C^\alpha_t C^0_x$ for some $\alpha \in (0,1)$. The motivation to consider such vector fields stems from phenomenological theories of fully developed hydrodynamic turbulence [13,26,31,42,55,62,66,70]. The Hölder regularity exponent $\alpha = \frac{1}{3}$ is singled out by the Onsager theory of *ideal turbulence* [36,39], and this exponent plays a prominent role in this manuscript.

For such a given incompressible vector field $u \in C^\alpha_{x,t}$, we consider the advection-diffusion equation

$$\partial_t \theta^\kappa + u\cdot\nabla\theta^\kappa = \kappa\Delta\theta^\kappa, \quad \theta^\kappa|_{t=0} = \theta_{\mathrm{in}}, \tag{1.1}$$

which models the evolution of a *passive scalar* field $\theta^\kappa = \theta^\kappa(t,x)\colon [0,\infty)\times\mathbb{T}^d \to \mathbb{R}$ in a background fluid with velocity u. The parameter $\kappa > 0$ represents molecular

Mathematics Subject Classification 2020: 35B27 (primary); 35K15 (secondary).
Keywords: anomalous diffusion, passive scalar, advection-diffusion equation, quantitative homogenization, iterated homogenization.

diffusivity, or, in non-dimensional form, an inverse Péclet number. Since we are interested in the behavior of the fields θ^κ as $\kappa \to 0$, it is sufficient to consider $\kappa \in (0, 1]$. The initial condition in (1.1) is given by $\theta_{\text{in}} = \theta_{\text{in}}(x)$, which is taken to have zero-mean on $\mathbb{T}^d$ (i.e., $\fint_{\mathbb{T}^d} \theta_{\text{in}} dx = 0$), and is assumed to be sufficiently smooth (for instance, $\theta_{\text{in}} \in H^1(\mathbb{T}^d)$). The advection-diffusion equation (1.1) is to be solved for $t \geq 0$; for simplicity, here we will only consider $t \in [0, 1]$. We note that if $u \in C^\alpha_{x,t}$ with $\alpha \in (0, 1)$, then for any $\kappa > 0$, the solution θ^κ of (1.1) is *uniquely defined*[1] and experiences parabolic smoothing: *in positive time*, we have $\theta^\kappa \in C^0_t C^{2,\alpha}_x \cap C^{1,\alpha/2}_t C^0_x$. Also, note that the solution θ^κ continues to have zero-mean in positive time, namely $\fint_{\mathbb{T}^d} \theta^\kappa(t, x) dx = 0$ for all $t \in (0, 1]$.

1.1. A rigorous definition of anomalous diffusion

The fundamental *energy balance* for the advection-diffusion equation is derived by multiplying (1.1) with θ^κ, and using that $\operatorname{div} u = 0$ to write

$$\partial_t \left(\frac{1}{2} |\theta^\kappa|^2 \right) + \operatorname{div} \left(u \frac{1}{2} |\theta^\kappa|^2 \right) - \kappa \Delta \left(\frac{1}{2} |\theta^\kappa|^2 \right) + \kappa |\nabla \theta^\kappa|^2 = 0.$$

Integrating the above relation on $[0, t] \times \mathbb{T}^d$ yields the energy balance[2]

$$\frac{1}{2} \|\theta^\kappa(t, \cdot)\|^2_{L^2(\mathbb{T}^d)} + \kappa \int_0^t \|\nabla \theta^\kappa(t', \cdot)\|^2_{L^2(\mathbb{T}^d)} dt' = \frac{1}{2} \|\theta_{\text{in}}\|^2_{L^2(\mathbb{T}^d)}. \tag{1.2}$$

We emphasize that (1.2) may be justified rigorously (with $=$ sign) under a very weak assumption on the incompressible velocity field u, namely $u \in L^\infty_t L^d_x$.

The uniform-in-κ a priori estimate on θ^κ in $L^\infty_t L^2_x$ provided by (1.2) shows that sequences $\{\theta^{\kappa_j}\}_{j \geq 0}$, with $\kappa_j \to 0$ as $j \to \infty$, have subsequential weak-* limits in $L^\infty_t L^2_x$. Moreover, the linearity of (1.1) implies that any such weak-* limit point $\overline{\theta}$ is an $L^\infty_t L^2_x$ weak solution of the transport equation

$$\partial_t \overline{\theta} + \operatorname{div}(u \overline{\theta}) = 0, \quad \overline{\theta}|_{t=0} = \theta_{\text{in}}. \tag{1.3}$$

[1] At lower regularity than the one considered in this paper, it is possible to construct non-unique weak solutions of the advection-diffusion equation (1.1). This was first achieved in [23, 58, 59], by treating the diffusion term as an error in an "intermittent convex-integration scheme" for the transport equation; this idea was introduced in [16, 19] for the three-dimensional Navier–Stokes equations. In the works [23, 58, 59], the velocity field u and the scalar field θ^κ are constructed simultaneously, resulting in low regularity for both. See also the recent work [60] for a non-uniqueness proof based on stochastic Lagrangians.

[2] The quantity $\|\theta^\kappa\|^2_{L^2(\mathbb{T}^d)}$ is the passive scalar's *variance* (recall, the mean of θ^κ is fixed to equal zero), but we may also think about it as an *energy*.

Since under weak convergence we have $\|\overline{\theta}(t,\cdot)\|_{L^2(\mathbb{T}^d)} \le \liminf_{\kappa\to 0}\|\theta^\kappa(t,\cdot)\|_{L^2(\mathbb{T}^d)}$ for $t \in [0,1]$, with (1.2) we obtain

$$\begin{aligned} 0 &\le \limsup_{\kappa\to 0} \kappa \int_0^t \|\nabla\theta^\kappa(t',\cdot)\|^2_{L^2(\mathbb{T}^d)} dt' \\ &= \frac{1}{2}\|\theta_{\mathsf{in}}\|^2_{L^2(\mathbb{T}^d)} - \liminf_{\kappa\to 0}\frac{1}{2}\|\theta^\kappa(t,\cdot)\|^2_{L^2(\mathbb{T}^d)} \\ &\le \frac{1}{2}\|\theta_{\mathsf{in}}\|^2_{L^2(\mathbb{T}^d)} - \frac{1}{2}\|\overline{\theta}(t,\cdot)\|^2_{L^2(\mathbb{T}^d)} \end{aligned} \tag{1.4}$$

for $t \in [0,1]$. The right-hand side of (1.4) depends only on the solution $\overline{\theta}$ of the transport equation (1.3) which arose in the weak-* limit.[3] The bound on the energy dissipation rate obtained in (1.4) shows that if *all* $L^\infty_t L^2_x$ *weak solutions* of the transport equation (1.3) have constant-in-time $L^2(\mathbb{T}^d)$-norm, then the right-hand side of (1.4) vanishes identically, and thus $\limsup_{\kappa\to 0}\kappa\int_0^t\|\nabla\theta^\kappa(t',\cdot)\|^2_{L^2(\mathbb{T}^d)}dt' = 0$. Such a situation occurs, for instance, if the solution of (1.3) is uniquely obtained via ODE theory under the Cauchy–Lipschitz assumption $u \in L^1_t W^{1,\infty}_x$. At a lower space integrability of the velocity gradient, the DiPerna–Lions theory [30] guarantees that if $u \in L^1_t W^{1,1}_x$, then all $L^\infty_t L^2_x$ weak solutions of (1.3) are *renormalized*, and thus they conserve their energy as a function of time.[4]

The above discussion shows that if $\nabla u \in L^1_{x,t}$, then by the DiPerna–Lions theory, we automatically have $\limsup_{\kappa\to 0}\kappa\int_0^t\|\nabla\theta^\kappa(t',\cdot)\|^2_{L^2(\mathbb{T}^d)}dt' = 0$ for $t \in [0,1]$. Turbulent flows u however do not possess the property that $\nabla u \in L^1_{x,t}$; instead, we expect u to have Onsager supercritical regularity, such as $u \in C^\alpha_{x,t}$ for some $\alpha \le \frac{1}{3}$ (see [18, 36, 39, 42] and the references therein). This leaves open the possibility that the right-hand side of (1.4) does not vanish identically, motivating the following definition.

Definition 1.1 (Anomalous diffusion). Fix an incompressible vector field $u \in C^0_{x,t}$. If for any $\theta_{\mathsf{in}} \in H^1(\mathbb{T}^d)$ with $\fint_{\mathbb{T}^d}\theta_{\mathsf{in}}(x)dx = 0$, there exists $0 < \varrho \le \frac{1}{2}$ such that the family of unique solutions $\{\theta^\kappa\}_{\kappa>0}$ of the Cauchy problem (1.1) satisfies

$$\limsup_{\kappa\to 0}\kappa\int_0^1\int_{\mathbb{T}^d}|\nabla\theta^\kappa(t',x)|^2 dx dt' \ge \varrho\|\theta_{\mathsf{in}}\|^2_{L^2(\mathbb{T}^d)}, \tag{1.5}$$

[3] For vector fields u whose regularity lies *below* $L^1_t W^{1,\infty}_x$, e.g., for $u \in L^1_t C^\alpha_x$ with $\alpha < 1$, the transport equations are expected to have non-unique solutions. Different subsequences $\kappa_j \to 0$ could in principle select different such solutions (see Theorem 1.2 below). Therefore, (1.4) should be written in terms of sequences $\kappa_j \to 0$, and for a *given* weak-* limit $\overline{\theta}$.

[4] See also the work of Ambrosio [2] for $u \in L^1_t \mathrm{BV}_x$.

we say the advection-diffusion equation associated to u exhibits *anomalous diffusion* on $[0, 1]$; equivalently, it exhibits *anomalous dissipation of the passive scalar's variance*.[5]

The anomalous diffusion associated to the time interval $[0, 1]$, as defined in Definition 1.1, may equivalently be introduced for any time interval $[0, t]$ with $t \in (0, 1]$. Moreover, given any sequence $\kappa_n \to 0$, one may find a further subsequence $\kappa_{n_j} \to 0$ along which the *time dissipation measure* $\mathcal{E}(dt)$ is the weak-* limit in the sense of bounded measures

$$\mathcal{E}(dt) := \lim_{j\to\infty} \kappa_{n_j} \|\nabla\theta^{\kappa_{n_j}}(t,\cdot)\|^2_{L^2(\mathbb{T}^d)}. \tag{1.6}$$

Here, we abuse notation, as clearly $\mathcal{E}$ also depends on the initial datum θ_{in}, and also on the chosen subsequence $\{\kappa_{n_j}\}_{j\geq 0}$. With (1.4) and (1.6), relation (1.5) becomes $\varrho\|\theta_{\text{in}}\|^2_{L^2(\mathbb{T}^d)} \leq \int_0^1 \mathcal{E}(dt) \leq \frac{1}{2}\|\theta_{\text{in}}\|^2_{L^2(\mathbb{T}^d)}$.

1.2. Predictions from hydrodynamic turbulence

It is widely accepted that if the vector field u represents a homogeneous isotropic turbulent velocity field, then solutions of (1.1) exhibit anomalous diffusion in the sense of Definition 1.1. This prediction was first discussed by Obukhov [62], who drew direct analogies to the anomalous dissipation of kinetic energy in a turbulent fluid, postulated by Kolmogorov's 1941 phenomenological theory [52]. Compelling numerical and experimental evidence for anomalous diffusion is presented in [31,66,67,70]. In this setting, one expects that the time dissipation measure $\mathcal{E}(dt)$ is non-atomic and is maybe even absolutely continuous with respect to Lebesgue measure on $[0, 1]$.

Under the ansatz that the advection-diffusion equation (1.1) associated to u exhibits anomalous diffusion (a so-called "experimental fact"), and assuming that u represents a homogeneous isotropic turbulent velocity field exhibiting *monofractal scaling in the inertial range*, Obukhov [62] and Corrsin [26] have furthermore predicted that the sequence $\{\theta^\kappa\}_{\kappa>0}$ retains uniform-in-κ Hölder regularity, with exponents which may be determined from scaling.

[5] For passive scalar transport, anomalous diffusion (1.5) is equivalent to Lagrangian spontaneous stochasticity [33]. Roughly speaking, this means that backward-in-time transition probabilities of the SDE used to represent the advection-diffusion equation do not converge as $\kappa \to 0$ to a Dirac mass around a deterministic Lagrangian trajectory.

In modern mathematical terms, in analogy with the Onsager conjecture for the incompressible Euler[6] equations [18, 25, 28, 29, 37, 39],[7] one may propose a *mathematical idealization* of the Obukhov–Corrsin prediction and postulate the dichotomy:

(i) If $u \in C^{\alpha}_{x,t}$ for some $\alpha \in (0,1)$ is divergence-free, and if the family of solutions $\{\theta^{\kappa}\}_{\kappa>0}$ of (1.1) is bounded uniformly in κ in $L^2_t C^{\bar{\alpha}}_x$ for some $\bar{\alpha} > \frac{1-\alpha}{2}$, then there is no anomalous diffusion, namely $\lim_{\kappa\to 0} \kappa \|\nabla\theta^{\kappa}\|^2_{L^2_t L^2_x} = 0$.

(ii) For each $\alpha \in \left(0, \frac{1}{3}\right)$,[8] there exists divergence-free $u \in C^{\alpha}_{x,t}$, such that anomalous diffusion holds in the sense of Definition 1.1, with a non-atomic time-dissipation measure $\mathcal{E}(dt)$. Moreover, for *all* smooth initial conditions θ_{in}, the solutions $\{\theta^{\kappa}\}_{\kappa>0}$ of (1.1) are bounded uniformly in κ in $L^2_t C^{\bar{\alpha}}_x$ for $\bar{\alpha} < \frac{1-\alpha}{2}$.

Part (i) of the Obukhov–Corrsin dichotomy is known and follows directly from the commutator estimate of Constantin, E, and Titi [25] (see, e.g., [32] or [24]).[9] As stated, part (ii) of the dichotomy remains open (see Section 1.3 for partial progress). The main purpose of our work [8] is to provide a new perspective on anomalous diffusion for the passive scalar equation, in the hope that these ideas will lead to a resolution of part (ii) of the Obukhov–Corrsin dichotomy.

1.3. Previous results

From a rigorous mathematical perspective, prior to our work [8], we are aware of only two (classes of) examples of incompressible vector fields u for which the associated passive scalar equation (1.1) exhibits (a form of) anomalous diffusion (1.5): the first is the stochastic Kraichnan model (introduced by Kraichnan in [53]), while the second is a class of deterministic singularly focusing mixing flows (introduced by Drivas, Elgindi, Iyer, and Jeong in [32]). We next discuss these examples, but do so *exceedingly briefly*. For a complete list of references and appropriate context, we refer the interested readers to the introductions of [8, 20, 32, 33, 36, 56].

Kraichnan proposed in [53] a model in which $u = u^{\nu}$ is a realization of a homogeneous isotropic stationary zero-mean Gaussian random field, which is delta-correlated

[6] A weak solution $(u, p) \in L^2_{x,t} \times L^1_{x,t}$ of the incompressible Euler equations solves the system $\partial_t u + \operatorname{div}(u \otimes u) + \nabla p = 0$, with $\operatorname{div} u = 0$, in the sense of space-time distributions.

[7] For the incompressible Euler equations, the Onsager dichotomy is a "theorem", except for the most important endpoint case $\alpha = \frac{1}{3}$; see [22, 34] for the sharpest *rigidity* statements and [17, 44, 45, 48, 61] for the latest *flexibility* statements in the Onsager theorem.

[8] The reason we do not include the full range $\alpha \in (0,1)$ here is discussed in Section 4 below.

[9] Specifically, one may modify the mollification+commutator argument of [25] to establish the bound $\kappa \int_0^1 \|\nabla\theta^{\kappa}(\cdot,t)\|^2_{L^2} dt \lesssim \kappa^{\alpha+2\bar{\alpha}-1/\alpha+1}(1 + \|u\|_{L^{\infty}_t C^{\alpha}_x})^{2(1-\bar{\alpha})/\alpha+1} \|\theta^{\kappa}\|^2_{L^{\infty}_t C^{\bar{\alpha}}_x}$.

in time and is colored in space in order to match the Kolmogorov 1941 phenomenological theory: the space covariance is fully smooth at length scales $\lesssim \nu^{1/2}$ (a UV cutoff serving as a proxy for the Kolmogorov dissipation scale) and behaves like a C^α function at length scales $\gtrsim \nu^{1/2}$ (a synthetic power spectrum in the inertial range). For such a stationary Gaussian random field u^ν, instead of (1.1) one studies the Stratonovich SDE $d\theta^{\kappa,\nu} + u^\nu \odot d\theta^{\kappa,\nu} = \kappa\Delta\theta^{\kappa,\nu}dt$. The main result is that in the joint limit $\nu, \kappa \to 0$, and for a.e. realization of the field u^ν, and any $\theta_{\text{in}} \in L^2(\mathbb{T}^d)$ of zero-mean, anomalous diffusion holds; see Bernard, Gawędzki, and Kupiainen [14], Le Jan and Raimond [54], or Falkovich, Gawędzki, and Vergassola [40]. A fully rigorous, complete, and concise proof of anomalous diffusion for the Kraichnan model (and its variants) may be found in the recent work of Rowan [64].

The main drawback of the Kraichnan model stems from the white-noise temporal correlation of the vector field u, which is indeed so rough that it is likely entirely responsible for the anomalous diffusivity; see the mathematical discussion in [64] and the numerical comparisons made in [68].

The first deterministic vector field u for which the dissipation anomaly (1.5) can be established rigorously was constructed by Drivas, Elgindi, Iyer, and Jeong [32]. These authors show that for any $\alpha \in [0, 1)$ and any dimension $d \geq 2$, there exists $u \in L^1([0,1]; C^\alpha(\mathbb{T}^d)) \cap L^\infty([0,1]; L^\infty(\mathbb{T}^d))$, such that $u(t,\cdot)$ is smooth for any $t < 1$, and such that for any $\theta_{\text{in}} \in H^2(\mathbb{T}^d)$ which is sufficiently close (in L^2) to an eigenfunction of the Laplacian, anomalous diffusion (1.5) holds, with a purely atomic time dissipation measure $\mathcal{E}(dt) = \rho(\theta_{\text{in}})\delta_{t=1}$. The construction in [32] is based on "quasi-self-similarly" (as $t \to 1^-$) accelerating an inviscid mixing dynamics [1] and treating the diffusion term as a perturbation. This is made possible by the singular nature of the "speed-up" of u with respect to time, and the fact that at each time $t < 1$ the field u has only one "active scale" with respect to space.

The ideas in [32] proved to be quite popular, leading to a number of further refinements [15, 21, 24, 35, 46, 50, 51]. Notably, Colombo, Crippa, and Sorella [24] gave examples of a smooth initial conditions θ_{in}, and of a divergence-free vector field $u \in C_t^0 C_x^\alpha$ (for any $\alpha < 1$), such that the associated family of solutions $\{\theta^\kappa\}_{\kappa>0}$ is bounded uniformly in κ in $L_t^2 C_x^{\overline{\alpha}}$ for any $\overline{\alpha} < \frac{1-\alpha}{2}$, consistent with the Obukhov–Corrsin dichotomy; moreover, (1.5) holds with $\mathcal{E}(dt) = \rho(\theta_{\text{in}})\delta_{t=1}$, and subsequential limits select different weak solutions of the transport equation.[10]

This result was further refined by Elgindi and Liss [35], who for any $\alpha \in (0, 1)$ provide a divergence-free vector field $u \in C^\infty([0,1]; C^\alpha(\mathbb{T}^d))$, such that for *all*

[10] At the lower regularity $u \in L^\infty_{x,t}$, the vanishing viscosity limit fails to select not just unique solutions of the transport equation, but also physically admissible ones, in the sense of non-increasing energy/entropy [47].

zero-mean and smooth initial conditions θ_{in} (instead of just *one initial condition*), anomalous diffusion holds with $\mathcal{E}(dt) = \rho(\theta_{\text{in}})\delta_{t=1}$; moreover, $\{\theta^\kappa\}_{\kappa>0}$ is bounded uniformly in κ in $L^2_t C^{\bar{\alpha}}_x$ for any $\bar{\alpha} < \frac{1-\alpha}{2}$. We also mention the work of Johanson–Sorella [51] who construct an *autonomous* divergence-free velocity field $u \in C^\alpha_x$ and a smooth zero-mean initial condition θ_{in}, such that (1.5) holds; the associated time dissipation measure $\mathcal{E}(dt)$ is purely absolutely continuous with respect to the Lebesgue measure on $[0, 1]$. Lastly, we note the recent paper of Hess-Childs and Rowan [46] who for any $\alpha \in (0, 1)$ construct $u \in C^0_t C^\alpha_x \cap C^{\alpha/1-\alpha}_t C^0_x$ such that for any zero-mean $\theta_{\text{in}} \in L^1(\mathbb{T}^2)$, the passive scalar θ^κ exhibits *asymptotic total dissipation* as $\kappa \to 0$, at $\{t = 1\}$; i.e., $\lim_{\kappa\to 0}\|\theta^\kappa(1, \cdot)\|_{L^1(\mathbb{T}^2)} = 0$.

The main drawbacks of the constructions in [24, 32, 35] are that the anomalous dissipation occurs at only a discrete set of times ($\mathcal{E}(dt)$ is proportional to $\delta_{t=1}$), the vector field u has only one active scale at each time $t \in [0, 1)$, and the advection-diffusion equation (1.1) is treated as a perturbation of the transport equation (1.3). Taken together, these properties are inconsistent with the observed properties of turbulent flows.

Note that the result in [35] *nearly solves* part (ii) of the Obukhov–Corrsin dichotomy discussed above, except for the singular nature of the time dissipation measure. In Section 4, we argue that this seemingly minor caveat (the atomic nature of $\mathcal{E}(dt)$), combined with the fact that anomalous diffusion is meant to hold for all smooth zero-mean initial conditions θ_{in}, is in fact fundamental to the Obukhov–Corrsin dichotomy, and we present a conjecture in this direction (see Conjecture 4.1).

1.4. Main result

Our goal in [8] was to build "minimally realistic examples" of "fluid-like" velocity fields u, for which anomalous diffusion holds in the sense of Definition 1.1,[11] and such that the time dissipation measure $\mathcal{E}(dt)$ defined in (1.6) is *non-atomic*.[12]

From a broader perspective, our goal was to highlight what we believe to be the "right reason" for which anomalous diffusion holds: it is not that the underlying transport dynamics exhibits *mixing*, which is accelerated towards a discrete set of times; rather, we show that as one "zooms out" from small scales to large scales, one may observe a backward cascade of renormalized "eddy diffusivities" taking shape, in such a way that the large-scale motion experiences $O(1)$ diffusivity independently of the vanishing molecular diffusivity (as $\kappa \to 0$). Of course, this picture is not at all original and is indeed widely shared by physicists, where it goes under the name

[11] In particular, (1.5) holds for *all* zero-mean smooth initial conditions.

[12] In particular, the anomalous dissipation occurs continuously in time.

renormalization group: see, for instance, Frisch [42, Section 9.6]. Our main result is as follows.

Theorem 1.2 ([8, Theorem 1.1]). *Fix $d \geq 2$ and $\alpha \in (0, \frac{1}{3})$.*[13] *There exists an incompressible vector field $u \in C^0([0,1]; C^\alpha(\mathbb{T}^d)) \cap C^\alpha([0,1]; C^0(\mathbb{T}^d))$, such that for every $\theta_{\text{in}} \in \dot{H}^1(\mathbb{T}^d)$, the family of solutions $\{\theta^\kappa\}_{\kappa>0}$ of the advection-diffusion equation (1.1) with velocity u exhibits anomalous diffusion in the sense of Definition 1.1, with a non-atomic time dissipation measure. Moreover, we have:*

- *The parameter $\varrho \in (0, \frac{1}{2}]$ appearing in (1.5) depends only on α, d, and a lower bound on the natural length scale of the initial datum, given by $L_{\text{in}} := \frac{\|\theta_{\text{in}}\|_{L^2}}{\|\nabla\theta_{\text{in}}\|_{L^2}}$. Specifically, for any $\varepsilon > 0$, we may find $c = c(d, \alpha, \varepsilon) > 0$ such that $\varrho = cL_{\text{in}}^{2(1+\alpha)/1-\alpha+\varepsilon}$.*
- *There exists a sequence $\kappa_j \to 0$, which depends on α and d but is independent of θ_{in}, along which the lower bound in (1.5) is realized; upon defining $\mathcal{K} = \bigcup_{j\geq 1}\left[\frac{1}{2}\kappa_j, 2\kappa_j\right]$, we have $\inf_{\kappa\in\mathcal{K}} \kappa\|\nabla\theta^\kappa\|^2_{L^2((0,1)\times\mathbb{T}^d)} \geq \varrho\|\theta_{\text{in}}\|^2_{L^2(\mathbb{T}^d)}$.*
- *There exists $\mu > 0$, such that for any sequence $\kappa_i \to 0$ with $\{\kappa_i\}_{i\geq 1} \subset \mathcal{K}$, the sequence $\{\theta^{\kappa_i}\}_{i\geq 1}$ is bounded uniformly in κ_i in $C^{0,\mu}([0,1]; L^2(\mathbb{T}^d))$, as $\kappa_i \to 0$. In particular, the anomalous dissipation occurs continuously in time: along this subsequence, we have that $\int_0^t \mathcal{E}(dt')$ is Hölder continuous in $t \in [0,1]$. See* [8, Remark 5.4].
- *There exist sequences $\{\kappa_i\}_{i\geq 1}, \{\kappa_i'\}_{i\geq 1} \subset \mathcal{K}$ such that as $\kappa_i, \kappa_i' \to 0$, the sequences $\{\theta^{\kappa_i}\}_{i\geq 1}$ and $\{\theta^{\kappa_i'}\}_{i\geq 1}$ converge strongly in $C_t^0 L_x^2$ to two* distinct *bounded weak solutions $\overline{\theta}$ and $\overline{\theta}'$ of the transport equation (1.3). This demonstrates the lack of a selection principle for vanishing diffusivity limits of the advection-diffusion equation towards the transport equation. See* [8, Proposition 5.5].
- *The constructed vector field u is "nearly" a weak solution of the incompressible Euler equations, in the sense of* [8, Proposition 2.8].
- *The class of vector fields u for which all the conclusions of this theorem hold is dense in the class of all smooth zero-mean divergence-free vector fields, with respect to the $C^0_{x,t}$ topology. See Remark 3.1 below.*

The construction of the vector field u postulated in Theorem 1.2 has many features in common with Nash-type convex-integration schemes for the incompressible Euler equations: multi-scale constructions with supergeometric scale separation, infinitely many nested active scales with "building blocks" that are stationary Euler

[13] The restriction $\alpha < \frac{1}{3}$ is intimately related to Conjecture 4.1 below.

solutions, and the usage of Lagrangian flows to advect small-scale features by large-scale flows. On the other hand, the analysis of the passive scalars θ^κ is based on ideas from quantitative homogenization. Specifically, we need to perform homogenization *iteratively across many length scales*. In this way, we observe the emergence of a sequence of effective "eddy" diffusivities, which are recursively defined by homogenizing "up the scales", from the small scale at which molecular diffusion acts, to the large scale of the periodic box. In this process, every homogenization step resembles Taylor [69] dispersion specialized to the case of a shear flow. An outline of the proof of Theorem 1.2 is provided in Section 3 below.

1.5. Anomalous diffusion via weak solutions of three-dimensional incompressible Euler

We wish to highlight a very important extension of Theorem 1.2, which was recently obtained by Burczak, Székelyhidi, and Wu in [20]. Whereas the vector field u in Theorem 1.2 may only be shown to be "nearly" a weak solution of the incompressible Euler equations on $\mathbb{T}^d$,[14] the authors of [20] have shown that anomalous dissipation (as in Theorem 1.2) holds for large classes of vector fields $u \in C^\alpha_{x,t}$ (with $\alpha < \frac{1}{3}$), that are true weak solutions of three-dimensional incompressible Euler equations.

Theorem 1.3 (Burczak, Székelyhidi, Wu [20]). *Fix $d = 3$. For any $\alpha \in \left(0, \frac{1}{3}\right)$, there exists a weak solution $u \in C^0([0,1]; C^\alpha(\mathbb{T}^3)) \cap C^\alpha([0,1]; C^0(\mathbb{T}^3))$ of the incompressible Euler equations, such that the advection-diffusion equation (1.1) with velocity u exhibits anomalous diffusion, precisely as in Theorem 1.2. Additionally, the family of all such weak solutions $u \in C^\alpha_{x,t}$ of the Euler equations is dense in the class of all solenoidal vector fields, with respect to the $C^0_t H^{-1}_x$ topology.*

The proof of Theorem 1.3 uses the concepts and analytical methodology of our work [8] and elegantly combines them with ideas from convex-integration/Nash schemes for the incompressible Euler equations (the technical implementation in [17] is of particular relevance). Notably, Burczak, Székelyhidi, and Wu succeed in [20] to interlace the backward energy cascade characteristic of convex-integration schemes, with the backward cascade of renormalized diffusivities which is underpinning iterated homogenization. At a technical level, [20] needs to contend with the fact that the correctors are not anymore explicit,[15] and with the fact that Isett's gluing

[14] See [8, Proposition 2.8].

[15] As opposed to [8], the solution of the cell problem in [20] is not explicit. Instead, the correctors are very well approximated by explicit functions. This is known to be sufficient, because the space-homogenization step is stable under small perturbations in the diffusion matrix.

technique [48] necessitates a more involved time-homogenization step.[16] Lastly, we note that the proof in [20] necessitates at least three space dimensions, due to the usage of non-intersecting Mikado flows [27]. It is conceivable that a two-dimensional result is attainable by combining the ideas in [8, 20] with the Newton–Nash scheme which was used to resolve the two-dimensional Onsager conjecture [45].

2. Background on quantitative homogenization

Before turning to the proof of Theorem 1.2, it is instructive to briefly recall the basic two-scale expansion in parabolic periodic homogenization, emphasizing the relation between the effective equation and the phenomenon of *advection-enhanced diffusion* (see Section 2.1), and then to briefly recall some of the basic themes of *quantitative homogenization*, highlighted in the context of the homogenization problem of a periodic shear flow, and leading to the Taylor diffusion formula (see Section 2.2).

2.1. The two-scale expansion in parabolic homogenization

We begin with a general presentation of the parabolic two-scale expansion in the periodic setting; this material is standard (see, e.g., [43, 49, 63]). We consider a $\mathbb{Z} \times \mathbb{Z}^d$-periodic, mean-zero, incompressible vector field $b = b(t, x)$ and for $\varepsilon, \kappa > 0$, consider the advection-diffusion equation

$$\partial_t \theta_\varepsilon - \kappa \Delta \theta_\varepsilon + \frac{1}{\varepsilon} b\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) \cdot \nabla \theta_\varepsilon = 0. \tag{2.1}$$

By introducing a stream matrix $\mathbf{s}$ for the vector field b (i.e., an anti-symmetric matrix such that $-\operatorname{div} \mathbf{s} = b$), the advection term in (2.1) may be expressed as a second-order term $\frac{1}{\varepsilon} b\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) \cdot \nabla \theta_\varepsilon = -\operatorname{div}\left(\mathbf{s}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) \nabla \theta_\varepsilon\right)$; in turn, this allows us to write (2.1) as

$$\partial_t \theta_\varepsilon - \operatorname{div}(\mathbf{a}^\varepsilon \nabla \theta_\varepsilon) = 0, \quad \text{where } \mathbf{a}^\varepsilon(t, x) := \kappa \mathrm{I}_d + \mathbf{s}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) = \mathbf{a}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right). \tag{2.2}$$

Note that (2.2) contains a diffusion matrix $\mathbf{a}^\varepsilon$ that oscillates fast (it depends on ε) and is not symmetric (since $\mathbf{s}$ is skew-symmetric). Homogenization theory quantifies the convergence as $\varepsilon \to 0$ (say with respect to the $L^2_{t,x}$ topology) of solutions θ_ε of (2.2) to solutions θ of the *effective equation*

$$\partial_t \theta - \operatorname{div}(\overline{\mathbf{a}} \nabla \theta) = 0, \tag{2.3}$$

[16] Time homogenization due to switching of flow maps was a difficulty already faced and resolved in [8]. The complication in [20] stems from the turning-on and turning-off of the fast oscillations in u, due to the so-called "gluing step".

where

- the homogenized/effective diffusion matrix $\overline{\mathbf{a}}$ is given by

$$\overline{\mathbf{a}}e = \langle\!\langle \mathbf{a}(e + \nabla\chi_e)\rangle\!\rangle, \quad e \in \mathbb{S}^{d-1},\ \mathbf{a} := \kappa \mathrm{I}_d + \mathbf{s}, \tag{2.4}$$

- $\langle\!\langle\cdot\rangle\!\rangle$ denotes the space-time average of a $\mathbb{Z}\times\mathbb{Z}^d$-periodic function,
- χ_e is the *corrector* with slope e; i.e., the unique $\mathbb{Z}\times\mathbb{Z}^d$-periodic solution of the *cell problem*

$$\partial_t\chi_e - \operatorname{div}(\mathbf{a}(e + \nabla\chi_e)) = 0, \quad \langle\!\langle\chi_e\rangle\!\rangle = 0. \tag{2.5}$$

The above definitions show that the symmetric part of the effective diffusion matrix $\overline{\mathbf{a}}$ is given by

$$\frac{1}{2}(\overline{\mathbf{a}} + \overline{\mathbf{a}}^t)_{ij} = \kappa\delta_{ij} + \kappa\,\langle\!\langle\nabla\chi_{e_i}\cdot\nabla\chi_{e_j}\rangle\!\rangle. \tag{2.6}$$

The second term on the right-hand side of (2.6) is positive (as a nonnegative definite matrix), and therefore the symmetric part of $\overline{\mathbf{a}}$ is larger than the original diffusion matrix $\kappa\mathrm{I}_d$. This effect is called the *enhancement of diffusivity due to advection*.[17]

The convergence $\|\theta_\varepsilon - \theta\|_{L^2_{t,x}} \to 0$ as $\varepsilon\to 0$, where θ_ε solves the microscopic diffusion equation (2.2) and θ is an appropriately chosen solution of the macroscopic or effective equation (2.3), is summarized informally as

"the operator $\partial_t - \operatorname{div}(\mathbf{a}^\varepsilon\nabla)$ *homogenizes* as $\varepsilon\to 0$ to the operator $\partial_t - \operatorname{div}(\overline{\mathbf{a}}\nabla)$."

A standard formalization of the above statement is achieved by proving that θ_ε is well approximated by its "two-scale expansion", which is defined in (2.7) below.

The *parabolic* two-scale expansion is constructed as follows. In addition to the corrector with slope e, χ_e (defined in (2.5)), and to the homogenized matrix $\overline{\mathbf{a}}$ (defined in (2.4)), we introduce the following:

- The space-homogenized matrix $\overline{\mathbf{a}}(t)$ is defined for $t\in\mathbb{R}$ by

$$\overline{\mathbf{a}}(t)e = \langle\mathbf{a}(t,\cdot)(e + \nabla\chi_e(t,\cdot))\rangle, \quad e\in\mathbb{S}^{d-1},$$

  where $\langle\cdot\rangle$ denotes the space average of a $\mathbb{Z}^d$-periodic function.
- This allows us to define the matrix

$$\mathbf{k}(t) := \int_0^t (\overline{\mathbf{a}}(t') - \overline{\mathbf{a}})dt',$$

  which by construction has zero mean in time, i.e., $\int_0^1 \mathbf{k}(t)dt = 0$.

[17] The enhancement of diffusivity due to advection, from the point of view of homogenization theory, is by now classical, and we do not offer a complete summary of the relevant literature here. We instead refer the reader to [41, 55, 63] and the references therein.

- Integrating (2.4) with respect to space alone implies that $\langle \partial_t \chi_e \rangle = 0$, and thus for each $t \in \mathbb{R}$ we may define the *time corrector* $h_e(t,\cdot)$ as the $\mathbb{Z}^d$-periodic $H^1(\mathbb{R}^d)$ solution of
$$-\Delta h_e(t,\cdot) = \partial_t \chi_e(t,\cdot), \quad \langle h_e(t,\cdot)\rangle = 0.$$
- The purpose of introducing the space-homogenized matrix $\overline{\mathbf{a}}(t)$ and the time corrector $h_e(t,\cdot)$ is to ensure that for each $t \in \mathbb{R}$ the vector field
$$g_e(t,\cdot) := \mathbf{a}(t,\cdot)(e + \nabla \chi_e(t,\cdot)) - \overline{\mathbf{a}}(t)e + \nabla h_e(t,\cdot)$$
satisfies $\operatorname{div} g_e(t,\cdot) = 0$ and $\langle g_e(t,\cdot)\rangle = 0$; in turn, this allows us to write $g_e(t,\cdot)$ as the divergence of a skew-symmetric matrix $\mathbf{m}_e(t,\cdot)$, which is the zero-mean solution of
$$-\Delta \mathbf{m}_{e,ij}(t,\cdot) = \partial_i g_{e,j}(t,\cdot) - \partial_j g_{e,i}(t,\cdot).$$

With this notation, the parabolic two-scale expansion is given by

$$\widetilde{\theta}_\varepsilon(t,x) := \theta(t,x) + \varepsilon \sum_{i=1}^d \partial_i \theta(t,x) \chi_{e_i}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) + \varepsilon^2 \sum_{i,j=1}^d \mathbf{k}_{ij}\left(\frac{t}{\varepsilon^2}\right)\partial_{ij}\theta(t,x), \tag{2.7}$$

where θ is a solution of (2.3). The function θ is determined only once we specify a domain and a parabolic boundary condition.

In order to show that $\widetilde{\theta}_\varepsilon$ is a good approximation of the solution θ_ε of (2.2) (again, by this we mean that one fixes a domain, and lets θ_ε match the values of θ on the corresponding parabolic boundary), we simply plug in the ansatz (2.7) into (2.2) and try to argue that the error we make is small. Using that θ solves (2.3), we arrive at

$$\partial_t \widetilde{\theta}_\varepsilon - \operatorname{div}(\mathbf{a}^\varepsilon \nabla \widetilde{\theta}_\varepsilon) = f_\varepsilon, \tag{2.8}$$

where the "error" f_ε is given explicitly as

$$\begin{aligned} f_\varepsilon(t,x) := \varepsilon \operatorname{div} &\left(\sum_{i=1}^d (h_{e_i} \mathrm{I}_d - \chi_{e_i}\mathbf{a} - \mathbf{m}_{e_i})\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right) \nabla \partial_{x_i}\theta(t,x)\right) \\ &+ \varepsilon \sum_{i=1}^d \left(\chi_{e_i}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right)\partial_{x_i}\partial_t\theta(t,x) + h_{e_i}\left(\frac{t}{\varepsilon^2}, \frac{x}{\varepsilon}\right)\partial_{x_i}\Delta\theta(t,x)\right) \\ &+ \varepsilon^2 \sum_{i,j=1}^d \mathbf{k}_{ij}\left(\frac{t}{\varepsilon^2}\right)\partial_{ij}\partial_t\theta(t,x). \end{aligned} \tag{2.9}$$

The proof of the statement "the operator $\partial_t - \operatorname{div}(\mathbf{a}^\varepsilon \nabla)$ *homogenizes* as $\varepsilon \to 0$ to the operator $\partial_t - \operatorname{div}(\overline{\mathbf{a}}\nabla)$", quantified by the convergence $\|\theta_\varepsilon - \theta\|_{L^2_{t,x}} \to 0$ as $\varepsilon \to 0$, now follows from (2.7), (2.8), and (2.9):

- (Good) bounds on the solution χ_e of the cell problem (2.5) imply (good) bounds for the time correctors h_e, the skew-symmetric matrix $\mathbf{m}_e$, and the matrix $\mathbf{k}$.
- Coupled with (good) derivative estimates for the macroscopic function θ, which solves the homogenized problem (2.3), the aforementioned bounds show that the forcing f_ε appearing on the right-hand side of (2.8) is $O(\varepsilon)$ (say in $L^2_t H^{-1}_x$).
- Additionally, note that by the linearity of (2.2), we have that

$$\partial_t(\widetilde{\theta}_\varepsilon - \theta_\varepsilon) - \operatorname{div}(\mathbf{a}^\varepsilon \nabla(\widetilde{\theta}_\varepsilon - \theta_\varepsilon)) = f_\varepsilon, \tag{2.10}$$

  and so by the standard parabolic energy estimate, $\widetilde{\theta}_\varepsilon - \theta_\varepsilon$ is $O(\varepsilon)$ (say in $L^2_{t,x}$).
- Since θ_ε match the values of θ on the parabolic boundary of the space-time domain, and $\widetilde{\theta}_\varepsilon$ is an $O(\varepsilon)$ perturbation from θ, the difference $\widetilde{\theta}_\varepsilon - \theta_\varepsilon$ is at most $O(\varepsilon)$ on the boundary (say in $L^\infty_{t,x}$).
- To conclude, note that from (2.7) and the above-mentioned bounds, we directly obtain that $\widetilde{\theta}_\varepsilon - \theta$ is $O(\varepsilon)$ (say in $L^2_{t,x}$) by standard energy estimates.

The above items sweep under the rug the fact that the aforementioned bounds depend in a crucial way on the ellipticity constant of the matrix $\mathbf{a} = \kappa \mathrm{I}_d + \mathbf{s}$, on the ellipticity ratio of $\mathbf{a}$, which quantifies the size of the skew part $\mathbf{s}$ to the size of the symmetric part $\kappa \mathrm{I}_d$. They also depend on the size of the derivatives of the macroscopic function θ, since these appear in the definition of f_ε in (2.9). These considerations are crucial for the quantitative homogenization estimates which we discuss next.

2.2. Quantitative homogenization for shear flows: Taylor diffusion

Stated as in Section 2.1, it appears that homogenization relies on *asymptotic scale separation* (the limit $\varepsilon \to 0$), which is a feature inconsistent with turbulence models.[18] Asymptotic scale separation is also inconsistent with the construction of the vector field u in Section 3.1. There, ε represents the ratio of two consecutive length scales, which for a fixed parameter m in the iteration corresponds to $0 < \frac{\varepsilon_{m+1}}{\varepsilon_m} < 1$. Note however that for $m = 100$, say, the ratio $\frac{\varepsilon_{101}}{\varepsilon_{100}}$ is fixed and cannot be sent to 0. This is why *quantitative homogenization* methods need to, and should, be used! Rather than arguing that some quantities converge to others in the limit of infinite scale separation, one needs to precisely quantify the length scales and time scales on which homogenization may be observed to have occurred, modulo small errors (which then need to be summed up across scales). Fortunately, quantitative homogenization rates for periodic coefficient fields are very explicit and quite well understood (see [49,65] for instance).

[18] See, e.g., the criticisms in [42, p. 225] and [55, p. 304].

We will next highlight this quantitative homogenization perspective for the operator $\partial_t - \kappa\Delta + u \cdot \nabla$, when u is a *two-dimensional shear flow*. For parameters $a, \varepsilon > 0$, let us consider

$$u = u_{a,\varepsilon} = 2\pi a\varepsilon \cos\left(\frac{2\pi x_2}{\varepsilon}\right) e_1, \tag{2.11}$$

so that $0 < \varepsilon \ll 1$ is the scale of spatial oscillation (note that u is time independent), and $a \gg 1$ is proportional to the Lipschitz norm of u (note that $\|\nabla u\|_{L^\infty(\mathbb{T}^2)} = 4\pi^2 a$). Since we are two space dimensions, $\langle u \rangle = 0$, and $\operatorname{div} u = 0$, we may write

$$u_{a,\varepsilon} = \nabla^\perp \psi_{a,\varepsilon}, \quad \text{where } \psi_{a,\varepsilon} = a\varepsilon^2 \sin\left(\frac{2\pi x_2}{\varepsilon}\right). \tag{2.12}$$

Associated to the stream function $\psi_{a,\varepsilon}$, we may associate the anti-symmetric matrix $\psi_{a,\varepsilon}\mathrm{I}_2^\perp$, where $\mathrm{I}_2^\perp$ is the (rotation) matrix with components $(\mathrm{I}_2^\perp)_{11} = (\mathrm{I}_2^\perp)_{22} = 0$ and $(\mathrm{I}_2^\perp)_{21} = -(\mathrm{I}_2^\perp)_{12} = 1$. Then, in analogy to (2.2), we may rewrite

$$\partial_t - \kappa\Delta + u_{a,\varepsilon} \cdot \nabla = \partial_t - \operatorname{div}(\mathbf{a}^\varepsilon \nabla) \tag{2.13}$$

where for compatibility of notation, we have defined

$$\mathbf{a}^\varepsilon(x) := \kappa \mathrm{I}_2 + \psi_{a,\varepsilon}(x)\mathrm{I}_2^\perp = \mathbf{a}\left(\frac{x}{\varepsilon}\right), \quad \text{with } \mathbf{a}(y) := \kappa \mathrm{I}_2 + a\varepsilon^2 \sin(2\pi y_2)\mathrm{I}_2^\perp. \tag{2.14}$$

Here, y is a placeholder for the fast space variable $\frac{x}{\varepsilon}$. In the example provided by (2.13)–(2.14), we may take advantage of the fact that $\mathbf{a}$ is independent of time and of the first space coordinate to deduce that the correctors χ_e (which solve the cell problem (2.5)) are also independent of time and of the first space coordinates. In fact, we may quickly verify that the solution of (2.5) is given by

$$\chi_{e_1}(y) = \frac{a\varepsilon^2}{2\pi\kappa} \cos(2\pi y_2), \quad \text{and} \quad \chi_{e_2}(y) = 0, \tag{2.15}$$

which is clearly $\mathbb{Z} \times \mathbb{Z}^2$-periodic. Plugging (2.15) into (2.4), we obtain

$$\overline{\mathbf{a}} = \kappa \mathrm{I}_2 + \begin{pmatrix} \frac{a^2\varepsilon^4}{2\kappa} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \kappa + \frac{1}{2}\kappa \cdot \left(\frac{a\varepsilon^2}{\kappa}\right)^2 & 0 \\ 0 & \kappa \end{pmatrix}. \tag{2.16}$$

The formula for the effective diffusion matrix $\overline{\mathbf{a}}$ given in (2.16) is called the *Taylor diffusion formula* [69] (specialized to the case of the shear flow in (2.11)); see also the work of Aris [3].

The Taylor diffusion formula for $\overline{\mathbf{a}}$ obtained in (2.16) is notable because $\overline{\mathbf{a}}$ is diagonal (note that $\mathbf{a}$ had a non-constant anti-symmetric part), by definition it is constant in space and in time, and most notably, the enhancement of diffusion only occurs in

the $e_1 \otimes e_1$ entry (a remnant of the fact that the u in (2.11) points in the e_1 direction but is independent of x_1). We also note that the quantity $\frac{a\varepsilon^2}{\kappa}$ which appears squared as a factor for the strength of the enhancement is dimensionless: a is the Lipschitz norm of a velocity field, so it scales as $(\text{time})^{-1}$, the parameter ε has units of (length), while κ is a molecular diffusivity, so it has units of $(\text{length})^2 \cdot (\text{time})^{-1}$.

At this point, we have simply derived an explicit formula for the homogenized matrix $\overline{\mathbf{a}}$, but we have not said anything *quantitative* about the statement "the operator $\partial_t - \operatorname{div}(\mathbf{a}^\varepsilon \nabla)$ *homogenizes* as $\varepsilon \to 0$ to the operator $\partial_t - \operatorname{div}(\overline{\mathbf{a}}\nabla)$." The following lemma provides such a "sample" quantitative result; we do not claim that the following result is the sharpest possible, it is not the most general either; we present it here for pedagogical value, as it provides useful intuition about the quantitative bounds that may be extracted from the parabolic two-scale expansion.

Lemma 2.1 (Quantitative homogenization and Taylor diffusion). *Let $\mathbf{a}^\varepsilon$ be defined by (2.14), and let $\overline{\mathbf{a}}$ be defined by (2.16). The operator $\partial_t - \operatorname{div}(\mathbf{a}^\varepsilon \nabla)$ homogenizes to the operator $\partial_t - \operatorname{div}(\overline{\mathbf{a}}\nabla)$, on*

$$\text{time scales} \gtrsim a^{-1} \cdot \left(1 + \frac{a\varepsilon^2}{\kappa}\right) \quad \text{and} \quad \text{space scales} \gtrsim \varepsilon \cdot \left(1 + \frac{a\varepsilon^2}{\kappa}\right). \tag{2.17}$$

More precisely:

- *Define θ to be the $\mathbb{T}^2$-periodic solution of the Cauchy problem for $\partial_t \theta - \operatorname{div}(\overline{\mathbf{a}}\nabla\theta) = 0$ (cf. (2.3)), with smooth initial datum $\theta|_{t=0} = \theta_{\text{in}}$.*
- *For $R > 0$, define the* parabolic cylinders adapted to $\overline{\mathbf{a}}$, *as*

$$Q_R = Q_R(t_0, x_0) = (t_0, x_0) + \left(-\frac{R^2}{\kappa}, 0\right] \times E_R(0),$$

  where E_R is the ellipsoid defined by $\{x; \kappa^{1/2}|\overline{\mathbf{a}}^{-1/2}x| \le R\} = \kappa^{-1/2}\overline{\mathbf{a}}^{1/2}B_R(0)$. In our case, due to (2.16), this ellipsoid has axes of length $\sim R\big(1 + \frac{a\varepsilon^2}{\kappa}\big)$ in the e_1 direction, and $\sim R$ in the e_2 direction.
- *For every such parabolic cylinder Q_R with $Q_{2R} \subseteq [0,1] \times \mathbb{T}^2$, let θ_ε be the solution of $\partial_t\theta_\varepsilon - \operatorname{div}(\mathbf{a}^\varepsilon\nabla\theta_\varepsilon) = 0$ in Q_R, with (parabolic) boundary data $\theta_\varepsilon \equiv \theta$ on $\partial_{\text{par}}Q_R$.*
- *Then, we have that*

$$\|\theta_\varepsilon - \theta\|_{L^\infty_t \underline{L}^2_x(Q_R)} \lesssim \left(1 + \frac{a\varepsilon^2}{\kappa}\right)\frac{\varepsilon}{R}\|\theta\|_{\underline{L}^2_{x,t}(Q_{2R})}. \tag{2.18}$$

Here, $L^\infty_t \underline{L}^2_x(Q_R)$ and $\underline{L}^2_{x,t}(Q_R)$ denote the volume-normalized norm (e.g., integrals $\int_{Q_R}$ are replaced by $\fint_{Q_R}$). In particular, when $R \gg \varepsilon\big(1 + \frac{a\varepsilon^2}{\kappa}\big)$, we see that

$\|\theta_\varepsilon - \theta\|_{L^\infty_t \underline{L}^2_x(Q_R)} \ll \|\theta\|_{\underline{L}^2_{x,t}(Q_{2R})}$, *which matches the heuristic claim regarding space scales in* (2.17).

Brief proof of Lemma 2.1. We wish to appeal to the parabolic two-scale expansion described in Section 2.1. We have already computed the correctors with slope e, namely χ_e, in (2.15). Since $\mathbf{a} = \mathbf{a}(y_2)$ and $\chi_e = \chi_e(y_2)$ are both independent of t, the space-homogenized matrix $\overline{\mathbf{a}}(t)$ is time independent and is equal to $\overline{\mathbf{a}}$ for all $t \in \mathbb{R}$. Hence, the matrix $\mathbf{k}(t) = 0$ for all $t \in \mathbb{R}$, and the time correctors also vanish identically, $h_e(t,\cdot) = 0$ for all $t \in \mathbb{R}$. Then, it follows that $g_{e_1}(t,y) = -\frac{a^2\varepsilon^4}{2\kappa}\cos(4\pi y_2)e_1$ and $g_{e_2}(t,y) = -a\varepsilon^2\sin(2\pi y_2)e_1$, so that the skew-symmetric matrices $\mathbf{m}_e(t,\cdot)$ are given explicitly by $\mathbf{m}_{e_1}(t,y) = \frac{a^2\varepsilon^4}{8\pi\kappa}\sin(4\pi y_2)\mathrm{I}_2^\perp$ and $\mathbf{m}_{e_2}(t,y) = -\frac{a\varepsilon^2}{2\pi}\cos(2\pi y_2)\mathrm{I}_2^\perp$.

Using the explicit computations in the previous paragraph, equation (2.3) for θ may also be written as

$$\partial_t\theta - \kappa\left(1 + \frac{a^2\varepsilon^4}{2\kappa^2}\right)\partial_{11}\theta - \kappa\partial_{22}\theta = 0. \tag{2.19}$$

Moreover, the parabolic two-scale expansion is then given by (2.7) as

$$\widetilde{\theta}_\varepsilon(t,x) := \theta(t,x) + \varepsilon\partial_1\theta(t,x)\chi_{e_1}\left(\frac{x_2}{\varepsilon}\right) = \theta(t,x) + \frac{a\varepsilon^3}{2\pi\kappa}\cos\left(\frac{2\pi x_2}{\varepsilon}\right)\partial_1\theta(t,x). \tag{2.20}$$

Additionally, the "error" term f_ε defined in (2.9) may be computed explicitly as

$$\begin{aligned}
f_\varepsilon(t,x) &= -\varepsilon\,\mathrm{div}\left((\chi_{e_1}\mathbf{a} + \mathbf{m}_{e_1})\left(\frac{x_2}{\varepsilon}\right)\nabla\partial_1\theta(t,x) + \mathbf{m}_{e_2}\left(\frac{x_2}{\varepsilon}\right)\nabla\partial_2\theta(t,x)\right) \\
&\quad + \varepsilon\chi_{e_1}\left(\frac{x_2}{\varepsilon}\right)\partial_1\partial_t\theta(t,x) \\
&= -\varepsilon\,\mathrm{div}\left(\frac{a\varepsilon^2}{2\pi}\cos\left(\frac{2\pi x_2}{\varepsilon}\right)\nabla\partial_1\theta(t,x) - \frac{a^2\varepsilon^4}{8\pi\kappa}\sin\left(\frac{4\pi x_2}{\varepsilon}\right)\nabla^\perp\partial_1\theta(t,x)\right) \\
&\quad + \varepsilon\,\mathrm{div}\left(\frac{a\varepsilon^2}{2\pi}\cos\left(\frac{2\pi x_2}{\varepsilon}\right)\nabla^\perp\partial_2\theta(t,x)\right) \\
&\quad + \varepsilon\partial_1\left(\frac{a\varepsilon^2}{2\pi\kappa}\cos\left(\frac{2\pi x_2}{\varepsilon}\right)\partial_t\theta(t,x)\right) \\
&=: \mathrm{div}\, v_\varepsilon(t,x),
\end{aligned} \tag{2.21}$$

for an implicitly defined vector v_ε, and where we recall that $\nabla^\perp = \mathrm{I}_2^\perp\nabla = (-\partial_2, \partial_1)$.

In order to conclude, we at last perform some PDE estimates. First, we need some standard bounds for θ which are just standard pointwise estimates for the heat

equation after a change of variables. These say that, for every $\ell \in \{0, 1, 2, \ldots\}$,

$$(\kappa^{-\frac{1}{2}} R)^{\ell} \|(\overline{\mathbf{a}}^{\frac{1}{2}} \nabla)^{\ell} \theta\|_{L^{\infty}_{x,t}(Q_R)} + (\kappa^{-\frac{1}{2}} R)^{2\ell} \|\partial_t^{\ell} \theta\|_{L^{\infty}_{x,t}(Q_R)} \leq C_{\ell} \|\theta\|_{\underline{L}^2_{x,t}(Q_{2R})}. \tag{2.22}$$

Next, we record a bound which is immediate from (2.21) and standard global energy estimates for (2.19), namely

$$\begin{aligned} \|v_\varepsilon\|_{L^2_{x,t}(Q_R)} &\lesssim \left(a\varepsilon^3 + \frac{a^2\varepsilon^5}{\kappa}\right) \|\partial_1 \nabla \theta\|_{L^2_{x,t}(Q_R)} + a\varepsilon^3 \|\partial_{22}\theta\|_{L^2_{x,t}(Q_R)} \\ &\quad + \frac{a\varepsilon^3}{\kappa} \|\partial_t \theta\|_{L^2_{x,t}(Q_R)} \\ &\lesssim \frac{\kappa\varepsilon}{R^2} \left(1 + \frac{a\varepsilon^2}{\kappa}\right) \|\theta\|_{L^2_{x,t}(Q_{2R})}, \end{aligned} \tag{2.23}$$

where the implicit constant in the $\lesssim$ symbol is universal. With this bound, we return to (2.10) and recall the parabolic problem

$$\partial_t(\widetilde{\theta}_\varepsilon - \theta_\varepsilon) - \operatorname{div}(\mathbf{a}^\varepsilon \nabla(\widetilde{\theta}_\varepsilon - \theta_\varepsilon)) = \operatorname{div} v_\varepsilon. \tag{2.24}$$

Note that (2.20) gives $(\widetilde{\theta}_\varepsilon - \theta_\varepsilon)(t,x) = \frac{a\varepsilon^3}{2\pi\kappa} \cos\left(\frac{2\pi x_2}{\varepsilon}\right) \partial_1 \theta(t,x)$ for $(t,x) \in \partial_{\mathrm{par}} Q_R$, and so $\widetilde{\theta}_\varepsilon - \theta_\varepsilon$ does not vanish identically on $\partial_{\mathrm{par}} Q_R$; instead, we have

$$\|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\|_{L^{\infty}_{x,t}(\partial_{\mathrm{par}} Q_R)} \lesssim \frac{a\varepsilon^3}{\kappa} \|\partial_1 \theta\|_{L^\infty(Q_R)} \leq \frac{\varepsilon}{R} \|\theta\|_{\underline{L}^2_{x,t}(Q_{2R})}.$$

The right-hand side is bounded by that of (2.18). Therefore, by the maximum principle, we can correct this discrepancy at the boundary (and just ignore it for our purposes).

We continue therefore by pretending that $\widetilde{\theta}_\varepsilon - \theta_\varepsilon$ vanishes on $\partial_{\mathrm{par}} Q_R$ and return to (2.24) to perform a standard energy estimate. We obtain

$$\|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\|^2_{L^\infty_t L^2_x(Q_R)} + \kappa \|\nabla(\widetilde{\theta}_\varepsilon - \theta_\varepsilon)\|^2_{L^2_{x,t}(Q_R)} \lesssim \kappa^{-1} \|v_\varepsilon\|^2_{L^2_{x,t}(Q_R)}.$$

We continue without considering the gradient term on the left-hand side and appeal to (2.20), (2.22), and (2.23) to obtain

$$\begin{aligned} \|\widetilde{\theta}_\varepsilon - \theta\|_{L^\infty_t \underline{L}^2_x(Q_R)} &\leq \|\widetilde{\theta}_\varepsilon - \theta_\varepsilon\|_{L^\infty_t \underline{L}^2_x(Q_R)} + \|\theta_\varepsilon - \theta\|_{L^\infty_t \underline{L}^2_x(Q_R)} \\ &\lesssim \kappa^{-\frac{1}{2}} (\kappa^{-1} R^2)^{\frac{1}{2}} \|v_\varepsilon\|_{\underline{L}^2_{x,t}(Q_R)} + \frac{\varepsilon}{R} \|\theta\|_{\underline{L}^2(Q_{2R})} \\ &\lesssim \left(1 + \frac{a\varepsilon^2}{\kappa}\right) \frac{\varepsilon}{R} \|\theta\|_{\underline{L}^2_{x,t}(Q_{2R})}. \end{aligned}$$

This concludes the proof of (2.18). ∎

3. An outline of the proof of Theorem 1.2

The proof consists of two parts: the construction of multi-scale vector field u (see Section 3.1), and the analysis of the passive scalars θ^κ via quantitative homogenization (see Sections 3.2–3.4). Throughout this section, it is instructive to keep in mind the intuition gained from the discussion in Section 2.

3.1. Construction of the vector field u

The proof of Theorem 1.2 is the most difficult in the lowest nontrivial dimension, namely $d = 2$, for the same geometric reasons that the Onsager conjecture was hardest to establish in the two-dimensional case [45]. We present the $d = 2$ construction next.

The advantage of the case $d = 2$ is notational: zero-mean divergence-free vector fields u are given as the rotated gradient of a scalar stream function ϕ (with associated anti-symmetric matrix $\mathrm{I}_2^\perp \phi$); as such, we avoid the notational care one must use when handling vectors, or when working with the curl operator.

We recursively construct a sequence of smooth, periodic vector fields $u_m = \nabla^\perp \phi_m$, and then define $u = \lim_{m\to\infty} u_m$; here, the limit is taken with respect to the $C^\alpha_{x,t}$ topology. We initialize the construction with $\phi_0 = 0$, so that $u_0 = 0$. For any $m \geq 1$, having constructed the smooth, periodic, incompressible vector field u_{m-1} with associated stream function ϕ_{m-1}, the induction step $m - 1 \mapsto m$ consists in the construction of a suitable stream function "increment" $\phi_m - \phi_{m-1}$. For this purpose, we make the following initial parameter choices:

- Let $\{\varepsilon_m\}_{m\geq 0}$ be a decreasing sequence of scales, with $\varepsilon_m \to 0$ as $m \to \infty$. We set $\varepsilon_0 = 1$, and $\varepsilon_1 = \Lambda^{-1}$ for a large integer $\Lambda \gg 1$, chosen at the end of the proof. At each step $m \geq 1$, we view ε_m as the "fast scale" of oscillation and ε_{m-1} as the "slow scale".[19] Based on previous experience with multi-scale constructions in turbulence (see, e.g., [29]), we expect that it will be technically (very!) convenient to consider the sequence ε_m to decay "slightly super-exponentially", meaning that
$$\varepsilon_m \simeq \varepsilon_{m-1}^q \quad \text{for some } 0 < q - 1 \ll 1 \tag{3.1}$$
for all $m \geq 2$. Here, the "$\simeq$" is not an "$=$" sign because we need ε_m^{-1} to be an integer for all $m \geq 0$ (in order to enforce $\mathbb{T}^2$-periodicity). How small the positive parameter $q - 1$ needs to be taken is chosen as part of the proof.

[19] The discussion in Section 2.2 refers to the fast scale ε and the slow scale 1. As such, in this new context it is the ratio $\frac{\varepsilon_m}{\varepsilon_{m-1}} < 1$ which plays the role of the small parameter ε in Section 2.2.

- We let $a_m \gg 1$ denote the Lipschitz norm (up to universal constants) of the stream function increment $\phi_m - \phi_{m-1}$. If we think of the increment $\phi_m - \phi_{m-1}$ as oscillating at frequency ε_m, then in order to ensure uniform boundedness of the velocity fields $\{u_m\}_{m\geq 0}$ in $C_t^0 C_x^\alpha$, it is natural and necessary to define

$$a_m := \varepsilon_m^{\alpha-1} \quad \text{for all } m \geq 0. \tag{3.2}$$

Note that since $\alpha < 1$, we have $a_m \to \infty$ as $m \to \infty$.

3.1.1. A naive attempt for all $0 < \alpha < 1$. With $\{\varepsilon_m\}_{m\geq 0}$ and $\{a_m\}_{m\geq 0}$ fixed according to (3.1)–(3.2), and keeping in mind the result of Lemma 2.1, a "naive" attempt at defining the stream function increment $\phi_m - \phi_{m-1}$ is as follows.

Based on (2.12), we are tempted to declare that the stream function increment is given by $a_m \varepsilon_m^2 \sin\left(\frac{2\pi x_2}{\varepsilon_m}\right)$. Note however that in this case the homogenized matrix (2.16) presents an enhancement only in the $e_1 \otimes e_1$ component, which is undesirable. The natural fix to this issue is to alternate the direction of shearing, by considering a stream function which sequentially in time alternates the stream functions $a_m \varepsilon_m^2 \sin\left(\frac{2\pi x_2}{\varepsilon_m}\right)$ and $a_m \varepsilon_m^2 \sin\left(\frac{2\pi x_1}{\varepsilon_m}\right)$, "glued" by a partition of unity in time, with characteristic time scale proportional to some $0 < \tau_m \ll 1$, to be determined. We would like, however, these shears acting in different directions not to interact/interfere with each other (see also [17,27,48]), and hence we leave room in which no shearing occurs, whenever we switch directions. As such, we define

$$\Psi_k(x) := \begin{cases} \sin(2\pi x_1), & k \equiv 1 \bmod 4, \\ \sin(2\pi x_2), & k \equiv 3 \bmod 4, \\ 0, & k \equiv 0 \bmod 2, \end{cases}$$

and we attempt to declare

$$\phi_m(t,x) - \phi_{m-1}(t,x) := \sum_{k\in\mathbb{Z}} \zeta\left(\frac{t}{\tau_m} - k\right) a_m \varepsilon_m^2 \Psi_k\left(\frac{x}{\varepsilon_m}\right), \tag{3.3}$$

where $\mathbf{1}_{[-\frac{1}{3},\frac{1}{3}]} \leq \zeta \leq \mathbf{1}_{[-\frac{2}{3},\frac{2}{3}]}$ is a C^∞ smooth unit-scale bump function with $\int_{\mathbb{R}} \zeta^2(t)dt = 1$, and which induces a partition of unity via $\sum_{k\in\mathbb{Z}} \zeta(\cdot - k) \equiv 1$.

By construction, the stream function increment in (3.3) is $4\tau_m$-periodic in time, and since the formula (2.4) for the homogenized matrix involves averaging in time, we expect that enhancement will occur in both the $e_1 \otimes e_1$ and $e_2 \otimes e_2$ components, as desired, with the "amount of enhancement" being 25% of the one in (2.16). Note however that in order to see this enhancement, Lemma 2.1 dictates that we "zoom out" in both time and space; in turn, this imposes restrictions on τ_m and ε_m.

The first condition in (2.17) dictates a constraint on the time scale τ_m, which determines the periodicity in time (inverse frequency) of the stream function increment in (3.3). We need the amount of time that each of the stream functions Ψ_k is active to be long enough to allow for homogenization to be observed; this dictates

$$\tau_m \gg a_m^{-1} \cdot \left(1 + \frac{a_m \varepsilon_m^2}{\kappa_m}\right) = a_m^{-1} + \frac{\varepsilon_m^2}{\kappa_m} \geq \frac{\varepsilon_m^2}{\kappa_m}. \tag{3.4}$$

The second condition in (2.17) dictates that the length scale corresponding to "slow oscillations" (meaning, the scale of oscillation of the "slow" stream function ϕ_{m-1}), is large enough for homogenization to be observed; this dictates

$$\varepsilon_{m-1} \gg \varepsilon_m \cdot \left(1 + \frac{a_m \varepsilon_m^2}{\kappa_m}\right) \geq \frac{a_m \varepsilon_m^3}{\kappa_m}. \tag{3.5}$$

With (3.4)–(3.5), Lemma 2.1 indicates that upon "zooming out" from scale ε_m to scale ε_{m-1},[20] the enhancement of diffusion is given by $\kappa \mathrm{I}_2 \mapsto \kappa \mathrm{I}_2 + \frac{1}{8}\kappa \cdot \left(\frac{a_m \varepsilon_m^2}{\kappa}\right)^2 \mathrm{I}_2$; the division by 4 when compared to (2.16) is due to time averaging.

If the above-described scenario can be justified rigorously, iteratively in m, then we can set up the parameters $\{\varepsilon_m\}_{m\geq 0}$, $\{a_m\}_{m\geq 0}$, and $\{\tau_m\}_{m\geq 0}$ such that upon "coarse graining" at scale ε_m,[21] the advection-diffusion equation (1.1) with vector field $u = \nabla^\perp \phi$, where ϕ is defined by summing over $m \geq 0$ the increments in (3.3), experiences an effective diffusivity κ_m (see Section 3.2), which satisfies the recursion relation

$$\kappa_{m-1} := \kappa_m + \frac{1}{8}\kappa_m \cdot \left(\frac{a_m \varepsilon_m^2}{\kappa_m}\right)^2. \tag{3.6}$$

The recursion (3.6) is to be "initialized" with $\kappa_M = \kappa$ for an integer $M \gg 1$ which is to be determined such that ε_M is proportional to the dissipation length-scale associated with the advection-diffusion equation (1.1) with a $C^\alpha_{x,t}$ vector field u. With this initialization, the recursion (3.6) is to be solved backwards: $M \to M-1 \to \cdots \to m \to m-1 \to \cdots \to 0$, with the hope that $\kappa = \kappa_M < \kappa_{M-1} < \cdots < \kappa_m < \kappa_{m-1} < \cdots < \kappa_0$, and that $\kappa_0 = O(1)$ as $\kappa \to 0$.

Let us next inquire whether the parameter choices (3.1), (3.2), together with the recursion (3.6), are consistent with the constraints (3.4), (3.5), and with the fact that

[20] It is this $m \mapsto m-1$ homogenization step which goes from small scales to larger ones, and this is the motivation for referring to the cascade of effective diffusivities as being "backwards".

[21] See [4, 6] for a general theory of *quantitative theory of coarse graining* for elliptic and parabolic PDEs with multi-scale oscillating coefficients.

$\kappa_0 = O(1)$. Assuming for simplicity that $\kappa = \kappa_M = \varepsilon_M^{1+\alpha+\gamma}$ for some $M \geq 1$,[22] then it is not hard to verify that the relation

$$\kappa_m \simeq \varepsilon_m^{1+\alpha+\gamma}, \quad \text{where } \gamma = (q-1)\frac{1+\alpha}{q+1}, \tag{3.7}$$

for all $M \geq m \geq 0$, is consistent with (3.6), where we recall that $0 < q - 1 \ll 1$. Moreover, setting $m = 0$ in (3.7) yields $\kappa_0 \simeq 1$, as desired, independently of the value of $M \gg 1$ (and hence of $\kappa \ll 1$).

The constraint (3.4) is then used to pick a time scale that satisfies

$$\tau_m \gg a_m^{-1} + \varepsilon_m^2 \kappa_m^{-1} = \varepsilon_m^{1-\alpha} + \varepsilon_m^{1-\alpha-\gamma} \simeq \varepsilon_m^{1-\alpha-\gamma}. \tag{3.8}$$

Such a choice is permissible[23] if $\tau_m \leq 1$ for all $m \geq 0$, i.e., as soon as $1 - \alpha - \gamma > 0$. In light of the definition of γ in (3.7), this is indeed possible *for all values* $\alpha \in (0, 1)$, by letting $1 < q < \frac{1}{\alpha}$. Note that as $\alpha \to 1^-$, this implies $q \to 1^+$.

3.1.2. What goes wrong with this construction? The discussion in Section 3.1.1 indicates that for any $\alpha \in (0, 1)$, we may construct a $C_{x,t}^{\alpha}$ vector field (with stream function obtained by summing (3.3) in $m \geq 1$), such that the sequence of renormalized diffusivities are compatible with the recursion relation (3.6), and the space-time separation is compatible with (3.4) and (3.5), as dictated by Lemma 2.1; additionally, the amount of diffusion experienced by the passive scalar when "coarse grained" at $O(1)$ scales is itself $O(1)$, independently of the value of the molecular diffusion $\kappa \ll 1$.

In discussing the above heuristic, *a subtle (and fatal) point was overlooked*: the Taylor-dispersion formula/homogenization of a shear flow discussed in Lemma 2.1 *assumed that the shear flow* $u = u_{a,\varepsilon}$ *had zero mean* when averaged over one periodic cell of the fast variable $y = \frac{x}{\varepsilon}$. Indeed, the $\cos(2\pi y_2)$ present in the definition of $u_{a,\varepsilon}$ (see (2.11)) ensures that this velocity field has zero mean over $\mathbb{T}^2(dy)$. It was the very fact that $\langle u_{a,\varepsilon} \rangle = 0$ that allowed us to write this velocity field as the rotated

[22] That is, we are letting $\kappa = \kappa_M \to 0$ along the designated sequence $\varepsilon_M^{1+\alpha+\gamma} \to 0$, as $M \to \infty$.

[23] The choice of a_m in (3.2), together with the definition (3.3), implies the uniform-in-m space regularity estimate of the velocity increments: $\|u_m - u_{m-1}\|_{C_t^0 C_x^{\alpha}} = \|\nabla^{\perp}(\phi_m - \phi_{m-1})\|_{C_t^0 C_x^{\alpha}} \simeq a_m \varepsilon_m^{1-\alpha} \simeq 1$. On the other hand, if we wish to estimate uniform-in-m time regularity of the velocity increments, we see that $\|u_m - u_{m-1}\|_{C_t^{\alpha} C_x^0} = \|\nabla^{\perp}(\phi_m - \phi_{m-1})\|_{C_t^{\alpha} C_x^0} \simeq \tau_m^{-\alpha} a_m \varepsilon_m \simeq \tau_m^{-\alpha} \varepsilon_m^{\alpha} \ll \varepsilon_m^{\alpha(\alpha+\gamma)}$, where in the last inequality we have used (3.8). For any $\alpha > 0$ and $q > 1$, we thus obtain $C_t^{\alpha} C_x^0$-regularity of our vector field for free.

gradient of a periodic stream function $\psi_{a,\varepsilon}$ (see (2.12)), which was at the core of our homogenization analysis.

Returning to the construction in Section 3.1.1, for each $m \geq 1$ we note that the velocity field u_m *does not have zero mean when averaged over a periodic cell of the fast variable* $\frac{x}{\varepsilon_m}$. Indeed, the "macroscopic" velocity field u_{m-1} *is essentially a constant* from the point of view of the "microscopic" velocity increment $u_m - u_{m-1}$, which is $(\varepsilon_m \mathbb{T})^2$-periodic. To be precise, recursion (3.3) shows that we have $\fint_{(\varepsilon_m \mathbb{T})^2} (u_m - u_{m-1})(t,\cdot) = 0$ for all $t \in \mathbb{R}$, while for any $1 \leq j \leq m-1$, we have[24] $\fint_{(\varepsilon_m \mathbb{T})^2} (u_{m-j} - u_{m-j-1})(t,\cdot) \simeq \varepsilon_{m-j}^{\alpha} \neq 0$ for $t \in \bigcup_{k \in 2\mathbb{Z}+1} \big[\tau_{m-j}\big(k - \frac{1}{3}\big), \tau_{m-j}\big(k + \frac{1}{3}\big)\big]$. Note that $\varepsilon_{m-j}^{\alpha} \gg \varepsilon_m^{\alpha}$ for any $1 \leq j \leq m-1$, so not only is the macroscopic velocity field u_{m-1} essentially a constant from the point of view of microscopic oscillations, in relative terms it is a *large constant*.

The issue is that the presence of a large constant drift velocity, superimposed on a zero-mean shear flow, is able to remove most of the enhancement of the diffusivity; this is so because generically this constant drift will point in an irrational/Diophantine direction, causing *averaging to overtake homogenization*. This is explained quantitatively in [8, Appendix A]. In the physics literature, this phenomenon—that a non-constant mean flow interferes and potentially destroys Taylor diffusion—is termed the "sweeping effect" [12,38,55]. To sum up, the construction presented in Section 3.1.1 does not work, at least not "as is".

3.1.3. Small-scale features are transported by the large-scale flow: $\alpha < \frac{1}{3}$. To overcome the issue raised at the end of Section 3.1.2, we recall one of the cornerstones of phenomenological turbulence theories, the observation that microscopic "eddies" are transported by the surrounding macroscopic flow, for a suitable amount of time [42].

Roughly speaking, the idea put forth in [8][25] is that the microscopic/fast shear flows should *not* be naively superimposed on top of the existing macroscopic/slow shear flows (as (3.3) dictates); *instead*, the microscopic/fast shear flows should be added in a frame of reference that moves according to the macroscopic/slow velocity field (which is to say, in the Lagrangian coordinates of this slow velocity), so that in this frame the shear flows continue to have zero mean (to leading order).

[24] In this case, we employ $\fint_0^{\varepsilon_m} \cos\big(\frac{2\pi x_1}{\varepsilon_{m-j}}\big) dx_1 = \varepsilon_{m-j}^{\alpha+1} (2\pi\varepsilon_m)^{-1} \sin\big(\frac{2\pi\varepsilon_m}{\varepsilon_{m-j}}\big) \simeq \varepsilon_{m-j}^{\alpha}$, since $\frac{\varepsilon_m}{\varepsilon_{m-j}} \ll 1$.

[25] In the context of the Onsager conjecture, this idea (a proto-version of it) was introduced first in [29] and played a major role in all subsequent works, leading to [17,44,45,48,61]. In the context of homogenization of advection-diffusion equations, a version of this idea appeared in [57].

For this purpose, for each $[0,1]\times\mathbb{T}^2$-periodic, divergence-free velocity field u_m, we define the associated flow maps $X_m = X_m(t,x;s)$ by solving the ODE

$$\frac{d}{dt}X_m(t,x;s) = u_m(t, X_m(t,x;s)), \quad X_m(s,x;s) = x,$$

where $t,s\in[0,1]$ and $x\in\mathbb{T}^2$. Here, we use the notation introduced in Section 3.1; in particular, for each $m\geq 1$, the velocity fields $u_m\in C^\infty_{x,t}$. We shall denote by $X_m^{-1} = X_m^{-1}(x,t;s)$ the corresponding inverse maps (sometimes referred to as the "back-to-labels map"), satisfying the relations $X_m^{-1}(t, X_m(t,x;s);s) = x$ and $X_m(t, X_m^{-1}(t,x;s);s) = x$.[26] Classical ODE theory dictates that the Lagrangian maps $x\mapsto X_m(\cdot,x;\cdot)$ and their inverse flows $x\mapsto X_m^{-1}(\cdot,x;\cdot)$ remain close to the identity map, in the sense that $|\nabla X_m(x,t;s)-\mathrm{I}_2| + |\nabla X_m^{-1}(x,t;s)-\mathrm{I}_2| \ll 1$, only for time intervals which are small multiples of the Courant–Friedrichs–Lax time, namely for $|t-s|\ll \|\nabla u_m\|^{-1}_{L^\infty_{x,t}}$. Thus, the usage of Lagrangian coordinates necessitates the introduction of a new time scale $\{\tau_m''\}_{m\geq 1}$, which needs to satisfy the relation[27]

$$\tau_m''\|\nabla u_m\|_{L^\infty_{x,t}} \ll 1 \quad \text{or, equivalently,} \quad \tau_m'' a_m = \tau_m''\varepsilon_m^{\alpha-1} \ll 1. \tag{3.9}$$

According to this new time scale, for every $\ell\in\mathbb{Z}$, we denote[28]

$$X_{m,\ell}(t,x) := X_m(t,x;\ell\tau_m'') \quad \text{and} \quad X^{-1}_{m,\ell}(t,x) := X_m^{-1}(t,x;\ell\tau_m''),$$

the forward and backward flows of u_m, initialized at time $\ell\tau_m''$.

With the above notation, for $m\geq 1$, we define the stream function increment by[29]

$$\phi_m(t,x)-\phi_{m-1}(t,x) := \sum_{k,\ell\in\mathbb{Z}} \zeta\left(\frac{t}{\tau_m}-k\right)\widehat{\zeta}\left(\frac{t}{\tau_{m-1}''}-\ell\right) a_m\varepsilon_m^2\,\Psi_k\left(\frac{X^{-1}_{m-1,\ell}(t,x)}{\varepsilon_m}\right), \tag{3.10}$$

[26] Note that the maps $x\mapsto X_m(t,x;s)$ and $x\mapsto X_m^{-1}(t,x;s)$ are volume preserving (since $\operatorname{div} u_m = 0$), while the maps $x\mapsto X_m(t,x;s)-x$ and $x\mapsto X_m^{-1}(t,x;s)-x$ are $[0,1]\times\mathbb{T}^2$-periodic (since so is u_m).

[27] We note here a difference with the notation in [8]: there, τ_m'' is chosen to be small with respect to the CFL time scale of u_{m-1}, i.e., $\tau_m'' a_{m-1}\ll 1$. In this note, we choose the slightly more intuitive notation, which relates τ_m'' to the CFL time scale of u_m, i.e., $\tau_m'' a_m\ll 1$; see (3.9).

[28] As before, we have a difference with the notation with respect to [8]: here we write $X_m(\cdot,\cdot;\ell\tau_m'')$, whereas in [8, equation (2.36)] we wrote $X_{m-1}(\cdot,\cdot;\ell\tau_m'')$.

[29] If one wishes to prove Theorem 1.2, the parameters a_m appearing in (3.10) need to become "slow functions" of space and time, oscillating on scales ε_{m-1} in x and τ_{m-1} in t. The purpose of these slow functions in a convex-integration scheme is to cancel (through the mean of their squares) a leftover Euler–Reynolds stress error. We refer the reader to [17, Section 5] and [20, Section 7.3] for these details.

where $\widehat{\zeta}$ is another C^∞ smooth unit-scale bump function, which smoothly approximates the cutoff function $\mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}$, and such that its integer shifts nearly form a partition of unity: $\sum_{\ell\in\mathbb{Z}} \widehat{\zeta}(\cdot-\ell) \simeq 1$.[30] Note that the cutoff functions $\zeta(\cdot-k)$, the parameters ε_m, a_m, τ_m, and also the alternating shears Ψ_k appearing in (3.10) are the same as in (3.3). At this stage, we only assume that the time scale τ''_{m-1} satisfies the smallness condition from (3.9), but as we shall see next, a largeness constraint emerges (see (3.11) below). With the new definition of the stream function increment in (3.10) replacing the previous naive attempt from (3.3), the velocity increment is defined again as $u_m - u_{m-1} = \nabla^\perp(\phi_m - \phi_{m-1})$.

The usefulness of (3.10) is that on the support of each $\widehat{\zeta}(\cdot/\tau''_{m-1}-\ell)$, the "Lagrangian increment" $(\phi_m - \phi_{m-1})(\cdot, X_{m-1,\ell}(\cdot,x))$ behaves as $a_m\varepsilon_m^2\Psi_k\big(\frac{x}{\varepsilon_m}\big)$, which has zero mean with respect to the fast space variable $\frac{x}{\varepsilon_m}$. This overcomes the issue raised at the end of Section 3.1.2. However, in order to use the same heuristics as in Section 3.1.1, leading to the recursion relation (3.6) and asymptotic (3.7) for the cascade of renormalized diffusivities, we need to ensure that the Lagrangian flow maps $X^{-1}_{m-1,\ell}(\cdot,\cdot)$ and the new time cutoffs $\widehat{\zeta}(\cdot/\tau''_{m-1}-\ell)$ appearing in (3.10) do not interfere with the quantitative homogenization picture painted in Section 3.1.1. This necessitates that the new periodic time oscillation which we have introduced (on time scale τ''_{m-1}) *does not interfere*[31] with the homogenization problem which produced (3.6); put differently, this necessitates that from the point of view of a function oscillating at time scale τ_m, functions varying on time scale τ''_{m-1} are essentially a constant, resulting in the new parameter constraint

$$\tau''_{m-1} \gg \tau_m. \tag{3.11}$$

Together, the constraint on the new time parameter τ''_m is

$$\tau_m \ll \tau''_{m-1} \ll a^{-1}_{m-1}. \tag{3.12}$$

We also recall that τ_m needs to be chosen to satisfy (3.4).

At last, we try to see if we can choose the parameters in order to satisfy all of the above-mentioned constraints. The two inequalities in (3.12) indicate that a permissible choice for τ''_{m-1} can be made if and only if τ_m is chosen to satisfy $a^{-1}_{m-1} \gg \tau_m$. In turn, the lower bound for τ_m established previously in (3.8) shows that such a permissible choice for τ_m is possible if and only if $a^{-1}_{m-1} \gg \varepsilon_m^{1-\alpha-\gamma}$, where we recall that γ

[30] For technical reasons, the time-cutoff functions ζ and $\widehat{\zeta}$ are chosen to satisfy a number of fine properties not mentioned here, in order to deal with the fast switching of shear flows and the slow switching of flow maps; see [8, Section 2.1] for precise descriptions.

[31] More than that, this new time oscillation must itself be homogenized!

is defined by (3.7). Recalling the definition of a_{m-1}, and the ε_m to ε_{m-1} relation from (3.1), we arrive at the parameter constraint

$$\varepsilon_m^{\frac{1-\alpha}{q}} \simeq \varepsilon_{m-1}^{1-\alpha} = a_{m-1}^{-1} \gg \varepsilon_m^{1-\alpha-\gamma}. \tag{3.13}$$

Inequality (3.13) may be satisfied for all $m \geq 1$ if $\varepsilon_1 = \Lambda^{-1}$ is chosen to be sufficiently small (to absorb the implicit constant in the $\gg$ symbol), and if $\frac{1-\alpha}{q} < 1-\alpha-\gamma$. Recalling that $\gamma = (q-1)\frac{1+\alpha}{q+1}$, that $\alpha \in (0,1)$ and $q > 1$, (3.13) thus necessitates

$$\frac{1-\alpha}{q} < 1-\alpha-\gamma \Leftrightarrow q\gamma < (q-1)(1-\alpha) \Leftrightarrow \frac{q(1+\alpha)}{q+1} < 1-\alpha \Leftrightarrow \alpha < \frac{1}{2q+1}.$$

Since $q > 1$, we arrive at the *Onsager-supercriticality constraint* $\alpha < \frac{1}{3}$ present in the statement of Theorem 1.2.

3.1.4. The construction of u: Order of choosing parameters. To sum up, for a given $\alpha \in \left(0, \frac{1}{3}\right)$, we first choose $q > 1$ so that $(2q+1)\alpha < 1$. Then, we define $\{\varepsilon_m\}_{m\geq 0}$ according to (3.1), and $\{a_m\}_{m\geq 0}$ according to (3.2). In light of the lower bound $\tau_m \gg \varepsilon_m^{1-\alpha-\gamma}$ (expected from (3.8)), and the upper bound $\tau_m \ll a_{m-1}^{-1}$ (implied by (3.12)), we choose $\tau_m^{-1} \gg 1$ to be an integer lying in the interval $(\varepsilon_{m-1}^{\alpha-1}, \varepsilon_m^{\alpha-1+\gamma})$.[32] With τ_m chosen, we then return to (3.12) and let $(\tau''_{m-1})^{-1}$ to be an integer lying in the interval (a_{m-1}, τ_m^{-1}).[33] With these choices, the stream functions $\{\phi_m\}_{m\geq 0}$ are defined by the telescoping relation (3.10). The parameter $\Lambda = \varepsilon_1^{-1} \gg 1$ is chosen last, to absorb all m-independent constants appearing in our estimates.

The vector field u is obtained as $u = \lim_{m\to\infty} u_m$, the limit being taken with respect to the $C^{\alpha'}_{x,t}$ topology for any $\alpha' < \alpha$. Here, $u_m = \nabla^\perp \phi_m$, $\phi_m = \sum_{1\leq m'\leq m}(\phi_m - \phi_{m-1})$, $\phi_0 = 0$, and the stream function increment is given by (3.10).

Remark 3.1 (Genericity of the vector fields u from Theorem 1.2). The choice ϕ_0 is arbitrary, made in [8] only for notational simplicity. The proof of Theorem 1.2 works "as is" if we replace ϕ_0 by any sufficiently smooth, zero-mean periodic stream function. Then, all we have to do is to ensure that $\varepsilon_1 = \Lambda^{-1}$ is chosen to be (much) smaller than the smallest scale of oscillation of this chosen smooth ϕ_0. By telescoping the available bounds on $\|\phi_m - \phi_{m-1}\|_{C^1_{x,t}}$ (see [8, Proposition 2.2]), we may then show that $\|\phi - \phi_0\|_{C^1_{x,t}} = \|u - \nabla^\perp\phi_0\|_{C^0_{x,t}} \leq C\Lambda^{-\alpha}$, where $C = C(\alpha, q) > 0$; this upper bound may be made arbitrarily small by choosing Λ sufficiently large, showing that the class of vector fields u for which Theorem 1.2 holds is in fact *dense*, in the

[32] The choice $\tau_m^{-1} \in \mathbb{N}$ is made in order to ensure $[0,1]$-time periodicity of the respective time cutoffs. For instance, $\tau_m^{-1} \simeq \varepsilon_m^{(\alpha-1)(q+1)+\gamma q/2q}$ is an allowable choice.

[33] For instance, an integer which satisfies $(\tau''_m)^{-1} \simeq \varepsilon_m^{(\alpha-1)(q+3)+\gamma q/4}$ is an allowable choice.

class of all smooth divergence-free zero-mean vector fields, with respect to the $C^0_{x,t}$ topology. A stronger form of genericity of the family of vector fields $u \in C^\alpha_{x,t}$ which induce anomalous diffusion is obtained in [20]; see Theorem 1.3 above.

One may show, see [8, Section 2.3], that the family $\{\phi_m\}_{m\geq 0}$ is uniformly bounded in $C^{1,\alpha}_{x,t}$, that for each m the stream function ϕ_m is real analytic in space (uniformly in time), with radius of analyticity proportional to ε_m, that the flow maps X_m and their inverses X_m^{-1} remain close to the identity on their natural time scale τ_m'', and that they are also real analytic with radius of analyticity proportional to ε_m. That is to say, the family of stream functions $\{\phi_m\}_{m\geq 0}$ behaves in the way we expect them to behave.

Remark 3.2 (Propagation of smoothness in the construction). In [8], we have chosen to prove the quantitative real analyticity of the stream functions ϕ_m, of the Lagrangian flows X_m, and of their inverses X_m^{-1}, simply because it is true: the Ψ_k are sine waves, which are certainly analytic functions; the Lagrangian flow maps of an analytic function are analytic, adding and composing analytic functions to each other preserves analyticity; one has to be careful however when estimating the radius of analyticity of the resulting objects [8, Appendix B]. This real analyticity is however not essential for the proof: all that is needed is "sufficiently high regularity", to justify certain high-order expansions in the proof. Instead of working with real-analytic building blocks Ψ_k (sine waves), it is sufficient to consider smooth ones, and then to add a "mollification step" to the proof, as is common in convex-integration schemes (see, e.g., [17, Section 2.4], [18, Section 5.3]). This perspective needs to be taken if one is to work with building blocks Ψ_k which are compactly supported in space, and this is precisely the perspective taken in the proof of Theorem 1.3; see [20, Section 7.1].

3.2. A backward cascade of eddy diffusivities and homogenized problems

With the vector field u constructed in Section 3.1, it remains to rigorously justify our scale-by-scale homogenization picture, which dictates that when "coarse grained" at scale ε_m, the advection-diffusion equation (1.1) with vector field u experiences an effective diffusivity κ_m, which satisfies the recursion relation (3.6), initialized with $\kappa_M = \kappa$ for a suitably chosen $M \geq 0$. This requires the introduction of a *permissible set of diffusivities* (the set $\mathcal{K}$ mentioned in Theorem 1.2), and a definition of "coarse graining" (at least, as it pertains to the analysis discussed here).

In spite of the inherent instability of the recursion formula $\kappa \mapsto \kappa + \frac{a_m^2 \varepsilon_m^4}{8\kappa}$, see (3.6), we have seen earlier in (3.7) that the relation $\kappa_m \simeq \varepsilon_m^{1+\alpha+\gamma}$ is consistent with stably solving (3.6). As such, we define the permissible set of diffusivities by

$$\mathcal{K} := \bigcup_{m\geq 0} \left[\frac{1}{2}\varepsilon_m^{1+\alpha+\gamma}, 2\varepsilon_m^{1+\alpha+\gamma} \right].$$

For any given $\kappa \in \mathcal{K}$, we may identify a unique integer $M \geq 0$ such that

$$\kappa \in \left[\frac{1}{2}\varepsilon_M^{1+\alpha+\gamma}, 2\varepsilon_M^{1+\alpha+\gamma}\right], \quad \text{and we define } \kappa_M := \kappa. \tag{3.14}$$

Then, for $m \in \{1, \dots, M\}$, we recursively define the *renormalized diffusivities*

$$\kappa_{m-1} := \kappa_m \left(1 + c_* \left(\frac{a_m \varepsilon_m^2}{\kappa_m}\right)^2\right). \tag{3.15}$$

Here, $c_* > 0$ is a universal constant, which may differ from the $\frac{1}{8}$ value presented in (3.6) depending on the L_t^2-normalization of the time cutoffs ζ and $\widehat{\zeta}$.

Note that (3.14)–(3.15) defines a *backward cascade of eddy diffusivities*,[34] which is $O(1)$ at the unit scale, in the sense that

$$\kappa = \kappa_M \ll \kappa_{M-1} \ll \cdots \ll \kappa_m \ll \kappa_{m-1} \ll \cdots \ll \kappa_0 \in \left[\frac{1}{2}, 2\right].$$

The above relation holds no matter how small the molecular diffusivity $0 < \kappa \ll 1$ is; except that as $\kappa \to 0$ we have $M \to \infty$, and thus in this limit we need to perform infinitely many renormalization/quantitative homogenization steps.

It is useful to keep in mind the relations

$$\frac{a_m \varepsilon_m^2}{\kappa_m} \simeq \varepsilon_m^{-\gamma} \gg 1, \quad \frac{\varepsilon_m^2}{\kappa_m \tau_m} \simeq \varepsilon_m^{\frac{(1-\alpha(1+2q))(q-1)}{2q(q+1)}} \ll 1,$$

which hold since $\kappa_m \simeq \varepsilon_m^{1+\alpha+\gamma} \simeq a_m \varepsilon_m^2 \cdot \varepsilon_m^{\gamma}$, since τ_m is given by the asymptotic in Footnote 32, and since $\alpha(2q+1) < 1$.

Having defined the permissible set of diffusivities $\mathcal{K}$, and the sequence $\{\kappa_m\}_{m=0}^M$ (see (3.14)–(3.15)), it thus remains to define what we mean by "coarse graining" the advection-diffusion equation (1.1) at scale ε_m. A precise and robust definition of this concept, in a much more general setting, is given in [4, 6]. For the purposes of proving Theorem 1.2, we are able to get away with a simpler perspective based on *iterated homogenization*, available since the construction of the field u relies solely on space-time periodic building blocks.

Fix $\kappa = \kappa_M \in \mathcal{K}$, and define θ^κ to be the solution of the advection-diffusion equation (1.1) with drift velocity u and smooth initial condition θ_{in}. Then, for each $m \in \{0, \dots, M\}$, define θ_m to be the solution of the advection-diffusion equation with the same initial datum, but with *drift velocity given by u mollified at scale* ε_m (which

[34] It is verified in [8, Lemma 3.4] that $\kappa_m \in \mathcal{K}$ for all $m \in \{0, \dots, M\}$.

in our context means that the velocity is taken to equal u_m), and with diffusivity parameter κ_m (as defined by (3.14)–(3.15)). That is, θ_m is defined as the solution of

$$\partial_t \theta_m + u_m \cdot \nabla \theta_m - \kappa_m \Delta \theta_m = 0, \quad \theta_m|_{t=0} = \theta_{\text{in}} \tag{3.16}$$

for each $m \in \{0, \dots, M\}$. With this notation, we say that

$$\textit{the operator } \partial_t + u \cdot \nabla - \kappa \Delta \textit{ ``coarse grains'' at scale } \varepsilon_m \textit{ to the operator } \partial_t + u_m \cdot \nabla - \kappa_m \Delta,$$

if we are able to show that we have

$$\frac{1}{C_*} \cdot \kappa_m \|\nabla \theta_m\|_{L^2_{t,x}}^2 \le \kappa \|\nabla \theta^\kappa\|_{L^2_{t,x}}^2 \le C_* \cdot \kappa_m \|\nabla \theta_m\|_{L^2_{t,x}}^2 \tag{3.17}$$

for some universal constant $C_* > 1$ (independent of m). If we are able to prove (3.17), then since $u_0 = 0$ and $\kappa_0 \simeq 1$, by the energy balance for (3.16) with $m = 0$, and appealing to the Poincaré inequality, we have $\kappa_0 \|\nabla \theta_0\|_{L^2_{t,x}}^2 \simeq \varrho \|\theta_{\text{in}}\|_{L^2}^2$ for some $\varrho \in \left(0, \frac{1}{2}\right]$; this would then prove Theorem 1.2. The proof of (3.17) is described next.

3.3. The heart of the matter: One step of the homogenization iteration

In order to establish (3.17), we take an iterative approach: using quantitative homogenization (akin to Section 2.2), for each $m \in \{1, \dots, M\}$ we show that the amount of energy dissipation experienced by θ_m is comparable to the amount of energy dissipation experienced by θ_{m-1}. In this "one step" of the homogenization process, we make an error, and we aim to prove that these errors are summable in m, leading to (3.17).

Lemma 3.3 (below) summarizes this "one step" of homogenization induction, and this is the key step in the proof of Theorem 1.2. Its proof consists of comparing the equations for θ_m and θ_{m-1} by homogenizing the oscillations at scale ε_m. The fact that we can only get away with worrying about one scale at each step of the proof is because the scales ε_m are supergeometric, and thus becoming more well separated as they get smaller: (3.1) shows that $\frac{\varepsilon_m}{\varepsilon_{m-1}} \simeq \varepsilon_{m-1}^{q-1} \to 0$ as $m \to \infty$. This is an undesirable and non-physical feature of our model, but it is necessary for such a "naive" homogenization procedure to work, since we rely on the summability (in m) of the errors we make. In more realistic examples, one should expect to see true geometric scale separation, and this necessitates examining the interaction between *many* scales at each step of a renormalization iteration, and dealing with errors which are always of order one. There has been some recent progress [4,6] in developing a general *quantitative theory of coarse graining* for elliptic and parabolic operators, which is based on quantitative homogenization theory and is capable of analyzing such models—using a strategy which is broadly similar to our proof of Theorem 1.2.

The precise statement of the main homogenization estimate in our induction is given in [8, Proposition 5.2]; here, we offer a simplified version (see also Remark 3.4).

Lemma 3.3 (The homogenization iteration). *Fix $\alpha \in \left(0, \frac{1}{3}\right)$. There exist a small parameter $\delta = \delta(\alpha) \approx (q-1)^2 > 0$ and a universal constant $C = C(\alpha) > 0$, such that if $\varepsilon_1 = \Lambda^{-1}$ is sufficiently small (in terms of α alone), then the following statement holds. Assume the zero-mean initial datum θ_{in} is real analytic, with analyticity radius $R_{\theta_{\text{in}}} \in (0,1]$ which satisfies $R_{\theta_{\text{in}}} \geq \varepsilon_1^{1+\gamma/2}$ (recall, γ is defined in (3.7)). For $\kappa \in \mathcal{K}$, define M via (3.14). Then, for all $m \in \{2, \dots, M\}$, the solutions of (3.16) satisfy*

$$\begin{aligned} &\|\theta_m - \theta_{m-1}\|^2_{L^\infty([0,1];L^2(\mathbb{T}^2))} + \kappa_m \|\nabla \theta_m - \nabla \widetilde{\theta}_m\|^2_{L^2([0,1]\times\mathbb{T}^2)} \\ &\qquad \leq C \varepsilon_{m-1}^{2\delta} \|\theta_{\text{in}}\|^2_{L^2(\mathbb{T}^2)} \end{aligned} \tag{3.18}$$

and

$$\left| \frac{\kappa_m \|\nabla \theta_m\|^2_{L^2_{t,x}}}{\kappa_{m-1} \|\nabla \theta_{m-1}\|^2_{L^2_{t,x}}} - 1 \right| \leq C \varepsilon_{m-1}^{\delta} \tag{3.19}$$

for a suitably defined function $\widetilde{\theta}_m$ (see (3.20) below). Here, $L^2_{t,x} = L^2([0,1] \times \mathbb{T}^2)$.

Remark 3.4 (Real-analytic data). We show in [8, Proposition 5.2] that Lemma 3.3 holds for all zero-mean real-analytic data θ_{in},[35] no matter how small their analyticity radius[36] $R_{\theta_{\text{in}}}$ is. The only difference is that the bounds in (3.18) and (3.19) hold only for $m \in \{m_* + 1, \dots, M\}$, where $m_* := \min\{m \in \mathbb{N} : m \geq 1,\ R_{\theta_{\text{in}}} \geq \varepsilon_m^{1+\gamma/2}\}$.

Discussion of the proof of Lemma 3.3. The proof of this lemma is the bulk of the paper [8]. It requires solving separate space- and a time-homogenization problems, due to the three different "fast" scales in the equation for θ_m: the smallest spatial scale is ε_m, which is the length scale of oscillation of the shear flows; the smallest time scale, on which the shear flows switch directions, is given by τ_m; lastly, we have the time scale τ''_{m-1} on which the Lagrangian flow maps $X_{m-1,\ell}$ must refresh. We should think of these three fast scales as being well separated, with the spatial scale ε_m being the smallest/fastest, and with $\tau_m \ll \tau''_{m-1}$ (see (3.11)).

The idea is the same as for the classical periodic case presented in Sections 2.1 and 2.2. We want to make a two-scale ansatz $\widetilde{\theta}_m$, similar to the one for $\widetilde{\theta}_\varepsilon$ in (2.7). Here, the "macroscopic" function is θ_{m-1}, and akin to (2.8)–(2.9), we want to plug this two-scale ansatz into the "microscopic" equation, which is the equation for θ_m

[35] Note that Theorem 1.2 is stated for $H^1(\mathbb{T}^d)$-smooth initial data θ_{in}, not real-analytic one. However, upon mollification any $H^1(\mathbb{T}^d)$ data can be well approximated by a real-analytic one, and this suffices for the proof to close; see Section 3.4.

[36] We say that $f : \mathbb{T}^d \to \mathbb{R}$ is real analytic with analyticity radius (at least) $R_f \in (0,1]$ if $\sup_{n \geq 0} \|\nabla^n f\|_{L^2(\mathbb{T}^d)} R_f^n (n!)^{-1} \leq \|f\|_{L^2(\mathbb{T}^d)}$.

(see (3.16)). The hope is that the resulting error is small in $L^2_t H^{-1}_x$, so that we can obtain good estimates by appealing to parabolic regularity theory, as in the proof of Lemma 2.1.

The initial "naive" idea would be to define $\widetilde{\theta}_m$ by the two-scale ansatz

$$\widetilde{\theta}_m = \theta_{m-1} + \sum_{k,\ell\in\mathbb{Z}} \xi\Big(\frac{\cdot}{\tau_m} - k\Big)\widehat{\xi}\Big(\frac{\cdot}{\tau''_{m-1}} - \ell\Big) \cdot \sum_{i\in\{1,2\}} (\chi_{m,k,e_i} \circ X^{-1}_{m-1,\ell})\partial_i\theta_{m-1} + H_m,$$

where the χ_{m,k,e_i} are fast periodic "correctors" with slope e_i, which are oscillating at scale ε_m and are obtained by solving a cell problem related to the shear flow Ψ_k; the H_m is responsible for homogenizing the time cutoffs, and the ξ, $\widehat{\xi}$ are unit-scale time cutoffs that are slightly wider than ζ and $\widehat{\zeta}$. Unfortunately, compared to the periodic setting, we face a number of additional complications, and for these reasons, the definition above is not precise enough and cannot work. First, our equation is not ε_m-periodic because of "macroscopic" perturbations from the scales $\varepsilon_{m-1}, \dots, \varepsilon_0$, including those caused by the Lagrangian flows. Moreover, we cannot just "freeze" the coefficients and make an error here; some of these macroscopic perturbations need to be homogenized along with the ε_m-scale oscillations. For these reasons, our two-scale ansatz is a bit more complicated.

The *actual* definition of $\widetilde{\theta}_m$ which is used in (3.18) is given in [8, (4.24)] as

$$\widetilde{\theta}_m = T_{m-1} + \sum_{k,\ell\in\mathbb{Z}} \xi\Big(\frac{\cdot}{\tau_m} - k\Big)\widehat{\xi}\Big(\frac{\cdot}{\tau''_{m-1}} - \ell\Big) \cdot \sum_{i\in\{1,2\}} \widetilde{\chi}_{m,k,e_i}(\nabla(T_{m-1} \circ X_{m-1,\ell}) \circ X^{-1}_{m-1,\ell})_i + \widetilde{H}_m, \tag{3.20}$$

where

- The "twisted correctors" $\widetilde{\chi}_{m,k,e_i}$ play the role of the fast periodic correctors with slope e_i. These are simply the periodic correctors for the horizontal and vertical shears at scale ε_m, denoted by χ_{m,k,e_i} (see [8, (3.6)]), evaluated along the inverse flow $X^{-1}_{m-1,k}$ (see [8, (4.26)]). Note that in (3.20) the twisted correctors $\widetilde{\chi}_{m,k,e_i}$, which are indexed by $k \in \mathbb{Z}$, are simply glued together in time by the partition of unity given by $\xi\big(\frac{\cdot}{\tau_m} - k\big)$; all other terms in (3.20) are indexed by $\ell \in \mathbb{Z}$.
- The $\widetilde{H}_m$ is an approximate time corrector; see the below discussion and [8, (4.21)].
- The function T_{m-1} (defined in [8, (4.13)]) is a perturbation of the function θ_{m-1}. It is chosen to be an *approximate* solution of the equation

$$\partial_t T_{m-1} + u_{m-1} \cdot \nabla T_{m-1} - \nabla \cdot (\mathbf{K}_m + \mathbf{s}_{m-1})\nabla T_{m-1} = 0, \quad T_{m-1}|_{t=0} = \theta_{\mathrm{in}}, \tag{3.21}$$

where

- The field $\mathbf{K}_m(t)$ (see [8, (3.26)]) is a time-oscillating field that oscillates between the homogenized matrices for the vertical and horizontal shears; it averages to $\kappa_{m-1}\mathrm{I}_2$, plus a negligible error.
- The field $\mathbf{s}_{m-1}$ is defined as

$$\mathbf{s}_{m-1} := \mathbf{K}_m \sum_{\ell\in\mathbb{Z}} \widehat{\xi}\left(\frac{\cdot}{\tau''_{m-1}} - \ell\right)(\nabla X_{m-1,\ell} \circ X^{-1}_{m-1,\ell} - \mathrm{I}_2).$$

  It is added to the diffusion matrix of the T_{m-1} equation in order to deal with "distortions" caused by the Lagrangian flows; these distortions are small, but cannot be neglected in the equation for θ_m.

Equation (3.21) is close to the equation for θ_{m-1} (see (3.16) with m replaced by $m-1$) in the sense that $\mathbf{s}_{m-1}$ is small (due to (3.9)) and the time-oscillating field $\mathbf{K}_m(t)$ will homogenize to $\kappa_{m-1}\mathrm{I}_2$, since time oscillations simply average.

A warning. For technical reasons, we do *not* define the function T_{m-1} to be a solution of equation (3.21)! The reason is that equation (3.21) does not possess sufficiently strong regularity estimates; this is because when compared to $\kappa_{m-1}\mathrm{I}_2$, the matrix $\mathbf{K}_m$ has ellipticity constants which are much worse. Our regularity estimates for T_{m-1} must be very sharp, since it is the "macroscopic" function here and, as we saw in Sections 2.1 and 2.2, higher derivatives of the macroscopic function show up in the quantitative homogenization estimates. To fix this problem, we choose T_{m-1} to be an approximate solution of (3.21) which we construct by hand, using a high-order iterative scheme. We cut off the iteration defining T_{m-1} at a finite stage, ensuring that the regularity estimates we need to hold are valid, while making a very small error in the equation for T_{m-1}. It is in these high-order estimates for T_{m-1} that we use the quantitative real-analyticity assumption of the initial data present in the statement of Lemma 3.3 (we refer the reader to [8, Section 4.3]).

Similarly, the function $\widetilde{H}_m$ in (3.20) plays the role of the time corrector (analogous to h_ϵ in Section 2.1). However, as for T_{m-1}, in order to ensure that we have good enough regularity estimates for $\widetilde{H}_m$, we actually construct it as an approximate time corrector using an iterative scheme, to ensure that it has better regularity properties.

With the definition of the ansatz (3.20) in hand, the argument in [8] proceeds by plugging $\widetilde{\theta}_m$ into the equation for θ_m and checking that the right-hand side is small. The error is quite explicit, although we do not write it here since it takes half a page to display the nine error terms: see [8, (5.18)–(5.26)]. It is then a tedious task to check that all these terms are suitably small in $L^2_t H^{-1}_x$, but after doing so we obtain the bounds (3.18)–(3.19) claimed in the lemma. ∎

3.4. The proof of Theorem 1.2: Anomalous diffusion

To conclude this section, we show how Lemma 3.3 (and Remark 3.4) implies that the family of solutions $\{\theta^\kappa\}_{\kappa\in\mathcal{K}}$ of the advection-diffusion equation (1.1) with velocity u exhibits anomalous diffusion in the sense of Definition 1.1, as promised in the statement of Theorem 1.2. For the proofs of the statement made in bullets three–six in Theorem 1.2, we refer the reader to [8].

Assume for the moment that the function θ_{in} is real analytic.[37] Denote by $R_{\theta_{\text{in}}} \in (0, 1]$ its analyticity radius, so that $\sup_{n\geq 0}\|\nabla^n\theta_{\text{in}}\|_{L^2(\mathbb{T}^d)}R_{\theta_{\text{in}}}^n(n!)^{-1} \leq \|\theta_{\text{in}}\|_{L^2(\mathbb{T}^d)}$. Define $m_* \geq 1$ such that $\varepsilon_{m_*}^{1+\gamma/2} \simeq R_{\theta_{\text{in}}}$; the precise definition of m_* is given in Remark 3.4. This definition, together with the relation $\kappa_m \simeq \varepsilon_m^{1+\alpha+\gamma}$ for all $m \geq 0$ (see (3.7)), gives

$$\kappa_{m_*} \simeq \varepsilon_{m_*}^{1+\alpha+\gamma} \simeq R_{\theta_{\text{in}}}^{\frac{2(1+\alpha+\gamma)}{2+\gamma}}. \tag{3.22}$$

Relation (3.22) shows that κ_{m_*} is a function of $R_{\theta_{\text{in}}}$ alone (it is independent of the molecular diffusivity parameter κ). By the energy inequality for the θ_{m_*} evolution (i.e., (3.16) with m replaced by m_*), and the Poincaré inequality, we deduce

$$\frac{1}{2}\frac{d}{dt}\|\theta_{m_*}(t,\cdot)\|_{L^2(\mathbb{T}^2)}^2 = -\kappa_{m_*}\|\nabla\theta_{m_*}(t,\cdot)\|_{L^2(\mathbb{T}^2)}^2 \leq -(2\pi)^2\kappa_{m_*}\|\theta_{m_*}(t,\cdot)\|_{L^2(\mathbb{T}^2)}^2.$$

In turn, the above estimate implies

$$\begin{aligned}\kappa_{m_*}\|\nabla\theta_{m_*}\|_{L^2([0,1]\times\mathbb{T}^2)}^2 &= \frac{1}{2}\|\theta_{\text{in}}\|_{L^2(\mathbb{T}^2)}^2 - \frac{1}{2}\|\theta_{m_*}(1,\cdot)\|_{L^2(\mathbb{T}^2)}^2\\ &\geq \frac{1}{2}\|\theta_{\text{in}}\|_{L^2(\mathbb{T}^2)}^2(1-e^{-8\pi^2\kappa_{m_*}})\\ &\geq 2\kappa_{m_*}\cdot\frac{1}{2}\|\theta_{\text{in}}\|_{L^2(\mathbb{T}^2)}^2.\end{aligned} \tag{3.23}$$

Here, we have used that without loss of generality $\kappa_{m_*} \leq \frac{1}{3}$ (since $m_* \geq 1$, and we can take Λ to be sufficiently large), and so $1 - e^{-8\pi^2\kappa_{m_*}} \geq 2\kappa_{m_*}$.

Next, take any $\kappa \in \mathcal{K}$, and define $M = M(\kappa) \geq 0$ according to (3.14), so that $\kappa = \kappa_M \simeq \varepsilon_M^{1+\alpha+\gamma}$; in particular, as $\kappa \to 0$ we have $M \to \infty$, and so $M > m_*$ for all sufficiently small κ (since m_* is only a function of $R_{\theta_{\text{in}}}$, which is fixed). By Remark 3.4, the bounds (3.18)–(3.19) hold for all m in the nontrivial range $\{m_* + 1, \dots, M\}$. For instance, by telescoping estimate (3.19), we obtain that for all $m \in \{m_*, \dots, M\}$

$$\prod_{m=m_*+1}^{M}(1 - C\varepsilon_{m-1}^{\delta}) \leq \frac{\kappa_m\|\nabla\theta_m\|_{L^2([0,1]\times\mathbb{T}^2)}^2}{\kappa_{m_*}\|\nabla\theta_{m_*}\|_{L^2([0,1]\times\mathbb{T}^2)}^2} \leq \prod_{m=m_*+1}^{M}(1 + C\varepsilon_{m-1}^{\delta}).$$

[37] At the end of the proof, we will use the fact that $H^1(\mathbb{T}^2)$ functions can be approximated by real-analytic ones, with an error that can be made quantitatively small in $L^2(\mathbb{T}^2)$.

The super geometric growth $\varepsilon_m \simeq \varepsilon_{m-1}^q$ with $q > 1$, together with the fact that $\varepsilon_1 = \Lambda^{-1}$ is chosen sufficiently small (in terms of α), implies

$$\frac{3}{4} \leq \prod_{m=2}^{\infty} (1 \pm C\varepsilon_{m-1}^{\delta}) \leq \frac{4}{3},$$

and thus for all $m \in \{m_*, \dots, M\}$, we have

$$\frac{3}{4}\kappa_{m_*}\|\nabla\theta_{m_*}\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \leq \kappa_M\|\nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \leq \frac{4}{3}\|\nabla\theta_{m_*}\|_{L^2([0,1]\times\mathbb{T}^2)}^2. \tag{3.24}$$

Estimate (3.24) is used with $M = m_*$ in the proof.

In view of the previously established lower bound (3.23), it remains to relate the term $\kappa_M\|\nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 = \kappa\|\nabla\theta_{M(\kappa)}\|_{L^2([0,1]\times\mathbb{T}^2)}^2$ appearing in (3.24), to the true energy dissipation rate appearing in (1.5), namely $\kappa\|\nabla\theta^\kappa\|_{L^2([0,1]\times\mathbb{T}^2)}^2$. In order to achieve this, we note that $\theta^\kappa - \theta_M$ satisfies the equation

$$(\partial_t - \kappa\Delta + u\cdot\nabla)(\theta^\kappa - \theta_M) = \operatorname{div}((\phi - \phi_M)\nabla\theta_M), \quad (\theta^\kappa - \theta_M)|_{t=0} = 0.$$

Since by construction we have $\|\phi - \phi_M\|_{L^\infty_{x,t}} \leq C\varepsilon_M^{(1+\alpha)q} \ll 1$ (which follows by telescoping the trivial bound for (3.10)), a simple energy estimate gives

$$\begin{aligned}&\|\theta^\kappa - \theta_M\|_{L^\infty([0,1];L^2(\mathbb{T}^2))}^2 + \kappa\|\nabla\theta^\kappa - \nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \\ &\quad\leq \frac{1}{8}\kappa\|\nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2.\end{aligned} \tag{3.25}$$

Combining (3.22), (3.23), (3.24), (3.25), and recalling that $\kappa = \kappa_M$, we deduce

$$\begin{aligned}\kappa\|\nabla\theta^\kappa\|_{L^2([0,1]\times\mathbb{T}^2)}^2 &\geq \kappa\|\nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 - \kappa\|\nabla\theta^\kappa - \nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \\ &\geq \frac{7}{8}\kappa_M\|\nabla\theta_M\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \geq \frac{7}{8}\cdot\frac{3}{4}\cdot\kappa_{m_*}\|\nabla\theta_{m_*}\|_{L^2([0,1]\times\mathbb{T}^2)}^2 \\ &\geq \kappa_{m_*}\cdot\frac{1}{2}\|\theta_{\mathsf{in}}\|_{L^2(\mathbb{T}^2)}^2 \geq C_* R_{\theta_{\mathsf{in}}}^{\frac{2(1+\alpha+\gamma)}{2+\gamma}}\cdot\frac{1}{2}\|\theta_{\mathsf{in}}\|_{L^2(\mathbb{T}^2)}^2.\end{aligned} \tag{3.26}$$

In order to complete the proof of anomalous diffusion for all zero-mean data in $H^1(\mathbb{T}^2)$, we need to address the real-analyticity assumption used in (3.26), which explicitly contains the analyticity radius $R_{\theta_{\mathsf{in}}}$.

For this purpose, for any $\theta_{\mathsf{in}} \in H^1(\mathbb{T}^2)$, we define $L_{\mathsf{in}} := \frac{\|\theta_{\mathsf{in}}\|_{L^2}}{\|\nabla\theta_{\mathsf{in}}\|_{L^2}}$. Then, we mollify θ_{in} at spatial scale λL_{in}, for a parameter λ which is to be determined; i.e., we define

$$\widetilde{\theta}_{\mathsf{in}} := \theta_{\mathsf{in}} * \Phi(\lambda^2 L_{\mathsf{in}}^2, \cdot),$$

where $\Phi(t,x)$ is the standard heat kernel (with unit diffusion coefficient). The function $\widetilde{\theta}_{\mathsf{in}}$ is real analytic, and its analyticity radius is given by $R_{\widetilde{\theta}_{\mathsf{in}}} = c\lambda L_{\mathsf{in}}$, for a

universal constant $c > 0$. Moreover, standard properties of the heat kernel Φ imply that

$$\begin{aligned}\|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} &\geq \|\widetilde{\theta}_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} \geq \|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} - C\lambda^2 L_{\mathrm{in}}^2 \|\nabla\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} \\ &\geq (1 - C\lambda^2)\|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)}. \end{aligned} \tag{3.27}$$

To conclude, we denote by $\widetilde{\theta}^\kappa$ the solution of the advection-diffusion equation (1.1) with initial condition $\widetilde{\theta}_{\mathrm{in}}$. By linearity, we have that $(\partial_t + u\cdot\nabla - \kappa\Delta)(\widetilde{\theta}^\kappa - \theta^\kappa) = 0$, and so by the standard energy balance for this difference (recall (1.2)), the triangle inequality, the bound (3.26) with θ^κ replaced by $\widetilde{\theta}^\kappa$ and θ_{in} replaced by $\widetilde{\theta}_{\mathrm{in}}$, and with the bounds in (3.27), we deduce

$$\begin{aligned}\kappa\|\nabla\theta^\kappa\|^2_{L^2([0,1]\times\mathbb{T}^2)} &\geq \kappa\|\nabla\widetilde{\theta}^\kappa\|^2_{L^2([0,1]\times\mathbb{T}^2)} - \kappa\|\nabla(\widetilde{\theta}^\kappa - \theta^\kappa)\|^2_{L^2([0,1]\times\mathbb{T}^2)} \\ &\geq C_* R_{\widetilde{\theta}_{\mathrm{in}}}^{\frac{2(1+\alpha+\gamma)}{2+\gamma}} \cdot \frac{1}{2}\|\widetilde{\theta}_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} - \frac{1}{2}\|\widetilde{\theta}_{\mathrm{in}} - \theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} \\ &\geq \frac{C_*}{2}(c\lambda L_{\mathrm{in}})^{\frac{2(1+\alpha+\gamma)}{2+\gamma}} \cdot \frac{1}{2}\|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)} - C\lambda^2 \cdot \frac{1}{2}\|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)}, \end{aligned} \tag{3.28}$$

assuming that $C\lambda^2 \leq \frac{1}{2}$. The bound (3.28) dictates that we choose λ small enough to ensure that the first term on the right-hand side dominates; this amounts to letting

$$\lambda \leq c_* L_{\mathrm{in}}^{\frac{1+\alpha+\gamma}{1-\alpha}}$$

for some universal constant $0 < c_* \ll 1$ which depends on c, C, and C_*. Inserting the above bound into (3.28) implies

$$\kappa\|\nabla\theta^\kappa\|^2_{L^2([0,1]\times\mathbb{T}^2)} \geq \underbrace{\left(\frac{C_*}{4}(cc_*)^{\frac{2(1+\alpha+\gamma)}{2+\gamma}}\right) \cdot L_{\mathrm{in}}^{\frac{2(1+\alpha+\gamma)}{1-\alpha}}}_{=:\varrho} \cdot \frac{1}{2}\|\theta_{\mathrm{in}}\|^2_{L^2(\mathbb{T}^2)}.$$

The dependence of the coefficient ϱ on L_{in} claimed in the first bullet of Theorem 1.2 now follows by letting $\varepsilon = \frac{2\gamma}{1-\alpha} = \frac{2(q-1)(1+\alpha)}{(q+1)(1-\alpha)} \to 0$ as $q \to 1$.

4. Extensions and open problems

4.1. Uniform Hölder regularity of the scalars

An important question which is not addressed by Theorem 1.2 is the *regularity* of the solutions θ^κ, in particular whether there is a uniform-in-κ Hölder estimate such as

$$\sup_{\kappa\in\mathcal{K}} \|\theta^\kappa\|_{L^2([0,1];C^{0,\bar{\alpha}}(\mathbb{T}^d))} < \infty. \tag{4.1}$$

If we could show (4.1) for $\bar{\alpha}$ arbitrarily close to $\frac{1-\alpha}{2}$, then this would resolve part (ii) of the Obukhov–Corrsin dichotomy discussed in Section 1.2.

To see why we should expect such an estimate, let us recall the Campanato characterization of the Hölder space $C^{0,\beta}$, which states that[38] for any $f \in H^1(\mathbb{T}^d)$,

$$[f]_{C^{0,\beta}(\mathbb{T}^d)} \simeq \sup_{r\in(0,1)} r^{1-\beta} \|\nabla(f * \eta_r)\|_{L^\infty(\mathbb{T}^d)}. \tag{4.2}$$

What we have shown in Theorem 1.2 (recall (3.24) and also (3.7)) is that

$$\|\nabla\theta_m\|_{L^2([0,1]\times\mathbb{T}^d)} \approx \kappa_m^{-\frac{1}{2}} \simeq \varepsilon_m^{-\frac{1+\alpha+\gamma}{2}},$$

where we ignore the dependence on the initial datum θ_{in} of the implicit constants in the $\approx$ symbol, and we recall that $\gamma > 0$ is a very small exponent (proportional to $q-1$, which can thus be made arbitrarily small in our construction of the vector field). While not explicitly stated, the estimate in Lemma 3.3 implies that, for $j < m$, we have

$$\kappa_j \|\nabla\theta_m * \eta_{\varepsilon_j',\tau_j'''} - \nabla\theta_j\|^2_{L^2([0,1]\times\mathbb{T}^d)} \lesssim \varepsilon_{j-1}^{2\delta}, \tag{4.3}$$

where $\varepsilon_j' = \sqrt{\varepsilon_{j+1}\varepsilon_j}$ and $\tau_j''' := \sqrt{\tau_j\tau_j''}$ are chosen so that $\varepsilon_j < \varepsilon_{j'} < \varepsilon_{j-1}$ and $\tau_j'' < \tau_j''' < \tau_j$, with some amount of separation between each of these scales; again, we ignore the dependence on the initial datum θ_{in} of the implicit constants in the $\lesssim$ symbol. In other words, we should consider $\nabla\theta_j \simeq \nabla\theta_m * \eta_{\varepsilon_j',\tau_j'''}$ if $j < m$. Therefore, the previous two displays imply the bound

$$\|\nabla(\theta_m * \eta_{\varepsilon_j',\tau_j'''})\|_{L^2([0,1]\times\mathbb{T}^d)} \approx \kappa_j^{-\frac{1}{2}} \simeq \varepsilon_j^{-\frac{1+\alpha+\gamma}{2}} \simeq (\varepsilon_j')^{-\frac{1+\alpha}{2}-O(\gamma)}.$$

If we could upgrade this bound from L^2 to L^∞,[39] then by (4.2) we would have the desired uniform Hölder estimate for θ_m with regularity exponent $\bar{\alpha} = 1 - \frac{1+\alpha}{2} - O(\gamma) = \frac{1-\alpha}{2} - O(\gamma)$.

The aforementioned "upgrade" amounts to proving the Lipschitz-type estimate

$$\|\nabla\theta_m\|_{L^2_t L^\infty_x([0,1]\times\mathbb{T}^d)} \approx \kappa_m^{-\frac{1}{2}}, \tag{4.4}$$

which states that the gradient field $\nabla\theta_m$ does not concentrate on sets of small measure: it is roughly the same size, everywhere in $\mathbb{T}^d$. It is very natural to expect such an estimate to be true. First, at an intuitive level, since (i) our vector field is built from

[38] We denote by $\{\eta_r\}_{r>0}$ a standard family of mollifiers, with η_r localized to scale $\simeq r$, and by $\{\eta_{r,s}\}_{r,s>0}$ a family space-time mollifiers at spatial scale r and time scale s.

[39] We would also need some uniform-in-time estimates to treat the time mollification.

periodic ingredients with supergeometric scale separation between successive scales (so that the main contribution of $\nabla\theta_m$ is the wiggles at scale ε_m), and (ii) we have shown that the solutions have expansions with periodic ingredients, there is no reason to expect any such "concentration" to occur.

Second, uniform Lipschitz estimates of exactly this type have played a central role in homogenization theory since the pioneering work of Avellaneda and Lin [10, 11] in the 1980s; they now go by the name *large-scale regularity theory* (see the monographs [7, 65] and the references therein). While the arguments of Avellaneda and Lin [10, 11] are based on compactness and apply only to equations with periodic coefficients and thus one active scale, a quantitative approach to large-scale regularity was later proposed in [9]; this method is more robust and applies to equations with non-periodic coefficients (such as almost periodic and random coefficients).

The way it works heuristically is that the regularity of the homogenized equation is "transferred" to the equation with oscillating coefficients using quantitative homogenization estimates, and an excess decay iteration. In our setting, the "homogenized equation" would be the equation for θ_{m-1} and the "equation with oscillating coefficients" would be the equation for θ_m. This sets us an induction *going down the scales* in which the regularity for θ_{m-1} is transferred to θ_m: a cascade of regularity, dual to the inverse cascade of homogenization. Such an iteration argument would obviously be very technical to implement and take a great deal of effort to write, but we have all the ingredients to implement it in our setting,[40] and we do not expect that formalizing it would require major new ideas beyond [8, 9].

As part of the induction argument that proves (4.4), we will also upgrade the estimate (4.3) to an L^∞-type estimate, and thus the uniform-in-m Hölder estimate for θ_m can be obtained.

One nice consequence of such a Hölder estimate would be that the condition in Theorem 1.2 that the initial data belong to $H^1(\mathbb{T}^d)$ would be removed, and the positive constant ϱ would be universal. This would then imply the exponential decay in time of the $L^\infty(\mathbb{T}^d)$-norm of our solutions θ^{κ_j}, uniformly along our subsequence κ_j.

4.2. Geometric separation of scales and intermittency

A major shortcoming of the vector field constructed in Section 3.1 is that the active scales $\{\varepsilon_m\}_{m\geq 0}$ are supergeometrically separated. This non-physical aspect was necessary for our proof of anomalous diffusion, based on iterative quantitative

[40] We would need a localized version of Lemma 3.3, which applies not just in the whole torus $\mathbb{T}^2$, but in all appropriately scaled space-time cylinders above the scale at which we expect homogenization to occur. Such an estimate can be obtained by an argument similar to that of Lemma 3.3; there will just be additional boundary layer terms to control.

homogenization, to work. If we want the methods in [8] to apply to more realistic physical models, we need to develop a more flexible method.

Therefore, we think that it is an important open problem to prove anomalous diffusion for a variant of our construction in which *the scales* $\{\varepsilon_m\}_{m\geq 0}$ *are geometrically separated*, i.e., $\varepsilon_m = \exp(-Cm)$.

If the scales are geometrically separated, then we would make *an error of order one at each scale* m—which we obviously cannot sum up over m! Moreover, we would have to deal with leading-order interactions between multiple scales at once. In contrast, the proof of Theorem 1.2 is based on comparing the equation for θ_{m-1} to the equation for θ_m; evidently, the scales ε_{m+1} and ε_{m-1} do not interact directly, which would not be the case if they were geometrically separated.

It is only due to the supergeometric scale separation that we should expect the Hölder estimate discussed in Section 4.1 above to be valid. A geometric separation of scales—which implies the interaction between many scales at once—should lead to concentration effects and therefore intermittency. We would expect in this case that the L^p norm of $\nabla\theta_m$ to be very different for different values of p.

At first glance, these issues may seem fatal to our whole strategy based on homogenization theory. However, a general "coarse-graining" theory, based on quantitative homogenization methods, has been recently developed [5, 6]. Based on this, and on an iterative quantitative homogenization argument, a superdiffusive central limit theorem was recently proved in [4] for a passive scalar model with a random vector field (with non-separated scales). There is therefore some hope that iterative quantitative homogenization methods as in [4, 8] may apply to more physically realistic models of scalar turbulence.

4.3. A regularity threshold for Euler solutions exhibiting anomalous diffusion?

We conjecture that the regularity threshold $\alpha = \frac{1}{3}$ present in Theorem 1.2 is not an artifact of our proof in [8], but rather, a delicate rigidity constraint imposed by the Lagrangian nature of turbulent diffusion. We propose the following problem.

Conjecture 4.1 (Anomalous diffusion + self advection = Onsager supercriticality). *Let* $d \in \{2, 3\}$. *Assume that* $u \in C^0([0,1]; C^\alpha(\mathbb{T}^d)) \cap C^\alpha([0,1]; C^0(\mathbb{T}^d))$ *is a weak solution of the incompressible Euler equations for some Hölder regularity exponent* $\alpha \in (0,1)$. *Furthermore, assume that for any initial condition* $\theta_{\text{in}} \in \dot{H}^1(\mathbb{T}^d)$, *the family of solutions* $\{\theta^\kappa\}_{\kappa>0}$ *of the passive scalar equation* (1.1) *with velocity field* u *displays anomalous diffusion continuously in time; i.e.,* (1.5) *holds and the time dissipation measure* $\mathcal{E}(dt)$ *from* (1.6) *is non-atomic. Then, we have that* $\alpha \leq \frac{1}{3}$.

We expect that the method of proof for resolving Conjecture 4.1 is more interesting than what the statement of the conjecture yields. Indeed, if one wishes to prove

that the conjecture is *false*, then one first needs to construct $C^\alpha_{x,t}$ weak solutions of the incompressible Euler equations for some $\alpha \in \left(\frac{1}{3}, 1\right)$. This question is widely open (in the absence of artificial forcing terms). Conversely, if one is to prove that the conjecture is *true*, then one needs to synthesize weak Euler solutions from the information gained by watching what the associated advection-diffusion equation does to arbitrary $H^1(\mathbb{T}^d)$ smooth scalar initial conditions, in the vanishing diffusivity limit; one of the clear enemies is the pressure term in the Euler system.

Funding. S.A. was supported by NSF grant DMS-2350340. V.V. was in part supported by the Collaborative NSF grant DMS-2307681 and a Simons Investigator Award.

References

[1] G. Alberti, G. Crippa, and A. L. Mazzucato, Exponential self-similar mixing by incompressible flows. *J. Amer. Math. Soc.* **32** (2019), no. 2, 445–490 Zbl 1456.35158 MR 3904158

[2] L. Ambrosio, Transport equation and Cauchy problem for BV vector fields. *Invent. Math.* **158** (2004), no. 2, 227–260 Zbl 1075.35087 MR 2096794

[3] R. Aris, On the dispersion of a solute in pulsating flow through a tube. *Proc. Roy. Soc. London Ser. A* **235** (1956), no. 1200, 67–77

[4] S. Armstrong, A. Bou-Rabee, and T. Kuusi, Superdiffusive central limit theorem for a Brownian particle in a critically-correlated incompressible random drift. 2024, arXiv: 2404.01115v2

[5] S. Armstrong and T. Kuusi, Elliptic homogenization from qualitative to quantitative. [v1] 2022, [v2] 2024, arXiv:2210.06488v2

[6] S. Armstrong and T. Kuusi, Renormalization group and elliptic homogenization in high contrast. *Invent. Math.* **242** (2025), no. 3, 895–1086 Zbl 08144389 MR 4978180

[7] S. Armstrong, T. Kuusi, and J.-C. Mourrat, *Quantitative stochastic homogenization and large-scale regularity*. Grundlehren Math. Wiss. 352, Springer, Cham, 2019, 518 pp. Zbl 1482.60001 MR 3932093

[8] S. Armstrong and V. Vicol, Anomalous diffusion by fractal homogenization. *Ann. PDE* **11** (2025), no. 1, article no. 2, 145 pp. Zbl 1559.35018 MR 4846255

[9] S. N. Armstrong and C. K. Smart, Quantitative stochastic homogenization of convex integral functionals. *Ann. Sci. Éc. Norm. Supér. (4)* **49** (2016), no. 2, 423–481 Zbl 1344.49014 MR 3481355

[10] M. Avellaneda and F.-H. Lin, Compactness methods in the theory of homogenization. *Comm. Pure Appl. Math.* **40** (1987), no. 6, 803–847 Zbl 0632.35018 MR 0910954

[11] M. Avellaneda and F.-H. Lin, L^p bounds on singular integrals in homogenization. *Comm. Pure Appl. Math.* **44** (1991), no. 8–9, 897–910 Zbl 0761.42008 MR 1127038

[12] M. Avellaneda and A. Majda, Mathematical models with exact renormalization for turbulent transport, II: fractal interfaces, non-Gaussian statistics and the sweeping effect. *Comm. Math. Phys.* **146** (1992), no. 1, 139–204 Zbl 0754.76046 MR 1163672

[13] G. K. Batchelor, Small-scale variation of convected quantities like temperature in turbulent fluid. I. General discussion and the case of small conductivity. *J. Fluid Mech.* **5** (1959), no. 1, 113–133 Zbl 0085.39701 MR 0100457

[14] D. Bernard, K. Gawędzki, and A. Kupiainen, Slow modes in passive advection. *J. Stat. Phys.* **90** (1998), no. 3–4, 519–569 Zbl 0932.76030 MR 1616898

[15] E. Bruè and C. De Lellis, Anomalous dissipation for the forced 3D Navier–Stokes equations. *Comm. Math. Phys.* **400** (2023), no. 3, 1507–1533 Zbl 1516.35302 MR 4595604

[16] T. Buckmaster, M. Colombo, and V. Vicol, Wild solutions of the Navier–Stokes equations whose singular sets in time have Hausdorff dimension strictly less than 1. *J. Eur. Math. Soc. (JEMS)* **24** (2022), no. 9, 3333–3378 Zbl 1493.35063 MR 4422213

[17] T. Buckmaster, C. De Lellis, L. Székelyhidi, Jr., and V. Vicol, Onsager's conjecture for admissible weak solutions. *Comm. Pure Appl. Math.* **72** (2019), no. 2, 229–274 Zbl 1480.35317 MR 3896021

[18] T. Buckmaster and V. Vicol, Convex integration and phenomenologies in turbulence. *EMS Surv. Math. Sci.* **6** (2019), no. 1–2, 173–263 Zbl 1440.35231 MR 4073888

[19] T. Buckmaster and V. Vicol, Nonuniqueness of weak solutions to the Navier-Stokes equation. *Ann. of Math. (2)* **189** (2019), no. 1, 101–144 Zbl 1412.35215 MR 3898708

[20] J. Burczak, L. Székelyhidi, Jr., and B. Wu, Anomalous dissipation and Euler flows. [v1] 2023, [v2] 2024, arXiv:2310.02934v2

[21] A. Cheskidov, Dissipation anomaly and anomalous dissipation in incompressible fluid flows. [v1] 2023, [v2] 2024, arXiv:2311.04182v2

[22] A. Cheskidov, P. Constantin, S. Friedlander, and R. Shvydkoy, Energy conservation and Onsager's conjecture for the Euler equations. *Nonlinearity* **21** (2008), no. 6, 1233–1252 Zbl 1138.76020 MR 2422377

[23] A. Cheskidov and X. Luo, Nonuniqueness of weak solutions for the transport equation at critical space regularity. *Ann. PDE* **7** (2021), no. 1, article no. 2, 45 pp. Zbl 1469.35001 MR 4199851

[24] M. Colombo, G. Crippa, and M. Sorella, Anomalous dissipation and lack of selection in the Obukhov–Corrsin theory of scalar turbulence. *Ann. PDE* **9** (2023), no. 2, article no. 21, 48 pp. Zbl 1536.35261 MR 4662772

[25] P. Constantin, W. E, and E. S. Titi, Onsager's conjecture on the energy conservation for solutions of Euler's equation. *Comm. Math. Phys.* **165** (1994), no. 1, 207–209 Zbl 0818.35085 MR 1298949

[26] S. Corrsin, On the spectrum of isotropic temperature fluctuations in an isotropic turbulence. *J. Appl. Phys.* **22** (1951), no. 4, 469–473 Zbl 0044.40601 MR 0047458

[27] S. Daneri and L. Székelyhidi, Jr., Non-uniqueness and h-principle for Hölder-continuous weak solutions of the Euler equations. *Arch. Ration. Mech. Anal.* **224** (2017), no. 2, 471–514 Zbl 1372.35221 MR 3614753

[28] C. De Lellis and L. Székelyhidi, Jr., The h-principle and the equations of fluid dynamics. *Bull. Amer. Math. Soc. (N.S.)* **49** (2012), no. 3, 347–375 Zbl 1254.35180 MR 2917063

[29] C. De Lellis and L. Székelyhidi, Jr., Dissipative continuous Euler flows. *Invent. Math.* **193** (2013), no. 2, 377–407 Zbl 1280.35103 MR 3090182

[30] R. J. DiPerna and P. L. Lions, Ordinary differential equations, transport theory and Sobolev spaces. *Invent. Math.* **98** (1989), no. 3, 511–547 Zbl 0696.34049 MR 1022305

[31] D. A. Donzis, K. R. Sreenivasan, and P. K. Yeung, Scalar dissipation rate and dissipative anomaly in isotropic turbulence. *J. Fluid Mech.* **532** (2005), 199–216 Zbl 1125.76346 MR 2262275

[32] T. D. Drivas, T. M. Elgindi, G. Iyer, and I.-J. Jeong, Anomalous dissipation in passive scalar transport. *Arch. Ration. Mech. Anal.* **243** (2022), no. 3, 1151–1180 Zbl 1508.35069 MR 4381138

[33] T. D. Drivas and G. L. Eyink, A Lagrangian fluctuation-dissipation relation for scalar turbulence. Part I. Flows with no bounding walls. *J. Fluid Mech.* **829** (2017), 153–189 Zbl 1460.76003 MR 3698051

[34] J. Duchon and R. Robert, Inertial energy dissipation for weak solutions of incompressible Euler and Navier-Stokes equations. *Nonlinearity* **13** (2000), no. 1, 249–255 Zbl 1009.35062 MR 1734632

[35] T. M. Elgindi and K. Liss, Norm growth, non-uniqueness, and anomalous dissipation in passive scalars. *Arch. Ration. Mech. Anal.* **248** (2024), no. 6, article no. 120, 28 pp. Zbl 1553.35159 MR 4829551

[36] G. Eyink, Onsager's 'ideal turbulence' theory. *J. Fluid Mech.* **988** (2024), article no. P1, 74 pp. Zbl 1540.76060 MR 4749943

[37] G. L. Eyink, Energy dissipation without viscosity in ideal hydrodynamics I. Fourier analysis and local energy transfer. *Phys. D* **78** (1994), no. 3–4, 222–240 Zbl 0817.76011 MR 1302409

[38] G. L. Eyink and D. Benveniste, Suppression of particle dispersion by sweeping effects in synthetic turbulence. *Phys. Rev. E* **87** (2013), no. 2, article no. 023011

[39] G. L. Eyink and K. R. Sreenivasan, Onsager and the theory of hydrodynamic turbulence. *Rev. Modern Phys.* **78** (2006), no. 1, 87–135 Zbl 1205.01032 MR 2214822

[40] G. Falkovich, K. Gawędzki, and M. Vergassola, Particles and fields in fluid turbulence. *Rev. Modern Phys.* **73** (2001), no. 4, 913–975 Zbl 1205.76133 MR 1878800

[41] A. Fannjiang and G. Papanicolaou, Convection enhanced diffusion for periodic flows. *SIAM J. Appl. Math.* **54** (1994), no. 2, 333–408 Zbl 0796.76084 MR 1265233

[42] U. Frisch, *Turbulence. The legacy of A. N. Kolmogorov*. Cambridge University Press, Cambridge, 1995, 296 pp. Zbl 0832.76001 MR 1428905

[43] J. Geng and Z. Shen, Convergence rates in parabolic homogenization with time-dependent periodic coefficients. *J. Funct. Anal.* **272** (2017), no. 5, 2092–2113 Zbl 1356.35031 MR 3596717

[44] V. Giri, H. Kwon, M. Novack, The L^3-based strong Onsager theorem. [v1] 2023, [v3] 2025, arXiv:2305.18509v3, to appear in *Ann. of Math. (2)*

[45] V. Giri and R.-O. Radu, The Onsager conjecture in 2D: a Newton–Nash iteration. *Invent. Math.* **238** (2024), no. 2, 691–768 Zbl 1556.35231 MR 4809443

[46] E. Hess-Childs and K. Rowan, A universal total anomalous dissipator. 2025, arXiv: 2501.18526v1

[47] L. Huysmans and E. S. Titi, Non-uniqueness & inadmissibility of the vanishing viscosity limit of the passive scalar transport equation. *J. Math. Pures Appl. (9)* **198** (2025), article no. 103685, 51 pp. Zbl 1561.35212 MR 4874276

[48] P. Isett, A proof of Onsager's conjecture. *Ann. of Math. (2)* **188** (2018), no. 3, 871–963 Zbl 1416.35194 MR 3866888

[49] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of differential operators and integral functionals*. Springer, Berlin, 1994, 570 pp. Zbl 0838.35001 MR 1329546

[50] C. J. P. Johansson and M. Sorella, Nontrivial absolutely continuous part of anomalous dissipation measures in time. *J. Differential Equations* **453, Part 4** (2026), article no. 113912, 60 pp. Zbl 08140317 MR 4983360

[51] C. J. P. Johansson and M. Sorella, Anomalous dissipation via spontaneous stochasticity with a two-dimensional autonomous velocity field. [v1] 2024, [v2] 2025, arXiv: 2409.03599v2, to appear in *Duke Math. J.*

[52] A. N. Kolmogorov, Local structure of turbulence in an incompressible viscous fluid at very high Reynolds number. *Dokl. Akad. Nauk. SSSR* **30** (1941), no. 4, 299–303 Zbl 0025.37602 MR 0004146

[53] R. H. Kraichnan, Small-scale structure of a scalar field convected by turbulence. *Phys. Fluids* **11** (1968), no. 5, 945–953 Zbl 0164.28904

[54] Y. Le Jan and O. Raimond, Integration of Brownian vector fields. *Ann. Probab.* **30** (2002), no. 2, 826–873 Zbl 1037.60061 MR 1905858

[55] A. J. Majda and P. R. Kramer, Simplified models for turbulent diffusion: theory, numerical modelling, and physical phenomena. *Phys. Rep.* **314** (1999), no. 4–5, 237–574 MR 1699757

[56] A. L. Mazzucato, Y. Feng, and C. Nobili, Enhanced dissipation by advection and applications to PDEs. *Appl. Math. Model.* **150, Part A** (2026), article no. 116310, 24 pp. Zbl 08129554 MR 4937374

[57] D. W. McLaughlin, G. C. Papanicolaou, and O. R. Pironneau, Convection of microstructure and related problems. *SIAM J. Appl. Math.* **45** (1985), no. 5, 780–797 Zbl 0622.76062 MR 0804006

[58] S. Modena and L. Székelyhidi, Jr., Non-uniqueness for the transport equation with Sobolev vector fields. *Ann. PDE* **4** (2018), no. 2, article no. 18, 38 pp. Zbl 1411.35066 MR 3884855

[59] S. Modena and L. Székelyhidi, Jr., Non-renormalized solutions to the continuity equation. *Calc. Var. Partial Differential Equations* **58** (2019), no. 6, article no. 208, 30 pp. Zbl 1428.35005 MR 4029736

[60] T. Moerschell and M. Sorella, Non-uniqueness of parabolic solutions for advection-diffusion equation. *J. Math. Pures Appl. (9)* **205** (2026), article no. 103777, 43 pp. Zbl 08132423 MR 4940033

[61] M. Novack and V. Vicol, An intermittent Onsager theorem. *Invent. Math.* **233** (2023), no. 1, 223–323 Zbl 1530.35200 MR 4601999

[62] A. M. Obukhov, The structure of the temperature field in a turbulent flow. *Izv. Akad. Nauk SSSR. Ser. Geograf. Geofiz.* **13** (1949), 58–69 MR 0034164

[63] G. A. Pavliotis and A. M. Stuart, *Multiscale methods. Averaging and homogenization*. Texts Appl. Math. 53, Springer, New York, 2008, 307 pp. Zbl 1160.35006 MR 2382139

[64] K. Rowan, On anomalous diffusion in the Kraichnan model and correlated-in-time variants. *Arch. Ration. Mech. Anal.* **248** (2024), no. 5, article no. 93, 47 pp. Zbl 1558.76061 MR 4801699

[65] Z. Shen, *Periodic homogenization of elliptic systems*. Oper. Theory Adv. Appl. 269, Birkhäuser, Cham, 2018, 291 pp. Zbl 1409.35007 MR 3838419

[66] B. I. Shraiman and E. D. Siggia, Scalar turbulence. *Nature* **405** (2000), no. 6787, 639–646

[67] K. R. Sreenivasan, The passive scalar spectrum and the Obukhov–Corrsin constant. *Phys. Fluids* **8** (1996), no. 1, 189–196 Zbl 1023.76551 MR 1367065

[68] K. R. Sreenivasan and J. Schumacher, Lagrangian views on turbulent mixing of passive scalars. *Philos. Trans. Roy. Soc. A* **368** (2010), no. 1916, 1561–1577 Zbl 1192.76021 MR 2596481

[69] G. I. Taylor, Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. Roy. Soc. London Ser. A* **219** (1953), no. 1137, 186–203

[70] Z. Warhaft, Passive scalars in turbulent flows. In *Annual review of fluid mechanics, Vol. 32*, pp. 203–240, Annu. Rev. Fluid Mech. 32, Annual Reviews, Palo Alto, CA, 2000 Zbl 0988.76042 MR 1744308

Scott Armstrong
Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, Warren Weaver Hall, New York, NY 10012, USA; Laboratoire Jacques-Louis Lions, Sorbonne Université, 4 Pl. Jussieu, 15-25.317, 75005 Paris, France; scotta@cims.nyu.edu

Vlad Vicol
Department of Mathematics, Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, Warren Weaver Hall, New York, NY 10012, USA;
vicol@cims.nyu.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/6

Positivity and non-commutativity

Anna Wienhard

Abstract. In this article, we revisit a new notion of positivity in real semisimple Lie groups that at the same time generalizes total positivity in split real Lie groups as well as positive Lie semigroups in Hermitian Lie groups of tube type. We shortly discuss the relationship with higher rank Teichmüller spaces and then focus on describing different aspects of positivity as well as open questions. In the second part, we describe a non-commutative perspective on Hermitian Lie groups of tube type that is suggested by positivity and leads to interesting applications, such as non-commutative generalizations of Markov numbers.

1. Introduction

In my plenary lecture "A beautiful world beyond hyperbolic geometry: Anosov representations and higher Teichmüller spaces," I described aspects of hyperbolic geometry through the lens of groups acting by isometries on hyperbolic spaces and gave a selective overview of recent developments. I focused particularly on Anosov representations of higher rank Teichmüller spaces. Anosov representations generalize many aspects of convex cocompact actions on hyperbolic spaces to more general discrete subgroups of Lie groups of higher rank. This leads to interesting group actions on higher rank symmetric spaces of non-compact type, such as the space of positive definite symmetric matrices or the Siegel upper half-space. Special families of Anosov representations are given by higher rank Teichmüller spaces. Higher rank Teichmüller spaces are connected components of the representation variety of the fundamental group $\pi_1(S)$ of a closed surface into a semisimple Lie group, which consist entirely of discrete and faithful representations. For this to make sense, recall that classical Teichmüller space can be realized as the space of marked hyperbolic structures on S. The holonomy representations of marked hyperbolic structures form connected components of the representation variety of $\pi_1(S)$ into the isometry group

Mathematics Subject Classification 2020: 22E15 (primary); 22E40 (secondary).
Keywords: positivity in Lie groups, discrete subgroups of Lie groups, higher Teichmüller spaces, cluster algebras, Markov numbers.

of the hyperbolic plane $\mathrm{PSL}(2,\mathbb{R})$. For $\mathrm{PSL}(2,\mathbb{R})$, Anosov representations of $\pi_1(S)$ are precisely the holonomy representations of hyperbolic structures on S. For other Lie groups, the existence of such higher rank Teichmüller spaces is in fact quite surprising, and in fact, they only exist for special families of Lie groups. We refer the reader who is interested in learning more on Anosov representations to Canary's informal lecture notes on Anosov representations [12] and to Kassel's article in the proceedings of the ICM [32], and those who want to get a view into higher rank Teichmüller space to the author's article in the proceedings of the ICM [42].

Higher rank Teichmüller spaces are intimately linked to positivity in Lie groups. In the case of split real Lie groups, the connection to Lusztig's total positivity has been made by Fock and Goncharov [15]. For Hermitian Lie groups of tube type, the connection to positivity has been made in the work of Burger, Iozzi, Wienhard [11]; characterizations in terms of causal structures and bi-invariant orders appeared in the work of Ben Simon, Burger, Hartnick, Iozzi, and Wienhard [5].

In recent years, Guichard and Wienhard [28, 29] introduced a new notion of positivity in real semisimple Lie groups, which at the same time generalizes Lusztig's total positivity (defined for split real Lie groups), and positive Lie semigroups in Hermitian Lie groups of tube type. There are four families of simple real Lie groups admitting a positive structure. Besides split real Lie groups and Hermitian Lie groups of tube type, one family is formed by indefinite orthogonal groups $\mathrm{SO}(p+1,p+k)$, where $p,k>1$. The other family consists of exceptional groups.

Positivity gives rise to a notion of positivity of configurations of flags in appropriate flag varieties associated to G. This can be used to define positive representations of surface groups $\pi_1(S)$ into G whenever G admits a positive structure. It has been conjectured by Guichard, Labourie, and Wienhard [42, Conjecture 19] that the set of positive representations always forms higher rank Teichmüller spaces. This has been proven recently by Beyrer, Guichard, Labourie, Pozzetti, and Wienhard [8, 9, 27]. Even though there are still several open questions on positive representations, this establishes that positivity provides the right underlying structure to explain the existence of higher rank Teichmüller spaces.

But positivity is much more. It opens a new perspective on the families of simple real Lie groups admitting a positive structure. In this perspective, non-commutative rings with involution play a crucial role. In this article, we want to give a glimpse of that new perspective and describe it in more detail for Hermitian Lie groups of tube type. Through the lens of positivity, Hermitian Lie groups of tube type of arbitrary rank look like Lie groups of type A_1 but defined over a non-commutative ring with involution. This opens up a fresh view, which leads to interesting new constructions, including new quantum groups, generalizations of Markov numbers, non-commutative trace identities, and many other things which are yet to be explored.

In the first part of this article, we review important aspects of positivity in Lie groups and state several open questions. In the second part, we illustrate the non-commutative perspective on Hermitian Lie groups of tube type and describe some of its consequences.

2. Positivity in Lie groups

In this section, we recall the notion of positivity in a real semisimple Lie group. It is a generalization of the notion of total positivity.

The theory of totally positive matrices arose in the beginning of the 20th century through the work of Schoenberg [41] and Gantmacher and Krein [19]. In the 1990s, the theory was generalized widely by Lusztig [34], who introduced total positivity in general split real semisimple Lie groups. Since then, total positivity plays an important role in representation theory and cluster algebras and has many relations to other areas in mathematics as well as to theoretical physics. The concept of positivity has been generalized by Guichard and Wienhard [28, 29] to more general real semisimple Lie groups which are not necessarily split. This notion of positivity is defined with respect to a subset Θ of the set of simple roots Δ. The classification of semisimple Lie groups that admit a positive structure can be reduced to simple Lie groups. There are four families of simple Lie groups that admit a positive structure. One family is the family of split real Lie groups with total positivity, which is positivity with respect to the full set Δ. The three other families carry a positive structure with respect to a proper subset Θ of Δ. As we will see below, when Θ is a proper subset of Δ, some non-commutativity enters the picture.

2.1. The four families

Let us shortly recall the notion of positivity. For more details, we refer the reader to [28, 29].

Consider a simple real Lie group G which is not necessarily split. A subset $\Theta \subset \Delta$ of the set of simple roots defines a standard parabolic subgroup P_Θ in G. (We choose the convention so that P_Δ is the minimal parabolic subgroup of G.) The group P_Θ is a semidirect product of a reductive Lie group L_Θ and its unipotent radical U_Θ. The Lie algebra $\mathfrak{u}_\Theta$ of U_Θ carries a natural action by L_Θ and can hence be decomposed into its L_Θ-irreducible pieces. For every simple root $\alpha \in \Theta$, there is an irreducible piece $\mathfrak{u}_\alpha$. When G is split and $\Delta = \Theta$, these would be precisely the root spaces $\mathfrak{g}_\alpha$. In that case, they are one dimensional. When Θ is a proper subset of Δ, then the $\mathfrak{u}_\alpha$ are in general not root spaces and not necessarily one dimensional.

The Lie group G is said to admit a positive structure with respect to Θ if it satisfies the following property.

For every $\alpha \in \Theta$, there exists a non-empty closed acute convex cone $c_\alpha \subset \mathfrak{u}_\alpha$ that is invariant under the action of L°_Θ on $\mathfrak{u}_\alpha$, where L°_Θ is the connected component of the identity.

When G is split, $\mathfrak{u}_\alpha = \mathfrak{g}_\alpha$ can be identified with $\mathbb{R}$. Then the cone c_α can be identified with $\mathbb{R}^{\geq 0} \subset \mathbb{R}$.

Theorem 2.1 ([29, Theorem 3.4]). *A simple Lie group G admits a positive structure with respect to Θ if and only if the pair (G, Θ) belongs to the following list:*

(1) *G is a split real Lie group, and $\Theta = \Delta$;*

(2) *G is Hermitian of tube type and of real rank r, and $\Theta = \{\alpha_r\}$, where α_r is the long simple restricted root;*

(3) *G is locally isomorphic to $\mathrm{SO}(p+1, p+k)$, $p > 1$, $k > 1$, and $\Theta = \{\alpha_1, \dots, \alpha_p\}$, where $\alpha_1, \dots, \alpha_p$ are the long simple restricted roots;*

(4) *G is the real form of F_4, E_6, E_7, or of E_8 whose reduced root system is of type F_4, and $\Theta = \{\alpha_1, \alpha_2\}$, where α_1, α_2 are the long simple restricted roots.*

The arguments that give the classification provide important additional information.

When Θ is a proper subset of Δ, there exists a unique $\alpha_\Theta \in \Theta$ for which $\mathfrak{u}_{\alpha_\Theta}$ is not equal to the root space $\mathfrak{g}_{\alpha_\Theta}$ and is a higher-dimensional vector space.

For all other elements α of $\Theta \setminus \{\alpha_\Theta\}$, we have that $\mathfrak{u}_\alpha$ is equal to $\mathfrak{g}_\alpha$, and this root space has to be of dimension one. Hence, in this case, if we identify $\mathfrak{u}_\alpha$ with $\mathbb{R}$, the cone c_α can be identified again with $\mathbb{R}^{\geq 0} \subset \mathbb{R}$.

2.2. A non-commutative view on the special piece $\mathfrak{u}_{\alpha_\Theta}$

For the special root α_Θ, the vector space $\mathfrak{u}_{\alpha_\Theta}$ carries an irreducible action of the reductive group L°_Θ, which preserves the convex cone $c_{\alpha_\Theta} \in \mathfrak{u}_{\alpha_\Theta}$ and acts transitively on its interior $\mathring{c}_{\alpha_\Theta}$. Such symmetric cones have been classified, and they are intimately related to Hermitian Lie groups of tube type; see [14] for more details.

Thus, let us first look at the case when G is Hermitian of tube type. In this case, there is just one special piece $\mathfrak{u}_{\alpha_\Theta}$. Then G is (up to isogeny) $\mathrm{Sp}(2n, \mathbb{R})$, $\mathrm{SU}(n, n)$, $\mathrm{SO}^*(4n)$, $\mathrm{SO}(2, q)$, or the exceptional group $\mathrm{E}_{7(-25)}$. The first three families of classical groups can be uniformly described: Let $\mathbb{K}$ be either $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$, and the space $V = \mathbb{K}^{2n}$ is a right vector space and is endowed with an anti-Hermitian form defined by

$$\omega_{\mathbb{K}}(v, v') = \sum_{i=1}^{n} \overline{x}_i y'_i - \overline{y}_i x'_i$$

with $v = (x_1, \dots, x_n, y_1, \dots, y_n)$ and $v' = (x'_1, \dots, x'_n, y'_1, \dots, y'_n)$. The subgroup G in $\mathrm{GL}_{\mathbb{K}}(V)$ of automorphisms of V stabilizing $\omega_{\mathbb{K}}$ is Hermitian of tube type; it is

$\mathrm{Sp}(2n,\mathbb{R})$, $\mathrm{SU}(n,n)$, or $\mathrm{SO}^*(4n)$ following that $\mathbb{K}$ is $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$, respectively. The Levi factor L_Θ in this case consists of the elements

$$\ell_A = \begin{pmatrix} A & 0 \\ 0 & {}^t\overline{A}^{-1} \end{pmatrix}, \quad A \in \mathrm{GL}_n(\mathbb{K}),$$

and the unipotent algebra $\mathfrak{u}_\Theta$ is equal to $\mathfrak{u}_{\alpha_\Theta}$. It consists of the matrices

$$u_M = \begin{pmatrix} 0 & M \\ 0 & 0 \end{pmatrix}, \quad M \in M_n(\mathbb{K}) \text{ such that } {}^t\overline{M} = M;$$

thus $\mathfrak{u}_{\alpha_\Theta}$ identifies with the space $\mathrm{Herm}(n,\mathbb{K})$ of Hermitian matrices over $\mathbb{K}$. Since $\ell_A u_M \ell_A^{-1} = u_{AM{}^t\overline{A}}$, we get that $\mathrm{Herm}^{\geq 0}(n,\mathbb{K})$ is the L_Θ-invariant cone.

For $\mathrm{SO}(2,k+1)$, let us consider the quadratic form $q_{2,k+1} = -x_1x_{k+3} + x_2x_{k+2} - \sum_{j=3}^{k+1} x_j^2$ on $V = \mathbb{R}^{k+3}$. The group $G = \mathrm{SO}(q_{2,k+1})$ is also a Hermitian Lie group of tube type. The unipotent algebra is composed of the matrices

$$u_v = \begin{pmatrix} 0 & {}^tvJ & 0 \\ 0 & 0 & v \\ 0 & 0 & 0 \end{pmatrix}, \quad v \in \mathbb{R}^{k+1},$$

where J is the matrix of the form $q_{1,k}$. One has in this case $\ell_{\lambda,g} u_v \ell_{\lambda,g}^{-1} = u_{\lambda g v}$, and the cone defined by $x_2 > 0$, $q_{1,k} > 0$ is invariant by L_Θ°.

The exceptional Hermitian simple Lie group of tube type is $\mathrm{E}_{7(-25)}$ (or EVII in Cartan's notation). In this case, the semisimple part of the group L_Θ is EIV $= \mathrm{E}_{6(-26)}$, and the action of L_Θ on $\mathfrak{u}_{\alpha_\Theta}$ is the action of the group of isomorphisms of $\mathrm{Herm}(3,\mathbb{O})$ in it; $\mathrm{Herm}^{\geq 0}(3,\mathbb{O})$ is a natural invariant cone (cf. [14, p. 213]).

In the general case, when $\Theta \neq \Delta$ and the cardinality of Θ is > 1, the description of the cone c_{α_Θ} reduces to the Hermitian case. When $G = \mathrm{SO}(p+1, p+k)$, then (using standard numbering for the simple roots) $\mathfrak{u}_{\alpha_\Theta} \cong (\mathbb{R}^{1,k}, q_{1,k})$ with $c_{\alpha_\Theta} = \{v \in \mathbb{R}^{1,k} \mid v_1 \geq 0,\ q_{1,k}(v) \geq 0\}$ ($q_{1,k}$ denotes here the standard quadratic form of signature $(1,k)$ on $\mathbb{R}^{1+k}$).

When G is one group in the exceptional family whose system of restricted roots is of type F_4, then $\mathfrak{u}_{\alpha_\Theta}$ and the cone c_{α_2} are given in Table 1.

There is a well-known connection between symmetric cones and Euclidean Jordan algebras (see, for example, [14]). The Lie algebra $\mathfrak{u}_{\alpha_\Theta}$ carries the structure of a Euclidean Jordan algebra, and the cone $c_{\alpha_\Theta} \in \mathfrak{u}_{\alpha_\Theta}$ that gives the positive structure is given as the cone of squares.

Here we want to stress a slightly different point of view. For the cases where $\mathfrak{u}_{\alpha_\Theta}$ identifies with the space $\mathrm{Herm}(n,\mathbb{K})$ of Hermitian matrices over $\mathbb{K}$, there is in fact an algebra $\mathcal{A}$ with an anti-involution σ such that $\mathfrak{u}_{\alpha_\Theta}$ identifies with $\mathcal{A}^\sigma$, the set of fix points of σ, and the cone is $c_{\alpha_\Theta} = \{a\sigma(a) \mid a \in \mathcal{A}\} \subset \mathcal{A}^\sigma$, and

G	$\mathfrak{u}_{\alpha_\Theta}$	c_{α_2}
$\mathrm{FI} = \mathrm{F}_{4(4)}$	$\mathrm{Sym}(3,\mathbb{R})$	$\mathrm{Sym}^{\geq 0}(3,\mathbb{R})$
$\mathrm{EII} = \mathrm{E}_{6(2)}$	$\mathrm{Herm}(3,\mathbb{C})$	$\mathrm{Herm}^{\geq 0}(3,\mathbb{C})$
$\mathrm{EVI} = \mathrm{E}_{7(-5)}$	$\mathrm{Herm}(3,\mathbb{H})$	$\mathrm{Herm}^{\geq 0}(3,\mathbb{H})$
$\mathrm{EIX} = \mathrm{E}_{8(-24)}$	$\mathrm{Herm}(3,\mathbb{O})$	$\mathrm{Herm}^{\geq 0}(3,\mathbb{O})$

Table 1. The nontrivial cone for groups in the exceptional family.

$\mathring{c}_{\alpha_\Theta} = \{a\sigma(a) \mid a \in A \text{ invertible}\}$. In fact, the case when $\mathfrak{u}_\alpha = \mathfrak{g}_\alpha$ also fits in this setup. In this case, $\mathcal{A} = \mathbb{R}$, and σ is the identity map.

The case of indefinite orthogonal groups, when $\mathfrak{u}_{\alpha_\Theta} \simeq \mathbb{R}^{(1,q+1)}$, a slightly modified picture holds where $\mathcal{A}$ is the Clifford group of a Clifford algebra. For the case when $\mathfrak{u}_{\alpha_\Theta}$ identifies with the space $\mathrm{Herm}(3,\mathbb{O})$, there is no ambient associative algebra, but many of the other features persist in that case as well; see, for example, [24] for a detailed discussion.

2.3. The unipotent semigroup

An important role in positivity is played by the non-negative and positive semigroups of the unipotent group U_Θ. The non-negative semigroup $U_\Theta^{\geq 0}$ is the sub-semigroup of U_Θ generated by elements $\exp(v)$, with $v \in c_\alpha$ for some $\alpha \in \Theta$. The positive semigroup $U_\Theta^{>0}$ is the interior of $U_\Theta^{\geq 0}$.

Similarly, one can define a non-negative semigroup $U_\Theta^{\mathrm{opp},\geq 0}$ and a positive semigroup $U_\Theta^{\mathrm{opp},>0}$ in U_Θ^{opp}. This gives rise to the non-negative semigroup $G_\Theta^{\geq 0}$ in G, which is the semigroup generated by $U_\Theta^{\geq 0}$, $U_\Theta^{\mathrm{opp},\geq 0}$, and L_Θ°. The positive semigroup $G_\Theta^{>0}$ is the interior of $G_\Theta^{\geq 0}$.

In fact, for the reader who is not happy with the definition of positivity, requiring the existence of invariant convex cones, the existence of such semigroups could alternatively be used as the definition of positivity.

Theorem 2.2 ([28]). *Let G be a connected simple Lie group. Suppose that there is $U^+ \subset U_\Theta$ such that U^+ is a closed L_Θ°-invariant semigroup of non-empty interior, which contains no nontrivial invertible element. Then G admits a positive structure with respect to Θ, and the semigroup U^+ contains the semigroup $U_\Theta^{\geq 0}$.*

If we further assume that the interior of U^+ is contained in the open Bruhat cell with respect to P_Θ^{opp}, then $U^+ = U_\Theta^{\geq 0}$.

Defining the positive semigroup $U_\Theta^{>0}$ as the interior of $U_\Theta^{\geq 0}$ is easy to state but does not provide much insight into how elements in this semigroup look.

In Lusztig's work, the positive semigroup $U_\Delta^{>0}$ is in fact defined by giving explicit parametrizations, proving that the image of the parametrizations does not depend on

the choices made. That the thus defined set coincides with the interior of $U_\Delta^{\geq 0}$ will be proved afterwards. The parametrizations are associated to reduced expressions of the longest element in the Weyl group, and the change of coordinates between parametrizations associated to different reduced expressions is given by an explicit formula involving only positive rational functions.

In the case of positive structures with respect to a proper subset Θ of Δ, similar results hold. However, the Weyl group has to be replaced by another Coxeter group, called the Θ-Weyl group. Every reduced expression of the longest word in the Θ-Weyl group gives rise to a parametrization of the positive semigroup $U_\Theta^{>0}$, and the change of coordinates between two different reparametrizations is determined by explicit formulas [29, 30].

2.4. The Θ-Weyl group

The Θ-Weyl group is a subgroup of the Weyl group W of G. It turns out to be the Weyl group of a different root system and lies at the heart of regarding the Lie groups admitting a positive structure with respect to Θ through the lens of non-commutativity.

To introduce the Θ-Weyl group, let us recall that the Weyl group W is generated by reflections (order two elements) s_α, $\alpha \in \Delta$. These elements are characterized by their action on the collection of roots or root spaces. The element s_α sends the root space $\mathfrak{g}_\alpha$ to $\mathfrak{g}_{-\alpha}$ and permutes all the other root spaces associated to positive roots.

The Θ-Weyl group $W(\Theta)$ is a subgroup of W generated by elements σ_α for all $\alpha \in \Theta$. For all $\alpha \in \Theta$ which are not connected to $\Delta \backslash \Theta$, i.e., when $\mathfrak{u}_\alpha$ is equal to $\mathfrak{g}_\alpha$, σ_α is just the reflection s_α. For the unique $\alpha_\Theta \in \Theta$ which is connected to Weyl $\Delta \backslash \Theta$, i.e., for which $\mathfrak{u}_{\alpha_\Theta}$ is a higher-dimensional vector space, the element σ_{α_Θ} is a specific word in the subgroup generated by the reflections associated to roots in $\{\alpha_\Theta\} \cup \Delta \backslash \Theta$. The precise word is not of importance to us here; the reader can consult [29]. The key point is that the action of the Θ-Weyl group on the decomposition of the Lie algebra

$$\mathfrak{g} = \bigoplus_{\alpha \in \langle \Theta \rangle} \mathfrak{u}_\alpha \oplus \mathfrak{l}_\Theta \oplus \bigoplus_{\alpha \in \langle \Theta \rangle} \mathfrak{u}_{-\alpha}$$

is completely analogous to the action of the classical Weyl group on the root decomposition

$$\mathfrak{g} = \bigoplus_{\alpha \in \langle \Delta \rangle} \mathfrak{g}_\alpha \oplus \mathfrak{g}_0 \oplus \bigoplus_{\alpha \in \langle \Delta \rangle} \mathfrak{g}_{-\alpha}.$$

In other words, elements σ_α act on the irreducible pieces $\mathfrak{u}_\alpha$ in the following way. The element σ_α sends $\mathfrak{u}_\alpha$ to $\mathfrak{u}_{-\alpha}$ and permutes the $\mathfrak{u}_\beta$ for all the other β in the positive span of Θ.

The Θ-Weyl group is itself an abstract Weyl group but of a different type than the Weyl group of G. Its rank is the cardinality of Θ, and it acts on the pieces $\mathfrak{u}_\alpha$ in the same way as the abstract Weyl group would act on the classical root spaces.

Here is a list of the Θ-Weyl groups:

(1) *Split case*: If G is a split real form and $\Theta = \Delta$, then $W(\Theta) = W$.

(2) A_1 *case*: If G is Hermitian of tube type and of real rank r and $\Theta = \{\alpha_r\}$, then the Weyl group is of type C_r, but the Θ-Weyl group is of type A_1.

(3) B_p *case*: If G is locally isomorphic to $\mathrm{SO}(p+1, p+k)$, $p > 1$, $k > 1$, and $\Theta = \{\alpha_1, \dots, \alpha_p\}$, then the Weyl group is of type B_{p+1}, but the Θ-Weyl group is of type B_p.

(4) G_2 *case*: If G is the real form of F_4, E_6, E_7, or of E_8 whose reduced root system is of type F_4, and $\Theta = \{\alpha_1, \alpha_2\}$, then the Weyl group is of type F_4, but the Θ-Weyl group is of type G_2.

The labels we assigned to the different cases reflect the fresh perspective we suggest when looking at the groups G admitting a positive structure with respect to Θ. We suggest considering G as Lie groups of type $W(\Theta)$ but defined over a (partially) non-commutative ring with an involution. We will illustrate this perspective below for the case of Hermitian Lie groups of tube type. In this case, $\Theta = \{\alpha_\Theta\}$, so the Θ-Weyl group is $\mathbb{Z}/2\mathbb{Z}$. Thus, we want to consider Hermitian Lie groups of tube type as groups of type A_1 or more precisely as symplectic groups Sp_2 over some non-commutative ring. This has been made precise in [2, 40] and leads to interesting new phenomena.

Before we illustrate this non-commutative perspective, we describe how the Θ-Weyl group appears in parametrizations of the positive semigroup and state some open questions on parametrizations of the non-negative semigroups.

2.5. Parametrizing the positive semigroup

We explain now how reduced expressions of the longest work in the Θ-Weyl group give parametrizations of the positive semigroups $U_\Theta^{>0}$.

Given any reduced expression $\boldsymbol{\gamma} = \sigma_{\gamma_1} \cdots \sigma_{\gamma_n}$ of the longest element $w_{\max}^\Theta \in W(\Theta)$, we define the map

$$F_{\boldsymbol{\gamma}} \colon \mathfrak{u}_{\gamma_1} \times \cdots \times \mathfrak{u}_{\gamma_n} \to U_\Theta$$

as the product of the exponential map on each factor, i.e.,

$$F_{\boldsymbol{\gamma}}(X_1, \dots, X_n) = \exp(X_1) \cdots \exp(X_n),$$

where $X_i \in \mathfrak{u}_{\gamma_i}$.

Since $\gamma_i \in \Theta$ for all i, we can restrict the map $F_{\boldsymbol{\gamma}}$ to the product of the cones $c_{\gamma_i} \subset \mathfrak{u}_{\gamma_i}$ as well as to their interiors $\mathring{c}_{\gamma_i}$.

Theorem 2.3 ([29, Section 10]). *The map $F_{\boldsymbol{\gamma}}$ enjoys the following properties:*

(1) *The map $F_{\boldsymbol{\gamma}}$ restricted to the product of open cones $\mathring{c}_{\gamma_1} \times \cdots \times \mathring{c}_{\gamma_n}$ is injective and $F_{\boldsymbol{\gamma}}(\mathring{c}_{\gamma_1} \times \cdots \times \mathring{c}_{\gamma_n}) = U_\Theta^{>0}$. Thus, $F_{\boldsymbol{\gamma}}$ gives a parametrization of the positive semigroup.*

(2) *The map $F_{\boldsymbol{\gamma}}$ restricted to the product of closed cones $c_{\gamma_1} \times \cdots \times c_{\gamma_n}$ is not injective, but proper, and $F_{\boldsymbol{\gamma}}(c_{\gamma_1} \times \cdots \times c_{\gamma_n}) = U_\Theta^{\geq 0}$.*

Note that in the case of split real Lie groups, all the cones $\mathring{c}_{\gamma_i} \subset c_{\gamma_i} \subset \mathfrak{u}_{\gamma_i}$ can be identified with $\mathbb{R}^{>0} \subset \mathbb{R}^{\geq 0} \subset \mathbb{R}$. Given two different reduced expressions $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$, the change of coordinates is given by positive rational functions [34]. Since any two different reduced expressions are related by a sequence of braid relations, it suffices to establish that for the braid relations in all split real Lie groups of rank 2. Berenstein and Zelevinsky [7] wrote explicit expressions for the change of coordinates. To derive these expressions, given a braid relation, they set up a system of polynomial equations in the universal enveloping algebra, which they then solve explicitly.

In the case when Θ is a proper subset of Δ, several of the cones in the domain of $F_{\boldsymbol{\gamma}}$ are cones in higher-dimensional vector spaces. By the above theorem, the image of $F_{\boldsymbol{\gamma}}$ is still independent of the reduced expression $\boldsymbol{\gamma}$ of the longest word in the Θ-Weyl group $W(\Theta)$. In [30], we give explicit expressions for the change of coordinates. These expressions are not positive rational functions, but surprisingly, they still preserve the positivity defined by the cones. In fact, associated to braid relations in the Θ-Weyl group, we derive an analogous system of "polynomial equations" as Berenstein and Zelevinsky, just that some of the variables are now not scalars, but take values in a vector space (or better, a Euclidean Jordan algebra), and the powers of such an element are interpreted in an appropriate way as taking a product or evaluating a linear form. For details, we refer the reader to [30].

What we would like to highlight here is that, in the split real case, the system of polynomial equations is associated to the type of the Weyl group, whereas in the general case of positive structures with respect to Θ, it is the same system of polynomial equations associated to the type of the Θ-Weyl group, where the variable associated to the special root in Θ should be thought of as elements in a Euclidean Jordan algebra.

Lusztig's foundational work on total positivity in Lie groups has spurred a lot of further research. Many of the results that have been established for total positivity in split real Lie groups lead to natural questions for positivity with respect to Θ in more general semisimple real Lie groups. In the following subsections, we mention some open questions, but literally almost any result you come across for total positivity can

lead one to ask what the corresponding statement would be for positivity with respect to Θ.

2.6. Parametrizations of the non-negative semigroups

The map $F_{\boldsymbol{\gamma}}$ associated to a reduced expression of the longest word in the Θ-Weyl group gives a parametrization of the positive semigroup $U_\Theta^{>0}$. This parametrization is equivariant with respect to the action of L_Θ°. Since the map $F_{\boldsymbol{\gamma}}$ is not injective when restricted to the product of closed cones $c_{\gamma_1} \times \cdots \times c_{\gamma_n}$, it does not give a parametrization of the non-negative semigroup $U_\Theta^{\geq 0}$.

To approach the problem of parametrizing the non-negative semigroup, it is instructive to look at the Bruhat decompositions of a Lie group G. Recall that G can be decomposed under the left-right multiplication by $P_\Delta^{\mathrm{opp}} \times P_\Delta^{\mathrm{opp}}$, where P_Δ^{opp} is the minimal standard opposite parabolic subgroup. These orbits are called the Bruhat cells and are indexed by elements of the Weyl group W. One can also consider decompositions with respect to the action of $P_\Delta^{\mathrm{opp}} \times P_\Theta^{\mathrm{opp}}$ and $P_\Theta^{\mathrm{opp}} \times P_\Theta^{\mathrm{opp}}$, where Θ is a subset of Δ. For these, the orbits are indexed by elements of $W/W_{\Delta-\Theta}$ and $W_{\Delta-\Theta} \setminus W/W_{\Delta-\Theta}$, respectively, where $W_{\Delta-\Theta}$ is the Weyl group of L_Θ.

(1) The group G is the disjoint union

$$G = \bigcup_{w \in W} P_\Delta^{\mathrm{opp}} w P_\Delta^{\mathrm{opp}}.$$

(2) The group G is the disjoint union

$$G = \bigcup_{x \in W/W_{\Delta\setminus\Theta}} P_\Delta^{\mathrm{opp}} x P_\Theta^{\mathrm{opp}}.$$

(3) The group G is the disjoint union

$$G = \bigcup_{a \in W_{\Delta\setminus\Theta}\setminus W/W_{\Delta\setminus\Theta}} P_\Theta^{\mathrm{opp}} a P_\Theta^{\mathrm{opp}}.$$

In the case of total positivity, when $\Theta = \Delta$, all these decompositions agree. The intersection $U_w^{\geq 0}$ of the non-negative semigroup $U^{\geq 0}$ with the Bruhat cell $P_\Delta^{\mathrm{opp}} w P_\Delta^{\mathrm{opp}}$ can be parametrized in a similar way as the positive semigroup [34]. Let $\boldsymbol{\gamma} = \gamma_1 \cdots \gamma_k$ be a reduced expression of an element $w \in W$, not necessarily the longest element. Then $F_{\boldsymbol{\gamma}}$ restricted to $(\mathring{c}_{\gamma_1} \times \cdots \times \mathring{c}_{\gamma_k})$ gives a parametrization of $U_w^{\geq 0}$. Clearly, $U^{\geq 0} = \bigcup_{w \in W} U_w^{\geq 0}$ is a disjoint union. Thus, this gives a parametrization of $U^{\geq 0}$ and in fact also a cell decomposition [39].

In the case when Θ is a proper subset of Δ, the task to parametrize the non-negative semigroup becomes more difficult. This becomes already apparent in the

case of Hermitian Lie groups of tube type, which is the non-commutative A_1-type situation, where $\Theta = \{\alpha\}$ consists of a single simple root α. Here the situation is simple; there is just one space $\mathfrak{u}_\alpha$ and cones $\mathring{c}_\alpha \subset c_\alpha \subset \mathfrak{u}_\alpha$. The exponential map $\mathfrak{u}_\alpha \to U_{\{\alpha\}}$ is an isomorphism, and therefore, the map $F : c_\alpha \to U^{\geq 0}_{\{\alpha\}}$ gives a parametrization.

However, the structure of the intersection of $U^{\geq 0}_{\{\alpha\}}$ with the different orbits in the above decompositions of G is already very interesting. This stems from the fact that the boundary of the cone $\partial c_\alpha = c_\alpha - \mathring{c}_\alpha$ is nontrivial. In the split real case, or when $\mathfrak{u}_\alpha$ is of dimension one, this boundary is just 0, but in general it has several strata, given by L°_Θ-invariant subsets. These strata are precisely the orbits $U^{\geq 0}_{\{\alpha\}} \cap P^{\mathrm{opp}}_\Theta a P^{\mathrm{opp}}_\Theta$ for $a \in W_{\Delta\setminus\Theta} \setminus W / W_{\Delta\setminus\Theta}$; see [29]. Note that this decomposition does not give you a cell decomposition as the individual strata have nontrivial topology.

The decompositions of $U^{\geq 0}_{\{\alpha\}}$ with respect to the other decompositions

$$G = \bigcup_{x \in W/W_{\Delta\setminus\Theta}} P^{\mathrm{opp}}_\Delta x P^{\mathrm{opp}}_\Theta \quad \text{and} \quad G = \bigcup_{w \in W} P^{\mathrm{opp}}_\Delta w P^{\mathrm{opp}}_\Delta$$

are more involved, as they are not invariant by L°_Θ. The paper [29] contains further information with respect to the first decomposition; see also [44,45] for further work in the case of the symplectic group. In particular, for the symplectic group, [44] gives a cell decomposition of the non-negative semigroup $U^{\geq 0}_{\{\alpha\}}$.

For $\mathrm{SO}(3, 2+k)$, where the Θ-Weyl group is of type B_p, an explicit parametrization of the non-negative semigroup $U^{\geq 0}_\Theta$ has been provided in [29, Section 11], but even in this case, the full structure of parametrizations of the non-negative semigroup is not known.

Open Questions 2.4. Here are several open questions regarding the non-negative semigroups $U^{\geq 0}_\Theta$.

(1) Describe the strata $U^{\geq 0}_{\Theta,a}$, $U^{\geq 0}_{\Theta,x}$, and $U^{\geq 0}_{\Theta,w}$ of the non-negative semigroup $U^{\geq 0}_\Theta$ with respect to the different decompositions of G above.
(2) Provide parametrizations of $U^{\geq 0}_\Theta$ and of the strata $U^{\geq 0}_{\Theta,a}$, $U^{\geq 0}_{\Theta,x}$, and $U^{\geq 0}_{\Theta,w}$.
(3) Find explicit cell decompositions of $U^{\geq 0}_\Theta$.

2.7. Positive parts of flag varieties

Totally positive and non-negative semigroups give rise to positive or non-negative parts in flag varieties. The positive Grassmannian has received a lot of interest not only in the mathematical community but also in theoretical physics; see Williams ICM talk [43] for an overview.

If G admits a positive structure with respect to Θ, the non-negative semigroup $G^{\geq 0}_\Theta$ and the positive semigroup $G^{>0}_\Theta$ also give rise to non-negative parts

and positive parts of flag varieties of the group G. These are introduced and studied in [30].

Let us give a short description here. Let $\mathcal{F} = G/P$ be a flag variety of G. The non-negative part of the flag variety $\mathcal{F}^{\geq 0}$ with respect to Θ is

$$\mathcal{F}^{\geq 0} := \{G_\Theta^{\geq 0} P\} = \{uPu^{-1} \mid u \in G_\Theta^{\geq 0}\},$$

where P is a parabolic subgroup compatible with P_Θ, i.e., such that $P \cap P_\Theta$ is a parabolic subgroup.

The positive part $\mathcal{F}^{>0}$ of the flag variety $\mathcal{F}$ is

$$\mathcal{F}^{>0} := \{G_\Theta^{>0} P\} = \{uPu^{-1} \mid u \in G_\Theta^{>0}\},$$

where P is a parabolic subgroup compatible with P_Θ. The non-negative part of the flag variety $\mathcal{F}^{\geq 0}$ is the closure of $\mathcal{F}^{>0}$ in $\mathcal{F}$.

If $P = P_\Theta$, the positive part of $\mathcal{F}$ is what is called a diamond in [27,29], and the non-negative part is the closure of a diamond.

In the case of the symplectic group (with the Θ-Weyl group being of type A_1), the positive and non-negative parts of the flag varieties have a nice geometric description; see [30]. The non-negative part of the Lagrangian Grassmannian is studied in detail in [44]. In general, the structure of the non-negative and positive parts of flag varieties when G is not necessarily split and the Θ-Weyl group is of rank at least 2 is not yet well understood.

Open Questions 2.5. Here are some open problems.

(1) Provide explicit parametrizations of the positive and non-negative parts of flag varieties.

(2) Give cell decompositions of the positive and non-negative parts of flag varieties. In the case of total positivity, for Grassmannians these are parametrized by positroid varieties.

(3) When G is a split real form, which admits another positive structure with respect to a proper subset Θ in Δ, describe precisely the relation between the positive and non-negative parts of the flag variety, with respect to Δ and with respect to Θ.

(4) The amplituhedron, originally presented in [4], is the image of the positive Grassmannian under a positive linear map. It plays a major role in the understanding of scattering amplitudes in gauge theories. Introduce a generalization of the amplituhedron for positive structures with respect to Θ.

In the totally positive case, the positive parts of flag varieties can be described by cluster coordinates, so we expect that the non-commutative cluster structures associated to groups admitting a positive structure, combined with a good understanding of

the situation when the Θ-Weyl group is of type A_1, should be useful in answering the above questions.

Thus, now we turn to describe in more detail the non-commutative perspective on Lie groups admitting a positive structure with respect to Θ, with a special focus on the case when the Θ-Weyl group is of type A_1. For this, we will first recall some aspects of the simple real Lie group of type A_1, namely, $\mathrm{SL}_2(\mathbb{R})$.

3. Aspects of $\mathrm{SL}_2(\mathbb{R})$

The Lie group $\mathrm{SL}_2(\mathbb{R})$ is the simplest real simple Lie group. It (and its Lie algebra) plays an important role in understanding the structure of any real simple Lie group. In this section, we just highlight some aspects of $\mathrm{SL}_2(\mathbb{R})$ which will play a role later when we consider Lie groups of type SL_2 over non-commutative rings.

3.1. $\mathrm{SL}_2(\mathbb{R})$ as $\mathrm{Sp}_2(\mathbb{R})$

Usually, we think of $\mathrm{SL}_2(\mathbb{R})$ as the subgroup of the general linear group $\mathrm{GL}_2(\mathbb{R})$ of determinant 1. However, here we would like to think of $\mathrm{SL}_2(\mathbb{R})$ as $\mathrm{Sp}_2(\mathbb{R})$, the subgroup invertible endomorphism preserving a symplectic form ω.

The symplectic form ω is a non-degenerate skew-symmetric bilinear form

$$\omega : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R},$$

defined by $\omega((x_1, x_2), (y_1, y_2)) = x_1 y_2 - x_2 y_1$.

Identifying linear endomorphism of $\mathbb{R}^2$ with 2×2 matrices, we get

$$\mathrm{Sp}_2(\mathbb{R}) = \left\{ g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \,\middle|\, \det g = ad - bc = 1 \right\} = \mathrm{SL}_2(\mathbb{R}).$$

3.2. The symmetric space

The group $\mathrm{Sp}_2(\mathbb{R})$ acts by isometries on the hyperbolic plane $\mathbb{H}^2$, which is the symmetric space associated to $\mathrm{Sp}_2(\mathbb{R})$. There are several different models of the hyperbolic plane, for example, the Poincaré disk model or the upper half-plane model.

The most natural model for the symmetric space of $\mathrm{Sp}_2(\mathbb{R})$ is the space $\mathcal{J}$ of ω-compatible almost complex structures on $\mathbb{R}^2$. An almost complex structure on $\mathbb{R}^2$ is a linear endomorphism, whose square is $-\mathrm{id}$. An almost complex structure J is ω-compatible if

$$\omega(\cdot, J\cdot) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$$

is a positive definite symmetric bilinear form.

If we consider $\mathbb{C}^2 = \mathbb{R}^2 \otimes \mathbb{C}$ the complexification of $\mathbb{R}^2$, any almost complex structure J can be extended $\mathbb{C}$-linearly to an endomorphism $J_{\mathbb{C}}$ of $\mathbb{C}^2$. This endomorphism now has two eigenspaces, V_J^+ and V_J^-, with eigenvalues $+i$ and $-i$, respectively. The symplectic form ω extends as well $\mathbb{C}$-linearly to a symplectic form $\omega_{\mathbb{C}}$ and gives rise to a non-degenerate Hermitian form $h(\cdot,\cdot) = i\omega_{\mathbb{C}}(\bar{\cdot},\cdot)$, where $\bar{\cdot}$ is the complex conjugation. The Hermitian form h is of signature $(1,1)$, and h restricted to V_J^+ is positive definite, h restricted to V_J^- is negative definite. This provides two embeddings $i^{\pm} : \mathcal{J} \to \mathbb{C}P^1$, with $\overline{i^+(J)} = i^-(J)$ and $i^+(\mathcal{J}) \cup i^-(\mathcal{J}) = \mathbb{C}P^1 \setminus \mathbb{R}P^1$.

The Poincaré model and the upper half-plane model arise now as the image of $i^+(\mathcal{J})$ with respect to two different affine charts for $\mathbb{C}P^1$. For the Poincaré model, consider the affine chart with respect to $V_{J_0}^-$, where J_0 is the standard complex structure $J_0(x_1, x_2) = (-x_2, x_1)$. For the upper half-plane model, consider the affine chart with respect to the complex line spanned by the vector e_2.

3.3. Trace identities

There are two invariant functions of 2×2 matrices: the trace (tr) and the determinant (det). By the classical Cayley–Hamilton theorem, any 2×2 matrix satisfies the relation

$$M^2 - \operatorname{tr}(M)M + \det(M)\mathrm{id} = 0.$$

Restricting to $M \in \mathrm{Sp}_2(\mathbb{R})$, where the determinant is 1, taking the trace of the Cayley–Hamilton relation, one obtains the trace identity

$$\operatorname{tr}(M^2) = \operatorname{tr}(M)^2 - 2.$$

This can be generalized and leads to the well-known SL_2-trace identity, going back to Fricke and Klein's work [18].

For any $M, N \in \mathrm{Sp}_2(\mathbb{R})$, we have

$$\operatorname{tr}(MN) + \operatorname{tr}(MN^{-1}) = \operatorname{tr}(M)\operatorname{tr}(N),$$

and for the commutator,

$$\operatorname{tr}(MNM^{-1}N^{-1}) = \operatorname{tr}(M)^2 + \operatorname{tr}(N)^2 + \operatorname{tr}(MN)^2 - \operatorname{tr}(M)\operatorname{tr}(N)\operatorname{tr}(MN) - 2. \tag{3.1}$$

Trace identities have played a major role in the study of discrete subgroups of $\mathrm{SL}_2(\mathbb{R})$ and the study of representation varieties. Since the trace is a conjugation invariant function, when S is a cylinder, $\pi_1(S) = \mathbb{Z}$, any representation $\pi_1(S) \to \mathrm{SL}_2(\mathbb{R})$ is essentially determined by the trace of a generator. For general surfaces, the traces of generators and some products of generators of the fundamental group give explicit coordinates for the character variety [38].

3.4. Markov numbers

The trace identities allow us to connect representations $\pi_1(S_{1,1}) \to \mathrm{SL}_2(\mathbb{R})$ that arise from hyperbolic structures on the once-punctured torus $S_{1,1}$ to Markov numbers.

The fundamental group $\pi_1(S_{1,1})$ is a free group on two generators A and B. Any representation $\pi_1(S_{1,1}) \to \mathrm{SL}_2(\mathbb{R})$ (up to conjugation) is completely determined by three coordinates: $a = \mathrm{tr}(A)$, $b = \mathrm{tr}(B)$, $c = \mathrm{tr}\, AB$. To make the connection with Markov numbers, one considers only representations which arise as holonomy representations of complete hyperbolic structures on $S_{1,1}$ of finite volume. For such representations, the trace of the commutator $[A, B]$ is -2.

Therefore, the trace identity becomes

$$a^2 + b^2 + c^2 - abc = 0. \tag{3.2}$$

This relates the study of hyperbolic structures on a punctured torus to the classical theory of Markov triples [37], which are of interest in number theory. Markov triples are integer solutions to

$$a^2 + b^2 + c^2 - 3abc = 0. \tag{3.3}$$

Note that these two equations are related; if $(3a, 3b, 3c)$ is a solution to the first equation, then (a, b, c) is a solution to the second equation.

The first surprising fact is that given a Markov triple (a, b, c), the triple $\left(\frac{b^2+c^2}{a}, b, c\right)$ is also a Markov triple. As proved by Markov, every Markov triple can be obtained from the "trivial" solution $(1, 1, 1)$ by repeatedly applying this rule at each entry.

Markov numbers are well studied in number theory [1, 20].

3.5. Cluster algebras

A cluster algebra is an algebra with a rich additional combinatorial structure. Informally, the algebra can be generated from an initial "seed" via a process called "mutation" which produces new seeds from old seeds. The family of seeds is called a cluster structure of the algebra. Let us shortly recall the cluster structure on the once-punctured torus $S_{1,1}$, which will give us a link to Markov numbers:

$$x_\gamma x_{\gamma'} = x_\alpha x_\eta + x_\beta x_\delta. \tag{3.4}$$

A seed for the cluster structure of the surface $S_{1,1}$ consists of an ideal triangulation T of $S_{1,1}$ along with a choice of cluster variable x_γ for each arc γ of T. Each seed $(T, \vec{x})$ can be mutated at an arc γ to produce a new seed $\mu_\gamma(T, \vec{x}) = (T', \vec{x'})$. The triangulation T' is obtained by removing γ and replacing it with the arc γ' corresponding to the opposite quadrilateral containing γ (Figure 1). We name the edges of the bounding quadrilateral cyclically by $\alpha, \beta, \eta, \delta$. The cluster variables in $\vec{x'}$ are the same as

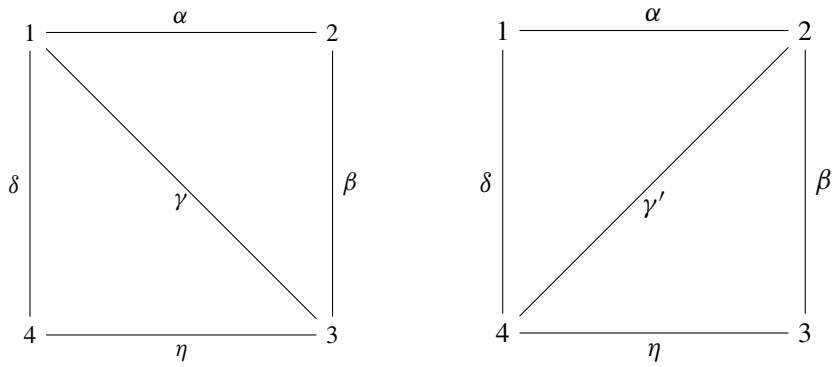


Figure 1. Mutation at the arc γ.

the variables of $\vec{x}$ except x_γ is replaced by $x_{\gamma'}$ which satisfies the exchange relation. The cluster algebra associated to the surface $S_{1,1}$ with initial seed $(T, x_1, \dots, x_{|T|})$ is the subalgebra of the ring of rational functions in the variables of the initial seed, $\mathbb{Q}(x_1, \dots, x_{|T|})$, generated by all cluster variables obtained by repeated mutation from the initial seed.

The cluster algebra associated to a triangulated surface can be geometrically realized as the coordinate ring of decorated Teichmüller space. The decorated Teichmüller space of $S_{1,1}$ is the moduli space of marked complete finite volume hyperbolic metrics on $S_{1,1}$ along with a decoration consisting of a choice of horocycle around each cusp. The cluster coordinate a_γ associated to an arc γ is the λ-length of the geodesic arc in the homotopy class of γ for the given hyperbolic metric. (See [17] for full details.) Importantly, [16] finds that Penner's decorated Teichmüller space is parametrized by the positive $\mathbb{R}$-points of this cluster algebra.

3.6. Cluster algebra and Markov triples

Considering the cluster structure on the once-punctured torus, we can now recover the Markov equation and Markov triples again.

Each Markov triple corresponds to a seed of this cluster algebra (Figure 2). The cluster mutation rule for this seed gives that the mutated coordinate a' satisfies the relation $aa' = b^2 + c^2$. This is exactly the mutation of Markov triples described above. There is a mutation invariant function on this cluster algebra $f(a, b, c) = 2\frac{a^2+b^2+c^2}{abc}$. When starting with the triple $(1, 1, 1)$, we see that $f(1, 1, 1) = 6$. Since this function is fixed by mutation, every other triple obtained by mutation also satisfies $f(a, b, c) = 6$, which is equivalent to the Markov equation. The fact that all triples obtained by mutation are positive integers is a consequence of the positivity of the cluster exchange relation.

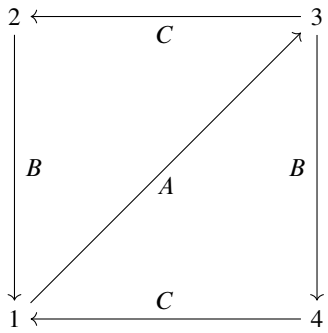


Figure 2. The Markov quiver and a seed associated to a triangulation of a punctured torus.

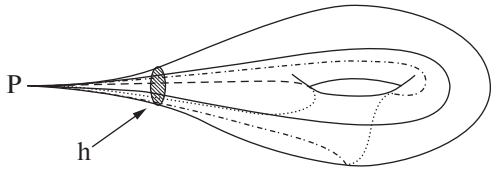


Figure 3. An ideal triangulation of a punctured torus with a horocycle. Picture taken from [13].

There is also a geometric interpretation of the mutation invariant function as the sum of "angles"

$$t_i^{jk} = \frac{a_{jk}}{a_{ij}a_{ki}}$$

in the triangulation. Each of these angles is the length of the horocycle at vertex i inside the triangle whose edges are labeled with the cluster variables a_{ij}, a_{jk}, a_{ki} (Figure 3). On the punctured torus, the total length of the horocycle is then given by the Markov function

$$f(a,b,c) = 2\left(\frac{a}{bc} + \frac{b}{ac} + \frac{c}{ab}\right) = 2\frac{a^2+b^2+c^2}{abc}. \tag{3.5}$$

The length of the horocycle is clearly independent of the chosen triangulation. Furthermore, the mutation of the Markov quiver leaves the adjacency of the triangles fixed. Thus the exact form of the function f is invariant under mutation.

4. Symplectic groups over non-commutative rings

We saw above that the Θ-Weyl group plays a key role in understanding the positive structures with respect to Θ. In fact, a lot of the structure we see in [29, 30] suggests that if G admits a positive structure with respect to Θ, it might heuristically be seen as a group of type $W(\Theta)$ over a (partly) non-commutative ring.

For Hermitian Lie groups of tube type, the Θ-Weyl group is of type A_1, and thus we would like to see a Hermitian Lie group of tube type as a group of type Sp_2 over a non-commutative ring. This leads to the definition of groups $\mathrm{Sp}_2(A,\sigma)$ over non-commutative rings A with involution σ in [2, 40], and in fact, many of the things we described in Section 3 can be generalized in this non-commutative setting. We first describe the construction of the groups and their symmetric space, before we discuss the relation with non-commutative cluster algebra, introduce non-commutative generalizations of Markov numbers, and construct $\mathrm{Sp}_2(A,\sigma)$ quantum groups.

We start with an involutive ring $(\mathcal{A},\sigma)$, where $\mathcal{A}$ is a (usually) non-commutative ring with unit and where every left inverse is also a right inverse and $\sigma\colon \mathcal{A}\to\mathcal{A}$ an anti-involution, i.e., $\sigma(AB)=\sigma(B)\sigma(A)$, and $\sigma(\sigma(A))=A$ for all $A,B\in\mathcal{A}$. We denote by $\mathcal{A}^\sigma$ the set of fixed points of σ. A key example of an involutive ring is $\mathrm{Mat}_n(k)$ the ring of $n\times n$ matrices over a field k with transpose as the anti-involution. Here the set of fixed points is the set of symmetric matrices. When $k=\mathbb{K}=\mathbb{C},\mathbb{H}$, and σ is the conjugate transpose, we have $\mathcal{A}^\sigma=\mathrm{Herm}(n,\mathbb{K})$.

We denote by $\mathrm{Mat}_2(\mathcal{A},\sigma)$ the ring of 2×2 matrices over $\mathcal{A}$.

We consider $\mathcal{A}^2=\mathcal{A}\times\mathcal{A}$ as an $\mathcal{A}$-module. Given

$$\Omega_1=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\in \mathrm{Mat}_2(\mathcal{A},\sigma),$$

we can define an $\mathcal{A}$-valued skew-symmetric right-bilinear form $\omega\colon \mathcal{A}^2\times\mathcal{A}^2\to\mathcal{A}$ by

$$\omega(\vec{v},\vec{w})=\sigma(\vec{v})^T\Omega_1\vec{w}.$$

This form is right bilinear in the sense that

$$\omega(\vec{v}B,\vec{w}A)=\sigma(B)\omega(\vec{v},\vec{w})A$$

for $A,B\in\mathcal{A}$.

The symplectic group is given by the elements of $\mathrm{Mat}_2(\mathcal{A})$ which preserve this form.

Definition 4.1. The symplectic group $\mathrm{Sp}_2(\mathcal{A},\sigma)$ is the set of right $\mathcal{A}$-linear endomorphisms of $\mathcal{A}^2$ preserving ω:

$$\mathrm{Sp}_2(\mathcal{A},\sigma)=\{f\in\mathrm{End}(A^2)\mid \forall\vec{v},\ \vec{w}\in\mathcal{A}^2:\ \omega(f(\vec{v}),f(\vec{w}))=\omega(\vec{v},\vec{w})\}.$$

In matrix form, this condition is equivalent to

$$\mathrm{Sp}_2(\mathcal{A},\sigma)=\{\mathrm{M}\in\mathrm{Mat}_2(\mathcal{A})\mid \sigma(\mathrm{M})^T\Omega_1\mathrm{M}=\Omega_1\}.$$

When $\mathcal{A} = \mathbb{R}$ and σ is the trivial involution, then $\mathrm{Sp}_2(\mathcal{A}, \sigma) \simeq \mathrm{Sp}_2(\mathbb{R})$. More generally, when $\mathcal{A} = \mathrm{Mat}_n(k)$ is the ring of $n \times n$ matrices over a field k, and σ the matrix transpose, we have $\mathrm{Sp}_2(\mathcal{A}, \sigma) \simeq \mathrm{Sp}(2n, k)$. In fact, the classical Hermitian Lie groups of tube type can be realized essentially as $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ for an appropriate choice of rings (see [2,40]); for the indefinite orthogonal groups $\mathrm{SO}(2, n)$, a small twist is necessary. When $\mathcal{A} = \mathrm{Mat}_n(\mathbb{C})$ and σ is the conjugate transpose, then $\mathrm{Sp}_2(\mathcal{A}, \sigma) \simeq \mathrm{U}_{(n,n)}$, and if $\mathcal{A} = \mathrm{Mat}_n(\mathbb{H})$ and σ is the conjugate transpose, then $\mathrm{Sp}_2(\mathcal{A}, \sigma) \simeq \mathrm{SO}^*_{(4n)}$. To realize the family of Hermitian Lie groups of tube type which are isogenous to $\mathrm{SO}(2, q)$, one has to adapt the construction a bit. In this case, $\mathcal{A}$ is the Clifford algebra associated to $\mathbb{R}^{1,q-1}$, and we have to consider a subgroup of $\mathrm{Sp}_2(\mathcal{A}, \sigma)$, where matrix coefficients take values in the Clifford group. For the exceptional group, one would like to take $\mathcal{A} = \mathrm{Mat}_3(\mathbb{O})$, but this is not a ring. Thus, we cannot describe the group as $\mathrm{Sp}_2(\mathcal{A}, \sigma)$, but most of the structural aspects of the group and the Lie algebra still survive and can be described nicely in terms of Euclidean Jordan algebras.

Note that the choice of anti-involution is crucial and dramatically affects the structure of $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ even if the underlying algebra is commutative. Consider $\mathcal{A} = \mathbb{C}$, the complex numbers. In the case of the trivial involution, the group $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ is simply $\mathrm{SL}_2(\mathbb{C})$, but in the case that the σ is complex conjugation, the group $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ can be computed to be $\mathrm{SL}_2(\mathbb{R}) \times \mathrm{U}(1)$.

Above, we described the symmetric space associated to $\mathrm{Sp}_2(\mathbb{R})$ as the space of compatible almost complex structures on $\mathbb{R}^2$ and derived several other models for the hyperbolic plane from that. Analogous constructions can be made for the groups $\mathrm{Sp}_2(\mathcal{A}, \sigma)$, starting with complex structure on $\mathcal{A}$-modules, where a complex structure on a right $\mathcal{A}$-module V is an $\mathcal{A}$-linear map $J : V \to V$ such that $J^2 = -\mathrm{id}$. The space of ω-compatible almost complex structures on $\mathcal{A}^2$ then gives a model of the symmetric space associated to $\mathrm{Sp}_2(\mathcal{A}, \sigma)$. From this model, one derives several other models, a projective model, a bounded (or precompact) model, and an upper half-space model. Thus, we realize the symmetric space as a hyperbolic plane over a non-commutative ring with involution. For details, we refer the reader to [2].

When specifying to the rings so that $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ is isomorphic to a classical Lie group G, this general construction provides nice models for the symmetric spaces of G. For classical Hermitian Lie groups of tube type, the projective model gives the Borel embedding, the precompact model gives the bounded symmetric domain model, and the upper half-space model gives the Siegel upper half-space model. But even in these cases, we obtain some new information. For example, the realization of the symmetric spaces as ω-compatible almost complex structures is usually only described for $\mathrm{Sp}(2n, \mathbb{R})$ but not for $\mathrm{SU}(n, n)$ or $\mathrm{SO}^*(4n)$. Also for the complexifications of Hermitian Lie groups of tube type, we get new models of the symmetric spaces that to our knowledge have not been described so far; see [2] for the details.

4.1. Trace identities

In [25], we derived generalizations of the $\mathrm{Sp}_2(\mathbb{R})$-trace identities for the groups $\mathrm{Sp}_2(\mathcal{A},\sigma)$.

For this, first we recall that $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is $\mathrm{Sp}_2(\mathcal{A},\sigma)$ if it satisfies the following equations:

$$\begin{aligned} \sigma(A)C - \sigma(C)A &= 0, \\ \sigma(B)D - \sigma(D)B &= 0, \\ \sigma(A)D - \sigma(C)B &= 1 = \sigma(D)A - \sigma(B)C. \end{aligned}$$

We extend σ to a map $\sigma\colon \mathrm{Mat}_2(\mathcal{A},\sigma) \to \mathrm{Mat}_2(\mathcal{A},\sigma)$ by setting

$$\sigma\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}\right) = \begin{bmatrix} \sigma(A) & \sigma(B) \\ \sigma(C) & \sigma(D) \end{bmatrix}.$$

Note that the map σ does not respect the ring structure on $\mathrm{Mat}_2(\mathcal{A},\sigma)$ in any way. In particular, for $\mathrm{M}, \mathrm{N} \in \mathrm{Mat}_2(\mathcal{A},\sigma)$, $\sigma(\mathrm{MN})$ is not equal to $\sigma(\mathrm{N})\sigma(\mathrm{M})$. The involution interacts however with the matrix transpose M^T

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^T = \begin{bmatrix} A & C \\ B & D \end{bmatrix}.$$

Note that when $\mathcal{A}$ is non-commutative, the matrix transpose also does not respect the multiplicative structure on $\mathrm{Mat}_2(\mathcal{A},\sigma)$, but the composite transform $\mathrm{M} \mapsto \sigma(\mathrm{M})^T$ is an anti-homomorphism.

We introduce $\mathcal{A}$-valued functions on $\mathrm{Mat}_2(\mathcal{A},\sigma)$, which we call the trace and determinant. When $\mathcal{A}$ is commutative and $\sigma = \mathrm{id}$, these are precisely the trace and determinants of 2×2 matrices. However, when σ is not the identity and $\mathcal{A}$ is potentially non-commutative, they have quite different properties.

The left determinant $\det_l\colon \mathrm{Mat}_2(\mathcal{A},\sigma) \to \mathcal{A}$ is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto \sigma(A)D - \sigma(C)B.$$

The right determinant $\det_r\colon \mathrm{Mat}_2(\mathcal{A},\sigma) \to \mathcal{A}$ is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto A\sigma(D) - B\sigma(C).$$

Note that $\det_r(\mathrm{M}) = \det_l(\sigma(\mathrm{M})^T)$. The trace $\mathrm{tr}\colon \mathrm{Mat}_2(\mathcal{A},\sigma) \to \mathcal{A}$ is given by

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \mapsto A + D.$$

We note that the trace function satisfies some interesting properties: For all M, Mat $N \in \mathrm{Mat}_2(\mathcal{A}, \sigma)$, we have

$$\mathrm{tr}(\sigma(\mathrm{MN})) = \mathrm{tr}(\sigma(\mathrm{N}) \cdot \sigma(\mathrm{M})).$$

Note that in general $\sigma(\mathrm{MN})$ is not equal to $\sigma(\mathrm{N})\sigma(\mathrm{M})$. If the involution σ is trivial and thus $\mathcal{A}$ is commutative, the above identity reduces to the classical commutativity property of the trace $\mathrm{tr}(\mathrm{MN}) = \mathrm{tr}(\mathrm{NM})$, which further implies the conjugation invariance of the trace function $\mathrm{tr}(\mathrm{NMN}^{-1}) = \mathrm{tr}(\mathrm{M})$. In general, the trace function is however not invariant.

If $\mathrm{M} \in \mathrm{Sp}_2(\mathcal{A}, \sigma)$, then $\mathrm{M}^{-1} = \left[\begin{smallmatrix} \sigma(D) & -\sigma(B) \\ -\sigma(C) & \sigma(A) \end{smallmatrix}\right]$, and $\mathrm{tr}(\mathrm{M}^{-1}) = \sigma(\mathrm{tr}(\mathrm{M}))$. This is familiar from $\mathrm{Sp}_2(\mathbb{R})$.

In [25], several trace identities are established.

(1) *Classical trace identity*: For $\mathrm{M}, \mathrm{N} \in \mathrm{Mat}_2(\mathcal{A})$, we have

$$\mathrm{tr}(\mathrm{Adj}(\mathrm{M})\mathrm{N}) + \mathrm{tr}(\sigma(\mathrm{M})\mathrm{N}) - \sigma(\mathrm{tr}(\mathrm{M}))\,\mathrm{tr}(\mathrm{N}) = 0,$$

where the adjoint matrix of $\mathrm{M} \in \mathrm{Mat}_2(\mathcal{A}, \sigma)$ is given by

$$\mathrm{Adj}(\mathrm{M}) = \begin{bmatrix} \sigma(D) & -\sigma(B) \\ -\sigma(C) & \sigma(A) \end{bmatrix}.$$

(2) *Cayley–Hamilton theorem*: An element $\mathrm{M} \in \mathrm{Mat}_2(\mathcal{A}, \sigma)$ lies in $\mathrm{Sp}_2(\mathcal{A}, \sigma)^l$ if and only if it satisfies

$$\sigma(\mathrm{M})\mathrm{M} - \sigma(\mathrm{tr}(\mathrm{M}))\mathrm{M} + \mathrm{id} = 0.$$

These trace identities are established in [25] by explicit computations, but we expect that there is a more natural underlying structure, which relies on considering a non-commutative algebra $\mathcal{A}$ with involution σ and modules over $\mathcal{A}$.

5. Relation to Manin matrices

There is an interesting connection to Manin matrices, introduced by Manin in [35,36]. They define some kind of SL_2 over non-commutative rings and play a role in quantizations of SL_2.

A matrix $\mathrm{M} = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathrm{Mat}_2(R)$ is Manin if it satisfies the Manin commutativity equations:

$$AC - CA = 0, \quad BD - DB = 0, \quad AD - CB = DA - BC.$$

Note that in the first two of Manin's commutativity equations, if we replace products xy with $\sigma(x)y$, we get precisely the first two conditions for a matrix to be in $\mathrm{Sp}_2(\mathcal{A},\sigma)$. If we add the additional constraint to a Manin matrix that $AD-CB=1$, then the third equation matches up as well.

Manin matrices can be characterized by their inverse: A matrix $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is Manin if and only if its left inverse is given by $\frac{1}{AD-CB}\left[\begin{smallmatrix} D & -B \\ -C & A \end{smallmatrix}\right]$ and also by satisfying the Cayley–Hamilton theorem. Additionally, a matrix $M=\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ is Manin if and only if it satisfies

$$M^2-\mathrm{tr}(M)M+\mathrm{id}=0.$$

If a matrix and its transpose are both Manin, then the entries are all forced to pairwise commute. In our situation, this condition simply means that $\mathrm{M}\in\mathrm{Sp}_2(\mathcal{A},\sigma)$. This makes $\mathrm{Sp}_2(\mathcal{A},\sigma)$ a more natural generalization of SL_2 over non-commutative rings than Manin matrices. For more similarities, we refer the reader to [25].

6. Quantum groups

The symplectic groups $\mathrm{Sp}_2(\mathcal{A},\sigma)$ can be quantized, and this quantization parallels the quantization of SL_2 [31,35,36].

Informally, the quantum SL_2 is a deformation in a parameter q of the algebra of matrix functions on SL_2. Formally, we fix a field k. Let $k_q=k(q)$ be the extension of k by an element q. Often, q is taken to be an element of k, although this introduces additional subtlety if q is a root of unity. The coordinate ring of the quantum plane $\mathcal{A}_q$ is the algebra $k_q[x_1,x_2]$ such that x_1 and x_2 q-commute, i.e., $x_1x_2=q^{-1}x_2x_1$. Intuitively, the quantization of $\mathrm{SL}_2(k)$ should act like the endomorphisms of the quantum plane.

The algebra of quantum matrix functions $\mathcal{O}(\mathrm{SL}_2)_q$ is the ring $k_q\{a,b,c,d\}/I$ where I is generated by the quantum SL_2 relations:

$$\begin{aligned} ac-q^{-1}ca&=0, & ab-q^{-1}ba&=0,\\ bd-q^{-1}db&=0, & cd-q^{-1}dc&=0,\\ ad-q^{-1}cb&=da-qbc, & bc&=cb,\\ ad-q^{-1}cb&=1. \end{aligned} \tag{6.1}$$

These equations can be obtained by adding the appropriate power of q to the equations for a matrix and its transpose to be Manin. To better compare with $\mathrm{Sp}_2(\mathcal{A},\sigma)$,

we can rewrite the quantum SL_2 relations as follows:

$$\begin{aligned} ac - q^{-1}ca &= 0, & ab - q^{-1}ba &= 0, \\ bd - q^{-1}db &= 0, & cd - q^{-1}dc &= 0, \\ ad - q^{-1}cb &= 1, & ad - q^{-1}bc &= 1, \\ da - qbc &= 1, & da - qcb &= 1. \end{aligned} \tag{6.2}$$

This algebra is actually a Hopf algebra. We will not address this here but refer the reader to [25,31] for more details.

To quantize $\mathrm{Sp}_2(\mathcal{A}, \sigma)$, we first add a quantum parameter q to our algebra $\mathcal{A}$. Formally, let $\mathcal{A}_q = \mathcal{A}[q^{\pm}]$ with the additional assumption that q commutes with all of $\mathcal{A}$. We then extend σ to $\mathcal{A}_q$ by setting $\sigma(q) = q^{-1}$.

Following the quantization of $\mathrm{SL}_2(\mathbb{R})$, we define a system of equations on the matrix entries of a 2×2 matrix specifying the quantum commutativity and quantum determinant of the elements.

The quantum symplectic equations for a matrix $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] \in \mathrm{Mat}_2(\mathcal{A}_q)$ are

$$\begin{aligned} \sigma(A)C - q^{-1}\sigma(C)A &= 0, & A\sigma(B) - q^{-1}B\sigma(A) &= 0, \\ \sigma(B)D - q^{-1}\sigma(D)B &= 0, & C\sigma(D) - q^{-1}D\sigma(C) &= 0, \\ \sigma(A)D - q^{-1}\sigma(C)B &= 1, & A\sigma(D) - q^{-1}B\sigma(C) &= 1. \end{aligned} \tag{6.3}$$

Note that we consider $\sigma(A)D - q^{-1}\sigma(C)D$ to be the quantum determinant.

Let us point out that these equations correspond both to standard symplectic equations and to rephrased quantum $\mathrm{SL}_2(\mathbb{R})$ equations (6.2). We stress that these equations do not (!) imply any commutativity between B and C.

Usually, when we define a quantum deformation of an algebra, we are constructing a non-commutative algebra out of a commutative one. Here we already start with a non-commutative algebra, so its quantization is a much less drastic change. In fact, the corresponding quantization can be seen simply as a change of symplectic form. The equations above are equivalent to M having a two-sided inverse and preserving the $\mathcal{A}_q$-valued skew-symmetric right-bilinear form $\omega_q : \mathcal{A}_q^2 \times \mathcal{A}_q^2 \to \mathcal{A}_q$ by

$$\omega_q(\vec{v}, \vec{w}) = \sigma(\vec{v})^T \Omega_q \vec{w},$$

with

$$\Omega_q = \begin{bmatrix} 0 & 1 \\ -q^{-1} & 0 \end{bmatrix} \in \mathrm{Mat}_2(\mathcal{A}_q).$$

The corresponding group $\mathrm{Sp}_2(\mathcal{A}, \sigma)_q$ acts on the left quantum plane, $(\mathcal{A}_\omega^2)_q$, which is the set of generic left q-commuting elements of $\mathcal{A}_q^2$:

$$(\mathcal{A}_\omega^2)_q = \{\vec{v} = (V_1, V_2) \in \mathcal{A}_q^2 \mid \vec{v} \text{ generic and } q^{-1}\sigma(V_2)V_1 = \sigma(V_1)V_2\}.$$

In fact, and contrary to the commutative situation, the group $\mathrm{Sp}_2(\mathcal{A},\sigma)_q$ has many elements if we assume that q has a square root $q^{1/2}$. This is very different from the usual quantization of $\mathrm{SL}_2(\mathbb{R})$ where the corresponding group is just $\left\{\left[\begin{smallmatrix} a & 0 \\ 0 & a^{-1}\end{smallmatrix}\right] \mid a \in \mathbb{R}\right\}$. Here $\mathrm{M} = \left[\begin{smallmatrix} A & B \\ C & D\end{smallmatrix}\right] \in \mathrm{Sp}_2(\mathcal{A},\sigma)$, then $\left[\begin{smallmatrix} A & q^{1/2}B \\ q^{-1/2}C & D\end{smallmatrix}\right] \in \mathrm{Sp}_2(\mathcal{A},\sigma)_q$.

One can further introduce the algebra of quantum symplectic matrix functions and endow it with an appropriate Hopf algebra-like structure. For this construction, we refer the reader to [25]. There the reader will also find a discussion of the relation to the classical algebra of quantum matrix functions when $\mathcal{A} = k$ is a field and σ is the trivial involution.

Open Question 6.1. Is there an interpretation of this quantization in terms of double Poisson brackets?

7. Non-commutative cluster algebras

In [6], Berenstein and Retakh introduced a non-commutative generalization of the A_1-cluster algebras associated to a surface.

For the non-commutative A_1-type cluster algebra associated to a surface S, to construct a seed, we start again with an ideal triangulation of the surface S. Each possible orientation of an edge $i \to j$ of the triangulation is assigned a variable X_{ij} in a non-commutative algebra. There is an additional condition for each triangle cyclically labeled i, j, k. Berenstein and Retakh introduce a non-commutative angle $T_i^{jk} = X_{ji}^{-1} X_{jk} X_{ik}^{-1}$. The seed has to satisfy the triangle condition: $T_i^{jk} = T_i^{kj}$. The non-commutative cluster algebra is then again grown through mutation. Each mutable edge $i \to k$ is contained in a unique quadrilateral with vertices (i, j, k, ℓ). The mutated seed contains the triangulation with $i \to k$ replaced with $j \to \ell$. The element assigned to $j \to \ell$ is given by the formula

$$X_{j\ell} = X_{jk} X_{ik}^{-1} X_{i\ell} + X_{ji} X_{ki}^{-1} X_{k\ell}.$$

We want to relate the variables associated to the same edge with different orientations, so we set $\sigma(X_{ij}) = X_{ji}$ and extend σ as an anti-involution. Using σ, we see that the triangle condition reads as

$$T_i^{jk} = T_i^{kj} = \sigma(T_i^{jk})$$

so that each non-commutative angle is in $C(S)^\sigma$.

The mutation rule is equivalent to the requirement that the angles are additive:

$$\begin{aligned} T_i^{j\ell} &= T_i^{jk} + T_i^{k\ell}, \\ X_{ji}^{-1} X_{j\ell} X_{i\ell}^{-1} &= X_{ji}^{-1} X_{jk} X_{ik}^{-1} + X_{ki}^{-1} X_{k\ell} X_{i\ell}^{-1}. \end{aligned}$$

The non-commutative cluster structure can also be used to parametrize the positive semigroups $U_\Theta^{>0}$ and $G_\Theta^{>0}$ for classical Hermitian Lie groups of tube type. This is discussed in [23, 25].

Recall that in the commutative A_1-cluster algebra associated to the surface S, we had a geometric realization by looking at parametrizations of the decorated Teichmüller space by Penner's λ-lengths. The decorated Teichmüller space can be realized as the positive points of the $\mathbb{R}$-points of the cluster algebra.

There is also a geometric realization of the non-commutative cluster algebra using positivity. In [3], the authors introduced non-commutative Λ-lengths for decorated representations $\pi_1(S) \to \mathrm{Sp}(2n, \mathbb{R})$ and show that these Λ-lengths give a geometric realization of the Berenstein–Retakh cluster algebra such that the appropriate positive points are precisely the space of decorated maximal representations, i.e., the space of positive representations with respect to the positive structure on $\mathrm{Sp}(2n, \mathbb{R})$ with respect to $\Theta = \{\alpha_n\}$. See also [25, 33, 40] for more general symplectic groups $\mathrm{Sp}_2(\mathcal{A}, \sigma)$ over non-commutative rings and [21, 33] for related work on GL_n over non-commutative rings.

These geometric realizations of the non-commutative cluster algebra of type A_1 provide non-commutative analogues of the $\mathcal{A}$-cluster variety which was introduced by Fock and Goncharov in the setting of split real Lie groups. In [3], the authors also provide $\mathcal{X}$-type coordinates, but they are not well adapted to the cluster structures. Goncharov and Shen [22] introduced $\mathcal{P}$-cluster varieties, which seem to be more appropriate for a non-commutative generalization.

Open Questions 7.1. We propose the following problems:

(1) Provide appropriate non-commutative $\mathcal{P}$-cluster varieties of type A_1.

(2) In [15], tropical points of cluster varieties play an important role. In the A_1-situation, they parametrize laminations on the surface. Provide appropriate tropical non-commutative $\mathcal{A}$-cluster varieties and $\mathcal{P}$-cluster varieties and geometrically describe what they parametrize.

(3) For split real Lie groups, Fock and Goncharov [15] formulated wide-ranging duality conjectures, which have been proved in [26]. It is an open question if there are any dualities for Lie groups admitting a positive structure with respect to Θ.

8. Non-commutative Markov numbers

We now use the non-commutative A_1-cluster algebra structure to obtain non-commutative generalizations of Markov numbers. For this, we replace the real numbers $\mathbb{R}$ by an involutive algebra $(\mathcal{A}, \sigma)$. For details, we refer the reader to [25]. We start with the quiver associated to the once-punctured torus; see Figure 4.

Now we do not have a horocycle and the length of the pieces of this horocycle to build our invariant function. We considered the lengths of horocycle arcs t_i^{jk}; they are now "upgraded" to the non-commutative angle T_i^{jk}, which appear in the triangle conditions of the non-commutative cluster algebra. We require these angles to be fixed by the anti-involution σ.

Explicitly, in A, B, C are the seed variables (see Figure 4); these angles are all cyclic permutations of

$$\sigma(A^{-1})B\sigma(C^{-1}) \in \mathcal{A}^\sigma \quad \sigma(A^{-1})C\sigma(B^{-1}) \in \mathcal{A}^\sigma.$$

The non-commutative mutation at the node A is given by the equation

$$A' = \sigma(C)A^{-1}\sigma(C) + B\sigma(A^{-1})B. \tag{8.1}$$

The generalized Markov function, which is invariant under mutations, for a seed (A, B, C) whose triangles are cyclically oriented, is

$$\begin{aligned} F(A,B,C) = {} & \sigma(A^{-1})B\sigma(C^{-1}) + \sigma(B^{-1})C\sigma(A^{-1}) + \sigma(C^{-1})A\sigma(B^{-1}) \\ & + \sigma(C^{-1})B\sigma(A^{-1}) + \sigma(A^{-1})C\sigma(B^{-1}) + \sigma(B^{-1})A\sigma(C^{-1}). \end{aligned} \tag{8.2}$$

The reader might check that if the involution σ is trivial and a, b, c commute, this simplifies F to $F(a,b,c) = 2\frac{a^2+b^2+c^2}{abc}$ which is the standard Markov function.

Now we can use the generalized Markov function to consider generalizations of Markov triples. In the classical case, we get Markov triples by starting with the seed

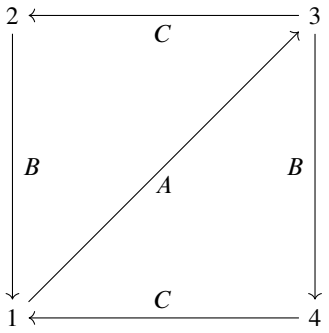


Figure 4. Seed associated to a triangulation of a punctured torus.

$(A, B, C) = (1, 1, 1)$, and by mutation, we get only integer solutions. There are two important things we have to take into account:

(1) *Integrality*: From the cluster theory, there is nothing special about the integers as a subring of $\mathbb{R}$. But the Laurent phenomenon for cluster algebras states that every number obtained by mutation can be expressed as a Laurent polynomial in the original triple (a, b, c). Thus, if (a, b, c) are units in any subring, all values obtained by mutation belong to that subring. Thus, for non-commutative generalizations, we are interested in starting with a seed that takes values in units of an appropriate subring of the integral elements in $(\mathcal{A}, \sigma)$. Recall that the integral elements of a ring $\mathcal{A}$ are the roots of monic polynomials with integer coefficients in $\mathcal{A}$.

(2) *Positivity*: Another subtlety when choosing an initial triple is that we require every element of the cluster algebra to be invertible. Simply choosing a triple of invertible elements is not enough to guarantee that all elements obtained by mutation are invertible. For example, in $\mathbb{C}$ with the trivial involution, mutation at 1 in the triple $(1, 1, i)$ results in the triple $(0, 1, i)$. A sufficient solution to avoid these triples is to require that all non-commutative angles in the initial seed are "positive" in the appropriate sense.
Over $\mathbb{R}$, it is clear what positive means, but over arbitrary involutive rings, it is less clear. We define a positive structure on $(\mathcal{A}, \sigma)$ to be a subset P of $\mathcal{A}^\sigma$ containing 1 and be closed under addition, inversion, and σ-congruence $(A \mapsto \sigma(B)AB)$ by invertible elements. Here we see the connection with the positive structure on Hermitian Lie groups of tube type.

Thus, putting these things together, we fix a subring R of integral elements of $\mathcal{A}$ and a positive structure P on $(\mathcal{A}, \sigma)$. An admissible seed (with respect to R and P) is then a triple $(a, b, c) \in \mathcal{A}^3$, where a, b, c are units in R and all the non-commutative angles are in P.

A simple way to get an admissible triple is to take a unit B in R, then the triple $(1, B, B^{-1})$ is admissible.

With this general setup, we now obtain several examples of non-commutative Markov triples.

8.1. Matrix Markov numbers

Consider $\mathcal{A} = \mathrm{Mat}(n, \mathbb{R})$ with σ being the standard matrix transpose. We choose $R = \mathrm{Mat}_n(\mathbb{Z})$ as the subring of integral elements to focus on. The positive structure P consists of the set of positive definite symmetric matrices.

There are many complicated examples of admissible triples; let us describe just the simplest one (see [25] for more examples). For this, let $B_t = \left[\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right]$ and consider

the seed $(1, B_t, B_t^{-1})$. Then the value of the Markov function is

$$\begin{bmatrix} 6+2t^2 & 0 \\ 0 & 6+2t^2 \end{bmatrix}.$$

The matrices we obtain are matrix Markov numbers. Then $\frac{1}{2}$ times the trace of each matrix Markov number is a polynomial with integral coefficients $p(t)$, where $p(0)$ is a classical Markov number. Thus, in this case, $\frac{1}{2}$ trace of each matrix Markov number can be considered as a one-dimensional deformation of classical Markov numbers.

8.2. Complex and dual Markov numbers

We also obtain interesting examples for commutative rings with nontrivial involutions σ. One interesting example is to consider the complex numbers $\mathbb{C}$ with the involution σ which is the complex conjugation; see Figure 5. Then $\mathcal{A}^\sigma = \mathbb{R}$. For the ring of integral elements R, we take the roots of unity. The positive structure is given by the positive real numbers.

A simple admissible seed triple is a triple $z_1, z_2, z_3 \in \mathbb{C}$ of three roots of unity whose product is 1. If these roots of unity are chosen to be nontrivial, the complex Markov numbers we get differ from the real Markov numbers. In this case, the mutation rule splits into a mutation rule for the norm, which is just the classical mutation rule for Markov triples, and a mutation rule for the arguments. Therefore, the norms of these complex Markov numbers are classical real Markov numbers. It is however interesting to understand the arguments of these complex Markov numbers. In [25], the following conjecture is made.

Conjecture 8.1. *If the initial seed triple consists of three roots of unity (z_1, z_2, z_3) such that n is the minimal integer with $z_i^n = 1$, then every triple of $\frac{a}{n}, \frac{b}{n}, \frac{c}{n}$ that sums to 1 is achieved as the argument of a complex Markov triple. Furthermore, if (z_1, z_2, z_3) are irrational angles, then the image of the arguments of Markov triples is dense in the circle.*

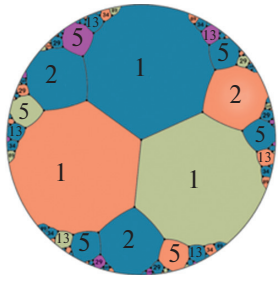


Figure 5. The Markov "tree" for complex Markov numbers.

Another interesting example is the dual numbers $\{a + \alpha\varepsilon \mid a, \alpha \in \mathbb{R},\ \varepsilon^2 = 0\}$, and the subring R of integral elements of the form $a + \alpha\varepsilon$ with $a, \alpha \in \mathbb{Z}$. The units of this subring are elements of the form $1 + \alpha\varepsilon$ for $\alpha \in \mathbb{Z}$. We can write dual numbers in the form $a(1 + \alpha\varepsilon)$ with $a \neq 0$. This is analogous to the polar form of a complex number.

If one considers the trivial involution, one obtains dual Markov numbers which have also been studied in [10]. The more interesting case is again when σ is the non-trivial involution given by the conjugation map $a + \alpha\varepsilon \mapsto a - \alpha\varepsilon$. In this case, the mutation rule splits again; see [25].

9. Beyond the A_1-case

In the above sections, we illustrated how the non-commutative perspective suggested by positivity leads to a fresh perspective on a variety of things. We mostly focused on the A_1-case which arises from the positive structure on Hermitian Lie groups of tube type with respect to a subset Θ of Δ consisting of one element. This already extends beyond just the family of Hermitian Lie groups of tube type. When considering the groups $\mathrm{Sp}_2(\mathcal{A}, \sigma)$, the trace identities, their quantizations, or the corresponding Markov triples, the non-commutative algebra $\mathcal{A}$ does not have to be of Hermitian type – any (non-commutative) ring $\mathcal{A}$ with ant-involution σ works. And actually, one does not have to stop at skew-symmetric bilinear forms but can also consider symmetric bilinear forms over $(\mathcal{A}, \sigma)$; see [2].

The phenomena we describe in fact suggest that there is a more fundamental approach to explain some of the observations by considering σ-commutative modules over $(\mathcal{A}, \sigma)$. More on this perspective will appear in forthcoming work with Dani Kaufmann.

Going back to the families of simple Lie groups admitting a positive structure, the next family is formed by the indefinite orthogonal groups (or indefinite spin groups), whose Θ-Weyl group is of type B_p. Here the positive structure suggests looking for a partially non-commutative structure. One can in fact introduce partially non-commutative cluster structures of type B_p. This leads to the introduction of non-commutative polygonal cluster algebras [23].

In [24], building on the work of Goncharov and Shen [22], we introduce more general partially non-commutative cluster structures associated to more general Θ-Weyl groups. These cluster structures can be used to parametrize the open semigroups $U_\Theta^{>0}$ and $G_\Theta^{>0}$, to give explicit positivity tests, and provide (partially) non-commutative cluster structure on the space of decorated positive representations.

Acknowledgments. I thank my collaborators for taking this mathematical journey with me. I also thank the Hector Fellow Academy for its support.

Funding. This work was partially supported by the European Research Council ERC-Advanced Grant 101018839 and by the Deutsche Forschungsgemeinschaft TRR 191.

References

[1] M. Aigner, *Markov's theorem and 100 years of the uniqueness conjecture. A mathematical journey from irrational numbers to perfect matchings*. Springer, Cham, 2013, 257 pp. Zbl 1276.00006 MR 3098784

[2] D. Alessandrini, A. Berenstein, V. Retakh, E. Rogozinnikov, and A. Wienhard, Symplectic groups over noncommutative algebras. *Selecta Math. (N.S.)* **28** (2022), no. 4, article no. 82, 119 pp. Zbl 1503.53111 MR 4482280

[3] D. Alessandrini, O. Guichard, E. Rogozinnikov, and A. Wienhard, Noncommutative coordinates for symplectic representations. *Mem. Amer. Math. Soc.* **300** (2024), no. 1504, 116 pp. Zbl 1560.22001 MR 4790264

[4] N. Arkani-Hamed and J. Trnka, The amplituhedron. *J. High Energy Phys.* **2014** (2014), no. 10, article no. 30, 33 pp. Zbl 1468.81075

[5] G. Ben Simon, M. Burger, T. Hartnick, A. Iozzi, and A. Wienhard, On order-preserving representations. *J. Lond. Math. Soc. (2)* **94** (2016), no. 2, 525–544 Zbl 1398.57043 MR 3556452

[6] A. Berenstein and V. Retakh, Noncommutative marked surfaces. *Adv. Math.* **328** (2018), 1010–1087 Zbl 1395.14001 MR 3771148

[7] A. Berenstein and A. Zelevinsky, Total positivity in Schubert varieties. *Comment. Math. Helv.* **72** (1997), no. 1, 128–166 Zbl 0891.20030 MR 1456321

[8] J. Beyrer, O. Guichard, F. Labourie, B. Pozzetti, and A. Wienhard, Positivity, cross-ratios and the Collar Lemma. 2024, arXiv:2409.06294v1

[9] J. Beyrer and M. B. Pozzetti, Positive surface group representations in $\mathrm{PO}(p,q)$. *J. Eur. Math. Soc. (JEMS)* (2024), DOI 10.4171/JEMS/1560

[10] N. Bonin and V. Ovsienko, A shadow Markov equation. *C. R. Math. Acad. Sci. Paris* **361** (2023), 1483–1489 Zbl 1552.11029 MR 4683325

[11] M. Burger, A. Iozzi, and A. Wienhard, Surface group representations with maximal Toledo invariant. *Ann. of Math. (2)* **172** (2010), no. 1, 517–566 Zbl 1208.32014 MR 2680425

[12] R. D. Canary, Anosov representations: informal lecture notes. 2021, https://websites.umich.edu/∼canary/lecnotespublic61125.pdf visited on 13 January 2026

[13] N. Drukker, J. Gomis, T. Okuda, and J. Teschner, Gauge theory loop operators and Liouville theory. *J. High Energy Phys.* **2010** (2010), no. 2, article no. 57, 62 pp. Zbl 1270.81134 MR 2672742

[14] J. Faraut and A. Korányi, *Analysis on symmetric cones*. Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York, 1994, 382 pp. Zbl 0841.43002 MR 1446489

[15] V. Fock and A. Goncharov, Moduli spaces of local systems and higher Teichmüller theory. *Publ. Math. Inst. Hautes Études Sci.* **103** (2006), no. 1, 1–211 Zbl 1099.14025 MR 2233852

[16] V. V. Fock and A. B. Goncharov, Dual Teichmüller and lamination spaces. In *Handbook of Teichmüller theory. Volume I*, pp. 647–684, IRMA Lect. Math. Theor. Phys. 11, European Mathematical Society, Zürich, 2007 Zbl 1162.32009 MR 2349682

[17] S. Fomin and D. Thurston, Cluster algebras and triangulated surfaces Part II: lambda lengths. *Mem. Amer. Math. Soc.* **255** (2018), article no. 1223, 97 pp. Zbl 07000309 MR 3852257

[18] R. Fricke and F. Klein, *Vorlesungen über die Theorie der automorphen Functionen*. Erster Band: Die gruppentheoretischen Grundlagen, B. G. Teubner, Leipzig, 1897, 634 pp. JFM 28.0334.01 MR 0183872

[19] F. R. Gantmacher and M. G. Krein, Sur les matrices oscillatoires. *C. R. Acad. Sci. Paris* **201** (1935), 577–579 Zbl 0012.28903

[20] A. Ghosh and P. Sarnak, Integral points on Markoff type cubic surfaces. *Invent. Math.* **229** (2022), no. 2, 689–749 Zbl 1510.11088 MR 4448994

[21] A. Goncharov and M. Kontsevich, Spectral description of non-commutative local systems on surfaces and non-commutative cluster varieties. In *Arithmetic and algebraic geometry. A mathematical tribute to Yuri Manin*, pp. 109–238, Simons Symp., Springer, Cham, 2024 Zbl 08107625 MR 4890446

[22] A. Goncharov and L. Shen, Quantum geometry of moduli spaces of local systems and representation theory. [v1] 2019, [v4] 2024, arXiv:1904.10491v4

[23] Z. Greenberg, D. Kaufman, M. Niemeyer, and A. Wienhard, Noncommutative polygonal cluster algebras. 2024, arXiv:2410.08813v1, to appear in *J. Lond. Math. Soc. (2)*

[24] Z. Greenberg, D. Kaufman, M. Niemeyer, and A. Wienhard, Work in progress. 2026, forthcoming

[25] Z. Greenberg, D. Kaufman, and A. Wienhard, SL_2-like properties of matrices over non-commutative rings and generalizations of Markov numbers. *Adv. Math.* **482, part A** (2025), article no. 110556, 62 pp. Zbl 08131458 MR 4968025

[26] M. Gross, P. Hacking, S. Keel, and M. Kontsevich, Canonical bases for cluster algebras. *J. Amer. Math. Soc.* **31** (2018), no. 2, 497–608 Zbl 1446.13015 MR 3758151

[27] O. Guichard, F. Labourie, and A. Wienhard, Positivity and representations of surface groups. *Forum Math. Pi* **14** (2026), article no. e6, 38 pp. MR 5029150

[28] O. Guichard and A. Wienhard, Positivity and higher Teichmüller theory. In *European Congress of Mathematics*, pp. 289–310, European Mathematical Society , Zürich, 2018 Zbl 1404.22034 MR 3887772

[29] O. Guichard and A. Wienhard, Generalizing Lusztig's total positivity. *Invent. Math.* **239** (2025), no. 3, 707–799 Zbl 1564.22006 MR 4861126

[30] O. Guichard and A. Wienhard, Positivity in Lie groups. 2026, forthcoming

[31] C. Kassel, *Quantum groups*. 1st edn., Grad. Texts in Math. 155, Springer, New York, 1995, 534 pp. Zbl 0808.17003 MR 1321145

[32] F. Kassel, Geometric structures and representations of discrete groups. In *Proceedings of the International Congress of Mathematicians (ICM 2018), Rio de Janeiro, Brazil, August 1–9, 2018*, pp. 1115–1151, Invited lectures II, World Scientific, Hackensack, NJ; Sociedade Brasileira de Matemática (SBM), Rio de Janeiro, 2018 Zbl 1447.57028 MR 3966802

[33] C. Kineider and E. Rogozinnikov, On partial abelianization of framed local systems. *Math. Ann.* **392** (2025), no. 1, 797–836 Zbl 08025866 MR 4887775

[34] G. Lusztig, Total positivity in partial flag manifolds. *Represent. Theory* **2** (1998), 70–78 (Published electronically) Zbl 0895.14014 MR 1606402

[35] Y. I. Manin, Some remarks on Koszul algebras and quantum groups. *Ann. Inst. Fourier (Grenoble)* **37** (1987), no. 4, 191–205 Zbl 0625.58040 MR 0927397

[36] Y. I. Manin, *Quantum groups and noncommutative geometry*. 2nd edn., CRM Short Courses, Springer, Cham, 2018, 125 pp. Zbl 1430.16001 MR 3839605

[37] A. Markoff, Sur les formes quadratiques binaires indéfinies. *Math. Ann.* **17** (1880), no. 3, 379–399 JFM 12.0143.02 MR 1510073

[38] C. Procesi, The invariant theory of $n \times n$ matrices. *Adv. Math.* **19** (1976), no. 3, 306–381 Zbl 0331.15021 MR 0419491

[39] K. Rietsch, An algebraic cell decomposition of the nonnegative part of a flag variety. *J. Algebra* **213** (1999), no. 1, 144–154 Zbl 0920.20041 MR 1674668

[40] E. Rogozinnikov, *Symplectic groups over noncommutative rings and maximal representations*. Ph.D. thesis, Heidelberg University, 2020

[41] I. Schoenberg, Über variationsvermindernde lineare Transformationen. *Math. Z.* **32** (1930), no. 1, 321–328 JFM 56.0106.06 MR 1545169

[42] A. Wienhard, An invitation to higher Teichmüller theory. In *Proceedings of the International Congress of Mathematicians (ICM 2018), Rio de Janeiro, Brazil, August 1–9, 2018*, pp. 1013–1039, Invited lectures II, World Scientific, Hackensack, NJ; Sociedade Brasileira de Matemática (SBM), Rio de Janeiro, 2018 Zbl 1447.32023 MR 3966798

[43] L. K. Williams, The positive Grassmannian, the amplituhedron, and cluster algebras. In *International Congress of Mathematicians (ICM 2022), Helsinki, Finland, July 6–14, 2022*, pp. 4710–4737, Vol. 6, Sections 12–14, EMS Press, Berlin, 2022 Zbl 1555.14105 MR 4680421

[44] K. Xie, Theta positivity in Lagrangian Grassmannian. *Int. Math. Res. Not. IMRN* **2025** (2025), no. 15, article no. rnaf240, 23 pp. Zbl 08085594 MR 4945123

[45] K. Xie and A. Wienhard, Work in progress. 2026, forthcoming

Anna Wienhard
Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, 04103 Leipzig, Germany;
wienhard@mis.mpg.de

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/7

Sign patterns of the Liouville function

Tamar Ziegler

Abstract. The last two decades have seen remarkable advances in our understanding of important arithmetic functions. We describe the intertwining of major developments in ergodic theory, number theory and additive combinatorics, leading to recent progress made towards the Chowla and Sarnak conjectures.

1. Introduction

The Liouville function $\lambda(n)$ is a very important arithmetic function. It takes the value -1 on the prime numbers and extends multiplicatively to a completely multiplicative function on the natural numbers. It is intimately related to the prime numbers. For example, the celebrated prime number theorem on the number of primes up to N is equivalent to a randomness statement about cancellation in the sum of the values of Liouville:

$$\sum_{n\leq N}\lambda(n) = o(N).$$

Furthermore, the Riemann hypothesis is equivalent to having square-root cancellation in the sum above:

$$\sum_{n\leq N}\lambda(n) = O_\varepsilon(N^{1/2+\varepsilon}).$$

One can view the above as a testament to "random behavior" of the Liouville function; indeed, square-root cancellation is exhibited by the classical random walk on the integer line.

The prime number theorem can be viewed as an instance of a general vague principle in analytic number theory concerning the randomness of the Liouville function, stating that the Liouville function should be orthogonal to any "structured"

Mathematics Subject Classification 2020: 11N37 (primary); 11B30, 37A44, 37A15 (secondary).
Keywords: Sarnak conjecture, Chowla conjecture, Gowers norms, Gowers–Host–Kra norms.

sequence ([16] Section 13.1). In 2010, P. Sarnak [24] suggested a formalization of this principle in the form of the following remarkable conjecture.

Conjecture 1.1 (Sarnak (2010)). *The Liouville function λ is orthogonal to any sequence arising from a topological dynamical system of zero entropy.*

A topological dynamical system is a pair (X, T), where X is a compact metric space, $T : X \to X$ a continuous map. The topological entropy of the system $h(X, T)$ is a measure of the complexity of the system – measuring the exponential rate of growth of the number of orbits. Any continuous function $f : X \to \mathbb{C}$ and any $x \in X$ give rise to a sequence $a(n) = f(T^n x)$. Sarnak's conjecture states that for any such sequence arising from a system of zero topological entropy, one should have

$$\sum_{n \leq N} \lambda(n) f(T^n x) = o(N).$$

We remark that if we endow the space (X, T) with an invariant Borel probability measure ν, then the above estimates hold ν-almost surely *without* the zero entropy condition; this is a classical result of Bourgain [2].

This wonderful conjecture may look very surprising, but as observed by Sarnak, it follows from a well-known old conjecture of Chowla [4] in number theory.

Conjecture 1.2 (Chowla (1965)). *For any k and any integers $0 \leq h_1 < \cdots < h_k$, we have*

$$\sum_{n \leq N} \mu(n + h_1) \cdots \mu(n + h_k) = o(N).$$

One can think of Chowla's conjecture as the multiplicative function version of a famous conjecture of Hardy and Littlewood, which predicts that given a k-tuple of integers $\mathcal{H} = \{h_1, \ldots, h_k\}$, there are infinitely many k-tuples

$$x + h_1, \ldots, x + h_k$$

such that all elements are simultaneously prime unless there is an obvious divisibility obstruction.

1.1. Entropy of the Liouville system

The Liouville function λ naturally gives rise to a topological dynamical system. Consider the sequence space $X = \{\pm 1\}^{\mathbb{N}}$. This is the set of all ± 1 valued sequences, with the product topology. This is a compact metric space. The topological space X is naturally endowed with a continuous map $T : X \to X$ – the left shift given by $(Tx)(n) = x(n + 1)$. We can consider the Liouville function λ as a point in X and

then consider $X_\lambda = \overline{\{T^n\lambda\}}_{n\in\mathbb{N}}$. Then (X_λ, T) is a topological dynamical system. Now λ arises naturally as a sequence in this system: indeed, consider the continuous function f_0 defined on X given by $f_0(x) = x(0)$. Then $f_0(T^n\lambda) = \lambda(n)$.

Lemma 1.3. *Sarnak's conjecture implies that $h(X_\lambda, T) > 0$.*

The argument for this is simple: suppose $h(X_\lambda, T) = 0$ and Sarnak's conjecture holds. Then one should have that for any continuous function f on X_λ and any x, we have that

$$\sum_{n\le N} \lambda(n) f(T^n x) = o(N).$$

In particular, this should hold for the function f_0 and $x = \lambda$. But for f_0 and $x = \lambda$, we have

$$\sum_{n\le N} \lambda(n) f_0(T^n\lambda) = \sum_{n\le N} \lambda(n)\lambda(n) = N,$$

a contradiction.

For closed shift-invariant subsets of X, the topological entropy measures the growth in the number of sign patterns. Let $Y \subset X$ be a closed shift-invariant subset. Denote

$$E_k(Y) = \{(y(n), \dots, y(n+k-1)) : n \in \mathbb{N},\ y \in Y\}.$$

Then E_k is the set of different blocks of sign patterns exhibited in elements $y \in Y$. Denote $S_k(Y) = |E_k(Y)|$. Then the entropy of (Y, T) is given by

$$h(Y, T) = \limsup \frac{\log S_k(Y)}{k}.$$

From Lemma 1.3, we get that Sarnak's conjecture implies that $S_k(X_\lambda)$ should grow exponentially. Namely, there exists $c > 0$ such that

$$S_k(X_\lambda) \gg 2^{ck}.$$

We are still quite far from this conjecture, though recent years have seen substantial progress. The current state of the art is the following theorem [22].

Theorem 1.4 (Matomäki–Radziwiłł–Tao–Teräväinen–Ziegler (2022)). *For any constant $A \ge 1$,*

$$S_k(X_\lambda) \gg_A k^A.$$

1.2. Important instances of Sarnak's conjecture

We examine some instances of Sarnak's conjecture.

1.2.1. The trivial system. In this case, X is a point and T the trivial map. Any continuous map here is just a constant map; thus Sarnak's conjecture in this case is simply the prime number theorem:

$$\sum_{n\leq N} \lambda(n) = o(N).$$

1.2.2. Periodic systems. Here $X = \mathbf{Z}/m\mathbf{Z}$, $Tx = x + 1$. In this case, any continuous map here is periodic, and Sarnak's conjecture is the prime number theorem in arithmetic progressions:

$$\sum_{n\equiv a(m), n\leq N} \lambda(n) = o(N).$$

1.2.3. Almost periodic system. Here $X = \mathbf{R}/\mathbf{Z}$, $Tx = x + \alpha$, $\alpha \in \mathbf{R}$. Sarnak's conjecture is reduced to a statement on correlation of λ with linear exponential functions and is a theorem of Davenport [6].

Theorem 1.5 (Davenport). *For any $\alpha \in \mathbb{R}$, any $m \in \mathbb{Z}$, we have*

$$\sum_{n\leq N} \lambda(n)e^{2\pi inm\alpha} = o(N).$$

1.2.4. Nilsystems. Here $X = G\Gamma$, G a nilpotent Lie group, Γ a lattice, $a \in G$, $Tg\Gamma = ag\Gamma$. Sarnak's conjecture in this case was proved by Green and Tao, and it played an important role in establishing asymptotics for linear equations in primes [12–14].

Theorem 1.6 (Green–Tao). *Let $F : X \to \mathbb{C}$ be a continuous function, $x \in G$, then*

$$\sum_{n\leq N} \lambda(n)F(a^n x\Gamma) = o(N).$$

1.2.5. The horocycle flow. Here $X = \mathrm{SL}_2(\mathbb{R})/\Gamma$, Γ a lattice, $Tx = ax$ where $a \in \mathrm{SL}_2(\mathbb{R})$. This was the first instance of Sarnak's conjecture where the dynamics is complex – the system is mixing but still of zero entropy. The conjecture in this case was verified by Bourgain, Sarnak, and Ziegler [3].[1]

Theorem 1.7 (Bourgain–Sarnak–Ziegler). *Let (X, T) be a horocycle flow, then μ is disjoint from (X, T), that is, given $x \in X$ and $f \in C(X)$ (if X is not compact, then take f continuous on the one-point compactification of X),*

$$\sum_{n\leq N} \mu(n)f(T^n x) = o(N).$$

[1] See [25] for a related result on orbits of the horocycle flow along the primes.

1.3. Joinings

A main observation in [3] is that one can use the multiplicative nature of λ to convert the problem into a dynamical problem regarding *joinings of dynamical systems*.

The key to turning Sarnak's conjecture into a purely dynamical question is the following proposition [3, 5, 17].

Proposition 1.8 (Daboussi–Delange, Kátai, Bourgain–Sarnak–Ziegler). *Let $x \in X$, and let $F : X \to \mathbb{C}$ be a continuous function, $x \in X$. Suppose for (almost) all p, q distinct primes, we have*

$$\sum_{n \le N} F(T^{pn}x)\overline{F(T^{qn}x)} = o(N).$$

Then

$$\sum_{n \le N} \lambda(n) F(a^n x \Gamma) = o(N).$$

Let us give a dynamical interpretation to the above proposition: fix $x \in X$, and consider the following sequence of Borel probability measures on $X \times X$:

$$\mu_N = \frac{1}{N} \sum_{n \le N} \delta_{(T^p \times T^q)^n (x,x)},$$

where δ_a, $a \in X$ is that δ-measure at $a \in X \times X$. The sequence μ_n has a subsequence μ_{n_k} that weakly converges to a Borel probability measure μ on $X \times X$. In particular, we have for any continuous function $F : X \to \mathbb{C}$

$$\frac{1}{N_k} \sum_{n \le N_k} F(T^{pnx})\overline{F(T^{qn}x)} \to \int F \otimes \overline{F} d\mu.$$

The measure μ is invariant under $T^p \times T^q$, and its marginals are $\pi_1\mu, \pi_2\mu$, which we can usually assume are well understood. Thus μ is a *joining* of the systems $(X, \pi_1\mu, T^p)$ and $(X, \pi_2\mu, T^q)$.

Many instances of Sarnak's conjecture have been verified via the classification of joinings of corresponding dynamical systems. In the case of the horocycle flow studies in [3] the key was the seminal work of M. Ratner classifying joinings of the horocycle flow [23].

1.4. Short averages of multiplicative functions and the entropy decrement argument

In 2016, K. Matomäki and M. Radziwiłł [21] made the first substantial progress towards Chowla's conjecture in many years. They proved the following theorem.

Theorem 1.9 (Matomäki and Radziwiłł). *For any function $H : \mathbb{N} \to \mathbb{N}$ satisfying $H \to \infty$ as $N \to \infty$, we have*

$$\sum_{n \leq N, h \leq H} \lambda(n)\lambda(n+h) = o(NH).$$

Shortly after, Tao [26] introduced a new entropy argument and combined with Theorem 1.9 was able to prove a logarithmic version of Chowla's conjecture for the case $k = 2$.

Theorem 1.10 (Tao). *For any integer $m > 0$,*

$$\sum_{n \leq N} \frac{\lambda(n)\lambda(n+m)}{n} = o(\log N).$$

Tao was also able to use the entropy argument to prove the equivalence between the logarithmic version of Chowla's conjecture and the logarithmic version of Sarnak's conjecture [27].[2]

Conjecture 1.11 (Log-Sarnak). *Let (X, T) be a topological dynamical system of zero entropy, and let $f : X \to \mathbb{C}$ be a continuous function. Then*

$$\sum_{n \leq N} \frac{\lambda(n) f(T^n x)}{n} = o(\log N).$$

Conjecture 1.12 (Log-Chowla). *For any k and any integers $0 \leq h_1 < \cdots < h_k$, we have*

$$\sum_{n \leq N} \frac{\mu(n+h_1)\cdots\mu(n+h_k)}{n} = o(\log N).$$

Theorem 1.13 (Tao). *The log-Sarnak and log-Chowla conjectures are equivalent.*

Theorem 1.13 relies on the new entropy argument, as well as results in additive combinatorics on solutions to linear equations in primes obtained in joint work with Green and Tao [12–14].

In a major advancement in 2018, Frantzikinakis and Host [7] were able to make substantial progress on the log-Sarnak conjecture.

Theorem 1.14 (Frantzikinakis–Host). *The log-Sarnak conjecture holds true for uniquely ergodic systems.*

[2] The logarithmic versions (Conjectures 1.11 and 1.12) are weaker than the original versions (Conjectures 1.1 and 1.2).

Theorems 1.13 and 1.14 are built on major advancements in ergodic theory and additive combinatorics, in particular in the development of higher-order Fourier analysis in both regimes.

1.5. Higher-order Fourier analysis

The origin of higher-order Fourier analysis can be traced back to the work of Furstenberg, providing an ergodic-theoretic proof of Szemerédi's theorem. In 1975, Szemerédi proved the following remarkable theorem.

Theorem 1.15. *Suppose $E \subset \mathbb{Z}$ has positive upper density, namely* $\limsup |E \cap [-N, N]|/(2N+1) = \delta > 0$. *Then E contains arbitrarily long arithmetic progressions.*

Szemerédi's proof used tools from graph theory. In 1977, Furstenberg introduced a completely new approach via ergodic theory. In the first part of the proof, Furstenberg introduces a correspondence principle, translating the combinatorial question into a multiple recurrence question in ergodic theory. Namely, he shows that Theorem 1.15 follows from the following multiple recurrence result.

Theorem 1.16 (Furstenberg). *Let $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ be a probability measure-preserving system. Let $A \in \mathcal{B}$ be of positive measure. Then for any integer $k > 0$, there exists an integer $n > 0$ such that*

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

We remark that the case $k = 1$ is the famous Poincaré recurrence theorem, which is one of the fundamental results in ergodic theory.

To prove this theorem, Furstenberg proves a structure theorem for general ergodic probability measure-preserving systems. One of the key new innovations in his proof was the study of probability measure-preserving systems via their maps to simpler systems with additional algebraic structure. This was a revolutionary approach at the time; prior to this, systems were traditionally studied via subsystems. Furstenberg shows the existence of a sequence of factors $\pi_k : \mathbf{X} \to Z_k(\mathbf{X})$ such that the systems $Z_k(\mathbf{X})$ have additional structure, and for any bounded functions $f_1, \ldots, f_k$,

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1 \circ T^n \cdots f_k \circ T^{nk} - \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}(f_1 \circ T^n \mid Z_{k-1}(\mathbf{X})) \cdots \mathbb{E}(f_k \circ T^{kn} \mid Z_{k-1}(\mathbf{X})) \longrightarrow 0,$$

where the convergence is in L^2. In particular, for $f_i = 1_A$, one can approximate the average

$$\frac{1}{N}\sum_{n=0}^{N-1}\mu(A\cap T^{-n}A\cap\cdots\cap T^{-kn}A)$$

by the average

$$\frac{1}{N}\sum_{n=0}^{N-1}\int \mathbb{E}(1_A \mid Z_{k-1}(\mathbf{X}))\mathbb{E}(1_A\circ T^n \mid Z_{k-1}(\mathbf{X}))\cdots\mathbb{E}(1_A\circ T^{kn} \mid Z_{k-1}(\mathbf{X}))d\mu.$$

In the case $k = 2$, corresponding to 3-term arithmetic progressions, the system $Z_1(\mathbf{X})$ is a *Kronecker system*: namely, a system of the form $\mathbf{Z} = (Z, \mathcal{D}, \nu, S)$ where Z is a compact abelian group, $\mathcal{D}$ the Borel sigma algebra, ν the Haar measure and $S : Z \to Z$ given by $Sz = z + t$ for some $t \in Z$. For a Kronecker system, $L^2(\mathbf{Z})$ is generated by the characters of the abelian group Z, and it follows that $L^2(Z_1(\mathbf{X}))$ is generated by the eigenfunctions of the T-action.

For higher k, the factors $Z_k(\mathbf{X})$ in Furstenberg's proof are towers of *isometric extensions* (a Kronecker system is an isometric extension of the trivial system), and this structure was sufficient for Furstenberg to prove that the lim inf of the sequence of averages in $(*)$ is positive, which implies Szemerédi's via a correspondence principle [8, 9].

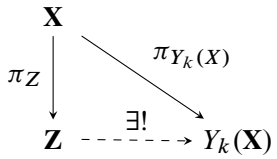
Theorem 1.17 (Furstenberg). *Let $\boldsymbol{X} = (X, \mathcal{B}, \mu, T)$ be a probability measure-preserving system. Let $A \in \mathcal{B}$ be of positive measure. Then*

$$\liminf \frac{1}{N}\sum_{n=0}^{N-1}\mu(A\cap T^{-n}A\cap\cdots\cap T^{-kn}A) > 0.$$

Naturally, one can then ask for a universal sequence of factors $Y_k(\mathbf{X})$, $\pi_{Y_k} : \mathbf{X} \to Y_k(\mathbf{X})$ so that for any factor $\pi_Z : \mathbf{X} \to \mathbf{Z}$ satisfying

$$\frac{1}{N}\sum_{n=0}^{N-1} f_1\circ T^n\cdots f_k\circ T^{nk} - \frac{1}{N}\sum_{n=0}^{N-1}\mathbb{E}(f_1\circ T^n \mid \mathbf{Z})\cdots\mathbb{E}(f_k\circ T^{kn} \mid \mathbf{Z}) \longrightarrow_{L^2} 0, \tag{1.1}$$

such a factor is called a *characteristic factor*; the factor map $\pi_{Y_k(X)}$ would factor through π_Z. Namely, we would have the following commuting diagram:



For $k = 1$, the universal characteristic factor $Y_1(\mathbf{X})$ is the same as $Z_1(\mathbf{X})$, the *Kronecker factor* of $\mathbf{X}$ described above, corresponding to the eigenfunctions of the T-action. A natural candidate for the universal characteristic factor $Y_k(\mathbf{X})$ would then be the factor corresponding to *k-order eigenfunctions*, namely functions defined inductively via the equation $Tf = \psi f$, where ψ is a $k-1$-order eigenfunction. However, in 1995, Furstenberg and Weiss [10] showed that this is not the case already for $k = 2$![3] They observed that there are 2-step nilsystems $\mathbf{X}$ that have no order 2 eigenfunctions but satisfy that $Y_2(\mathbf{X}) = \mathbf{X}$. This remarkable observation of Furstenberg and Weiss is one of the cornerstones of higher-order Fourier analysis.

The universal characteristic factors associated with the question of arithmetic progressions were classified and shown to have an algebraic structure associated to dynamics on nilmanifolds [15, 30].

Theorem 1.18 (Host–Kra, Ziegler). *The universal k-characteristic factor $Y_k(\boldsymbol{X})$ is a k-step pro-nilsystem.*

Another cornerstone is the work of Gowers [11] on the combinatorial front, providing a new quantitative proof of Szemerédi's theorem, generalizing the Fourier analytic approach of Roth in his study of 3-term progressions. In his proof, Gowers introduces a sequence of norms, the *Gowers uniformity norms* defined as follows: let G be a finite abelian group, $f : G \to \mathbb{C}$, then

$$\|f\|_{U_k} = (\mathbb{E}_{x,h_1,\dots,h_k} \Delta_{h_1} \cdots \Delta_{h_k} f)^{1/2^k},$$

where $\Delta_h f(x) := f(x+h)\overline{f(x)}$, and for a finite set S, $\mathbb{E}_{x\in S}$ denotes $1/|S| \sum_{x\in S}$. The Gowers uniformity norms are indeed norms for $k > 1$ and are monotone in k: $\|f\|_{U_1} \le \|f\|_{U_2} \le \cdots$. Most importantly, they dominate the count of arithmetic progressions: if $f_i : G \to \mathbb{C}$ are 1-bounded functions, namely $|f_i| \le 1$, then

$$\mathbb{E}_{x,y\in G} f_0(x) f_1(x+y) \cdots f_k(x+ky) \le \min \|f_i\|_{U_k}.$$

In particular, if $A \subset G$, and $|A| = \delta|G|$, then the average number of $k+1$-term arithmetic progressions in A satisfies

$$\mathbb{E}_{x,y\in G} 1_A(x) 1_A(x+y) \cdots 1_A(x+ky) = \delta^{k+1} + O_k(\|1_A - \delta\|_{U_k}).$$

This naturally leads to the question: when is $\|1_A - \delta\|_{U_k}$ small? Or alternatively, what can one say about a 1-bounded function $f : G \to \mathbb{C}$ satisfying $\|f\|_{U_k} \ge \eta$?

[3] This was later shown to be true for the action of $\oplus\mathbb{F}_p$ [1].

In his landmark paper, Gowers proved a "local" structure theorem for the U_k-norms for the cyclic group $G = \mathbb{Z}/N\mathbb{Z}$ and used it to obtain the first quantitative bounds for Szemerédi's theorem.

A global structure theorem for the U_k-norms was proved by Green, Tao and Ziegler [14].

Theorem 1.19 (Green–Tao–Ziegler). *Let $f : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$ be with $|f| \leq 1$. Then if $\|f\|_{U_k} \geq \delta$, then $\mathbb{E} f(x)p(x) \geq c(\delta)$ for some $< k$-step nilsequence p of complexity which is bounded in terms of δ.*

The theorems mentioned above have had a great impact on major questions in ergodic theory, number theory and additive combinatorics in the last decade. In particular, Theorem 1.19 played a pivotal role in the proof of the Hardy–Littlewood conjectures for systems of linear equations of bounded complexity in the prime numbers. A more detailed exposition of these results can be found in [31].

We finally remark that there has been remarkable progress in recent years in extending the results mentioned above to other group actions and in obtaining alternative proofs to the inverse theorems with quantitative bounds, leading in particular to improved quantitative bounds on Szemerédi's theorem [18–20].

1.6. Local Gowers uniformity of the Liouville function

We return to the log-Chowla conjecture, Conjecture 1.12. The inverse theorem for the Gowers norms, Theorem 1.19, together with the entropy decrement argument of Tao reduces the log-Chowla conjecture to the following.

Conjecture 1.20 (Local Gowers uniformity of the Liouville function). *For any $\varepsilon > 0$ and $H \leq (\log N)^{\varepsilon}$, we have*

$$\sum_{n \leq N} \sup_{g} \left| \sum_{h \leq H} \lambda(n+h)g(n+h) \right| = o(NH),$$

where the supremum is taken over bounded complexity k-step nilsequences.

Recent work of Matomäki, Radziwiłł, Tao, Teräväinen and Ziegler [22] makes substantial progress towards Conjecture 1.20. Indeed, we obtain the conjecture in the range $H = N^{\varepsilon}$, and Theorem 1.4 follows as a consequence.

For $k = 1$, Conjecture 1.20 takes the following form.

Conjecture 1.21 (Local Fourier uniformity of the Liouville function). *For any $\varepsilon > 0$ and $H \geq (\log N)^{\varepsilon}$, we have*

$$\sum_{n \leq N} \sup_{\alpha} \left| \sum_{h \leq H} \lambda(n+h) \exp(2\pi i \alpha h) \right| = o(NH).$$

For this special case, it was demonstrated in [22] that one can take $H = \exp((\log N)^{3/4+\varepsilon})$. In [28], Walsh improves this to $H = \exp((\log N)^{1/2+\varepsilon})$, and in [29], Walsh shows that conditional on GRH in the range $(\log N)^{\psi(N)}$ where $\psi(N) \to \infty$ arbitrarily slow in N.

Sarnak's and Chowla's conjectures appear to be very hard and still far away. However, their logarithmic versions may be within reach, at least conditioned on the Riemann hypothesis, and, as mentioned, would in particular imply the positive entropy conjecture.

We remark that the above questions can be asked for a larger collection of arithmetic functions. Of special interest are other *multiplicative functions*. Many of the results obtained for the Liouville function hold true for general multiplicative functions that behave well on arithmetic progressions.

Funding. This work was partially supported by ISF grant 2112/20.

References

[1] V. Bergelson, T. Tao, and T. Ziegler, An inverse theorem for the uniformity seminorms associated with the action of $\mathbb{F}_p^\infty$. *Geom. Funct. Anal.* **19** (2010), no. 6, 1539–1596 Zbl 1189.37007 MR 2594614

[2] J. Bourgain, Pointwise ergodic theorems for arithmetic sets. *Publ. Math. Inst. Hautes Études Sci.* **69** (1989), no. 1, 5–41 Zbl 0705.28008 MR 1019960

[3] J. Bourgain, P. Sarnak, and T. Ziegler, Disjointness of Moebius from horocycle flows. In H. Farkas, R. Gunning, M. Knopp, and B. Taylor (eds.), *From Fourier analysis and number theory to Radon transforms and geometry*, pp. 67–83, Dev. Math. 28, Springer, New York, 2013 Zbl 1336.37030 MR 2986954

[4] S. Chowla, *The Riemann hypothesis and Hilbert's tenth problem*. Math. Appl. 4, Gordon and Breach Science, New York, 1965, 119 pp. Zbl 0133.30003 MR 0177943

[5] H. Daboussi and H. Delange, On multiplicative arithmetical functions whose modulus does not exceed one. *J. Lond. Math. Soc. (2)* **26** (1982), no. 2, 245–264 Zbl 0499.10052 MR 0675168

[6] H. Davenport, On some infinite series involving arithmetical functions (II). *Q. J. Math.* **8** (1937), no. 1, 313–320 Zbl 0017.39101

[7] N. Frantzikinakis and B. Host, The logarithmic Sarnak conjecture for ergodic weights. *Ann. of Math. (2)* **187** (2018), no. 3, 869–931 Zbl 1400.11129 MR 3779960

[8] H. Furstenberg, Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Anal. Math.* **31** (1977), no. 1, 204–256 Zbl 0347.28016 MR 0498471

[9] H. Furstenberg, Y. Katznelson, and D. Ornstein, The ergodic theoretical proof of Szemerédi's theorem. *Bull. Amer. Math. Soc. (N.S.)* **7** (1982), no. 3, 527–552 Zbl 0523.28017 MR 0670131

[10] H. Furstenberg and B. Weiss, A mean ergodic theorem for $\frac{1}{N}\sum_{n=1}^{N} f(T^n x)g(T^{n^2}x)$. In V. Bergelson, P. March, and J. Rosenblatt (eds.), *Convergence in ergodic theory and probability (Columbus, OH, 1993)*, pp. 193–228, Ohio State Univ. Math. Res. Inst. Publ. 5, De Gruyter, Berlin, 1996 Zbl 0869.28010 MR 1412607

[11] W. T. Gowers, A new proof of Szemerédi's theorem. *Geom. Funct. Anal.* **11** (2001), no. 3, 465–588 Zbl 1028.11005 MR 1844079

[12] B. Green and T. Tao, Linear equations in primes. *Ann. of Math. (2)* **171** (2010), no. 3, 1753–1850 Zbl 1242.11071 MR 2680398

[13] B. Green and T. Tao, The Möbius function is strongly orthogonal to nilsequences. *Ann. of Math. (2)* **175** (2012), no. 2, 541–566 Zbl 1347.37019 MR 2877066

[14] B. Green, T. Tao, and T. Ziegler, An inverse theorem for the Gowers $U^{s+1}[N]$-norm. *Ann. of Math. (2)* **176** (2012), no. 2, 1231–1372 Zbl 1282.11007 MR 2950773

[15] B. Host and B. Kra, Nonconventional ergodic averages and nilmanifolds. *Ann. of Math. (2)* **161** (2005), no. 1, 397–488 Zbl 1077.37002 MR 2150389

[16] H. Iwaniec and E. Kowalski, *Analytic number theory*. Amer. Math. Soc. Colloq. Publ. 53, American Mathematical Society, Providence, RI, 2004, 615 pp. Zbl 1059.11001 MR 2061214

[17] I. Kátai, A remark on a theorem of H. Daboussi. *Acta Math. Hungar.* **47** (1986), no. 1–2, 223–225 Zbl 0607.10034 MR 0836415

[18] J. Leng, A. Sah, and M. Sawhney, Improved bounds for Szemerédi's theorem. 2024, arXiv:2402.17995v2

[19] J. Leng, A. Sah, and M. Sawhney, Quasipolynomial bounds on the inverse theorem for the Gowers $U^{s+1}[N]$-norm. 2024, arXiv:2402.17994v3

[20] F. Manners, Quantitative bounds in the inverse theorem for the Gowers U^{s+1}-norms over cyclic groups. [v1] 2018, [v2] 2024, arXiv:1811.00718v2

[21] 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

[22] K. Matomäki, M. Radziwiłł, T. Tao, J. Teräväinen, and T. Ziegler, Higher uniformity of bounded multiplicative functions in short intervals on average. *Ann. of Math. (2)* **197** (2023), no. 2, 739–857 Zbl 1521.11059 MR 4543441

[23] M. Ratner, Rigidity of horocycle flows. *Ann. of Math. (2)* **115** (1982), no. 3, 597–614 Zbl 0506.58030 MR 0657240

[24] P. Sarnak, Three lectures on the Möbius function randomness and dynamics, 2010

[25] P. Sarnak and A. Ubis, The horocycle flow at prime times. *J. Math. Pures Appl. (9)* **103** (2015), no. 2, 575–618 Zbl 1307.11048 MR 3298371

[26] T. Tao, The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi* **4** (2016), article no. e8, 36 pp. Zbl 1383.11116 MR 3569059

[27] T. Tao, Equivalence of the logarithmically averaged Chowla and Sarnak conjectures. In C. Elsholtz and P. Grabner (eds.), *Number theory — Diophantine problems, uniform distribution and applications*, pp. 391–421, Springer, Cham, 2017 Zbl 1426.11111 MR 3676413

[28] M. N. Walsh, Phase relations and pyramids. *J. Eur. Math. Soc. (JEMS)* (2025), DOI 10.4171/JEMS/1613

[29] M. N. Walsh, Stability under scaling in the local phases of multiplicative functions. *Invent. Math.* **241** (2025), no. 1, 325–362 Zbl 08060591 MR 4921576

[30] T. Ziegler, Universal characteristic factors and Furstenberg averages. *J. Amer. Math. Soc.* **20** (2007), no. 1, 53–97 Zbl 1198.37014 MR 2257397

[31] T. Ziegler, Linear equations in primes and dynamics of nilmanifolds. In S. Y. Jang, Y. R. Kim, D.-W. Lee, and I. Yie (eds.), *Proceedings of the International Congress of Mathematicians, Vol. II (Seoul 2014)*, pp. 569–589, Kyung Moon Sa, Seoul, 2014 Zbl 1373.11009 MR 3728628

Tamar Ziegler
Einstein Institute of Mathematics, Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, 9190401 Jerusalem, Israel; tamar.ziegler@mail.huji.ac.il

Invited lectures

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/8

A discontinuous Galerkin method for the three-dimensional heterodimer model with application to prion-like proteins' dynamics

Paola F. Antonietti, Mattia Corti, and Giacomo Lorenzon

Abstract. Neurocognitive disorders, such as Alzheimer's and Parkinson's, have a wide social impact. These proteinopathies involve misfolded proteins accumulating into neurotoxic aggregates. Mathematical and computational models describing the prion-like dynamics offer an analytical basis to study the diseases' evolution and a computational framework for exploring potential therapies. This work focuses on the heterodimer model in a three-dimensional setting, a reactive-diffusive system of non-linear partial differential equations describing the evolution of both healthy and misfolded proteins. We investigate traveling wave solutions and diffusion-driven instabilities as a mechanism of neurotoxic pattern formation. For the considered mathematical model, we propose a space discretization, relying on the discontinuous Galerkin method on polytopal/polyhedral grids, allowing high-order accuracy and flexible handling of the complicated brain's geometry. Further, we present a priori error estimates for the semi-discrete formulation, and we perform convergence tests to verify the theoretical results. Finally, we conduct simulations using realistic data on a three-dimensional brain mesh reconstructed from medical images.

1. Introduction

Neurodegenerative disorders involve inevitable cognitive decline, often diagnosed late after substantial brain deterioration [46]. Many of these conditions, classified as proteinopathies, share a common neurodegenerative mechanism [24]. Despite varied symptoms, they exhibit similar molecular behaviors: misfolded proteins act as templates, causing further protein misfolding, neurotoxic aggregation, and cell loss [37, 44]. Diseases such as Alzheimer's, Parkinson's, and amyotrophic lateral sclerosis follow this pattern, with specific proteins showing stereotypical progression [23]. In Parkinson's disease, α-synuclein lesions first appear in the dorsal motor

Mathematics Subject Classification 2020: 65N22 (primary); 65N30 (secondary).
Keywords: discontinuous Galerkin method, heterodimer model, Parkinson's disease, neurodegeneration.

and anterior olfactory nuclei before spreading [8]. This resembles prion diseases, suggesting a shared molecular mechanism [22], with axonal transport potentially aiding this spread [22].

In Parkinson's disease, α-synuclein monomers form insoluble Lewy bodies [43]. Targeting misfolding and aggregation could offer therapeutic avenues [18, 23]. However, detecting α-synuclein remains a considerable challenge since current methods are postmortem and PET imaging ligands are lacking [25, 42]. Advances in detecting α-synucleinopathies, including Parkinson's disease, are limited [42]. Consequently, mathematical and numerical modeling is essential for understanding prion-like disease mechanisms [11].

Physics-based models have recently begun shedding light on neurodegenerative processes [11, 47], correlating molecular scales with symptoms [48]. The heterogeneity of spatial and temporal scales involved poses a challenge for clinical observations, but mathematical and computational models offer analytical tools for tailoring personalized therapies. Aggregate models often lack accuracy in depicting prion concentration evolution, as spatial domain description is critical for estimating disease progression [31]. The prion paradigm is essential for designing physics-based models that consistently describe neurodegenerative disorder traits [47]. Several models have been proposed to understand prion kinetics, including monomeric [12] and polymeric [21] seeding models. This work focuses on a simpler kinetic model describing the conversion from native to misfolded proteins at a fixed rate [47]. Despite the difficulty in quantifying diffusion and transport mechanisms *in vivo* [40], the heterodimer model offers a quantitative framework for prion-like propagation [2, 47]. Frequently used models in the literature include the Smoluchowski and Fisher–Kolmogorov models [16]. The Smoluchowski model, while well suited to describe polymeric fragmentation and aggregation, is complex, prompting the use of simpler models with high spatial detail for intricate domains like the brain. The Fisher–Kolmogorov model, a well-known non-linear reaction-diffusion equation, effectively describes population dynamics as an extension of the logistic equation with a diffusive term. Numerous analytical [32, 33] and numerical [13, 14, 47] studies are available. A historical overview of mathematical models for prion-like dynamics is provided in [11], with a comprehensive survey in [16].

This work focuses on the heterodimer model [37], aiming to model molecular interactions and describe macroscopic prion-like propagation across complex spatial domains such as the human brain. In particular, we extend and validate the work of [2] in a three-dimensional, patient-specific setting, demonstrating that our model not only replicates but also refines these findings. Indeed, our approach considers both native and misfolded protein fields, simplifying the Smoluchowski approach while enhancing the infection mechanism description compared to the Fisher–Kolmogorov model. The heterodimer model explicitly includes physical coefficients representing

molecular kinetic processes (production, destruction, and conversion) [16], crucial for developing effective therapeutic interventions. The bistability of the dynamical system justifies seeking traveling wave-like solutions, for which a closed expression in linear approximation can be obtained, and a velocity estimate derived, generalizing considerations in [32]. Moreover, we validate our results thanks to Braak staging theory, and we delve further into the sensitivity analysis results found in the literature.

For the numerical approximation of the heterodimer model, finite element methods are commonly used [16, 31], alongside reduced-order network diffusion models [31, 45]. Recently, discontinuous Galerkin (DG) formulations of the heterodimer model have been proposed on polytopal/polyhedral grids [2]. These partial differential equations typically admit wavefront propagating solutions [33], necessitating high-order approximating schemes capable of handling complex geometries with locally varying discretization parameters. This justifies the use of a polytopal discontinuous (PolyDG) approach [4], employing arbitrarily shaped elements and enabling mesh construction from clinical images through agglomeration techniques [5]. For time integration, the θ-method scheme is adopted, with the non-linear term treated semi-implicitly. Stability results and a priori convergence error estimates are extensively studied by [2].

This work is organized as follows. In Section 2, we present the heterodimer model and discuss some theoretical results. In Section 3, we derive the polytopal discontinuous Galerkin formulation and present theoretical results concerning stability and a priori error estimates. Also, we discuss the fully discrete formulation. In Section 4, we present numerical results verifying the theoretical convergence rates, and in Section 5 we deal with a realistic simulation of α-synuclein spreading in a three-dimensional domain obtained from magnetic resonance imaging and tractography data. In Section 6, we conclude and discuss future developments.

2. Mathematical modeling of protein spreading

The mathematical modeling of the prion-like pathogenesis kinetics establishes a coherent connection between different spatial and temporal scales, including microscopic biochemical reactions and the accumulation of misfolded proteins across extensive brain regions, causing neuronal demise followed by the disease's symptomatic expressions. While the prion conversion mechanism remains poorly understood, a straightforward approach to unravel the interactions between native and prion proteins is the heterodimer model, initially proposed by Prusiner [35, 36]. According to this model, a prion molecule interacts with a healthy prion protein, forming a heterodimer. Subsequently, the prion protein undergoes conversion into its misfolded variant, and the resultant polymer fragments into two singular misfolded prions

that perpetuate as infectious seeds, propagating the infection [28]. The heterodimer model offers one of the simplest descriptions of the proteins' kinetic in their native state c and misfolded one q. As suggested in [16], the conversion rate between them is globally described by a coefficient k_{12}, representing the single-step reaction

$$c + q \xrightarrow{k_{12}} q + q,$$

in which the rate k_{12} is proportional to both the concentrations of native and misfolded proteins [47].

2.1. Strong formulation

The dynamics of the aggregate concentration of healthy and misfolded proteins within a bounded, open domain $\Omega \subset \mathbb{R}^N$, $N = 2, 3$, with Lipschitz boundary, over a time span $(0, T]$, $T \in \mathbb{R}^+$ is described by a system of non-linear partial differential equations (2.1) denoted by heterodimer model [16]. It reads

$$\begin{cases} \dfrac{\partial c}{\partial t} = \nabla \cdot (\mathbf{D}\nabla c) - k_1 c - k_{12} c q + k_0, & \text{in } \Omega \times (0, T], \\ \dfrac{\partial q}{\partial t} = \nabla \cdot (\mathbf{D}\nabla q) - \widetilde{k}_1 q + k_{12} c q, & \text{in } \Omega \times (0, T]. \end{cases} \tag{2.1}$$

The aggregate concentration of native proteins and of those that are misfolded are represented here by the functions $c, q : \Omega \times (0, T] \to \mathbb{R}^+$, respectively. A comprehensive mathematical model is obtained by imposing the set of natural border and initial conditions. Coefficients $k_0, k_1, \widetilde{k}_1, k_{12}$ and data are all non-negative and sufficiently regular. In particular, the diffusion tensor $\mathbf{D}$ is chosen to be uniformly elliptic.

In specific brain regions, the modeled protein is highly expressed at a rate of $k_0 \geq 0$. Clearance can occur through various biological pathways. This elimination mechanism is described by the clearance rate coefficient $k_1 \geq 0$. The overall conversion from healthy to misfolded proteins happens at a rate of $k_{12} \geq 0$, as previously described. A similar clearance process also interests the misfolded proteins. Notice that, being stereoisomers of the healthy ones, their three-dimensional conformation is different, and so is their chemical functionality. In particular, their solubility is strongly affected. The clearance coefficient $\widetilde{k}_1 \geq 0$ is thus in general different from k_1. The diffusion term in (2.1) describes the nature of the spreading for both healthy and misfolded proteins from the macroscopic perspective. We take into account two contributions, unlike in [31]: extracellular and axonal diffusion. The first one is concentration gradient-driven and is assumed to be isotropic; typically, it is associated with neuronal proximity [17]. The second mechanism of spread is anisotropic, occurring along anatomically defined paths called axons [41].

Let $\boldsymbol{a} : \Omega \subset \mathbb{R}^N \to \mathbb{R}^N : |\boldsymbol{a}(\boldsymbol{x})| = 1 \ \forall \boldsymbol{x} \in \Omega$ be the local axonal direction at $\boldsymbol{x}$ in Ω, with $|\cdot|$ being the Euclidean norm in $\mathbb{R}^N$. The diffusion tensor is then defined as

$$\mathbf{D}(\boldsymbol{x}) = d_{\text{ext}}\mathbf{I} + d_{\text{axn}}\boldsymbol{a}(\boldsymbol{x}) \otimes \boldsymbol{a}(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega,$$

with $d_{\text{ext}} > 0$, and with $\mathbf{I}$ being the identity. Notice that being the axonal diffusion faster than the extracellular one, typically $d_{\text{axn}} > d_{\text{ext}} \geq 0$ is assumed [47].

The bistability of the system suggests the following assumption.

Assumption 2.1 (Data assumptions). To have physically relevant solutions with equilibria lying in the first quadrant of the phase space (c, q), the inequality $k_0 k_{12} > k_1 \widetilde{k}_1 > 0$ must be satisfied.

Remark 2.2. Under Assumption 2.1, no bifurcations can be observed in system (2.1). Namely, no qualitative changes in the system's dynamics emerge as its parameters are varied.

We observe that, under Assumption 2.1, the equilibrium point $(\widetilde{k}_1/k_{12}, k_0/\widetilde{k}_1 - k_1/k_{12})$ is always stable. Under the same condition, $(k_0/k_1, 0)$ is instead always unstable and, in particular, it is a saddle point. This is coherent with the disease's phenomenological point of view, since from the time of the onset, the disease always evolves inevitably toward a pathological condition. This motivates the search for a solution in the form of a traveling wave. Also, we obtain an explicit expression for the minimum velocity of the traveling wave solution that reads

$$v_{\min} = 2\sqrt{d_{\text{ext}}\widetilde{k}_1\left(\frac{k_0 k_{12}}{k_1 \widetilde{k}_1} - 1\right)}. \tag{2.2}$$

According to [33], the traveling wavefront's velocity can be derived as a function of the initial solution. In particular, if the wave profile has compact support, the solution will travel with minimum speed $v_{\min}$ defined as in (2.2). Further, it can be shown that the microstructure of the domain, and so the diffusion tensor, together with the front shape, influences the propagation velocity.

3. Polytopal discontinuous Galerkin semi-discrete formulation

In this section, we present the polytopal discontinuous Galerkin semi-discrete formulation of model (2.1), along with the stability and a priori error analysis [2]. We then derive the fully discrete formulation based on the discontinuous Galerkin method.

3.1. PolyDG semi-discrete formulation

Discontinuous Galerkin methods on polytopal grids are a family of finite element methods for the approximation of partial differential equations that can support arbitrarily shaped mesh elements [3, 19, 39] and extend the classical paradigm of discontinuous Galerkin methods [6,10]. The idea is to look for the numerical solution in a discrete space of polynomials, possibly discontinuous across mesh elements. We introduce some preliminary results and definitions to derive the PolyDG formulation. Let $\mathcal{T}_h$ be a partition of Ω. Its elements $K \in \mathcal{T}_h$ are a finite number of non-empty and disjoint polyhedra. In what follows, for each of them, $|K|$ is the element's measure, namely the N-dimensional Lebesgue measure of K, $h_K = \mathrm{diam}(K)$ is the element's diameter, and $h = \max\{h_K : K \in \mathcal{T}_h\}$ is the global mesh size parameter.

We will refer to mesh interfaces as the intersection of the $(N-1)$-dimensional facets of two neighboring elements $K, K' \in \mathcal{T}_h$. In dimension $N = 3$, each interface consists of a general polygon. By assumption, we suppose that each polygon may be decomposed into a set of co-planar triangles. We will refer to $\mathcal{F}_h$ as the collection of all the triangles of all the sub-triangulations of each interface. We will use the term face to refer to one of the elements in $\mathcal{F}_h$ [4]. Internal faces are denoted as $\mathcal{F}_h^I = \{F \in \mathcal{F}_h : F \subset \Omega\}$ and the boundary ones as $\mathcal{F}_h^B = \mathcal{F}_h \setminus \mathcal{F}_h^I$. The latter can be further subdivided into Dirichlet boundary faces $\mathcal{F}_h^D = \{F \in \mathcal{F}_h : F \subset \partial\Omega_D\}$ and Neumann boundary faces $\mathcal{F}_h^N = \{F \in \mathcal{F}_h : F \subset \partial\Omega_N\}$. Here it is assumed that each face F belongs entirely either to $\mathcal{F}_h^D$ or to $\mathcal{F}_h^N$. We also require $\mathcal{T}_h$ to satisfy some regularity conditions, following the work of [2, 9]. This allows us for exploiting the trace, the inverse, and the discrete Gagliardo–Nirenberg inequalities, and eventually deriving the upcoming estimates.

Assumption 3.1. We assume that the family of meshes $\{\mathcal{T}_h\}_h$ is uniformly polytopic regular.

The following assumption will allow us to state the inverse trace estimate [9]. For more details, we refer to [4]. Also, we require the next assumption that will be instead useful [9] for the approximation results presented in the next sections.

Definition 3.2. A covering $\mathcal{T}_\# = \{T_K\}$ related to the polytopic mesh $\mathcal{T}_h$ is a set of shape-regular N-dimensional simplices T_K such that for each $K \in \mathcal{T}_h$, there exists a $T_K \in \mathcal{T}_\#$ such that $K \subsetneq T_K$.

From now on, the binary operator $\lesssim$ denotes the following relation: $\alpha \lesssim \beta$ if and only if there exists $C = C(m, p) > 0 : \alpha \leq C\beta$, where p is the degree of the polynomial approximation and m are the model parameters. Notice that C is independent of the mesh size h.

Assumption 3.3. There exists a covering $\mathcal{T}_\#$ of $\mathcal{T}_h$ such that

$$\max_{K\in\mathcal{T}_h} \text{card}\{K' \in \mathcal{T}_h : K' \cap T_K \neq \varnothing, T_K \in \mathcal{T}_\# \text{ s.t. } K \subset T_K\} \lesssim 1,$$

and $h_{T_K} \lesssim h_K$ for each pair $K \in \mathcal{T}_h$ and $T_K \in \mathcal{T}_\#$ with $K \subset T_K$.

For the subsequent analysis, we require some more definitions and results. We define the broken Sobolev space as follows:

$$W = H^1(\Omega, \mathcal{T}_h) = \{w \in L^2(\Omega) : w|_K \in H^1(\Omega)\ \forall K \in \mathcal{T}_h\},$$

and the DG-broken version

$$W_h^{\text{DG}} = \{w \in L^2(\Omega) : w|_K \in \mathbb{P}_p(K)\ \forall K \in \mathcal{T}_h\},$$

where $\mathbb{P}_p(K)$ denotes the space of polynomials of total degree $p \geq 1$ on K. We will also make use of two standard trace-related operators: the jump and average operators, applicable to both scalar- and vector-valued functions. These will be referred to as $[\![\cdot]\!]$ and $\{\!\{\cdot\}\!\}$, respectively. For further details, refer to [7].

3.1.1. PolyDG semi-discrete approximation of the heterodimer model. Here we present the semi-discrete formulation of model (2.1). Let $c_h^0, q_h^0 \in W_h^{\text{DG}}$ be the $L^2(\Omega)$-projection of the initial conditions c^0 and q^0 onto W_h^{DG}, respectively. The semi-discrete problem reads as follows.

Given the initial conditions (c_h^0, q_h^0), for each $t \in (0, T]$ find $(c_h(t), q_h(t)) \in W_h^{\text{DG}} \times W_h^{\text{DG}}$ so that

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial c_h}{\partial t} w_h + \mathcal{A}(c_h, w_h) + r_L(c_h, w_h) + r_N(c_h, q_h, w_h) = F_c(w_h) & \forall w_h \in W_h^{\text{DG}}, \\ \displaystyle\int_\Omega \frac{\partial q_h}{\partial t} v_h + \mathcal{A}(q_h, v_h) + \tilde{r}_L(q_h, v_h) - r_N(q_h, c_h, v_h) = F_q(v_h) & \forall v_h \in W_h^{\text{DG}}. \end{cases} \tag{3.1}$$

The bilinear and trilinear forms are defined, for a fixed triangulation $\mathcal{T}_h$, for any $u_h, \varphi_h, \phi_h \in W_h^{\text{DG}}$ in the following way:

$$\begin{aligned} \mathcal{A}(u_h, \varphi_h) = &\int_\Omega (\mathbf{D}\nabla_h u_h) \cdot \nabla_h \varphi_h - \sum_{F \in \mathcal{F}_h^I \cup \mathcal{F}_h^D} \int_F [\![u_h]\!] \cdot \{\!\{\mathbf{D}\nabla_h \varphi_h\}\!\} \\ &+ \sum_{F \in \mathcal{F}_h^I \cup \mathcal{F}_h^D} \int_F \gamma_F [\![u_h]\!] \cdot [\![\varphi_h]\!] - [\![\varphi_h]\!] \cdot \{\!\{\mathbf{D}\nabla_h u_h\}\!\}, \end{aligned}$$

$$r_L(u_h,\varphi_h)=\int_\Omega k_1 u_h\varphi_h,\quad \widetilde{r}_L(u_h,\varphi_h)=\int_\Omega \widetilde{k}_1 u_h\varphi_h,$$
$$r_N(u_h,\varphi_h,\phi_h)=\int_\Omega k_{12} u_h\varphi_h\phi_h,$$
$$F_c(\varphi_h)=\int_\Omega f_c\varphi_h,\quad F_q(\varphi_h)=\int_\Omega f_q\varphi_h,$$

where ∇_h is the element-wise gradient, and γ_F is the discontinuity penalization function defined as

$$\gamma_F=\gamma_0\begin{cases}\max\{\{d^K\}_{\mathrm{H}},\{k^K\}_{\mathrm{H}}\}\dfrac{p^2}{\{h\}_{\mathrm{H}}}, & \text{on } F\in\mathcal{F}_h^I,\\ \max\{d^K,k^K\}\dfrac{p^2}{h}, & \text{on } F\in\mathcal{F}_h^D.\end{cases}$$

The coefficient γ_0 is a constant parameter that should be chosen sufficiently large to ensure the stability of the discrete formulation, $d^K=\|\sqrt{\mathbf{D}|_K}\|^2$, and $k^K=\|(1+k_{12}|_K)(k_1|_K+\widetilde{k}_1|_K)\|$. Finally, $\{\cdot\}_{\mathrm{H}}$ is the harmonic average operator defined as $\{v\}_{\mathrm{H}}=(2v^+v^-)/(v^++v^-)$.

3.2. Semi-discrete PolyDG formulation's analysis

We recall some useful definitions to prove the problem's well-posedness, the error estimates, and the scheme stability. For sufficiently regular functions, we can define

$$\|w\|_{\mathrm{DG}}=\|\sqrt{\mathbf{D}}\nabla_h w\|+\|\gamma_F^{1/2}[\![w]\!]\|_{\mathcal{F}_h^I\cup\mathcal{F}_h^D},\tag{3.2}$$
$$|||w|||_{\mathrm{DG}}=\|w\|_{\mathrm{DG}}+\|\gamma_F^{-1/2}\{\!\{\mathbf{D}\nabla_h w\}\!\}\|_{\mathcal{F}_h^I\cup\mathcal{F}_h^D},\tag{3.3}$$

where $\nabla_h v$ is the element-wise gradient, namely $\nabla_h v|_K=\nabla v|_K$. These two norms are well defined and also equivalent in the space of discontinuous functions W_h^{DG}, provided the regularity Assumptions 3.1 and 3.3 on the mesh [2] hold. Two more norms will be useful in our analysis:

$$\|w(t)\|_\varepsilon{}^2=\|w(t)\|^2+\int_0^t\|w(s)\|_{\mathrm{DG}}{}^2\mathrm{d}s,$$
$$|||w(t)|||_\varepsilon{}^2=\|w(t)\|^2+\int_0^t|||w(s)|||_{\mathrm{DG}}{}^2\mathrm{d}s$$

for $t>0$. Under the same regularity Assumptions 3.1 and 3.3, the bilinear form $\mathcal{A}$ is coercive [9] with respect to DG-norm (3.2) and continuous with respect to (3.3), provided that the penalty parameter is large enough.

The stability estimate for the DG-discrete solution is proven in [2]. Here we omit the proof and just recall that the solution in the energy norm $\|\cdot\|_\varepsilon$ is bounded by data,

provided that γ_0 is large enough, and under the regularity Assumptions 3.1 and 3.3 on the mesh. The detailed analysis of the error estimates can be found in [2]. For our purpose, the following result is sufficient.

Theorem 3.4 (Error bounds [2]). *Let $\mathcal{T}_h$ be the partition of Ω, so that it induces a quasi-uniform grid. If the exact solution $c(t), q(t) \in H^{p+1}(\Omega)\ \forall t \in (0,T]$, with p polynomial degree of approximation, then*

$$|||c(t) - c_h(t)|||_\varepsilon^2 + |||q(t) - q_h(t)|||_\varepsilon^2 \lesssim h^p,$$

provided that the penalty parameter γ_0 is large enough.

In this section, we present the algebraic formulation of model (2.1), and the fully discrete formulation.

3.2.1. Algebraic form. Let $N_h = \dim W_h^{\mathrm{DG}}$ be the dimension of the DG space. Fix a basis $\{\varphi_j\}_{j=1}^{N_h}$. Let $\boldsymbol{C}, \boldsymbol{Q} : \mathcal{T}_h \times (0,T] \to \mathbb{R}^{N_h}$ be the vectors of all the coefficients of the solution expansion in W_h^{DG}. More precisely, the solution expansion with respect to the chosen basis reads as

$$c_h(\boldsymbol{x},t) = \sum_{j=1}^{N_h} C_j(t)\varphi_j(\boldsymbol{x}), \quad q_h(\boldsymbol{x},t) = \sum_{j=1}^{N_h} Q_j(t)\varphi_j(\boldsymbol{x}).$$

The algebraic form of problem (3.1) follows:

$$\begin{cases} \mathbf{M}\dfrac{\partial \boldsymbol{C}(t)}{\partial t} + \mathbf{A}\boldsymbol{C}(t) + \mathbf{R}_L\boldsymbol{C}(t) + \mathbf{R}_N(\boldsymbol{C}(t))\boldsymbol{Q}(t) = \boldsymbol{F}_c(t), \quad t \in (0,T], \\ \mathbf{M}\dfrac{\partial \boldsymbol{Q}(t)}{\partial t} + \mathbf{A}\boldsymbol{Q}(t) + \widetilde{\mathbf{R}}_L\boldsymbol{Q}(t) - \mathbf{R}_N(\boldsymbol{Q}(t))\boldsymbol{C}(t) = \boldsymbol{F}_q(t), \quad t \in (0,T], \\ \boldsymbol{C}(0) = \boldsymbol{C}^0, \quad \boldsymbol{Q}(0) = \boldsymbol{Q}^0, \end{cases} \tag{3.4}$$

where $\boldsymbol{C}^0$ and $\boldsymbol{Q}^0$ are the expansion coefficients in the chosen basis of the projected initial data. In system (3.4), $\mathbf{M}, \mathbf{A}, \mathbf{R}_L, \widetilde{\mathbf{R}}_L, \mathbf{R}_N \in \mathbb{R}^{N_h \times N_h}$, whereas $\boldsymbol{F}_c, \boldsymbol{F}_q \in \mathbb{R}^{N_h}$. The component-wise definitions are given by

$$\mathrm{M}_{ij} = (\varphi_i, \varphi_j), \quad \mathrm{A}_{ij} = \mathcal{A}(\varphi_i, \varphi_j), \quad \mathrm{R}_{L,ij} = r_L(\varphi_i, \varphi_j), \quad \widetilde{\mathrm{R}}_{L,ij} = \widetilde{r}_L(\varphi_i, \varphi_j),$$
$$\mathrm{R}_{\mathrm{N}}(\psi(t))_{ij} = r_N(\psi(t), \varphi_i, \varphi_j), \quad F_{c,i} = F_c(\varphi_i), \quad F_{q,i} = F_q(\varphi_i),$$

with i, j ranging in $\{1, \dots, N_h\}$.

3.2.2. Time discretization. Let $\{t_i\}_{i=0}^{N_T}$ be the partition of the time interval $[0,T]$, so that $0 = t_0 < t_1 < \cdots < t_{N_T} = T$ and $\Delta t = t_{k+1} - t_k$, $\forall k \in \{0, \dots, N_T - 1\}$. Denote with $(\boldsymbol{C}^0, \boldsymbol{Q}^0) = (\boldsymbol{C}(0), \boldsymbol{Q}(0))$ the initial datum and define $(\boldsymbol{C}^k, \boldsymbol{Q}^k) \approx$

$(\boldsymbol{C}(t_k), \boldsymbol{Q}(t_k))$, $\forall k \in \{1, \dots, N_T\}$. Now we discretize the time derivative with a classic forward scheme and introduce the theta-method scheme [38] to integrate in t.

Given the initial datum $(\boldsymbol{C}^0, \boldsymbol{Q}^0)$, find $(\boldsymbol{C}^k, \boldsymbol{Q}^k)$, $\forall k = 1, 2, \dots, N_T$, that solve system (3.5):

$$\begin{cases} \left(\dfrac{\mathbf{M}}{\Delta t} + \theta\mathbf{A} + \theta\mathbf{R}_L\right)\boldsymbol{C}^k + \mathbf{R}^{\theta}_{c,N}\boldsymbol{Q}^k = \boldsymbol{F}^{\theta}_c, \\ \left(\dfrac{\mathbf{M}}{\Delta t} + \theta\mathbf{A} + \theta\widetilde{\mathbf{R}}_L\right)\boldsymbol{Q}^k - \mathbf{R}^{\theta}_{q,N}\boldsymbol{C}^k = \boldsymbol{F}^{\theta}_q, \end{cases} \tag{3.5}$$

where the forcing term is defined, and the non-linear term is linearized according to the chosen method.

- Implicit Euler method $\theta = 1$

$$\mathbf{R}^{\theta=1}_{c,N} = \mathbf{R}_N(\boldsymbol{C}^{k-1}) = \boldsymbol{C}^{k-1}, \quad \mathbf{R}^{\theta=1}_{q,N} = \mathbf{R}_N(\boldsymbol{Q}^{k-1}) = \boldsymbol{Q}^{k-1},$$
$$\boldsymbol{F}^{\theta=1}_c = \boldsymbol{F}_c(t_k) + \frac{\mathbf{M}}{\Delta t}\boldsymbol{C}^{k-1}, \quad \boldsymbol{F}^{\theta=1}_q = \boldsymbol{F}_q(t_k) + \frac{\mathbf{M}}{\Delta t}\boldsymbol{Q}^{k-1}.$$

- Crank–Nicolson method $\theta = 1/2$

$$\mathbf{R}^{\theta=1/2}_{c,N} = \frac{3}{4}\mathbf{R}_N(\boldsymbol{C}^{k-1}) - \frac{1}{4}\mathbf{R}_N(\boldsymbol{C}^{k-2}) = \frac{3\boldsymbol{C}^{k-1} - \boldsymbol{C}^{k-2}}{4},$$
$$\mathbf{R}^{\theta=1/2}_{q,N} = \frac{3}{4}\mathbf{R}_N(\boldsymbol{Q}^{k-1}) - \frac{1}{4}\mathbf{R}_N(\boldsymbol{Q}^{k-2}) = \frac{3\boldsymbol{Q}^{k-1} - \boldsymbol{Q}^{k-2}}{4},$$
$$\boldsymbol{F}^{\theta=1/2}_c = \frac{\boldsymbol{F}_c(t_k) + \boldsymbol{F}_c(t_{k-1})}{2} - \frac{3\boldsymbol{C}^{k-1} - \boldsymbol{C}^{k-2}}{4}\boldsymbol{Q}^{k-1} + \left(\frac{\mathbf{M}}{\Delta t} - \frac{\mathbf{A} + \mathbf{R}_L}{2}\right)\boldsymbol{C}^{k-1},$$
$$\boldsymbol{F}^{\theta=1/2}_q = \frac{\boldsymbol{F}_q(t_k) + \boldsymbol{F}_q(t_{k-1})}{2} - \frac{3\boldsymbol{Q}^{k-1} - \boldsymbol{Q}^{k-2}}{4}\boldsymbol{C}^{k-1} + \left(\frac{\mathbf{M}}{\Delta t} - \frac{\mathbf{A} + \widetilde{\mathbf{R}}_L}{2}\right)\boldsymbol{Q}^{k-1}.$$

Notice that the Crank–Nicolson scheme is defined for $k \geq 2$. Thus, the first iteration should nonetheless be performed with the Euler implicit method.

4. Three-dimensional numerical results: Verification

This section focuses on validating the numerical results against theoretical predictions. Firstly, we have reproduced the results of [2] in two dimensions. Here we just discuss the three-dimensional generalization.

4.1. Convergence rates verification

For the first test case, we set $\Omega = (0,1)^N$, $N = 2,3$. Diffusion is set isotropic and the final time is set sufficiently small. Since the nuanced interplay between the diverse contributions of the reaction term and the diffusion term may introduce heightened numerical challenges, here we discuss a realistic test case. The detailed setting of the simulations follows in Table 1. From the numerical viewpoint, refer to Table 2. For these test cases, we used uniform cubical cell grids. We chose the exact solution as

$$c(\boldsymbol{x},t) = \cos(t)\sum_{i=1}^{N}\cos(2\pi x_i), \quad q(\boldsymbol{x},t) = \left(\prod_{i=1}^{N}\cos(6\pi x_i) + 2\right)e^{-t},$$

in $\Omega \subset \mathbb{R}^N \times (0,T]$, and set the right-hand side and the boundary data accordingly.

4.1.1. Realistic coefficients, idealized domain. Figures 1(a) and 1(b) show the computed errors in $|||\cdot|||_\varepsilon$, $\|\cdot\|_{L^2(\Omega)}$, and $\|\cdot\|_{\mathrm{DG}}$ norms as a function of h for $p = 1,2,3,4$, whereas Figure 1(c) shows the computed errors in $|||\cdot|||_\varepsilon$ as a function of p for $h = 1/4$, $\Delta t = 5\cdot 10^{-6}$. Notably, Figure 1(a) shows that the theoretical convergence rate is consistently attained, in some cases even outperformed. Both Figures 1(a) and 1(b) depict errors that have been normalized to highlight the convergence rate. Also, from Figure 1(c), we observe an exponential convergence for the error. Notice that this case is not covered by the theoretical bounds of Section 3.2. The error for the component $q(\boldsymbol{x},t)$ is always greater than that of $c(\boldsymbol{x},t)$ because its exact representation has a much higher frequency of oscillation, and thus it is more challenging to be accurately reproduced numerically.

Parameter	Value	Dimension	Parameter	Value	Dimension
d_{axn}	0.0	$[\mathrm{mm}^2\ \mathrm{years}^{-1}]$	k_0	0.6	$[\mu\mathrm{g}^{-1}\ \mathrm{years}^{-1}\ \mathrm{mm}^{-3}]$
d_{ext}	8.0	$[\mathrm{mm}^2\ \mathrm{years}^{-1}]$	k_1	0.5	$[\mathrm{years}^{-1}]$
k_{12}	1.0	$[\mu\mathrm{g}^{-1}\ \mathrm{mm}^3\ \mathrm{years}^{-2}]$	$\widetilde{k}_1$	0.3	$[\mathrm{years}^{-1}]$

Table 1. Test case of Section 4.1.1: Physical parameters, taken from [2].

Temporal discretization	Time step	$\Delta t = 5\cdot 10^{-6}$
	Final time	$T = 5\cdot 10^{-5}$
	Theta method	$\theta = 1/2$
DG parameters	Interior penalty parameter	$\eta = 1$
	Penalty parameter	$\gamma_0 = 10$

Table 2. Test case of Section 4.1.1: Numerical parameters.

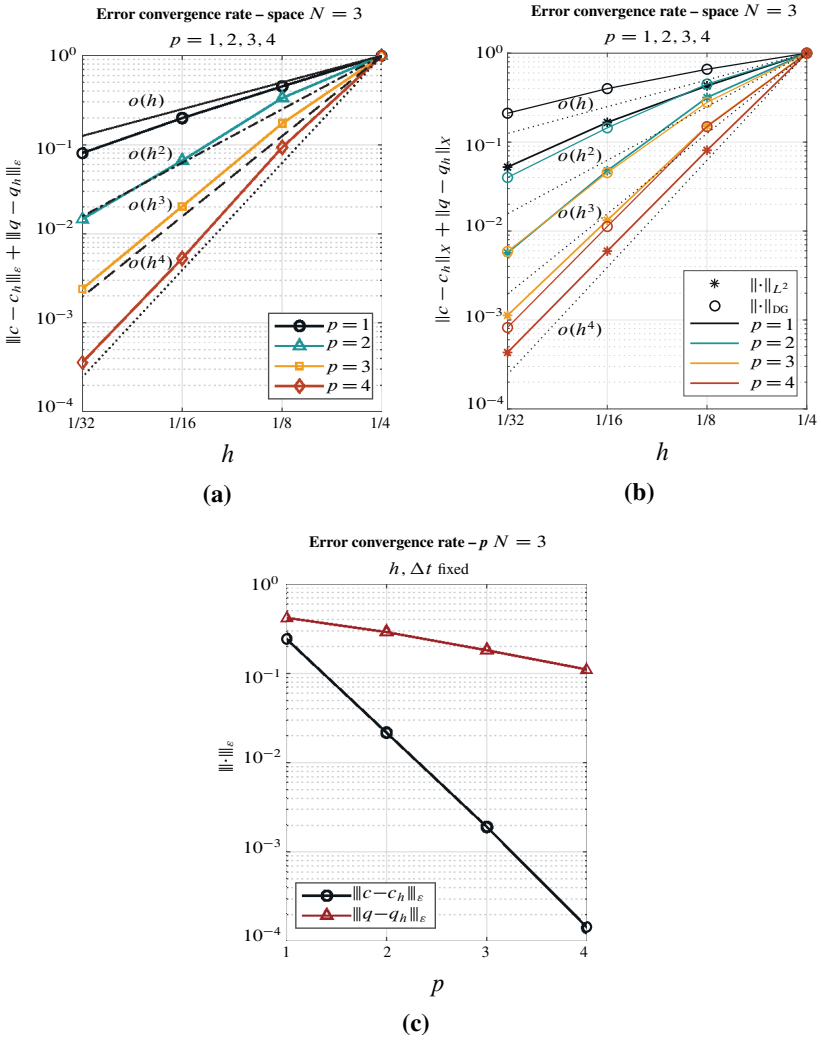


Figure 1. Test case of Section 4.1.1. (a) Computed errors in the $|||\cdot|||_\varepsilon$ norm versus the mesh size h (loglog scale) for different polynomial approximation degrees $p = 1, 2, 3, 4$. (b) Computed errors in the $\|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{\mathrm{DG}}$ norms versus the mesh size h (loglog scale) for different polynomial approximation degrees $p = 1, 2, 3, 4$. (c) Computed errors in the $|||\cdot|||_\varepsilon$ norm versus the polynomial degree of approximation (semilog scale) for $h = 1/4$, $\Delta t = 5 \cdot 10^{-6}$.

4.1.2. Realistic coefficients, realistic domain. We perform a convergence test on a realistic domain, based on data taken from magnetic resonance imaging (MRI) within the OASIS-3 database [26]. The mesh of the brain has been reconstructed as explained in [2, 13] and it is reported in Figure 2. Since the mesh is fixed, we only check that the



Figure 2. 3D domain partition.

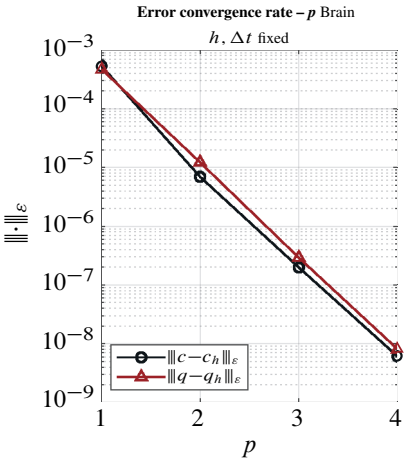


Figure 3. Test case of Section 4.1.2. Computed errors in the $|||\cdot|||_\varepsilon$ norm versus the polynomial degree of approximation (semilog scale) for $\Delta t = 5 \cdot 10^{-6}$.

error gets lower for a higher degree of polynomial approximation. Figure 3 shows the computed errors in $|||\cdot|||_\varepsilon$ as a function of p for $\Delta t = 5 \cdot 10^{-6}$ fixed. An exponential convergence rate is clearly observable. Moreover, one can notice that the value of the error norm $|||\cdot|||_\varepsilon$ is similar for the two solution components $c(\boldsymbol{x}, t)$ and $q(\boldsymbol{x}, t)$ since they both have the same frequency of oscillation.

5. Numerical simulation of α-synuclein spreading

In this section, we analyze the numerical results obtained from the simulation of the diffusion of the α-synuclein protein [43] within the above-mentioned three-dimensional brain mesh. For the details, refer to [14, 29]. Thanks to the high-order DG method used, the triangulation obtained is coarsened, following the work of [5]. This way, the computational load is kept low. Diffusion-weighted imaging is used to obtain the connectome wiring, as explained in [14], following the work of [29].

As for the initial condition of the simulation, the seeding region is inferred from postmortem histological analysis of infected brains [15], and it is located in the dorsal motor nucleus. Since the patient under analysis is healthy, the initial concentration $c_0(\boldsymbol{x})$ is set equal to the healthy component of the unstable equilibrium $\boldsymbol{e}_2$. Thus, $c_0(\boldsymbol{x}) = k_0/k_1 = 1.2$. As for the boundary conditions, it is assumed that no flux is allowed across the brain boundary. This translates into null Neumann boundary conditions for both healthy and misfolded proteins, namely $\partial\Omega$ coincides with Γ_N.

The specific choice of the parameters of model (2.1) is reported in Table 1. Some studies have tried to measure each parameter of Prusiner's model [20, 27, 30, 34], and others have investigated the contribution of each component through sensitivity analyses [40, 47]. All these considerations lead to the choices shown in Table 1. In particular, the ratio $d_{\text{axn}}/d_{\text{ext}} = 10$ is assumed as a modeling choice, following [40]. As shown in [16, Section 2.2], the heterodimer model can be reduced to the Fisher–Kolmogorov model, with a single partial differential equation. This can be done assuming the healthy protein concentration constant, and thus focusing solely on the description of the misfolded protein field evolution. The resulting model comprises a diffusive and a reaction term, multiplied by a constant, say α, which can be expressed in terms of the coefficients of the heterodimer model:

$$\alpha = k_{12}\frac{k_0}{k_1} - \widetilde{k}_1.$$

Through a sensitivity analysis, [40, 47] argue that a high value of α effectively describes the underlying phenomenon, and thus set $\alpha = 0.9/\text{years}$. The reactive coefficients are thus not independent. The work of [2, 14] independently confirms the appropriateness of this choice.

From the numerical perspective, we employ the MUltifrontal Massively Parallel sparse direct Solver [1] (MUMPS). This solver is specifically tailored for addressing large sparse systems to solve algebraic equations on distributed memory parallel computers. The time step is fixed at $\Delta t = 1/120$ years, while we set the final simulation time at $T = 30$ years. We employ a Crank–Nicolson time integration scheme, and we fix the polynomial order of approximation to $p = 2$.

As discussed in Section 2, the system is expected to evolve from the unstable initial equilibrium point to the stable pathological equilibrium $\boldsymbol{e}_1 = (0.3, 1.5)$. Figure 4 shows the temporal evolution of healthy and misfolded proteins across the domain, with a temporal resolution of 5 years, accurately reaching the expected equilibrium. Figure 5 depicts instead the numerical solution obtained over the medial sagittal section of the brain mesh. The results are in alignment with the ones presented in [14], which were obtained starting from the same data but exploiting the Fisher–Kolmogorov model. Instead, a major discrepancy can be noticed with the spreading pattern of two-dimensional models. For example, consider [2]: not simply connected brain sections strongly affect protein diffusion (see Figure 5). Lateral ventricles would represent an obstacle to protein propagation in two dimensions. However, not simply connected points in two-dimensional can be reached with continuity in the three-dimensional brain mesh. Obviously, this affects the overall spreading of the wave solution. Also, the evolution time is of the order of 25–30 years, which is comparable to the medical references [8], and similar to other simulation results [2, 14, 16]. Furthermore, diffusion directions are globally coherent with the medical literature

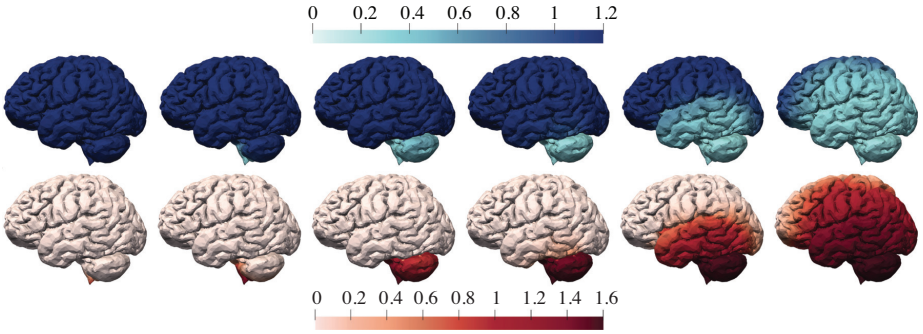


Figure 4. Test case from Section 5. Snapshot of the computed solution $c_h(\boldsymbol{x}, t)$ (blue) and $q_h(\boldsymbol{x}, t)$ (red) for $t \in \{0, 5, 10, 15, 20, 25\}$ years.

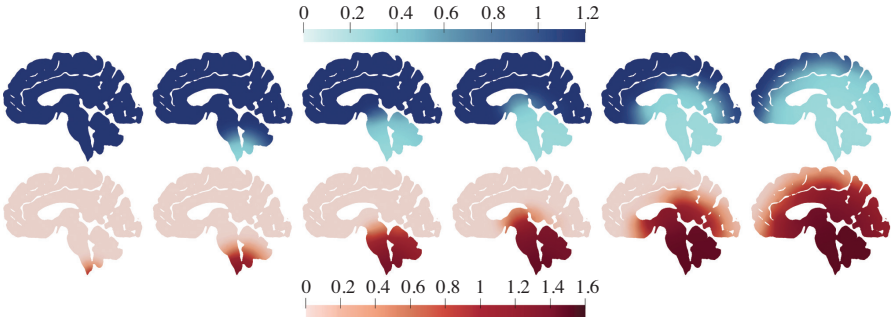


Figure 5. Test case from Section 5. Snapshot of the computed solution $c_h(\boldsymbol{x}, t)$ (blue) and $q_h(\boldsymbol{x}, t)$ (red) for $t \in \{0, 5, 10, 15, 20, 25\}$ years in a sagittal medial section.

since they reproduce the stereotypical pattern described in [8]. For instance, there is no cerebral cortex involvement in the absence of lesions in the brain stem, and early disease stages affect only the *medulla oblongata* and the *pontine tegmentum*.

From the phenomenological point of view, *nigral* and extranigral induction sites vary in their susceptibility to disease-related changes and follow a well-described temporal sequence of degeneration [8]. Starting from the motor nucleus, the misfolded protein concentration locally increases in the seeding area. Subsequently, it invades the lower brain stem along the axonal network, then spreads firstly into the mesocortex and eventually into the neocortex, coherently with Braak's staging theory [8]. The concentration of healthy proteins evolves almost symmetrically, giving insight into cell loss.

5.1. Validation versus Braak's stages

Refer to Figures 6, 7, 8, 9, and to Figure 10: the former depicts biomarker abnormality curves $B_i^{\alpha}(t)$ from the prion-like spreading of α-synuclein across different brain

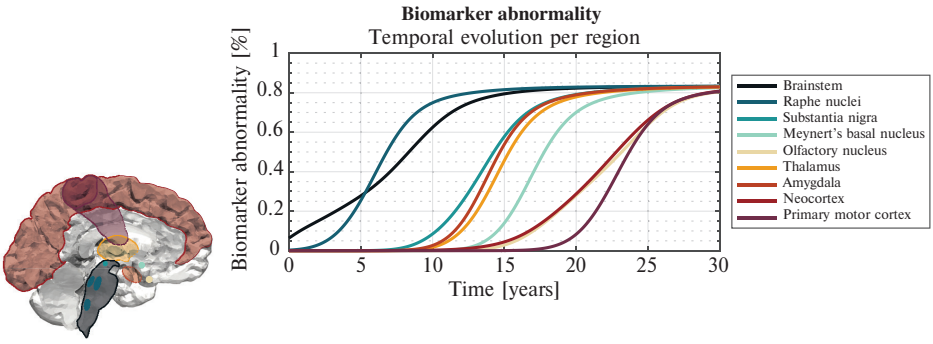


Figure 6. Temporal evolution of biomarker curves for different brain regions of clinical interest [8].

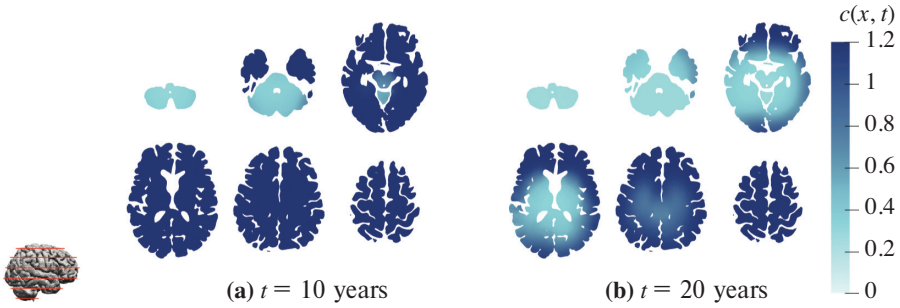


Figure 7. Test case from Section 5. Evaluation of healthy protein loss at 10 and 20 years, observed across six transversal sections within the three-dimensional domain shown on the left.

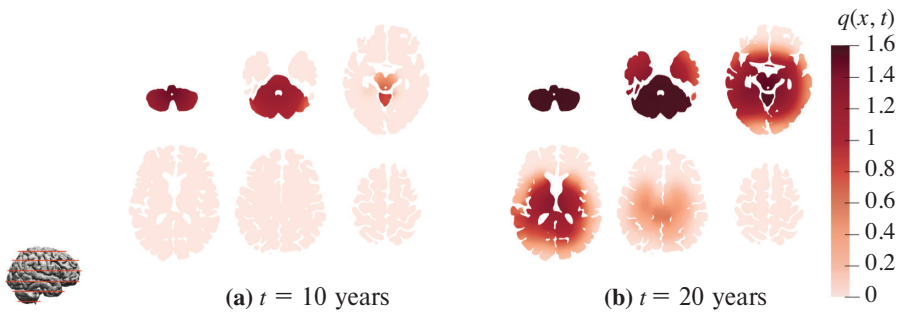


Figure 8. Test case from Section 5. Evaluation of misfolded protein diffusion at 10 and 20 years, observed across six transversal sections within the three-dimensional domain shown on the left.

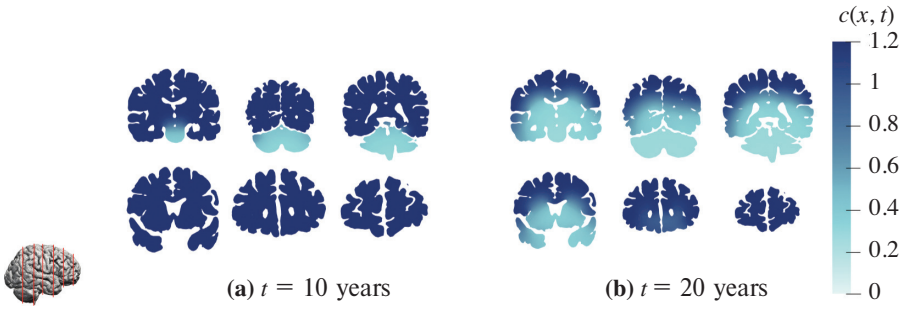


Figure 9. Test case from Section 5. Evaluation of healthy protein loss at 10 and 20 years, observed across six coronal sections within the three-dimensional domain shown on the left.

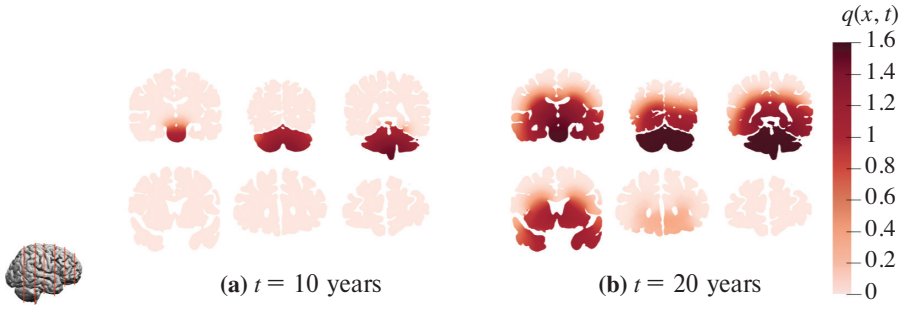


Figure 10. Test case from Section 5. Evaluation of misfolded protein diffusion at 10 and 20 years, observed across six coronal sections within the three-dimensional domain shown on the left.

regions $R_i \subset \Omega$ of clinical interest [8], where we define $B_i^{\alpha}(t)$ as

$$B_i^{\alpha}(t) = \frac{\int_{R_i} q(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}}{\int_{R_i} c(\boldsymbol{x}, t) + q(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x}}.$$

All of them are smooth sigmoidal curves, in agreement with clinical biomarkers of neurodegenerative disease progression. It is also possible to identify the six different Braak's stages [8]. During stage 1, the lower brainstem is affected: firstly, the dorsal motor nucleus of the vagus nerve in the medulla oblongata, and then the pons reach a critical concentration value. After, during stage 2, the raphe nuclei are also affected, and the misfolded proteins travel up the brainstem into the pontine tegmentum. During stage 3, the *substantia nigra* is reached: the *pars compacta*, together with the basal nucleus of Meynert, starts showing misfolding. Stage 4 coincides with the mesocortex invasion, while the neocortex is still healthy. However, the amygdala is now compromised, as the anterior olfactory nucleus is. These two stages represent the start of sleep and motor disturbances. Stage 5 also involves the neocortex,

thus the temporal, parietal, and frontal lobes. The substantia nigra is almost completely compromised. In stage 6, misfolded proteins have fully invaded the neocortex, affecting also, for example, the primary motor cortex, and other sensory areas. Cognitive impairment is now severe. The olfactory cortex appears delayed by several years when compared to clinical observations. This discrepancy arises because, in our simulation, we consider only the medulla oblongata as the seeding region. Consequently, the infection of the olfactory cortex occurs significantly later than the average observed in patients, where misfolded proteins are detected early in the olfactory cortex due to its proximity to the olfactory lobe.

The substantia nigra impairment is a crucial point in the disease progression. Indeed, at this stage, symptoms become evident [8]. Consequently, this region is particularly significant in the study of Parkinson's disease evolution. By establishing a critical concentration threshold of $q_{\text{crit}} = 0.2$, at which point a brain region is deemed compromised, we can determine the temporal duration of each Braak stage in our simulation. Consistent with [8], we align the onset of stage 3 and stage 5, with substantia nigra and neocortex compromise, respectively. Additionally, by considering the data in [8, Table 2], we can validate the temporal evolution of the biomarker in the substantia nigra, $B^{\alpha}_{\text{SN}}(t)$. Figure 11 illustrates the curve $B^{\alpha}_{\text{SN}}(t)$ obtained from the numerical simulation of this section alongside the averaged values of 110 Parkinson's disease patients at each Braak stage. The result is promising, yet it is heavily constrained by the semi-quantitative nature of the literature-acquired data.

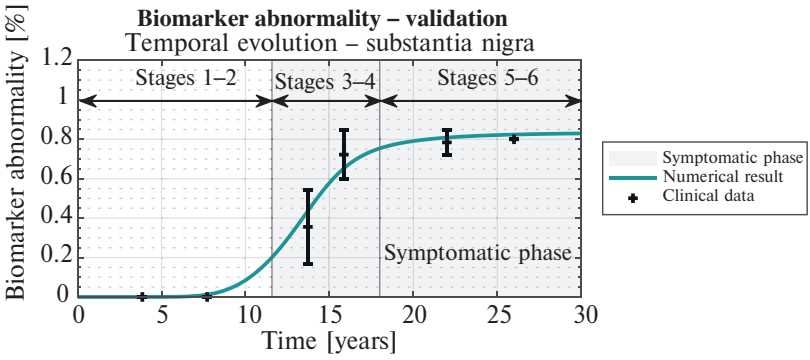


Figure 11. Test case of Section 5.1. Validation for the biomarker abnormality in the substantia nigra brain region. The green line represents the biomarker abnormality curve computed from the numerical results obtained from the simulation of Section 5. Clinical data are taken from [8, Table 2]. The severity of the disease is mapped on the range $[0, 0.8]$, and for each stage, we take the average value and show the symmetric interval with amplitude twice the standard deviation.

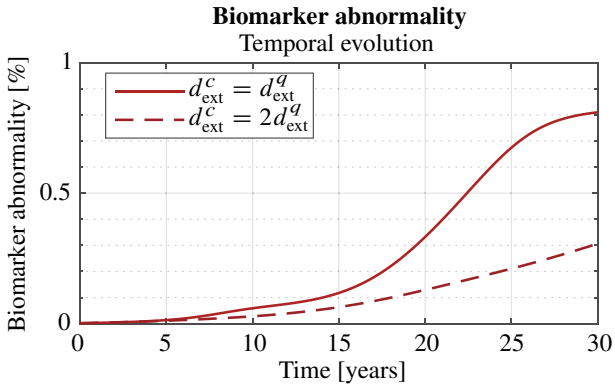


Figure 12. Test case of Section 5.2. Sensitivity analysis for two different mechanisms of extracellular diffusion.

5.2. Sensitivity analysis

This section briefly discusses the results of a sensitivity analysis study that explores two parameter variations not covered in the literature.

Time step length. First, we address the time step length. Even though a longer Δt does not cause spurious oscillations in the numerical solution representation, it induces a much lower spreading over the domain. A proper choice of Δt is in the order of 10^{-2} years.

Diffusion coefficients. Second, we study the relative contribution of extracellular diffusion of healthy and misfolded proteins: indeed, the different steric hindrances might induce different gradient-driven velocities. Increasing the extracellular diffusion of the healthy component induces a faster spreading of the healthy proteins within the domain. Thus, it takes a longer time to locally accumulate misfolded proteins and cause the disease. As an instance, the choice $d^c_{\text{ext}} = 2d^q_{\text{ext}} = 16\,\text{mm}^2\,\text{years}^{-1}$ induces a temporal delay for the biomarker curve of over 10 years. Figure 12 depicts the temporal evolution of the global biomarker abnormality curve versus the period of 30 years for the two choices $d^c_{\text{ext}} = d^q_{\text{ext}}$ and $d^c_{\text{ext}} = 2d^q_{\text{ext}}$. In the latter case, the saturation time is reached much later than expected according to the medical literature. Therefore, deviating from clinical observations over several years, it does not appear to be a good modeling choice.

6. Conclusions

In this study, we applied the heterodimer model to the study of the diffusion of the α-synuclein protein throughout the brain in Parkinson's disease. We discussed

some theoretical results concerning the heterodimer model, and we derived its polytopal discontinuous Galerkin formulation. Convergence tests in two and three dimensions confirm the theoretical results for implicit and semi-implicit treatments of the non-linear term in trivial domains. Furthermore, we conducted a convergence analysis with respect to the mesh size h and the order of approximation p. Lastly, we performed a realistic simulation of α-synuclein protein propagation across a three-dimensional brain mesh obtained from structural magnetic resonance imaging. Our results have been validated by comparison with clinical data from the literature, showing good agreement. Additionally, biomarker curves correspond well with the available clinical data. The choice of using the polytopal discontinuous Galerkin method has been justified by both the high rate of convergence achieved in numerical tests and the effective description of propagating fronts in the three-dimensional domain. Indeed, our results align well with the postmortem histological analyses and Braak's staging theory. The simulation also underscored the importance of detailed spatial simulations and the limitations of two-dimensional ones on brain sections. Moreover, it represents a step forward compared to the Fisher–Kolmogorov model, offering insights into the concentration of healthy proteins within the brain during the pathology.

The computational framework enables the study of potential biomarkers for neurodegeneration not in vivo. This approach could lead to benefits in terms of early intervention in therapy. The mechanistic description of the phenomenon enables examination of the relative significance of individual contributions, thereby facilitating the customization of therapy approaches. Furthermore, the complete spatial description also allows one to study a priori the timing and the location of interventions. In the future, it is reasonable to envision patient-specific therapy based on quantitative models for prion-like diseases.

Acknowledgments. The brain MRI images were provided by OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

Funding. P.F.A. and M.C. received funding from the European Union (ERC SyG, NEMESIS, project number 101115663). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. P.F.A. was partially supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data, and Quantum Computing, funded by the European Union – NextGenerationEU.

P.F.A. and M.C. are members of the INdAM Research Group GNCS. The present research is part of the activities of Dipartimento di Eccellenza 2023–2027.

References

[1] P. R. Amestoy, I. S. Duff, and J.-Y. L'Excellent, Multifrontal parallel distributed symmetric and unsymmetric solvers. *Comput. Methods Appl. Mech. Engrg.* **184** (2000), no. 2–4, 501–520 Zbl 0956.65017

[2] P. F. Antonietti, F. Bonizzoni, M. Corti, and A. Dall'Olio, Discontinuous Galerkin approximations of the heterodimer model for protein–protein interaction. *Comput. Methods Appl. Mech. Engrg.* **431** (2024), article no. 117282, 17 pp. Zbl 1550.92001 MR 4785276

[3] P. F. Antonietti, A. Cangiani, J. Collis, Z. Dong, E. H. Georgoulis, S. Giani, and P. Houston, Review of discontinuous Galerkin finite element methods for partial differential equations on complicated domains. In G. R. Barrenechea, F. Brezzi, A. Cangiani, and E. H. Georgoulis (eds.), *Building bridges: connections and challenges in modern approaches to numerical partial differential equations*, pp. 281–310, Lect. Notes Comput. Sci. Eng. 114, Springer, Cham, 2016 Zbl 1357.65251 MR 3585793

[4] P. F. Antonietti, C. Facciolà, P. Houston, I. Mazzieri, G. Pennesi, and M. Verani, High-order discontinuous Galerkin methods on polyhedral grids for geophysical applications: seismic wave propagation and fractured reservoir simulations. In D. A. Di Pietro, L. Formaggia, and R. Masson (eds.), *Polyhedral methods in geosciences*, pp. 159–225, SEMA SIMAI Springer Ser. 27, Springer, Cham, 2021 Zbl 1470.86002 MR 4287607

[5] P. F. Antonietti, N. Farenga, E. Manuzzi, G. Martinelli, and L. Saverio, Agglomeration of polygonal grids using graph neural networks with applications to multigrid solvers. *Comput. Math. Appl.* **154** (2024), 45–57 Zbl 1538.65487 MR 4669421

[6] P. F. Antonietti, S. Giani, and P. Houston, *hp*-Version composite discontinuous Galerkin methods for elliptic problems on complicated domains, *SIAM J. Sci. Comput*. **35** (2013), no. 3, A1417–A1439 Zbl 1284.65163 MR 3061474

[7] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39** (2002), no. 5, 1749–1779 Zbl 1008.65080 MR 1885715

[8] H. Braak, K. Del Tredici, U. Rüb, R. A. I. de Vos, E. N. H. Jansen Steur, and E. Braak, Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging* **24** (2003), no. 2, 197–211

[9] A. Cangiani, Z. Dong, E. H. Georgoulis, and P. Houston, *hp*-Version discontinuous Galerkin methods on polygonal and polyhedral meshes. SpringerBriefs Math., Springer, Cham, 2017, 131 pp. Zbl 1382.65307 MR 3729265

[10] A. Cangiani, E. H. Georgoulis, and P. Houston, *hp*-Version discontinuous Galerkin methods on polygonal and polyhedral meshes, *Math. Models Methods Appl. Sci*. **24** (2014), no. 10, 2009–2041 Zbl 1298.65167 MR 3211116

[11] F. Carbonell, Y. Iturria-Medina, and A. C. Evans, Mathematical modeling of protein misfolding mechanisms in neurological diseases: a historical overview. *Front. Neurol.* **9** (2018), article no. 37, 16 pp.

[12] F. E. Cohen, K.-M. Pan, Z. Huang, M. Baldwin, R. J. Fletterick, and S. B. Prusiner, Structural clues to prion replication. *Science* **264** (1994), no. 5158, 530–531

[13] M. Corti, F. Bonizzoni, and P. F. Antonietti, Structure preserving polytopal discontinuous Galerkin methods for the numerical modeling of neurodegenerative diseases. *J. Sci. Comput.* **100** (2024), no. 2, article no. 39, 24 pp. Zbl 1551.65066 MR 4761432

[14] M. Corti, F. Bonizzoni, L. Dede', A. M. Quarteroni, and P. F. Antonietti, Discontinuous Galerkin methods for Fisher–Kolmogorov equation with application to α-synuclein spreading in Parkinson's disease. *Comput. Methods Appl. Mech. Engrg.* **417, part A** (2023), article no. 116450, 17 pp. Zbl 1539.65116 MR 4645591

[15] D. W. Dickson, Neuropathology of Parkinson disease. *Parkinsonism Relat. Disord.* **46** (2018), no. Suppl. 1, S30–S33

[16] S. Fornari, A. Schäfer, M. Jucker, A. Goriely, and E. Kuhl, Prion-like spreading of Alzheimer's disease within the brain's connectome. *J. Roy. Soc. Interface* **16** (2019), no. 159, article no. 20190356, 12 pp.

[17] B. Frost, R. L. Jacks, and M. I. Diamond, Propagation of Tau misfolding from the outside to the inside of a cell. *J. Biol. Chem.* **284** (2009), no. 9, 12845–12852

[18] M. Hasegawa, T. Nonaka, and M. Masuda-Suzukake, Prion-like mechanisms and potential therapeutic targets in neurodegenerative disorders. *Pharmacol. Ther.* **172** (2017), 22–33

[19] J. S. Hesthaven and T. Warburton, *Nodal discontinuous Galerkin methods. Algorithms, analysis, and applications*. 1st edn., Texts Appl. Math. 54, Springer, New York, 2008, 502 pp. Zbl 1134.65068

[20] M. Iljina, G. A. Garcia, M. H. Horrocks, L. Tosatto, M. L. Choi, K. A. Ganzinger, A. Y. Abramov, S. Gandhi, N. W. Wood, N. Cremades, C. M. Dobson, T. P. J. Knowles, and D. Klenerman, Kinetic model of the aggregation of alpha-synuclein provides insights into prion-like spreading. *Proc. Natl. Acad. Sci. USA* **113** (2016), no. 9, E1206–E1215

[21] J. T. Jarrett and P. T. Lansbury, Jr., Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* **73** (1993), no. 6, 1055–1058

[22] M. Jucker and L. C. Walker, Pathogenic protein seeding in Alzheimer disease and other neurodegenerative disorders. *Ann. Neurol.* **70** (2011), no. 4, 532–540

[23] M. Jucker and L. C. Walker, Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature* **501** (2013), no. 7465, 45–51

[24] M. Jucker and L. C. Walker, Propagation and spread of pathogenic protein assemblies in neurodegenerative diseases. *Nat. Neurosci.* **21** (2018), no. 10, 1341–1349

[25] Š. Korat, N. S. R. Bidesi, F. Bonanno, A. Di Nanni, A. N. N. Hoàng, K. Herfert, A. Maurer, U. M. Battisti, G. D. Bowden, D. Thonon, D. Vugts, A. D. Windhorst, and M. M. Herth, Alpha-synuclein PET tracer development—an overview about current efforts. *Pharmaceuticals* **14** (2021), no. 9, article no. 847, 45 pp.

[26] P. J. LaMontagne, T. L. S. Benzinger, J. C. Morris, S. Keefe, R. Hornbeck, C. Xiong, E. Grant, J. Hassenstab, K. Moulder, A. G. Vlassenko, M. E. Raichle, C. Cruchaga, and D. Marcus, OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. 2019, medRxiv, DOI 10.1101/2019.12.13.19014902

[27] Z. R. Li, G. R. Liu, and D. Mi, Quantifying the parameters of Prusiner's heterodimer model for prion replication. *Phys. A* **346** (2005), no. 3–4, 459–474

[28] A. Liebert, B. Bicknell, and R. Adams, Prion protein signaling in the nervous system—a review and perspective. *Signal Transduct. Insights* **3** (2014), article no. STI.S12319, 22 pp.

[29] K.-A. Mardal, M. E. Rognes, T. B. Thompson, and L. M. Valnes, *Mathematical modeling of the human brain. From magnetic resonance images to finite element simulation*. 1st edn., Simula SpringerBriefs Comput. 10, Springer, Cham, 2022, 118 pp. Zbl 1501.92001

[30] J. Masel, V. A. A. Jansen, and M. A. Nowak, Quantifying the kinetic parameters of prion replication. *Biophys. Chem.* **77** (1999), no. 2–3, 139–152

[31] F. Matthäus, Diffusion versus network models as descriptions for the spread of prion diseases in the brain. *J. Theoret. Biol.* **240** (2006), no. 1, 104–113 Zbl 1447.92095 MR 2227720

[32] J. D. Murray, Spatial pattern formation with reaction diffusion systems. In J. D. Murray (ed.), *Mathematical biology. II: Spatial models and biomedical applications*, pp. 71–140, Interdiscip. Appl. Math. 18, Springer, New York, 1993 Zbl 1006.92002

[33] J. D. Murray, *Mathematical biology. I. An introduction*. 3rd edn., Interdiscip. Appl. Math. 17, Springer, New York, 2002, 551 pp. Zbl 1006.92001 MR 1908418

[34] G. Perrino, C. Wilson, M. Santorelli, and D. di Bernardo, Quantitative characterization of α-synuclein aggregation in living cells through automated microfluidics feedback control. *Cell Rep.* **27** (2019), no. 3, 916–927

[35] S. B. Prusiner, Novel *proteinaceous* infectious particles cause scrapie. *Science* **216** (1982), no. 4542, 136–144

[36] S. B. Prusiner, Molecular biology of prion diseases. *Science* **252** (1991), no. 5012, 1515–1522

[37] S. B. Prusiner, Prions. *Proc. Natl. Acad. Sci. USA* **95** (1998), no. 23, 13363–13383 [Nobel lecture]

[38] A. Quarteroni, *Numerical models for differential problems*. MS&A. Model. Simul. Appl. 16, Springer, Cham, 2017, 692 pp. Zbl 1436.65003 MR 3702005

[39] B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*. Frontiers Appl. Math. 35, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008, 178 pp. Zbl 1153.65112 MR 2431403

[40] A. Schäfer, J. Weickenmeier, and E. Kuhl, The interplay of biochemical and biomechanical degeneration in Alzheimer's disease. *Comput. Methods Appl. Mech. Engrg.* **352** (2019), 369–388 Zbl 1441.74132 MR 3949624

[41] W. W. Seeley, R. K. Crawford, J. Zhou, B. L. Miller, and M. D. Greicius, Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62** (2009), no. 1, 42–52

[42] R. Smith, F. Capotosti, M. Schain, T. Ohlsson, E. Vokali, J. Molette, T. Touilloux, V. Hliva, I. K. Dimitrakopoulos, A. Puschmann, J. Jögi, P. Svenningsson, M. Andréasson, C. Sandiego, D. S. Russell, P. Miranda-Azpiazu, C. Halldin, E. Stomrud, S. Hall, K. Bratteby, E. T. L'Estrade, R. Luthi-Carter, A. Pfeifer, M. Kosco-Vilbois, J. Streffer, and O. Hansson, The α-synuclein PET tracer [18F] ACI-12589 distinguishes multiple system atrophy from other neurodegenerative diseases. *Nat. Commun.* **14** (2023), no. 1, article no. 6750, 12 pp.

[43] M. G. Spillantini, M. L. Schmidt, V. M.-Y. Lee, J. Q. Trojanowski, R. Jakes, and M. Goedert, α-Synuclein in Lewy bodies. *Nature* **388** (1997), no. 6645, 839–840

[44] B. E. Stopschinski and M. I. Diamond, The prion model for progression and diversity of neurodegenerative diseases. *Lancet. Neurol.* **16** (2017), no. 4, 323–332

[45] T. B. Thompson, P. Chaggar, E. Kuhl, and A. Goriely, Protein-protein interactions in neurodegenerative diseases: a conspiracy theory. *PLOS Comput. Biol.* **16** (2020), no. 10, article no. e1008267, 41 pp.

[46] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff, D. Ames, C. C. Rowe, and C. L. Masters, for the Australian Imaging Biomarkers and Lifestyle (AIBL) Research Group, Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. *Lancet. Neurol.* **12** (2013), no. 4, 357–367

[47] J. Weickenmeier, M. Jucker, A. Goriely, and E. Kuhl, A physics-based model explains the prion-like features of neurodegeneration in Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis. *J. Mech. Phys. Solids* **124** (2019), 264–281

[48] J. Weickenmeier, E. Kuhl, and A. Goriely, Multiphysics of prionlike diseases: progression and atrophy. *Phys. Rev. Lett.* **121** (2018), no. 15, article no. 158101, 6 pp.

Paola F. Antonietti
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy; paola.antonietti@polimi.it

Mattia Corti
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy; mattia.corti@polimi.it

Giacomo Lorenzon
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy; giacomo.lorenzon@polimi.it

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/9

Dualities in convex geometric analysis and beyond

Shiri Artstein-Avidan

Abstract. After a short introduction to asymptotic geometric analysis, we will discuss duality theory in the context of asymptotic geometric analysis and describe recent results with roots in optimal transport theory, as well as their connections to some of the older open problems in the theory.

1. Short introduction: Asymptotic geometric analysis

Asymptotic geometric analysis (AGA) lies in the meeting point of analysis and geometry. This field, "born" in the 1980s, originally sought to understand infinite-dimensional Banach spaces by studying their finite-dimensional sections and projections. One starting point of the field is the pioneering work of Vitali Milman [68], who provided a quantitative proof for Dvoretzky's theorem [36], a cornerstone result which states that every infinite-dimensional Banach space has subspaces of dimension n which are ε-close to Euclidean, of any dimension n and any ε-closeness. Milman showed that this result holds not only for infinite-dimensional spaces but also for spaces of sufficiently large dimension, depending on ε and on the dimension of the subspace required. Furthermore, he showed that in this regime, with the correct choice of a Euclidean structure, this holds for a large measure of subspaces. He introduced one of the first important tools in this field, which is the concentration of measure phenomenon (see [9, Chapter 3]), roughly saying that a 1-Lipschitz function on the sphere S^{n-1} has to be close to its median on a set of very large measure. This phenomenon, that he used to give a probabilistic proof of Dvoretzky's theorem, continues to play a major role not only in the study of high-dimensional convex geometry but in many other areas as well, such as analysis, combinatorics, probability and computer science (see, e.g., [58, 83]).

Since a normed space is defined by its unit ball, which is a centrally symmetric convex body, the study of AGA naturally involves convexity theory. Within AGA, the

Mathematics Subject Classification 2020: 52A05 (primary); 06D50, 52A20, 52A30 (secondary).
Keywords: duality, convex bodies, asymptotic geometric analysis.

focus is usually on asymptotic questions, particularly as the dimension becomes large. This shifted the types of questions considered from those in classical convexity theory. Nowadays, AGA has applications in many fields, a good example being questions in so-called "big data", where large dimensions are ubiquitous. An exciting aspect of the field is that it combines methods from many different areas in mathematics, ranging from topology to advanced analysis, geometry, probability and combinatorics. At the same time, due to the very basic nature of the topics, tools developed in AGA are applicable in wide areas of mathematics as well.

Let us list shortly the basic concepts which will be relevant in this short note. We work in $\mathbb{R}^n$ where we will be concerned with closed convex sets, or bodies (when the sets are assumed to be compact and have non-empty interior). Since originally the sets considered were unit balls of norms, central symmetry is sometimes assumed, but usually we shall consider convex bodies without requiring central symmetry, and then it is sometimes useful to assume that their center of mass is at the origin. The notion of volume, namely n-dimensional Lebesgue measure in $\mathbb{R}^n$, will be our way of measuring the size of a set (and we denote it by $\mathrm{Vol}(\cdot)$). There are various other ways to measure the size of convex bodies, ranging from classical geometric measures such as surface area or width, to more specialized ones like various capacities and eigenvalues. For the purpose of this note, the classical Lebesgue measure will be sufficient. A basic operation on subsets of $\mathbb{R}^n$ which combines two sets is that of Minkowski addition. For $A, B \subseteq \mathbb{R}^n$, it is defined by $A + B = \{x + y : x \in A,\ y \in B\}$. One easily checks that if A and B are convex, then so is $A + B$, and that $A + A = 2A$ when A is convex. This operation appears naturally and plays an important role in many different scientific fields. For example, in robotics, when an object is to traverse some region with obstacles, the permissible locations for the center of mass of the robot are the complement of the Minkowski sum of the obstacles with (the minus of) the robot.

We have not yet described the main object we address in this note, which is duality or polarity. But even before introducing dualities, we can with the above concepts already state two very fundamental theorems from AGA which will also play a role later in this text and can serve to complete this short introduction of the reader to this field. The first is the well-known Brunn–Minkowski inequality from [34,70]; see also [80].

Theorem 1.1 (Brunn–Minkowski inequality). *For any measurable $A, B, C \subset \mathbb{R}^n$ with $A + B \subset C$, it holds that*

$$\mathrm{Vol}(C)^{1/n} \geq \mathrm{Vol}(A)^{1/n} + \mathrm{Vol}(B)^{1/n}. \tag{1.1}$$

(Usually one assumes that $A + B = C$ is measurable and omits C from the statement of the theorem.) Note that no convexity was assumed, and indeed elementary

proofs for the inequality exist which can be explained to an interested middle school student. Applications of Theorem 1.1 are remarkably broad; for example, it can be used to give a one-line proof of the isoperimetric inequality. Whole books can be (and have been) written on the consequences of this inequality; see [9, 10, 43, 80].

Another fundamental fact for AGA, which can be stated using only volume and Minkowski addition, is Minkowski's polynomiality theorem [70] which we formulate next. The case of two bodies, one of which is the Euclidean ball, is due to Steiner [82].

Theorem 1.2 (Minkowski's polynomiality theorem). *Let $n \geq 1, m \geq 1$. Assume that $K_1, \ldots, K_m \subset \mathbb{R}^n$ are compact convex sets, and let $(t_i)_{i=1}^m \subset \mathbb{R}^+$. The volume of the Minkowski combination $\sum_{i=1}^m t_i K_i$ is a homogeneous polynomial of degree n in the variables (t_i) and, moreover, the coefficient of any monomial depends only on the bodies corresponding to indices appearing in the monomial. One writes*

$$\mathrm{Vol}\Big(\sum_{i=1}^m t_i K_i\Big) = \sum_{i_1,\ldots,i_n=1}^m \Big(\prod_{j=1}^n t_{i_j}\Big) V(K_{i_1}, \ldots, K_{i_n}), \quad t_i \geq 0, \tag{1.2}$$

and the coefficients $V(K_{i_1}, \ldots, K_{i_n})$, which are symmetric with respect to permutations of the arguments, are called the mixed volumes of the bodies.

When there are only two bodies involved, the coefficients are given by the mixed volume of j copies of K_1 and $(n-j)$ copies of K_2. We denote these quantities by $V(K_1[j], K_2[n-j])$. In such a case, we have for $s, t \geq 0$

$$\mathrm{Vol}(tK_1 + sK_2) = \sum_{j=0}^n \binom{n}{j} t^j s^{n-j} V(K_1[j], K_2[n-j]).$$

In particular, one easily sees that $V(K[n]) = \mathrm{Vol}(K)$ and that for the Euclidean ball B_2^n the quantity $nV(K[n-1], B_2^n)$ is simply the surface area of K as it is the derivative of $f(t) = \mathrm{Vol}(K + tB_2^n)$ at $t = 0$.

Observe that if we consider the Minkowski sum of K and $-K = \{-x : x \in K\}$, the Brunn–Minkowski inequality implies $2^n \mathrm{Vol}(K) \leq \mathrm{Vol}(K + (-K))$, with equality when K is centrally symmetric. In this case, there is a counterpart: the Rogers–Shephard inequality [77] which states that

$$\mathrm{Vol}(K + (-K)) \leq \binom{2n}{n} \mathrm{Vol}(K) \tag{1.3}$$

with equality attained for a simplex (the convex hull of $n+1$ affinely independent points). This inequality is in the reverse direction of the one in Theorem 1.1, and it should be mentioned that such reverse theorems are of great interest, as we shall see in the next sections. Among such theorems are the reverse isoperimetric inequality [25]

and the reverse Brunn–Minkowski inequality [67] (see also [9, Chapter 8]), both of which require the bodies involved to be in a certain position (or, in other words, a reasonable choice of the coordinate system). The Rogers–Shephard inequality above is invariant under linear transformations and thus does not require a "position".

Note that by Theorem 1.2, we may expand the volume of $K - K$ (and rewrite the binomial coefficient as a sum) so that inequality (1.3) takes the form

$$\sum_{j=0}^{n}\binom{n}{j}V(K[j],-K[n-j]) \le \sum_{j=0}^{n}\binom{n}{j}^2 \mathrm{Vol}(K).$$

This form brings us to the first conjecture to be mentioned in this note, which is due to Godbersen [47] (formulated independently in [50]). The conjecture is that for any convex body $K \subset \mathbb{R}^n$ and any $1 \le j \le n$, it holds that

$$V(K[j],-K[n-j]) \le \binom{n}{j}\mathrm{Vol}(K), \tag{1.4}$$

with equality only for K being a simplex. The conjecture is verified only in some specific cases, and the current best upper bounds stem from [6] or [19]. In the latter, we show that for any $\lambda \in [0,1]$, it holds (with equality for simplices) that

$$\sum_{j=0}^{n}(1-\lambda)^j\lambda^{n-j}V(K[j],-K[n-j]) \le \mathrm{Vol}(K), \tag{1.5}$$

any by taking just the jth term and optimizing over λ, one gets the bound $\binom{n}{j}$ up to some factor $\sim\sqrt{n}$. This short introduction was meant to show that even with very few definitions, deep theorems can be stated as well as long-standing open problems.

2. Duality and polarity

Duality is a central theme in mathematics, with many manifestations across different fields. It represents a very general meta-mathematical concept with several influential realizations. Examples include duality of normed spaces (associating a normed space with the space of linear functionals on it, with the operator norm), duality in linear programming and duality for convex functions, associating to a convex function φ its Legendre transform

$$\mathcal{L}\varphi(y) = \sup_x(\langle y,x\rangle - \varphi(x)). \tag{2.1}$$

The latter is analogous to the relationship between Lagrangians and Hamiltonians in mechanics. It seems that in every area of mathematics, some special and useful

dualities appear. In the context of convex bodies, duality can be understood through polarity. The polar of a convex body K is given by

$$K^\circ = \Big\{ y \in \mathbb{R}^n : \sup_{x \in K} \langle x, y \rangle \leq 1 \Big\}. \tag{2.2}$$

Usually one assumes that K contains the origin in its interior, as otherwise the polar is not compact. This notion, for centrally symmetric convex bodies, coincides with the duality in (finite-dimensional) normed spaces alluded to above, under the usual identification of a space with its unit ball. We mention that with a duality at hand (we will shortly explain what we mean by a "duality"), self-dual objects are of special interest. For normed spaces self-dual objects are Hilbert spaces, for convex functions the only self-dual function is $\|x\|_2^2/2$ and for set polarity the self-polar set is B_2^n, the Euclidean unit ball.

Bringing polarity into the convex geometry picture gives rise to many exciting new possibilities. As a first step, the volume product is introduced and is defined by $\mathrm{Vol}(K)\mathrm{Vol}(K^\circ)$. A classical result by Blaschke and Santaló [28, 29, 79] states that the volume product is maximized when K is the Euclidean ball, namely the self-dual object in the class (or any of its linear images, a.k.a. ellipsoids). This is true when K is centrally symmetric, and when it is not, one should make sure that the origin is not too close to the boundary of K (which would cause the dual body to be very large). For example, the inequality $\mathrm{Vol}(K)\mathrm{Vol}(K^\circ) \leq \mathrm{Vol}(B_2^n)^2$ holds whenever the center of mass of K is at the origin (see [64] and also [9, Chapter 1]).

A long-standing open problem is to determine the convex body which minimizes the volume product. Mahler's conjecture [61] is that the centered simplex is yet again the solution, namely that the volume product is minimized by simplices, and that considering only centrally symmetric convex sets (we will refer to this as the "centrally symmetric Mahler conjecture") the minimal value for the volume product is that of the cube $[-1, 1]^n$ and its dual $\mathrm{conv}(\pm e_i)$, which is $4^n/n!$. The cube is not unique in obtaining this volume product, and there is a family of polytopes called "Hanner polytopes" all of which have the same volume product and are conjectured to be the set of minimizers. Mahler's conjecture in dimension 2 was proved in his original manuscript [61]. The case of $n = 3$ was recently solved by Iriyeh and Shibata [52]; see also [39, 42]. In higher dimensions, the conjecture is open, but has been verified in many special cases; see, e.g., [48, 65, 75, 78].

It is worthwhile to note that taking power $1/n$ of the volumes (which is rather natural as volume is n-homogeneous, which is why these powers appear in Theorem 1.1, for example), the upper bound $(\mathrm{Vol}(B_2^n))^{2/n}$ and the conjectured lower bound $4/(n!)^{1/n}$ are of the same order. In [32], a lower bound for the volume product of the form $c^n\,\mathrm{Vol}(B_2^n)^2$ was obtained and was sufficient for many applications (it is also very intimately related to the reverse form of the Brunn–Minkowski

inequality mentioned above). Similar lower bounds in the general case were obtained in [45, 56, 72].

Mahler's conjecture, surprisingly, turns out to be intimately connected with a certain conjecture in symplectic geometry [85]. The conjecture pertains to symplectic capacities, which are a symplectically invariant way of measuring the size of a set in $\mathbb{R}^{2n}$. Many capacities exist (a function is called a capacity if it is monotone, symplectic invariant and equals πr^2 for both a Euclidean ball of radius r and the complex cylinder $rB_2^2 \times \mathbb{R}^{2n-2} \subset \mathbb{R}^{2n}$), and among them there is a "largest" one and a "smallest" one, called the Gromov width and the ball capacity, respectively; see [51] for details. Viterbo's conjecture was that, among sets of fixed volume, the largest capacity is attained for the Euclidean ball

$$\frac{\mathrm{cap}(T)}{\mathrm{cap}(B_2^{2n})} \leq \left(\frac{\mathrm{Vol}(T)}{\mathrm{Vol}(B_2^{2n})}\right)^{1/n}.$$

In [11] we showed that Viterbo's conjecture for certain centrally symmetric bodies (namely, $T = K \times K^\circ$ when $\mathbb{R}^{2n}$ is identified as $\mathbb{C}^n = \mathbb{R}^n \times i\mathbb{R}^n$) is equivalent to Mahler's conjecture. (As a side note, we mention that the introduction of convexity, which is a linear notion, into symplectic geometry is quite unnatural. Nevertheless, it turns out to be quite fruitful; for instance, we proved in [18] a Brunn–Minkowski inequality for certain important capacities.) Very recently, a counterexample to Viterbo's conjecture was found by Haim-Kislev and Ostrover [49]. The example is not centrally symmetric and therefore does not directly relate to Mahler's conjecture. In the flavor of AGA, we also mention here our result from [17] where we showed that up to a universal constant C, which does not depend on the dimension nor on the bodies involved, an isoperimetric inequality of Viterbo-type does hold, namely $\mathrm{cap}(T)/\mathrm{cap}(B_2^{2n}) \leq C\,\mathrm{Vol}(T)^{1/n}/\mathrm{Vol}(B_2^{2n})^{1/n}$, improving the constant n which was established by Viterbo in [85].

Another direction in which questions about volume product have been addressed is that of functional versions for the geometric inequality. This is a general theme in AGA, where many definitions, theorems and conjectures have functional counterparts. These analogies have proven to be very useful, where in many cases the geometric inequalities are special cases of the functional ones, but within the realm of functions, we have many more tools at our disposal. For an overview of this direction of research, see [10, Chapter 9]. For this analogy, usually a convex body is replaced by a lower semi-continuous convex function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and Minkowski addition is replaced by infimal convolution (which corresponds to Minkowski addition of the epi-graphs in $\mathbb{R}^{n+1}$ which are themselves closed convex sets). The "volume" of φ is given by $\int \exp(-\varphi)$. Under this analogy, Brunn–Minkowski inequality has a counterpart which is called the Prékopa–Leindler inequality [60, 74] (it too does not

require assumptions about convexity)

$$\int h \geq \left(\int f\right)^{1-\lambda} \left(\int g\right)^{\lambda} \tag{2.3}$$

whenever $h((1-\lambda)x + \lambda y) \geq f(x)^{1-\lambda} g(y)^{\lambda}$ for all $x, y \in \mathbb{R}^n$.

To replace duality in this analogy, we will consider the Legendre transform (2.1), which is, as mentioned, the classical duality for convex functions [76]. We will readily explain in which senses this is a natural choice and what it can be replaced with. Nevertheless, under this analogy, a version of the Blaschlke–Santaló inequality was proved in [24] for even functions (which are the analogues of centrally symmetric bodies):

$$\int \exp(-\varphi) \int \exp(-\mathcal{L}\varphi) \leq \left(\int \exp(-|x|^2/2) dx\right)^2 = (2\pi)^n. \tag{2.4}$$

Again we see that the maximizer is the self-dual function. In [12], this was generalized to functions which are not necessarily even, in which case "choosing the right origin" amounts, in the dual, to adding some linear function to φ. Many other proofs of the functional version of the Santaló inequality have appeared since then; see [40, 59]. Very recently an analytical approach based on heat flow was obtained, where it is proven that the functional volume product increases along the heat flow [71]. Regarding an analogue for Mahler's conjecture, such were given in [41]. Some partial results resembling those for bodies have been obtained; see [42] for a full account.

3. Characterizing duality

In the paper [40], along with providing a simple proof that the functional Santaló inequality follows from the geometric one, the authors gave a more general form of a functional inequality of "Santaló-type", which to us served as a motivation to try and understand what is the "correct" form of a functional duality, and more generally, why are the specific forms of duality used in convex geometry, $K \mapsto K^\circ$ and $\varphi \mapsto \mathcal{L}\varphi$, so central and useful. The very basic properties of these two operations are that they reverse the natural partial order in their class (set-inclusion for convex bodies and point wise inequality for functions) and that both are bijections (in fact, they are not only bijections but actually involutions).

Motivated by this, Böröczky and Schneider [31] showed that, up to obvious linear changes, $K \mapsto K^\circ$ given in (2.2) is the only order-reversing isomorphism on the class of convex bodies. This result opened the floodgates to the idea that other classical operations are characterized by a small subset of their properties (for some such

papers, see, e.g., [1,2,13,37,44,69]). One of the first results we proved in this direction pertained to the functional duality which we used in the functional Santaló inequality. In [15], we proved that for the class of lower semi-continuous convex functions on $\mathbb{R}^n$, the Legendre transform (2.1) is the only order-reversing isomorphism, up to obvious linear degrees of freedom coming from a choice of coordinate system.

However, it turned out that making a small modification to the class, which was still enough for all of the geometric applications we needed, the picture changes dramatically. Indeed, in [14, 16] we considered the class of *geometric convex functions*. These are non-negative lower semi-continuous convex functions which vanish at the origin. We showed that, for this subclass, there are (up to obvious linear changes) precisely two essentially different order-reversing isomorphisms. One of these is given by the Legendre transform, as the subclass is invariant under $\mathscr{L}$, but it turned out that there is yet another essentially different order-reversing isomorphisms. We call this second order-reversing isomorphism "polarity for functions", and it can be defined by

$$\mathscr{A}\psi(y) = \sup_{x:\psi(x)>0} \left(\frac{\langle x, y\rangle - 1}{\psi(x)} \right), \tag{3.1}$$

where one should make sense of the cases where one divides by 0. This curious new player has a very geometric definition, from which the above definition can be derived. Consider the epi-graph of ψ, which is an unbounded, closed, convex set in $\mathbb{R}^{n+1}$ which lies in the upper half-space and includes the origin. The epi-graph of $\mathscr{A}\psi$ is simply the reflection (with respect to the last, special coordinate) of the set $(\mathrm{epi}(\psi))^\circ$. So, in some sense, the functional polarity $\mathscr{A}$ is more closely related to the usual polarity for convex bodies than the Legendre transform $\mathscr{L}$. There have been several interesting developments with respect to this new operation $\mathscr{A}$. In [23], we proved Mahler- and Santaló-type inequalities revolving around $\mathscr{A}$, and more was proved in [38, 46]. In [8], we found a convincing analogue of the Prékopa–Leindler inequality, which, if we write (2.3) as

$$\mathrm{Vol}(\mathscr{L}((1-\lambda)\mathscr{L}\varphi + \lambda\mathscr{L}\psi)) \geq (\mathrm{Vol}(\varphi))^{1-\lambda}(\mathrm{Vol}(\psi))^{\lambda}$$

(which can be done, since $\mathscr{L}$ maps addition to infimum convolution), then our analogue can be written as

$$\mathrm{Vol}(\mathscr{A}((1-\lambda)\mathscr{A}\varphi + \lambda\mathscr{A}\psi)) \geq ((1-\lambda)(\mathrm{Vol}(\varphi))^{-1} + \lambda(\mathrm{Vol}(\psi))^{-1})^{-1}.$$

More generally, without any convexity assumptions, we proved

$$\int_{\mathbb{R}^n} h \geq \left((1-\lambda)\left(\int_{\mathbb{R}^n} f \right)^{-1} + \lambda \left(\int_{\mathbb{R}^n} g \right)^{-1} \right)^{-1}, \tag{3.2}$$

whenever $h((1-t)x+ty) \geq \min\{f(x)^{1-t/1-\lambda}, g(y)^{t/\lambda}\}$ for all $t \in (0,1)$ and for all $x, y \in \mathbb{R}^n$. In a different direction, we studied the differential properties of $\mathcal{A}$ in [20,22]. We showed it can be used to linearize certain families of differential equations, and we discussed how it arises as a cost-induced functional transform from a certain non-standard cost on $\mathbb{R}^n$ which we call the polar cost. We discovered that the transform $\mathcal{A}$ has appeared in the book [76] from 1970, but received very little attention since then, and in fact was also present in some form in earlier works on duality of norms [66]. The "transportation of measure" point of view, which helped us think of $\mathcal{A}$ as induced by a cost function on $\mathbb{R}^n$, was an undercurrent that influenced what materialized next.

4. Optimal transport and cost-induced dualities

Given a symmetric cost function on $X \times X$ and two probability measures $\mu, \nu \in \mathcal{P}(X)$, a question going back to Monge asks whether there exists a "good" transport map, namely a function $T : X \to X$ satisfying $\mu(T^{-1}A) = \nu(A)$ for all measurable $A \subseteq X$. The relaxation due to Kantorovich [53,54] has to do with transport plans. A transport plan is a probability measure $\gamma \in \mathcal{P}(X \times X)$ with marginals μ and ν, and we denote $\gamma \in \Pi(\mu, \nu)$. A transport plan always exists (e.g., $\mu \times \nu$) and the main questions regard optimal plans, where the cost of a plan γ is given by

$$C(\gamma) = \int c(x,y) d\gamma,$$

and the total cost of transporting μ to ν is given by $C(\mu,\nu) = \inf\{C(\gamma) : \gamma \in \Pi(\mu,\nu)\}$.

For an overview of transportation theory, see [84]. For the quadratic cost $c(x,y) = |x-y|^2$ in $\mathbb{R}^n$, it is well known that transport maps exist and have a special form. More precisely [33,63].

Theorem 4.1 (Brenier–McCann). *Let μ, ν be Borel probability measures on $\mathbb{R}^n$. If μ is absolutely continuous with respect to the Lebesgue measure, then there exists a transport map between μ and ν which is given by the gradient of a convex function. Moreover, it is the optimal transport plan with respect to quadratic cost.*

Theorem 4.1 is a very important tool in AGA, and one can use it to prove the Prékopa–Leindler inequality (2.3), the Brunn–Minkowski inequality (1.1), concentration of measure inequalities and much more. It is also intimately connected with the Legendre transform; for example, one can easily check that if $\nabla\varphi$ is the optimal map from μ to ν, then $\nabla\mathcal{L}\varphi = (\nabla\varphi)^{-1}$ is the optimal from ν back to μ. This

happens, of course, since the quadratic cost in effect defines the Legendre transform (after some cosmetic manipulations).

To see more clearly how a cost function defines a functional transform, we need to recall the Kantorovich duality theorem [53, 54] which states that in the case of a lower semi-continuous cost function, the infimum in the definition of $C(\mu,\nu)$ is attained and is given by

$$C(\mu,\nu)=\sup\left\{\int_X \varphi d\mu+\int_Y \psi d\nu : (\varphi,\psi) \text{ admissible}\right\}. \tag{4.1}$$

Here (φ,ψ) is called admissible if $\varphi\in L_1(X,d\mu)$, $\psi\in L_1(X,d\nu)$, and for all x,y, $\varphi(x)+\psi(y)\le c(x,y)$. This gives rise to the functional cost transform as follows. Given a function $\varphi: X\to[-\infty,\infty]$, its c-transform $\varphi^c: X\to[-\infty,\infty]$ is given by

$$\varphi^c(y)=\inf_{x\in X}\{c(x,y)-\varphi(x)\}. \tag{4.2}$$

In other words, φ^c is the maximal admissible partner of φ. The c-transform reverses the pointwise order on functions and can be seen to be an involution on its image class. In other words, $\varphi^{ccc}=\varphi^c$, namely letting $\mathcal{S}=\{\varphi^c:\varphi: X\to\mathbb{R}\cup\{\pm\infty\}\}$, the mapping $\varphi\to\varphi^c$ is an order-reversing isomorphism on $\mathcal{S}$. For the "quadratic" cost function $c(x,y)=-\langle x,y\rangle$, we get that the class $\mathcal{S}$ consists of upper semi-continuous concave functions and $\psi^c=-(\mathcal{L}(\psi))$. (We write "quadratic" with quotation marks since formally the quadratic costs is $|x-y|^2$; however, the only interaction term is $-\langle x,y\rangle$, so that these two costs give rise to the same optimal plans and maps between measures.) The so-called "polar" cost $-\ln(\langle x,y\rangle-1)$ is similarly associated, up to a logarithmic change of variables, to the class of geometric convex functions and the $\mathcal{A}$-transform.

In dealing with the polar cost, one immediately notices that it attains the value $+\infty$. We call such costs "non-traditional", and some of the classical facts from optimal transport become more complicated for such costs. For example, even the existence of a finite-cost plan is not guaranteed. In [21], we provide necessary conditions, and slightly stronger sufficient conditions, for the existence of such a plan. For our discussion, it is only important to note the following (easily verifiable) necessary condition on μ and ν (which are probability measures, and let us assume $X=\mathbb{R}^n$). For there to exist a finite-cost map between μ and ν, it is necessary that for every subset $A\subset\mathbb{R}^n$,

$$\mu(A)+\nu(A^c)\le 1, \tag{4.3}$$

where $A^c=\{y\in\mathbb{R}^n: c(x,y)=\infty\ \forall x\in A\}$. Indeed, this condition can be rephrased as $\mu(A)\le\nu(B)$ where $B=\{y: c(x,y)<\infty \text{ for some } x\in A\}$. If a plan has finite cost, then almost every $x\in A$ has to be mapped to some y with $c(x,y)<\infty$, namely to B, so clearly a necessary condition is $\mu(A)\le\nu(B)$. We see that a non-traditional

cost function naturally defines a kind of "duality" on sets, essentially associated with its finiteness set $\{(x, y) : c(x, y) < \infty\}$.

Concentrating on subsets of $\mathbb{R}^n$, we call a map $T : P(\mathbb{R}^n) \to P(\mathbb{R}^n)$ (here use $P(X)$ to denote all subsets of X) an order-reversing quasi-involution if it reverses the partial order of inclusion, $A \subseteq B \Rightarrow TA \supseteq TB$, and when it is applied twice, we have an inclusion of the form $T(TA) \supseteq A$ for all $A \in P(\mathbb{R}^n)$. Any order-reversing quasi-involution is an involution on its image class $\{TA : A \subset \mathbb{R}^n\}$. For example, the polarity mapping $K \mapsto K^\circ$ given in (2.2) is such a mapping, and its image set consists of closed convex subsets of $\mathbb{R}^n$ which include the origin. Any symmetric subset of $S \subset \mathbb{R}^n \times \mathbb{R}^n$ (here symmetric means $(x, y) \in S \Leftrightarrow (y, x) \in S$) induces an order-reversing quasi-involution, by $A \mapsto \{y : (x, y) \in S \ \forall y \in A\}$ (as explained after (4.3)). The polarity mapping corresponds to the set $\{(x, y) : \langle x, y\rangle \leq 1\}$. The image class corresponding to S will be sets which are intersections of images of singletons $\{x\}$, namely intersections of "fibers" of S of the form $\{y : (x, y) \in S\}$. In this way, convex sets which include the origin appear as intersections of closed half-spaces which include the origin.

This point of view allows for the construction of many more set dualities, some of which seem quite natural and have not been studied in depth. Such is the "polar polarity", which one gets by switching the sign on the inequality in the set used for usual polarity, namely taking $S = \{(x, y) : \langle x, y\rangle \geq 1\}$. The image class here will consist of unbounded convex sets, which are intersections of half-spaces which do not include the origin. These sets always have a unique point closest to the origin, and, properly rotated, can be viewed as epi-graphs of certain functions, a structure the set duality respects. For some Santaló-type inequalities for this class, see [21]. In the same paper, we also showed that all quasi-dualities on $P(\mathbb{R}^n)$ are of this form, namely to each quasi-duality there corresponds a symmetric set $S \subset \mathbb{R}^n \times \mathbb{R}^n$ such that the duality is given by $A \mapsto \{y : (x, y) \in S \ \forall y \in A\}$.

5. Ball duality and sets of constant width

An interesting example of a quasi-duality on $\mathbb{R}^n$ is the one associated with the set $S = \{(x, y) : |x - y| \leq 1\}$. (One could take any norm, but the Euclidean one will already be quite interesting.) The quasi-duality is thus the mapping taking a set A to the intersection of the unit balls centered at A. The image class is of course a subset of the class of convex sets (when we might want to exclude the empty set and the whole of $\mathbb{R}^n$ which arise when these balls do not intersect, respectively, when we take an empty intersection). This subclass can be characterized in many different ways. It is the class of convex bodies with all principal curvatures (at smooth points) at least 1.

At the same time, it can be described as the class of summands of B_2^n [80, Theorem 3.2.2]. In fact, the duality mapping on this class, which we denote $K \mapsto K^c$, simply switches, up to sign, between the pair of summands; that is, K is mapped to the unique body K^c satisfying $K - K^c = B_2^n$. This class has been studied in different contexts and under several names: 1-convex bodies, hyperconvex bodies, ball bodies, spindle-convex bodies, ball summands, etc. and is linked to many classical problems, including the Kneser–Poulsen conjecture [26, 55, 73] and the reverse isoperimetric problem [30, 35]. Let us briefly discuss its connection to another open problem that has seen some recent advances.

Note that the fixed points of this duality are precisely bodies K for which $K - K = B_2^n$. These are the so-called "bodies of constant width 1", that is, bodies for which, when one considers two parallel hyperplanes supporting the body, the distance between the two is always 1, regardless of the direction of the hyperplane. Equivalently, the support function of the body, $h_K(u) = \sup_{x \in K} \langle x, u \rangle$, satisfies $h_K(u) + h_K(-u) = |u|$ (so, it is the support function of $B_2^n/2$ plus an odd function). Bodies of constant width have fascinated the scientific community for generations. One major open problem regarding bodies of constant width is called the Blaschke–Lebesgue problem. It asks, in any given dimension, which body of constant width 1 has minimal volume (the fact that the ball has maximal volume very easily follows from Urysohn's inequality, for example). In dimension $n = 2$, the answer is known and is the Reuleaux triangle (these are the original works of Blaschke and Lebesgue [27, 57]). The problem is open for $n \geq 3$, where in $\mathbb{R}^3$ the minimizers are conjectured to be the two Meissner bodies, which are of equal volume (see, e.g., [62]).

In the spirit of AGA, Schramm [81] asked whether the minimal volume of a body of constant width 1 behaves asymptotically like the maximal value, which is $\mathrm{Vol}(B_2^n/2)$, or is exponentially smaller than this value. He provided what is still the best lower bound for the volume of a body of constant width 1, which is $\left(\sqrt{3} - 1\right)^n \mathrm{Vol}(B_2^n/2)$. Very recently, Arman, Bondarenko, Nazarov, Prymak and Radchenko [3] constructed a body of constant width 1 which has indeed volume exponentially smaller than $\mathrm{Vol}(B_2^n/2)$. Nevertheless, the value of the exponent for the lower bound of the volume is still unknown, let alone exact solutions or even conjectured minimizers to the Blaschke–Lebesgue problem in dimension above 3.

It is pleasant to connect this problem to one mentioned in the introduction. Recall the Rogers–Shephard inequality (1.3), where the constant $\binom{2n}{n}$ was sharp for simplices. Since for a body of constant width 1 we have $K - K = B_2^n$, the Blaschke–Lebesgue problem is equivalent to finding the sharp constant in (1.3) for K of constant width 1. In dimension $n = 3$, the problem is also equivalent to maximizing $V(K, K, -K)$, which is called the (first) Godbersen coefficient, again for this subclass of bodies.

The study of ball duality brings some new components into the picture, and we are able to get quite a lot of insight into these problems, constructing families of bodies of constant width, understanding the boundary structure of these bodies and some of their extremal properties [4, 5, 7].

Funding. Support for this work was provided by the European Union's Horizon 2020 research and innovation programme (ERC grant agreement no. 770127), by ISF grant no. 784/20 and by the BSF (grant no. 2020329).

References

[1] S. Alesker, S. Artstein-Avidan, D. Faifman, and V. Milman, A characterization of product preserving maps with applications to a characterization of the Fourier transform. *Illinois J. Math.* **54** (2010), no. 3, 1115–1132 Zbl 1272.42004 MR 2928347

[2] S. Alesker, S. Artstein-Avidan, and V. Milman, A characterization of the Fourier transform and related topics. In A. Alexandrov, A. Baranov, and S. Kislyakov (Eds.), *Linear and complex analysis. Dedicated to V. P. Havin on the occasion of his 75th birthday*, pp. 11–26, Adv. Math. Sci. Appl. 63, American Mathematical Society, Providence, RI, 2009 Zbl 1184.42009 MR 2500506

[3] A. Arman, A. Bondarenko, F. Nazarov, A. Prymak, and D. Radchenko, Small volume bodies of constant width. *Int. Math. Res. Not. IMRN* **2025** (2025), no. 4, article no. rnaf020, 7 pp. Zbl 08103115 MR 4863831

[4] S. Artstein-Avidan, A. Chor, and D. Florentin, Isometries of the class of ball-bodies. *Colloq. Math.* **178** (2025), no. 2, 231–239 Zbl 08096800 MR 4937268

[5] S. Artstein-Avidan, A. Chor, and D. I. Florentin, A full classification of the isometries of the class of ball-bodies. *Bull. London Math. Soc.* **57** (2025), no. 12, 3691–3698 Zbl 08140774 MR 5007369

[6] S. Artstein-Avidan, K. Einhorn, D. I. Florentin, and Y. Ostrover, On Godbersen's conjecture. *Geom. Dedicata* **178** (2015), no. 1, 337–350 Zbl 1335.52005 MR 3397498

[7] S. Artstein-Avidan and D. I. Florentin, An in-depth study of ball-bodies. 2025, arXiv: 2505.09200v3

[8] S. Artstein-Avidan, D. I. Florentin, and A. Segal, Functional Brunn–Minkowski inequalities induced by polarity. *Adv. Math.* **364** (2020), article no. 107006, 19 pp. Zbl 1439.44005 MR 4062256

[9] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman, *Asymptotic geometric analysis, Part I*. Math. Surveys Monogr. 202, American Mathematical Society, Providence, RI, 2015, 451 pp. Zbl 1337.52001 MR 3331351

[10] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman, *Asymptotic geometric analysis, Part II*. Math. Surveys Monogr. 261, American Mathematical Society, Providence, RI, 2021, 645 pp. Zbl 1496.52001 MR 4390799

[11] S. Artstein-Avidan, R. Karasev, and Y. Ostrover, From symplectic measurements to the Mahler conjecture. *Duke Math. J.* **163** (2014), no. 11, 2003–2022 Zbl 1330.52004 MR 3263026

[12] S. Artstein-Avidan, B. Klartag, and V. Milman, The Santaló point of a function, and a functional form of the Santaló inequality. *Mathematika* **51** (2004), no. 1–2, 33–48 Zbl 1121.52021 MR 2220210

[13] S. Artstein-Avidan, H. König, and V. Milman, The chain rule as a functional equation. *J. Funct. Anal.* **259** (2010), no. 11, 2999–3024 Zbl 1203.39014 MR 2719283

[14] S. Artstein-Avidan and V. Milman, A new duality transform. *C. R. Math. Acad. Sci. Paris* **346** (2008), no. 21–22, 1143–1148 Zbl 1171.26007 MR 2464254

[15] S. Artstein-Avidan and V. Milman, The concept of duality in convex analysis, and the characterization of the Legendre transform. *Ann. of Math. (2)* **169** (2009), no. 2, 661–674 Zbl 1173.26008 MR 2480615

[16] S. Artstein-Avidan and V. Milman, Hidden structures in the class of convex functions and a new duality transform. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 4, 975–1004 Zbl 1221.26018 MR 2800482

[17] S. Artstein-Avidan, V. Milman, and Y. Ostrover, The M-ellipsoid, symplectic capacities and volume. *Comment. Math. Helv.* **83** (2008), no. 2, 359–369 Zbl 1138.53060 MR 2390048

[18] S. Artstein-Avidan and Y. Ostrover, A Brunn–Minkowski inequality for symplectic capacities of convex domains. *Int. Math. Res. Not. IMRN* **2008** (2008), no. 13, article no. rnn044, 31 pp. Zbl 1149.52006 MR 2436562

[19] S. Artstein-Avidan and E. Putterman, On unbalanced difference bodies and Godbersen's conjecture. *Proc. Amer. Math. Soc.* **153** (2025), no. 12, 5345–5359 Zbl 08125110 MR 4989647

[20] S. Artstein-Avidan and Y. A. Rubinstein, Differential analysis of polarity: polar Hamilton–Jacobi, conservation laws, and Monge Ampère equations. *J. Anal. Math.* **132** (2017), no. 1, 133–156 Zbl 1375.35009 MR 3666808

[21] S. Artstein-Avidan, S. Sadovsky, and K. Wyczesany, A zoo of dualities. *J. Geom. Anal.* **33** (2023), no. 8, article no. 238, 40 pp. Zbl 07695034 MR 4588159

[22] S. Artstein-Avidan, S. Sadovsky, and K. Wyczesany, Optimal measure transportation with respect to non-traditional costs. *Calc. Var. Partial Differential Equations* **62** (2023), no. 1, article no. 35, 39 pp. Zbl 1504.49061 MR 4519640

[23] S. Artstein-Avidan and B. A. Slomka, A note on Santaló inequality for the polarity transform and its reverse. *Proc. Amer. Math. Soc.* **143** (2015), no. 4, 1693–1704 Zbl 1321.52012 MR 3314082

[24] K. Ball, *Isometric problems in l_p and sections of convex sets*. Ph.D. dissertation, Trinity College, Cambridge, 1986

[25] K. Ball, Volume ratios and a reverse isoperimetric inequality. *J. Lond. Math. Soc. (2)* **44** (1991), no. 2, 351–359 Zbl 0694.46010 MR 1136445

[26] K. Bezdek, From the Kneser–Poulsen conjecture to ball-polyhedra. *European J. Combin.* **29** (2008), no. 8, 1820–1830 Zbl 1156.52007 MR 2463159

[27] W. Blaschke, Konvexe Bereiche gegebener konstanter Breite und kleinsten Inhalts. *Math. Ann.* **76** (1915), no. 4, 504–513 JFM 45.0731.04 MR 1511839

[28] W. Blaschke, *Vorlesungen über Differentialgeometrie und geometrische Grundlagen von Einsteins Relativitätstheorie II. Affine Differentialgeometrie*. Springer, Berlin, 1923, 262 pp. JFM 49.0499.01

[29] W. Blaschke, Über affine Geometrie VII. Neue Extremeigenschaften von Ellipse und Ellipsoid. In *Gesammelte Werke. Band 3*. Thales-Verlag, Essen, 1985 Zbl 0656.53003 MR 0909487

[30] A. A. Borisenko, Reverse isoperimetric inequality in two-dimensional Alexandrov spaces. *Proc. Amer. Math. Soc.* **145** (2017), no. 10, 4465–4471 Zbl 1381.53128 MR 3690629

[31] K. J. Böröczky and R. Schneider, A characterization of the duality mapping for convex bodies. *Geom. Funct. Anal.* **18** (2008), no. 3, 657–667 Zbl 1168.52002 MR 2438994

[32] J. Bourgain and V. D. Milman, New volume ratio properties for convex symmetric bodies in $\mathbf{R}^n$. *Invent. Math.* **88** (1987), no. 2, 319–340 Zbl 0617.52006 MR 0880954

[33] 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

[34] H. Brunn, Über Ovale und Eiflächen. *Monatsh. Math. Phys.* **1** (1887), 254–274 JFM 19.0615.01

[35] K. Drach and K. Tatarko, Reverse isoperimetric problems under curvature constraints. 2023, arXiv:2303.02294v1

[36] A. Dvoretzky, Some results on convex bodies and Banach spaces (in Russian). *Matematika* **8** (1964), no. 1, 73–102. English translation: *Proc. Internat. Sympos. Linear Spaces (Jerusalem, 1960)* (1961), 123–160 Zbl 0119.31803 MR 0139079

[37] D. Florentin, V. D. Milman, and R. Schneider, A characterization of the mixed discriminant. *Proc. Amer. Math. Soc.* **144** (2016), no. 5, 2197–2204 Zbl 1341.15004 MR 3460178

[38] D. I. Florentin and A. Segal, A Santaló-type inequality for the $\mathcal{J}$ transform. *Commun. Contemp. Math.* **23** (2021), no. 1, article no. 1950090, 18 pp. Zbl 1453.52009 MR 4169247

[39] M. Fradelizi, A. Hubard, M. Meyer, E. Roldán-Pensado, and A. Zvavitch, Equipartitions and Mahler volumes of symmetric convex bodies. *Amer. J. Math.* **144** (2022), no. 5, 1201–1219 Zbl 1498.52010 MR 4494180

[40] M. Fradelizi and M. Meyer, Some functional forms of Blaschke–Santaló inequality. *Math. Z.* **256** (2007), no. 2, 379–395 Zbl 1128.52007 MR 2289879

[41] M. Fradelizi and M. Meyer, Some functional inverse Santaló inequalities. *Adv. Math.* **218** (2008), no. 5, 1430–1452 Zbl 1153.52003 MR 2419928

[42] M. Fradelizi, M. Meyer, and A. Zvavitch, Volume product. In *Harmonic analysis and convexity*, pp. 163–222, Adv. Anal. Geom. 9, De Gruyter, Berlin, 2023 Zbl 1532.52001 MR 4654479

[43] R. J. Gardner, The Brunn–Minkowski inequality. *Bull. Amer. Math. Soc. (N.S.)* **39** (2002), no. 3, 355–405 Zbl 1019.26008 MR 1898210

[44] R. J. Gardner, L. Parapatits, and F. E. Schuster, A characterization of Blaschke addition. *Adv. Math.* **254** (2014), 396–418 Zbl 1291.52011 MR 3161103

[45] A. Giannopoulos, G. Paouris, and B.-H. Vritsiou, The isotropic position and the reverse Santaló inequality. *Israel J. Math.* **203** (2014), no. 1, 1–22 Zbl 1307.52003 MR 3273430

[46] S. Gilboa, A. Segal, and B. A. Slomka, The scaled polarity transform and related inequalities. *J. Fixed Point Theory Appl.* **27** (2025), no. 3, article no. 73, 22 pp. Zbl 08084671 MR 4938963

[47] C. Godbersen, *Der Satz vom Vektorbereich in Räumen beliebiger Dimensionen*. Ph.D. thesis, Göttingen, 1938 Zbl 0020.07702

[48] Y. Gordon, M. Meyer, and S. Reisner, Zonoids with minimal volume-product—a new proof. *Proc. Amer. Math. Soc.* **104** (1988), no. 1, 273–276 Zbl 0663.52003 MR 0958082

[49] P. Haim-Kislev and Y. Ostrover, A counterexample to Viterbo's conjecture. [v1] 2024, [v3] 2025, arXiv:2405.16513v3

[50] A. Hajnal and E. Makai, Research problems. *Period. Math. Hungar.* **5** (1974), no. 4, 353–354 MR 1553587

[51] H. Hofer and E. Zehnder, *Symplectic invariants and Hamiltonian dynamics*. Birkhäuser Adv. Texts Basler Lehrbücher, Birkhäuser, Basel, 1994, 341 pp. Zbl 0805.58003 MR 1306732

[52] H. Iriyeh and M. Shibata, Symmetric Mahler's conjecture for the volume product in the 3-dimensional case. *Duke Math. J.* **169** (2020), no. 6, 1077–1134 Zbl 1439.52007 MR 4085078

[53] L. Kantorovitch, On the transfer of masses. *Dokl. Akad. Nauk* **37** (1942), 227–229 Zbl 0061.09705

[54] L. V. Kantorovich, On a problem of Monge [reprint from *Uspekhi Mat. Nauk* **3** (1948), no. 2, 225–226]. *Zap. Nauchn. Semin. POMI* **312** (2004), 15–16. English translation: *J. Math. Sci.* **133** (2006), no. 4, 1383 Zbl 1080.49508

[55] M. Kneser, Einige Bemerkungen über das Minkowskische Flächenmaß. *Arch. Math. (Basel)* **6** (1955), no. 5, 382–390 Zbl 0065.04001 MR 0073679

[56] G. Kuperberg, From the Mahler conjecture to Gauss linking integrals. *Geom. Funct. Anal.* **18** (2008), no. 3, 870–892 Zbl 1169.52004 MR 2438998

[57] H. Lebesgue, Sur le problème des isopérimètres et sur les domaines de largeur constante. *Bull. Soc. Math. France* **42** (1914), 72–76

[58] M. Ledoux, *The concentration of measure phenomenon*. Math. Surveys Monogr. 89, American Mathematical Society, Providence, RI, 2001, 181 pp. Zbl 0995.60002 MR 1849347

[59] J. Lehec, A direct proof of the functional Santaló inequality. *C. R. Math. Acad. Sci. Paris* **347** (2009), no. 1–2, 55–58 Zbl 1159.52010 MR 2536749

[60] L. Leindler, On a certain converse of Hölder's inequality. II. *Acta Sci. Math. (Szeged)* **33** (1972), no. 3–4, 217–223 Zbl 0245.26011 MR 2199372

[61] K. Mahler, Ein Minimalproblem für konvexe Polygone. *Mathematica* **7** (1939), 118–127 Zbl 0020.05002

[62] H. Martini, L. Montejano, and D. Oliveros, *Bodies of constant width. An introduction to convex geometry with applications*. Birkhäuser, Cham, 2019, 486 pp. Zbl 1468.52001 MR 3930585

[63] 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

[64] M. Meyer and A. Pajor, On the Blaschke–Santaló inequality. *Arch. Math. (Basel)* **55** (1990), no. 1, 82–93 Zbl 0718.52011 MR 1059519

[65] M. Meyer and S. Reisner, Shadow systems and volumes of polar convex bodies. *Mathematika* **53** (2006), no. 1, 129–148 Zbl 1118.52015 MR 2304056

[66] V. D. Milman, A certain transformation of convex functions and a duality of the β and δ characteristics of a B-space (in Russian). *Dokl. Akad. Nauk SSSR* **187** (1969), no. 1, 33–35. English translation: *Sov. Math.* **10** (1969), 789–792 Zbl 0188.43503 MR 0256138

[67] V. D. Milman, Inégalité de Brunn–Minkowski inverse et applications à la théorie locale des espaces normés. *C. R. Acad. Sci. Paris Sér. I Math.* **302** (1986), no. 1, 25–28 Zbl 0604.52003 MR 0827101

[68] V. D. Milman, The concentration phenomenon and linear structure of finite-dimensional normed spaces. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986)*, pp. 961–975, American Mathematical Society, Providence, RI, 1987 Zbl 0698.46009 MR 0934298

[69] V. D. Milman and R. Schneider, Characterizing the mixed volume. *Adv. Geom.* **11** (2011), no. 4, 669–689 Zbl 1242.52011 MR 2852926

[70] H. Minkowski, *Geometrie der Zahlen*. Teubner, Leipzig and Berlin, 1910 JFM 41.0239.03

[71] S. Nakamura and H. Tsuji, The functional volume product under heat flow. *J. Eur. Math. Soc. (JEMS)* (2025), DOI 10.4171/JEMS/1720

[72] F. Nazarov, The Hörmander proof of the Bourgain–Milman theorem. In *Geometric aspects of functional analysis*, pp. 335–343, Lecture Notes in Math. 2050, Springer, Heidelberg, 2012 Zbl 1291.52014 MR 2985302

[73] E. T. Poulsen, Problem 10. *Math. Scand.* **2** (1954), 345–347

[74] A. Prékopa, Logarithmic concave measures with application to stochastic programming. *Acta Sci. Math. (Szeged)* **32** (1971), 301–316 Zbl 0235.90044 MR 0315079

[75] S. Reisner, Zonoids with minimal volume-product. *Math. Z.* **192** (1986), no. 3, 339–346 Zbl 0578.52005 MR 0845207

[76] R. T. Rockafellar, *Convex analysis*. Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970, 451 pp. Zbl 0193.18401 MR 0274683

[77] C. A. Rogers and G. C. Shephard, The difference body of a convex body. *Arch. Math. (Basel)* **8** (1957), no. 3, 220–233 Zbl 0082.15703 MR 0092172

[78] J. Saint-Raymond, Sur le volume des corps convexes symétriques. In *Initiation Seminar on Analysis: G. Choquet-M. Rogalski-J. Saint-Raymond, 20th Year: 1980/1981*, Exp. No. 11, Publ. Math. Univ. Pierre et Marie Curie 46, Univ. Paris VI, Paris, 1981, 25 pp. Zbl 0531.52006 MR 0670798

[79] L. A. Santaló, An affine invariant for convex bodies of n-dimensional space. *Port. Math.* **8** (1949), no. 4, 155–161 Zbl 0038.35702 MR 0039293

[80] R. Schneider, *Convex bodies: the Brunn–Minkowski theory*. 2nd expanded edn., Encyclopedia Math. Appl. 151, Cambridge University Press, Cambridge, 2014, 736 pp. Zbl 1287.52001 MR 3155183

[81] O. Schramm, On the volume of sets having constant width. *Israel J. Math.* **63** (1988), no. 2, 178–182 Zbl 0657.52003 MR 0968537

[82] J. Steiner, Über parallele Flächen. *J. Reine Angew. Math.* **19** (1840), 1–20

[83] R. Vershynin, *High-dimensional probability. An introduction with applications in data science*. Camb. Ser. Stat. Probab. Math. 47, Cambridge University Press, Cambridge, 2018, 284 pp. Zbl 1430.60005 MR 3837109

[84] C. Villani, *Optimal transport. Old and new*. Grundlehren Math. Wiss. 338, Springer, Berlin, 2009, 973 pp. Zbl 1156.53003 MR 2459454

[85] C. Viterbo, Metric and isoperimetric problems in symplectic geometry. *J. Amer. Math. Soc.* **13** (2000), no. 2, 411–431 Zbl 0964.53050 MR 1750956

Shiri Artstein-Avidan
Department of Mathematics, Tel Aviv University, Haim Lebanon, 69978 Tel Aviv, Israel;
shiri@tauex.tau.ac.il

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/10

The continuum problem: Hilbert's first problem revisited

David Asperó and Ralf Schindler

Abstract. The continuum problem has recently seen a revival, brought about by recent advances involving strong forcing axioms and other natural principles extending the standard axiomatization for set theory. In this article, we survey these developments. This is a slightly edited version of the slides we used in the talk that we gave at the 9 ECM.

1. Cantor's paradise

Set theory is the study of sets (of mathematical objects) under the membership relation $\in$. It provides a particularly simple language to model (all?) mathematical notions.

Example. Following Kuratowski, we can make sense of the notion of an ordered pair using the language of set theory by defining

$$(a, b) = \{\{a\}, \{a, b\}\}.$$

It is then easy to check that for all a, a', b, b',

$$(a, b) = (a', b') \iff a = a' \quad \text{and} \quad b = b'.$$

Once we have ordered pairs, we can define relations as sets of ordered pairs, functions as relations f such that $(a, b), (a, b') \in f \Rightarrow b = b'$, and so on.

We can define $0 = \emptyset$, $1 = \{0\}$, $2 = \{0, 1\}$, $3 = \{0, 1, 2\}$, and so on, and we can also define the general notion of natural number. (We can define a natural number as an ordinal which is either the empty set or some successor ordinal, and such that each of its members is either the empty set or a successor ordinal. Both the notions of ordinal and of successor ordinal are purely set theoretic.)

Mathematics Subject Classification 2020: 03E50 (primary); 03E57, 03E55, 03E60 (secondary).
Keywords: the continuum problem, forcing axioms, axiom (*), natural axioms.

We can then construct $\mathbb{Z}$ from $\mathbb{N}$ in the standard way, as well as $\mathbb{Q}$ from $\mathbb{Z}$, and $\mathbb{R}$ from $\mathbb{Q}$ (as Dedekind cuts of rationals or classes of Cauchy sequences of rationals).

And so on.

Given sets X and Y, we write

- $|X| \leq |Y|$ if there is an injective function $f : X \to Y$,
- $|X| = |Y|$ if $|X| \leq |Y|$ and $|Y| \leq |X|$,[1] and
- $|X| < |Y|$ if $|X| \leq |Y|$ but $|Y| \not\leq |X|$.

In 1873, Georg Cantor famously proved that although $\mathbb{N}$ and $\mathbb{R}$ are both infinite, $|\mathbb{N}| < |\mathbb{R}|$. Here is a more general version of this theorem.

Theorem 1.1 (Cantor's theorem). *For every set X, $|X| < |\mathcal{P}(X)|$, where $\mathcal{P}(X) = \{Y : Y \subseteq X\}$.*

Proof. Clearly $|X| \leq |\mathcal{P}(X)|$ (take the function sending $y \in X$ to $\{y\} \in \mathcal{P}(X)$).

To see that $|\mathcal{P}(X)| \not\leq |X|$, suppose that there were an injection from $\mathcal{P}(X)$ into X. There would then be a surjection $f : X \to \mathcal{P}(X)$. Let

$$A = \{y \in X : y \notin f(y)\}.$$

If $A = f(y)$, then

$$y \in f(y) \iff y \notin f(y),$$

which is absurd. Hence $A \notin \text{range}(f)$ and f cannot be a surjection after all. ∎

It is not difficult to see that $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})|$. By Cantor's theorem, we then have that

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| = |\mathbb{R}| < |\mathcal{P}(\mathcal{P}(\mathbb{N}))| < |\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))| < \cdots.$$

Thus, we have a hierarchy of bigger and bigger infinities.

Ordinals form a natural extension of the natural number sequence. They look like this: $0, 1, 2, \ldots \quad \omega = \aleph_0 = \mathbb{N}, \omega + 1, \ldots \qquad \omega_1 = \aleph_1, \quad \ldots$

Here, ω, also known as $\aleph_0$, is the first infinite *cardinal*.

Also, $\aleph_1$ is the first ordinal above $\aleph_0$ which is not bijective with $\aleph_0$; it is the second infinite cardinal. Similarly, $\aleph_2$ is the first ordinal above $\aleph_1$ which is not bijective with $\aleph_1$; it is the third infinite cardinal.

All these cardinals can be proved to exist. In fact, we can prove that for every ordinal α, the α-th infinite cardinal exists. This cardinal is denoted by $\aleph_\alpha$.

[1] By a theorem of Cantor–Schroder–Bernstein, this is equivalent to having a bijection from X onto Y.

Assuming the axiom of choice (which was soon to be regarded as one of the core truths about sets and is an indispensable tool in mathematics), every set is either finite or in bijection with some $\aleph_\alpha$. In this case, the unique $\aleph_\alpha$ such that $|X| = \aleph_\alpha$ is the *cardinality of* X.

The cardinality $|\mathbb{R}| = |\mathcal{P}(\mathbb{N})| = |\mathcal{P}(\aleph_0)|$ is also known as $2^{\aleph_0}$. The following is now a natural question.

The continuum problem. $2^{\aleph_0} = \aleph_1$? $2^{\aleph_0} = \aleph_2$? In general, which is the exact cardinality of $\mathbb{R}$? In other words: Exactly how many reals are there?

By Cantor's theorem, $\aleph_0 < 2^{\aleph_0}$, and Cantor conjectured that $2^{\aleph_0}$ is the least possible value compatible with this inequality.

Cantor's continuum hypothesis. CH is the statement $2^{\aleph_0} = \aleph_1$.

Deciding the truth value of the continuum hypothesis, and in general solving the continuum problem, was ♯1 on Hilbert's famous list of open problems for the ICM meeting in 1900; see [7].

A more fundamental question is the following.

Question 1.2. What counts as a solution to the continuum problem?

Early work in set theory actually gave rise to contradictions (!).

(Russell's paradox): Let $\mathcal{R}$ be the collection of all sets x such that $x \notin x$. If $\mathcal{R}$ is a set, then

$$\mathcal{R} \in \mathcal{R} \iff \mathcal{R} \notin \mathcal{R}.$$

Hence, working under the assumption that every well-defined collection of objects forms a set leads to a contradiction. That is, the early naive set theory was inconsistent.

On the other hand, the intuitions behind set theory looked *natural*, and they seemed to have the potential to serve as a foundation for mathematics. A natural next move was to retreat to a more modest, hopefully *consistent*, but still powerful, "true" theory of sets.

The standard way to make sense of this is to work in the framework of first order logic. That is, we want to put together an $\mathcal{L}_\in$-theory T for set theory which is

(1) hopefully consistent,

(2) powerful, and

(3) hopefully "true".

The language $\mathcal{L}_\in$ of set theory is the first-order language with a symbol for $\in$. And T is meant to be a list (maybe infinite) of truths about sets.

Some examples of such truths are the following:

- The axiom of extensionality: $(\forall x)(\forall y)(x = y \leftrightarrow (\forall z)(z \in x \leftrightarrow z \in y))$
  (Two sets are equal if and only if they have the same elements.)
- The power set axiom: $(\forall x)(\exists y)(\forall z)(z \in y \leftrightarrow (\forall w)(w \in z \rightarrow w \in x))$
  (If X is a set, then the power set of X, $\mathcal{P}(X)$, is also a set.)

The standard theory T is known as ZFC (Zermelo–Fraenkel set theory with the axiom of choice).

ZFC proves that the (set-theoretic) universe, denoted by V, can be naturally constructed as

$$V = \bigcup_{\alpha \in \mathrm{Ord}} V_\alpha.$$

Here, $(V_\alpha : \alpha \in \mathrm{Ord})$ is defined by:

(1) $V_0 = \emptyset$.

(2) $V_{\alpha+1} = \mathcal{P}(V_\alpha)$.

(3) If α is a limit ordinal, then $V_\alpha = \bigcup_{\beta<\alpha} V_\beta = \{x : x \in V_\beta \text{ for some } \beta < \alpha\}$.

The fact that $V = \bigcup_{\alpha \in \mathrm{Ord}} V_\alpha$ tells us that sets are orderly generated in stages indexed by the ordinals. Sets appearing at a stage α are formed of sets having appeared at stages before α. This is a nice incarnation of Cantor's intuition that sets are "collections of already existing objects".

The hope now was to show that ZFC proves that $2^{\aleph_0} = \aleph_1$ or to show that ZFC proves that $2^{\aleph_0} > \aleph_1$ or in fact to show that ZFC decides the exact value of $2^{\aleph_0}$. This is of course what would have constituted a solution of the continuum problem.

"Would have", because:

- In 1938, Kurt Gödel proved that if ZFC is consistent, then ZFC + $2^{\aleph_0} = \aleph_1$ is also consistent [5].
- In 1963, Paul Cohen proved that if ZFC is consistent, then ZFC + $2^{\aleph_0} = \aleph_2$ is also consistent [2].

To prove the consistency of ZFC + $2^{\aleph_0} = \aleph_1$ (assuming the consistency of ZFC), Gödel defined a minimal subuniverse L of V, called *the constructible universe*, and showed that L satisfies ZFC + $2^{\aleph_0} = \aleph_1$.

We have $L = \bigcup_{\alpha \in \mathrm{Ord}} L_\alpha$, where

(1) $L_0 = \emptyset$.

(2) For every ordinal α, $L_{\alpha+1} = \mathrm{Def}(L_\alpha)$, where $\mathrm{Def}(L_\alpha)$ is the set of all subsets of L_α which are definable over L_α by some formula (possibly with parameters).

(3) If α is a limit ordinal, then $L_\alpha = \bigcup_{\beta<\alpha} L_\beta$.

To prove the consistency of ZFC + $2^{\aleph_0} = \aleph_2$, Cohen devised the method of *forcing*. This is a very general method which, given a model M of (enough axioms of) ZFC, enables us to build an outer universe $M[G] \supseteq M$ of M satisfying (enough axioms of) ZFC. The object $M[G]$ is called a *generic extension of* M.

If we choose the *generic object* G carefully, we may be able to arrange that $M[G]$ satisfies some interesting statement, like $2^{\aleph_0} = \aleph_2$. Cohen won the Fields Medal for this.

Using the method of forcing, one can show, given the consistency of ZFC, that each of the following is consistent:

- ZFC + $2^{\aleph_0} = \aleph_2$
- ZFC + $2^{\aleph_0} = \aleph_3$
- ZFC + $2^{\aleph_0} = \aleph_{273453453667889}$
- ZFC + $2^{\aleph_0} = \aleph_{\omega+1}$
- ZFC + $2^{\aleph_0} = \aleph_{\omega_1}$[2]
- ...

Is this the end of the story? Do these consistency results show that the continuum problem is in fact a pseudo-problem? A formalist, who takes our official theory ZFC as the only source of set-theoretic "truth", would in fact answer Yes.

But this is not the only reasonable position. In fact, the formalist position faces very serious problems. We will now present a way to see this.

2. Trouble with the formalist position

Taking a question like "How many real numbers are there?" to have no answer commits the formalist to denying there is any such thing as "the set-theoretic universe $\bigcup_{\alpha \in \mathrm{Ord}} V_\alpha$".

For the formalist, all there is to say are things like

$$\text{"}T \text{ proves } \varphi\text{"}$$

or

$$\text{"If } T \text{ is consistent, then it does not prove } \varphi\text{"},$$

for T = ZFC or T being some given extension of ZFC, and for some sentence φ.

[2] On the other hand, ZFC + $2^{\aleph_0} = \aleph_\omega$ is inconsistent.

These proof-theoretic statements are ultimately just arithmetical statements (via standard coding of the relevant syntactical notions into arithmetic). So the formalist seems to be at least committed to the existence of $V_\omega = \bigcup_{n\in\omega} V_n$.

But arithmetic is in fact dependent on the higher V_αs. For example, whether or not a simple (and rather natural) purely arithmetical combinatorial statement known as the Paris–Harrington theorem holds depends on whether or not there are infinite sets in a way which makes it hard to justify this statement and extensions thereof without making reference to objects which are beyond V_ω.

So $V_{\omega+1}$ seems indispensable.

But the properties of $V_{\omega+1}$ depend crucially on what happens at V_α for higher αs. For example, Borel determinacy holding depends on the existence of V_α for all $\alpha < \omega_1$ [4].

So the universe should in fact contain V_α for all $\alpha < \omega_1$. But the statement "There is no $X \subseteq \mathbb{R}$ with $\aleph_0 < |X| < |\mathbb{R}|$" (which is equivalent to CH in ZFC) is decided already in $V_{\omega+2}$, so it should be decided!

Where should we stop and why?

3. The realist position: Natural axioms

If the set-theoretic universe V is real, then the continuum problem has a definite solution. The fact that ZFC does not solve the continuum problem means that we need to supplement ZFC with *natural* axioms solving this problem.

The search for natural axioms supplementing ZFC is also known as *Gödel's programme*. Gödel indeed made the above point [6]. Although he had proved the consistency of CH, he suspected the CH to be false *in reality* and was hoping that natural axioms would eventually settle the issue.

Adopting $2^{\aleph_0} = \aleph_1$, or $2^{\aleph_0} = \aleph_{27}$, as a new axiom seems surely ad hoc, therefore unnatural. How do we decide if an axiom is natural?

We need general criteria to assess the naturalness of axioms. They should address the question: What virtues do we want our theory of sets to have?

3.1. Strong axioms of infinity

Large cardinal (LC) axioms are an (open-ended) hierarchy of axioms asserting the existence of very high cardinals with strong properties. They realize the idea that "the universe is large".

They form a hierarchy of stronger and stronger theories: Given an LC axiom A, ZFC + A proves the consistency of ZFC, and in fact of ZFC + A' for any weaker large cardinal axiom A'.

The stronger A is, the more daring ZFC + A is (i.e., more likely to be inconsistent). We can prove the consistency of LC axioms only by working in a strictly stronger LC theory.

Arguably, the weakest large cardinal axiom is "ω exists". The next natural large cardinal axiom is "There is an inaccessible cardinal". The following are some classical large cardinal axioms.

$$\begin{aligned}\omega \;&<\; \text{inaccessible} \;<\; \text{weakly compact} \;<\; \text{measurable}\\ &<\; \text{strong} \;<\; \text{Woodin} \;<\; \text{superstrong} \;<\; \text{supercompact}\\ &<\; \text{huge} \;<\; \text{2-huge} \;<\; \cdots .\end{aligned}$$

An empirical fact is that every natural mathematical theory ever considered can be interpreted relative to ZFC + A for some LC axiom A. In fact, for theories T naturally occurring in mathematics, we can often show that T is *equiconsistent* with ZFC + A for some LC axiom A (i.e., we can prove that T is consistent if and only if ZFC + A is consistent).

A natural axiom should therefore be compatible with all consistent large cardinal axioms.

3.2. Invariance with respect to forcing

Forcing is our prime method for proving the independence of some given statement φ from some base set theory; in other words, to prove that if T is consistent, then φ and its negation $\neg\varphi$ are both consistent with T. Therefore, if we want our set theory T to be strong, we better have that T neutralizes the effects of forcing as much as possible, in the sense of proving, for as many sentences φ as possible, that the truth value of φ cannot be changed by forcing.

The axiom $V = L$ neutralizes the effects of forcing for silly reasons: Nontrivial generic extensions do not satisfy $V = L$.

However, $V = L$ is *not* compatible with LC axioms. In fact, if $V = L$, then there are no measurable cardinals.

Quite remarkably, large cardinals in the region of Woodin cardinals also neutralize the effects of forcing to a suitable extent (and are of course compatible with all LC axioms).

Theorem 3.1 (Woodin, mid-1980s). *Suppose there are arbitrarily large Woodin cardinals. Then the following are equivalent for every sentence φ in the language of set theory:*

(1) *φ is true in $L(\mathbb{R})$.*

(2) *It can be forced that φ is true in $L(\mathbb{R})$.*

The model $L(\mathbb{R})$ is the $\subseteq$-minimal subuniverse of ZF (=ZFC without the axiom of choice) containing all the reals and all the ordinals. The model $L(\mathbb{R})$ is where all of classical analysis takes place.

Hence, if there are arbitrarily large Woodin cardinals, classical analysis is immune to the forcing method.

3.3. Maximality with respect to forcing: Forcing axioms

The structure $\mathbb{P} = (P, \leq_{\mathbb{P}})$ is a *partial order* if P is a set and $\leq_{\mathbb{P}}$ is a relation on P which is transitive, anti-symmetric, and reflexive on P.

Given a partial order $\mathbb{P} = (P, \leq_{\mathbb{P}})$ (a.k.a. *forcing notion*), $D \subseteq P$ is a *dense* subset of $\mathbb{P}$ if for every $p \in P$ there is some $q \in D$ such that $q \leq_{\mathbb{P}} p$.

A set $G \subseteq P$ is a *filter* if

- for every $q \in G$ and $p \in P$, if $q \leq_{\mathbb{P}} p$, then $p \in G$;
- for all $q_1, q_2 \in G$, there is some $q \in G$ such that $q \leq_{\mathbb{P}} q_1$ and $q \leq_{\mathbb{P}} q_2$.

If M is some model such that $\mathbb{P} \in M$, we say that G is $\mathbb{P}$*-generic over* M if $G \cap D \neq \emptyset$ for every dense subset D of $\mathbb{P}$ such that $D \in M$.

Theorem 3.2 (Cohen). *Suppose M is a transitive model of (enough of)* ZFC. *Let $\mathbb{P} = (P, \leq_{\mathbb{P}}) \in M$ be a forcing notion and let $G \subseteq P$ be $\mathbb{P}$-generic over M. Then:*

(1) *If $\mathbb{P}$ is* non-atomic*, meaning that for every $p \in \mathbb{P}$ there are $q_0 \leq_{\mathbb{P}} p$ and $q_1 \leq_{\mathbb{P}} p$ such that q_0 and q_1 are $\mathbb{P}$-incompatible (i.e., there is no $q \in \mathbb{P}$ with $q \leq_{\mathbb{P}} q_0$ and $q \leq_{\mathbb{P}} q_1$), then $G \notin M$.*

(2) *There is a $\subseteq$-minimal model $M[G]$ of (enough of)* ZFC *such that*

(a) *$M \subseteq M[G]$ and*

(b) *$G \in M[G]$.*

Example. Let $\mathbb{P}$ be the set of finite tuples $(x_0, x_1, \ldots, x_{n-1})$ of 0s and 1s. Given such tuples $\vec{\sigma}_1, \vec{\sigma}_2$, let us set $\vec{\sigma}_2 \leq_{\mathbb{P}} \vec{\sigma}_1$ if and only if $\vec{\sigma}_1$ in an initial segment of $\vec{\sigma}_2$.

Let G be a filter of $\mathbb{P}$. If G meets $D_n = \{\vec{\sigma} \in \mathbb{P} : \text{length}(\vec{\sigma}) > n\}$ for every $n \in \mathbb{N}$, then $c = \bigcup G$ is an infinite sequence $(x_n : n \geq 0)$ of 0s and 1s.

If G meets $\{\vec{\sigma} \in \mathbb{P} : \text{there are } 2n, 2n+2 < \text{length}(\vec{\sigma}), \vec{\sigma}(2n) \neq \vec{\sigma}(2n+2)\}$, then it is not the case that all consecutive even entries of c are equal.

If G meets

$$\{\vec{\sigma} \in \mathbb{P} : \text{there are } 2n+1, 2n+3 < \text{length}(\vec{\sigma}), \vec{\sigma}(2n+1) \neq \vec{\sigma}(2n+3)\},$$

then it is not the case that all consecutive odd entries of c are equal.

If G meets $\{\vec{\sigma} \in \mathbb{P} : \text{there is } n \text{ with } n + 10000 < \text{length}(\vec{\sigma}) \text{ such that } \vec{\sigma}(n) = \vec{\sigma}(n+1) = \cdots = \vec{\sigma}(n + 10000) = 0\}$, then c has somewhere 10001 consecutive entries all taking value 0.

...

If G is sufficiently generic, then c is a very chaotic sequence. In fact, if G is generic over some model M, then c avoids every regularity pattern expressible within M.

The forcing $\mathbb{P}$ in the example above is the simplest possible non-atomic forcing. It is called *Cohen forcing*. And if G is $\mathbb{P}$-generic over a model M, c is a *Cohen real over* M.

By (1) in Cohen's theorem, there are no $\mathbb{P}$-generic filters over V (whenever $\mathbb{P}$ is non-atomic).

On the other hand, we have the following.

Fact 3.3 (ZFC). *If* $|M| = \aleph_0$*, then for every forcing notion* $\mathbb{P} \in M$ *there is a* $\mathbb{P}$*-generic filter* G *over* M.

Proof. Let $(D_n : n \in \mathbb{N})$ enumerate all dense subsets of $\mathbb{P}$ in M. Let $q_0 \in D_0$. Since D_1 is dense, we may find $q_1 \in D_1$ such that $q_1 \leq_{\mathbb{P}} q_0$. In general, since D_{n+1} is dense, we may find $q_{n+1} \in D_{n+1}$ such that $q_{n+1} \leq_{\mathbb{P}} q_n$. Then

$$G = \{p \in P : q_n \leq_{\mathbb{P}} p \text{ for some } n\}$$

is a $\mathbb{P}$-generic filter over M. ∎

We can now naturally ask: Is it possible to find strengthening of this fact applying to larger collections of dense sets?

It turns out that it is consistent with ZFC to have the answer to be Yes in some cases; for example, if we restrict the class of forcings in some suitable way and if, for example, "larger" is interpreted as "of size $\aleph_1$". These assertions are known as *forcing axioms*. Their consistency was traditionally shown by forcing "many times" with forcing notions in the relevant class.

Let us now introduce a standard piece of notation for forcing axioms for meeting $\aleph_1$-many dense sets.

Definition 3.4. Given a class Γ of forcing notions, FA(Γ) is the following assertion.

Suppose $\mathbb{P}$ is a forcing notion in Γ. If $\mathcal{D}$ is a collection of $\aleph_1$-many dense subsets of $\mathbb{P}$, then there is a filter $G \subseteq \mathbb{P}$ such that $G \cap D \neq \emptyset$ for all $D \in \mathcal{D}$.

The first forcing axiom ever considered (1970, Solovay–Tennenbaum, Martin [8,9]) was Martin's axiom at ω_1, MA_{ω_1}, which is

$$\text{FA}(\{\mathbb{P} : \mathbb{P} \text{ has the countable chain condition}\}).$$

Here, $\mathbb{P}$ has the countable chain condition if and only if every subset of $\mathbb{P}$ consisting of pairwise incompatible conditions is at most countable.

The strongest possible forcing axiom of this sort (i.e., applying to the widest possible class of forcing notions), known as Martin's maximum (MM), was isolated and proved consistent by Foreman–Magidor–Shelah in 1984 assuming the consistency of ZFC+ "There is a supercompact cardinal" [3]. MM is

$$\mathrm{FA}(\{\mathbb{P} : \text{forcing with } \mathbb{P} \text{ preserves all stationary subsets of } \omega_1\}).$$

Here, a subset S of ω_1 is stationary if and only if $S \cap C \neq \emptyset$ for every $C \subset \omega_1$ which is unbounded in ω_1 and closed below ω_1 in the order topology on the ordinals.

The existence of "partially generic" filters given by forcing axioms ensures that many of the facts that would hold in the corresponding generic extensions actually hold in V. Thus, forcing axioms realize the following "maximality idea":

Any statement (of the right syntactical form) that could possibly hold (by forcing with a forcing notion in the relevant class) is actually true.

Also, the wider the class Γ is, the more "democratic" $\mathrm{FA}(\Gamma)$ is (in the sense of not discriminating between possible generic extensions).

Given a set X, $\mathrm{TC}(X) = X \cup \bigcup X \cup \bigcup\bigcup X \cup \cdots = X \cup \{a : a \in b \in X \text{ for some } b\} \cup \{a : a \in b \in c \in X \text{ for some } b, c\} \cup \cdots$.

Let

$$H(\omega_2) = \{X : |\mathrm{TC}(X)| < \aleph_2\}$$

Thus, $H(\omega_2)$ is the collection of all sets which are "small relative to $\aleph_2$".

The object $H(\omega_2)$ is a set which is a model of a convenient fragment of ZFC.

Many natural statements, like CH and $\neg$ CH, are decided in $H(\omega_2)$.

Forcing axioms have many consequences at the level of $H(\omega_2)$. In particular, MM implies all of the following.

(1) (Suslin's hypothesis) If $(L, \leq)$ is a Dedekind-complete dense linear order such that $|\mathcal{I}| \leq \aleph_0$ whenever $\mathcal{I}$ is a collection of pairwise disjoint intervals of L, then $(L, \leq)$ is order-isomorphic to the real line with the usual order.

(2) $2^{\aleph_0} = 2^{\aleph_1} = \aleph_2$.

(3) (Moore) There is a set of reals X and a Countryman line[3] C such that whenever $(L, \leq)$ is a linear order such that $\aleph_0 < |L|$, L contains some suborder order-isomorphic to one of the following:

[3] A linear order $(C, \leq)$ is a Countryman line if its square $C \times C$ is a union of countably many chains. In other words, there is a decomposition $C \times C = \bigcup_{n<\omega} X_n$ of the Cartesian product $C \times C$ into countably many pieces X_n such that for each n and for all $(x_0, y_0), (x_1, y_1) \in X_n$, $x_0 \leq x_1$ if and only if $y_0 \leq y_1$.

- X;
- ω_1;
- ω_1^* (the reverse of ω_1);
- C;
- C^* (the reverse of C).

The consistency proof of MM shows in fact that the following enhanced form of this axiom is consistent. Its formulation makes use of the notion of $\mathcal{P}$-names[4] for objects in the forcing extension as well as of the forcing relation $\Vdash_{\mathcal{P}}$.[5]

MM^{++}: For every forcing notion $\mathcal{P}$ preserving stationary subsets of ω_1, every collection $\mathcal{D}$ of $\aleph_1$-many dense subsets of $\mathcal{P}$, and every collection $\{\dot{S}_i : i < \omega_1\}$ of $\mathcal{P}$-names for stationary subsets of ω_1, there is a filter $G \subseteq \mathcal{P}$ such that

- $G \cap D \neq \emptyset$ for all $D \in \mathcal{D}$ and
- for each i, $(\dot{S}_i)_G = \{\nu < \omega_1 : (\exists p \in G)\, p \Vdash_{\mathcal{P}} \nu \in \dot{S}_i\}$ is stationary.

MM^{++} in fact seemed to decide *all* questions about $H(\omega_2)$ modulo forcing.

3.4. The axiom of determinacy

Let A be a set of sequences $(x_i : i \geq 0)$ of natural numbers. Consider the following game G_A between two players, I and II, who alternate picking natural numbers x_i:

I	x_0		x_2		...	
II		x_1		x_3		...

Thus, player I starts the run of the game by picking some $x_0 \in \mathbb{N}$. Then player II responds by picking some $x_1 \in \mathbb{N}$. Then player I moves some $x_2 \in \mathbb{N}$. And so on.

- Player I wins this run of the game if $(x_i : i \geq 0) \in A$.
- Player II wins if $(x_i : i \geq 0) \notin A$.

We say that A *is determined* if either player I or player II has a winning strategy in G_A (i.e., a strategy ensuring a win for that player no matter how the other player makes their moves).

[4] Any object in a generic extension $V[G]$ of V by a forcing $\mathcal{P} \in V$ is the interpretation, by the generic filter G, of some $\mathcal{P}$-name $\dot{x} \in V$.

[5] The forcing relation $\Vdash_{\mathcal{P}}$ is a relation, definable in V from $\mathcal{P}$, which captures forcibility in the following sense: for every $p \in \mathcal{P}$, every formula $\varphi(x_0, \ldots, x_n)$, and any $\mathcal{P}$-names $\dot{a}_0, \ldots, a_n$ in V, we have $p \Vdash_{\mathcal{P}} \varphi(\dot{a}_0, \ldots, \dot{a}_n)$ if and only if $V[G] \vDash \varphi((\dot{a}_0)_G, \ldots, (\dot{a}_n)_G)$ for every $\mathcal{P}$-generic filter G over V such that $p \in G$, where for each i, $(\dot{a}_i)_G$ is the interpretation of $\dot{a}_i$ by G.

The *axiom of determinacy* (AD) is the statement: "Every set A of sequences of natural numbers is determined".

AD contradicts the axiom of choice.

On the other hand, ZF + AD gives a remarkably rich theory. In particular, it provides a very fine analysis of $L(\mathbb{R})$. Also, ZF + AD arguably gives the correct structure theory for the definable sets of reals.[6]

Theorem 3.5 (Martin–Steel, Woodin, mid-1980s). *Suppose there are infinitely many Woodin cardinals with a measurable cardinal above them all. Then* AD *holds in* $L(\mathbb{R})$.

We will now see a simple application of AD (due to Martin).

A sequence $(b_n)_{n\geq 0}$ *computes* another sequence $(a_n)_{n\geq 0}$ (written $(a_n)_{n\geq 0} \leq_{\mathrm{Tu}} (b_n)_{n\geq 0}$) if and only if there is a computer program (technically, a Turing machine) which outputs $(a_n)_{n\geq 0}$ upon input of $(b_n)_{n\geq 0}$.

The sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ are *Turing equivalent* (written $(a_n)_{n\geq 0} \equiv_{\mathrm{Tu}} (b_n)_{n\geq 0}$) if

- $(a_n)_{n\geq 0} \leq_{\mathrm{Tu}} (b_n)_{n\geq 0}$ and
- $(b_n)_{n\geq 0} \leq_{\mathrm{Tu}} (a_n)_{n\geq 0}$.

A set C is a *Turing cone* if there is some $(a_n)_{n\geq 0}$ such that

$$C = \{(b_n)_n : (a_n)_n \leq_{\mathrm{Tu}} (b_n)_n\}.$$

Then $(a_n)_{n\geq 0}$ is the *base of* C.

Now, let A be a *Turing invariant* set of sequences (i.e., if $(a_n)_{n\geq 0} \in A$ and $(b_n)_{n\geq 0} \equiv_{\mathrm{Tu}} (a_n)_{n\geq 0}$, then also $(b_n)_{n\geq 0} \in A$).

- If player I has a winning strategy σ in G_A and $\vec{a} = (a_n)_{n\geq 0}$ codes σ, then $C \subseteq A$, where C is the Turing cone with base $\vec{a}$. ($\vec{a} \leq_{\mathrm{Tu}} \vec{b} \Rightarrow \vec{b} \equiv_{\mathrm{Tu}} \sigma * \vec{b} \in A \Rightarrow \vec{b} \in A$.)
- If player II has a winning strategy τ in G_A and $\vec{a} = (a_n)_{n\geq 0}$ codes τ, then $C \cap A = \emptyset$, where C is the Turing cone with base $\vec{a}$. (Argue as above.)

Here, $\sigma * \vec{b}$ is the play of G_A resulting from player I moving according to the strategy σ and player II playing the members of $\vec{b}$ in increasing order. (And we would define $\vec{b} * \tau$ similarly if τ is a strategy for player II.)

[6] Definable in the sense of being *projective* (equivalently, being definable in the structure $H_{\omega_1} = \{x : |\mathrm{TC}(x)| < \aleph_1\}$ from parameters). This is a natural restricted notion of definability which has in particular the feature that all sets of reals falling under it are in $L(\mathbb{R})$. Being in $L(\mathbb{R})$ can itself be rendered as a natural notion of definability.

Hence, if AD holds, then every Turing invariant A is either large (i.e., it contains a Turing cone) or small (i.e., it is disjoint from a Turing cone).

Note: If C_1 is the Turing cone with base $(a_n)_{n\geq 0}$ and C_2 is the Turing cone with base $(b_n)_{n\geq 0}$, then $C \subseteq C_1 \cap C_2$, where C is the Turing cone with base $(a_0, b_0, a_1, b_1, a_2, b_2, \ldots)$.

This means that for any finitely many Turing invariant properties $P_1, \ldots, P_n$ of sequences, there are choices $Q_i \in \{P_i, \neg P_i\}$, for $i \leq n$, such that the set of sequences satisfying Q_i for all i is large (in particular nonempty).

In fact, this is true for any countable collection P_i (for $i \in \mathbb{N}$) of properties.

4. The axiom (∗)

In the 1990s, Woodin defined and studied the following axiom (∗) [10].

Definition 4.1. (∗) is the conjunction of (1) and (2).

(1) AD holds in $L(\mathbb{R})$. (This follows from large cardinals by Theorem 3.5.)

(2) There is a $\mathbb{P}_{\max}$-generic filter G over $L(\mathbb{R})$ such that $L(\mathcal{P}(\omega_1)) = L(\mathbb{R})[G]$.

Here, $\mathbb{P}_{\max}$ is a certain *homogeneous* forcing notion in $L(\mathbb{R})$ (i.e., with the property that $L(\mathbb{R})[G_1]$ and $L(\mathbb{R})[G_2]$ satisfy the same sentences whenever G_1 and G_2 are $\mathbb{P}_{\max}$-generic filters over $L(\mathbb{R})$), definable over $L(\mathbb{R})$ without parameters.

Assuming the theory of $L(\mathbb{R})$ is frozen under forcing (which follows from large cardinals), it follows that the $L(\mathcal{P}(\omega_1))$s of any two models of (∗) obtained by forcing satisfy exactly the same sentences.

Also, (∗) implies the following very strong maximality principle (relative to all generic extensions).

Theorem 4.2 (Woodin). *Suppose* (∗) *holds and there are arbitrarily large Woodin cardinals. For every parameter-free* Π_2 *sentence* φ *(*$= (\forall x)(\exists y)\psi(x, y)$*, where all quantifiers in* $\psi(x, y)$ *are restricted), the following are then equivalent:*

(1) φ *is true in* $H(\omega_2)$.

(2) *In some forcing extension, it holds that* φ *is true in* $H(\omega_2)$.

Remark 4.3. In (2), the forcing extension need not preserve stationary subsets of ω_1.

In addition, (∗) implies that $L(\mathcal{P}(\omega_1))$ is obtained from $L(\mathbb{R})$ by adding to it any subset of ω_1 not in $L(\mathbb{R})$. Given that AD makes $L(\mathbb{R})$ into a "canonical" subuniverse, this means that (∗) makes $L(\mathcal{P}(\omega_1))$ into a larger "canonical" subuniverse.

The above facts rendered $(*)$ a very appealing axiom. However, in order for $(*)$ to be truly natural, it would have to be compatible with all consistent LC axioms. The main question can therefore be formulated as follows.

Question 4.4. Is $(*)$ compatible with all consistent LC axioms? Is $(*)$ even forcible over V (assuming enough large cardinals)?

4.1. $\mathrm{MM}^{++} \Rightarrow (*)$

In 2018, we answered this question.

Theorem 4.5 (Asperó–Schindler [1]). MM^{++} *implies* $(*)$.

In particular, if there is a supercompact cardinal κ, then there is a forcing notion of cardinality κ forcing MM^{++}, and therefore forcing $(*)$. All large cardinals there might be above κ are preserved in the generic extension.

This theorem renders $(*)$ a truly natural axiom. And it makes the observed completeness of MM^{++} for the theory of $L(\mathcal{P}(\omega_1))$ under forcing into a mathematical fact, thus rendering MM^{++} even more natural.

And both $(*)$ and MM^{++} were known to imply $2^{\aleph_0} = \aleph_2$.

5. High forcing axioms

Given the success of classical forcing axioms (for meeting $\aleph_1$-many dense sets) culminating in MM^{++}, it is natural to enquire whether a comparable theory can be developed for higher forcing axioms (i.e., for meeting more than $\aleph_1$-many dense sets).

The answer is no. For example:

- Π_2 maximality for $H(\omega_3)$ fails: CH and $\neg$ CH are both expressible over $H(\omega_3)$ by Π_2 sentences, and both can be forced.
- There are serious problems with building models of strong high forcing axioms. In fact, many forcing axiom candidates for meeting families of $\aleph_2$-many dense, even for fairly modest classes Γ of forcing notions, are just false.

The conclusion is that the success of MM^{++} cannot be replicated, at the level of high forcing axioms, to yield a natural axiom implying, for example, $2^{\aleph_0} = \aleph_3$.

This completes the argument for $2^{\aleph_0} = \aleph_2$.

Still, there are interesting open questions in the higher forcing axioms camp.

Question 5.1. Do reasonable LC axioms imply that there is no partition $\mathcal{P} \in L(\mathbb{R})$ of $\mathbb{R}$ into $\aleph_3$-many pieces?

A positive answer to this question would show that sufficiently strong large cardinals block the existence of counterexamples to $2^{\aleph_0} \leq \aleph_2$ which are *effectively given*, in a natural precise sense, and would thus contribute to the naturalness of the statement $2^{\aleph_0} \leq \aleph_2$. It should be noted that, by a result of Woodin, it is consistent to have, together with any consistent large cardinal axiom, that there is a partition $\mathcal{P} \in L(\mathbb{R})$ of $\mathbb{R}$ into $\aleph_2$-many pieces. In fact, the existence of such a partition follows from both MM and $(*)$.

Question 5.2. Do reasonable LC axioms imply that one can force some Σ_2 statement A which is complete for the theory of $H(\omega_3)$ modulo forcing preserving A?

The point of Question 5.2 is that, by Theorem 4.5, $(*)$ is such an axiom for $H(\omega_2)$.

Question 5.3. Is there any Π_2 sentence σ such that the following holds?

(1) ZFC proves that if $H(\omega_3)$ satisfies σ, then $2^{\aleph_0} = \aleph_3$.

(2) For some reasonable LC axiom A, ZFC $+A$ proves that it is forcible that $H(\omega_3)$ satisfies σ.

This question is motivated by two facts. The first fact is that Π_2 sentences about a suitable $H(\kappa)$ are precisely the type of sentences that forcing axioms are designed to imply.[7] The second fact is that, for the classical case, the proof that a forcing axiom FA(Γ) implies $2^{\aleph_0} = \aleph_2$ proceeds usually by showing that FA(Γ) implies that $H(\omega_2)$ satisfies some Π_2 sentence σ, where this statement in turn implies $2^{\aleph_0} = \aleph_2$.

6. A competing view

Woodin has championed an alternative view [11]. This is the content of *the Ultimate-L programme*. It is an ambitious version of the *inner model programme*, one of whose main goals is to construct an L-like subuniverse accommodating all possible LC axioms.

The axiom $V =$ Ultimate-L would provide a complete picture modulo forcing for the entire universe. And it implies $2^{\aleph_0} = \aleph_1$.

On the downside, $V =$ Ultimate-L is a difficult axiom to work with. And it is not clear at present to what extent the programme can be implemented.

[7] With κ being ω_2 in the classical case (i.e., in the case of forcing axioms for meeting $\aleph_1$-many dense sets).

At any rate, the implications and ramifications of the programme are extremely deep.

7. Back to MM^{++}: A challenge

By our theorem, MM^{++} implies that $L(\mathcal{P}(\omega_1))$ is a homogeneous extension of the AD-model $L(\mathbb{R})$ and hence a canonical model. And by forcing with homogeneous forcing notions over stronger AD-models, it is possible to obtain models of MM$^{++}(2^{\aleph_0})$ (i.e., MM^{++} for forcings of size $2^{\aleph_0}$) and even stronger fragments of MM^{++}. A major challenge nowadays is therefore the following.

Challenge. The following problems look hard.

(1) Obtain a model of full MM^{++}, or of stronger fragments of MM^{++}, by (homogeneous) forcing over strong models of the axiom of determinacy.

(2) Obtain a model of MM^{++} in which $(*)^{++}$ holds. The axiom $(*)^{++}$ says that there is $\Gamma \subseteq \mathcal{P}(\mathbb{R})$ and a $\mathbb{P}_{\max}$-generic filter G over $L(\mathbb{R}, \Gamma)$ such that $\mathcal{P}(\mathbb{R}) \subseteq L(\mathbb{R}, \Gamma)[G]$.

Remark 7.1. In the presence of large cardinals (e.g., a proper class of Woodin cardinals), the ability to obtain a model of $(*)^{++}$ by set forcing would give a positive answer to Question 5.2.

Funding. Ralf Schindler was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044 390685587, Mathematics Münster: Dynamics – Geometry – Structure.

References

[1] D. Asperó and R. Schindler, Martin's Maximum^{++} implies Woodin's axiom $(*)$. *Ann. of Math. (2)* **193** (2021), no. 3, 793–835 Zbl 1496.03203 MR 4250390

[2] P. J. Cohen, The independence of the continuum hypothesis. *Proc. Natl. Acad. Sci. USA* **50** (1963), no. 6, 1143–1148 Zbl 0192.04401 MR 0157890

[3] M. Foreman, M. Magidor, and S. Shelah, Martin's maximum, saturated ideals, and non-regular ultrafilters. Part I. *Ann. of Math. (2)* **127** (1988), no. 1, 1–47 Zbl 0645.03028 MR 0924672

[4] H. M. Friedman, Higher set theory and mathematical practice. *Ann. Math. Logic* **2** (1971), no. 3, 325–357 Zbl 0215.32702 MR 0284327

[5] K. Gödel, The consistency of the axiom of choice and of the generalized Continuum-Hypothesis. *Proc. Natl. Acad. Sci. USA* **24** (1938), no. 12, 556–557 Zbl 0020.29701

[6] K. Gödel, What is Cantor's continuum problem? *Amer. Math. Monthly* **54** (1947), no. 9, 515–525 Zbl 0038.03003 MR 0023780

[7] D. Hilbert, *Mathematische Probleme. Vortrag, gehalten auf dem internationalen Mathematiker-Congress zu Paris 1900*, pp. 253–297, Nachr. K. Ges. Wiss. Göttingen, Math. Phys. Klasse 3, 1900 JFM 31.0068.03

[8] D. A. Martin and R. M. Solovay, Internal Cohen extensions. *Ann. Math. Logic* **2** (1970), no. 2, 143–178 Zbl 0222.02075 MR 0270904

[9] R. M. Solovay and S. Tennenbaum, Iterated Cohen extensions and Souslin's problem. *Ann. of Math. (2)* **94** (1971), no. 2, 201–245 Zbl 0244.02023 MR 0294139

[10] W. H. Woodin, *The axiom of determinacy, forcing axioms, and the nonstationary ideal*. De Gruyter Ser. Log. Appl. 1, De Gruyter, Berlin, 1999, 934 pp. Zbl 0954.03046 MR 1713438

[11] W. H. Woodin, In search of ultimate-L the 19th Midrasha Mathematicae lectures. *Bull. Symb. Log.* **23** (2017), no. 1, 1–109 Zbl 1420.03134 MR 3632568

David Asperó
School of Engineering, Mathematics, and Physics, University of East Anglia, Norwich, Norfolk, UK; d.aspero@uea.ac.uk

Ralf Schindler
Department of Mathematics and Computer Science, Institute for Mathematical Logic and Fundamental Research, Universität Münster, Einsteinstr. 62, 48149 Münster, Germany; rds@uni-muenster.de

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/11

Self-affine sets and measures – A survey

Balázs Bárány

Abstract. Self-affine sets can be expressed as affine copies of themselves. The dimension theory, when the defining mappings are similarities, is relatively well understood, but the understanding of the general case is far from being complete. Here, we provide an overview from the most basic definitions of the topic to the most recent results.

1. Introduction

There is no widely accepted definition of fractals. However, most authors would call a set fractal if, on infinitely many scales, its smaller parts resemble the whole. The self-affine sets provide a model for that phenomenon. This class includes several well-known fractal sets, like the triadic-Cantor set, the Sierpiński gasket, the von Koch curve, and the graph of Takagi's function. One of the central problems of geometric measure theory is determining the "size" of these objects in several aspects, for example, finding the Hausdorff or box-counting dimension. This paper aims to introduce the most basic definitions and to survey the most recent results of the topic.

1.1. Basic definitions

Let $\mathcal{I}$ be a finite set of indices, and let $\Phi = \{f_i(x) = A_i x + t_i\}_{i\in\mathcal{I}}$ be a finite tuple of invertible contractive affine mappings on $\mathbb{R}^d$. That is, $A_i \in \mathrm{GL}_d(\mathbb{R})$ such that $\|A_i\| < 1$ for every $i \in \mathcal{I}$ and $t_i \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the induced matrix norm by the usual Euclidean norm on $\mathbb{R}^d$. Hutchinson [33] showed (in a much higher generality) that there exists a unique non-empty compact set X such that

$$X = \bigcup_{i\in\mathcal{I}} f_i(X). \tag{1.1}$$

Mathematics Subject Classification 2020: 28A80 (primary); 28A78 (secondary).
Keywords: box-counting dimension, Hausdorff dimension, self-affine sets, self-affine measures.

We call the set X *self-affine*. In the special case, when all the affine mappings of Φ are similarities, that is, $f_i(x) = \lambda_i O_i x + t_i$, where $\lambda_i \in (0,1)$ is the scaling, $O_i \in O(d,\mathbb{R})$ is an orthogonal matrix and $t_i \in \mathbb{R}^d$, then we call the set X *self-similar*.

The construction of X relies on an iterative application of the maps in Φ, so we call Φ an *iterated function system* or IFS for short. Sometimes, we also call the set X the *attractor* of the IFS Φ. To construct X, take $R > 0$ sufficiently large such that every map in Φ maps the closed ball $B(0,R)$ to itself, that is, $f_i(B(0,R)) \subseteq B(0,R)$. For instance, the choice $R = \max_{i\in\mathcal{I}} \|t_i\|(1 - \|A_i\|)^{-1}$ is sufficient. Then for every $n \geq 1$, let

$$X_n = \bigcup_{i_1\in\mathcal{I}} \cdots \bigcup_{i_n\in\mathcal{I}} f_{i_1} \circ \cdots \circ f_{i_n}(B(0,R)). \tag{1.2}$$

It is easy to see that the set X_n is non-empty, compact and $X_{n+1} \subseteq X_n$ for every $n \geq 1$. Hence, $X = \bigcap_{n=1}^{\infty} X_n$ is a non-empty compact set by Cantor's intersection theorem, and X clearly satisfies (1.1). The uniqueness of the set X relies on simple properties of the Hausdorff metric on compact subsets of $\mathbb{R}^d$. For details, see, for example, the book of Bárány, Simon and Solomyak [8, Theorem 1.3.6]. For some examples of self-similar and self-affine sets, see Figure 1.

Let us now consider the following Markovian random walk on X, which is also called the *chaos game*, introduced by Barnsley [9]. Let $\mathbf{p} = (p_i)_{i\in\mathcal{I}}$ be a probability vector (i.e., $p_i \geq 0$ for every $i \in \mathcal{I}$ and $\sum_{i\in\mathcal{I}} p_i = 1$). Starting from an initial point $Y_0 \in X$, let us choose a function f_{i_1} randomly from Φ according to $\mathbf{p}$ and jump to the point $Y_1 = f_{i_1}(Y_0)$. It continues iteratively generating a sequence of random points $Y_2, Y_3, \ldots$, where $Y_n = f_{i_n}(Y_{n-1})$ and the indices i_n are independently and identically distributed (i.i.d.) according to $\mathbf{p}$. This process has a unique stationary measure; more precisely, there exists a unique compactly supported Borel probability measure ν such that

$$\nu = \sum_{i\in\mathcal{I}} p_i (f_i)_* \nu, \tag{1.3}$$

where $f_*\nu$ denotes the pushforward of the measure ν with respect to the map f. For a proof, see, for example, Edgar's book [14]. We call the measure ν a *self-affine measure*, and if Φ consists only of similarities, then we call ν a *self-similar measure*.

The attractor X of Φ has a strong connection with symbolic coding. Namely, every point in X can be associated with infinite sequences formed by the symbols in $\mathcal{I}$ (associated with the functions of Φ). That is, let $\Sigma = \mathcal{I}^{\mathbb{N}}$ be the set of one-sided infinite sequences, called *symbolic space*. For a finite word $\mathtt{j} \in \Sigma_* = \bigcup_{n=0}^{\infty} \mathcal{I}^n$, let us define the cylinder set $[\mathtt{j}]$ by $[j_1, \ldots, j_n] = \{\mathtt{i} \in \Sigma : i_1 = j_1, \ldots, i_n = j_n\}$. The natural Borel σ-algebra on Σ is the one which is induced by the cylinder sets. For brevity, we introduce the following notation: for a finite word $\mathtt{j} = (j_1, \ldots, j_n) \in \Sigma_*$, let $f_{\mathtt{j}} = f_{j_1} \circ \cdots \circ f_{j_n}$ be the iterates of the maps of Φ in the prescribed order by $\mathtt{j}$. We will also adapt this notation later for the product of matrices and real numbers.

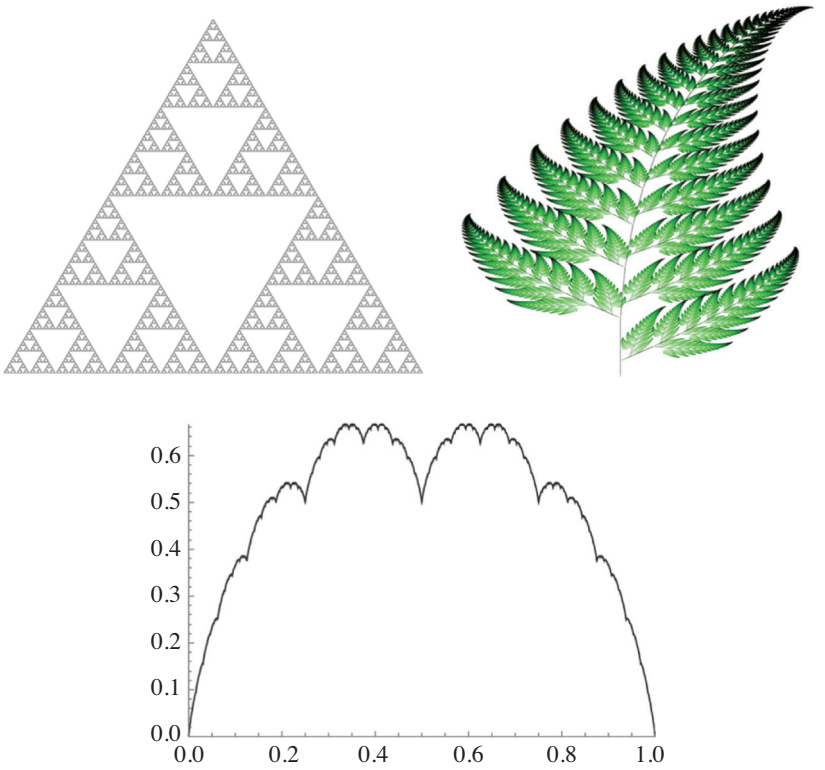


Figure 1. Examples of self-affine sets: the Sierpiński gasket, Barnsley's fern (source: Wikipedia) and Takagi's nowhere differentiable function graph.

For every $\mathtt{i} = (i_1, i_2, \ldots) \in \Sigma$, we define the *symbolic coding map* as

$$\pi(\mathtt{i}) = \lim_{n\to\infty} f_{i_1} \circ \cdots \circ f_{i_n}(0).$$

The set Σ is a compact metric space equipped with the natural metric $d(\mathtt{i}, \mathtt{j}) = 2^{-|\,\mathtt{i}\wedge\mathtt{j}\,|}$, where $|\,\mathtt{i}\wedge\mathtt{j}\,| = \min\{n \geq 0 : i_{n+1} \neq j_{n+1}\}$, and the map $\pi\colon \Sigma \to X$ is Hölder continuous and surjective but not necessarily injective.

Denote $\sigma(i_1, i_2, \ldots) = (i_2, i_3, \ldots)$ as the left-shift operator on Σ. As simple as this mapping is, its dynamic structure is equally rich, which plays an essential role in dynamical systems theory. It is easy to see that

$$\pi(\mathtt{i}) = f_{i_1}(\pi(\sigma\,\mathtt{i})). \tag{1.4}$$

Hence, the left-shift operator and its properties benefit the study of self-affine sets. We call a Borel probability measure μ on Σ left-shift invariant if $\mu = \sigma_*\mu$, that is, $\mu(A) = \mu(\sigma^{-1}A)$ for every Borel set $A \subseteq \Sigma$. We call a left-shift invariant probability

measure μ ergodic if $\mu(A) = 0$ or $\mu(A) = 1$ for every invariant set $A = \sigma^{-1}A$. One of the main advantages of ergodicity is Birkhoff's ergodic theorem, which roughly states that the time average equals the space average. That is, for every $f\colon \Sigma \to \mathbb{R}$ μ-integrable map, $\frac{1}{n}\sum_{k=0}^{n-1} f(\sigma^k \mathtt{i}) \to \int f(\mathtt{j})d\mu(\mathtt{j})$ as $n \to \infty$ for μ-almost every $\mathtt{i}$. One of the most basic left-shift invariant ergodic measures is the family of Bernoulli measures. Namely, for a probability vector $\mathbf{p} = (p_i)_{i\in\mathcal{I}}$, let μ be the measure such that

$$\mu([i_1, \dots, i_n]) = p_{i_1} \cdots p_{i_n}.$$

The measure μ on Σ is σ-invariant and ergodic; furthermore, $\pi_*\mu = \nu$, where ν is the self-affine measure defined in (1.3). For an introduction to ergodic theory, see, for example, Walter's book [62].

For an ergodic left-shift invariant measure μ on Σ, let us denote the entropy of μ by h_μ. That is,

$$h_\mu = \lim_{n\to\infty} \frac{-1}{n} \sum_{\mathtt{i}\in\mathcal{I}^n} \mu([\mathtt{i}]) \log \mu([\mathtt{i}]) = \lim_{n\to\infty} \frac{-1}{n} \log \mu([\mathtt{i}\,|_n]) \quad \text{for } \mu\text{-almost every } \mathtt{i},$$

where $\mathtt{i}\,|_n$ denotes the first n symbols of $\mathtt{i} \in \Sigma$. The second equality follows from the Shannon–McMillan–Breiman theorem. Note that if μ is a Bernoulli measure with probability vector $\mathbf{p} = (p_i)_{i\in\mathcal{I}}$, $h_\mu = -\sum_{i\in\mathcal{I}} p_i \log p_i$.

One of the cardinal questions of fractal geometry is determining the dimension of self-affine sets and measures. Let us introduce the dimension concepts we will study.

1.2. Dimension concepts

In this note, we will be particularly interested in the Hausdorff (or covering) dimension and the Minkowski (or box-counting) dimension of sets. For a bounded set $E \subset \mathbb{R}^d$, let $N_\delta(E)$ be the minimal number of balls with radius $\delta > 0$ needed to cover E. We define the *lower and upper box-counting dimensions* by

$$\underline{\dim}_B(E) = \liminf_{\delta\to 0} \frac{\log N_\delta(E)}{-\log \delta} \quad \text{and} \quad \overline{\dim}_B(E) = \limsup_{\delta\to 0} \frac{\log N_\delta(E)}{-\log \delta}.$$

If the lower and upper Minkowski dimensions of E coincide, then we say that the Minkowski dimension of E exists, and we denote the common value by $\dim_B(E)$.

Before we define the Hausdorff dimension, we define the *s-dimensional (outer) Hausdorff measure* of a set $E \subset \mathbb{R}^d$ by

$$\mathcal{H}^s(E) = \sup_{\delta>0} \mathcal{H}^s_\delta(E), \quad \text{where } \mathcal{H}^s_\delta(E) = \inf\Big\{\sum_i |U_i|^s : E \subset \bigcup_i U_i \text{ and } |U_i| < \delta\Big\}.$$

Let $\dim_H(E)$ be the *Hausdorff dimension* of the set E defined as

$$\dim_H(E) = \inf\{s > 0 : \mathcal{H}^s(E) = 0\}.$$

For the basic properties of the Minkowski and Hausdorff dimensions, see Falconer's book [15]. Both dimension concepts measure how well a set fills the available space via effective covers with balls in some proper sense. We must mention here that there are other essential dimension concepts, like the Tricot (packing) and Assouad dimensions. Both play a significant role in geometric measure theory. For their definitions and basic properties, we direct the reader to Falconer's book [15] and Fraser's book [26].

We also introduce dimension concepts for measures via local dimension. Let ν be a Borel probability measure on $\mathbb{R}^d$. We define the *upper and lower local dimensions* of ν at the point x by

$$\overline{d}_\nu(x) = \limsup_{\delta\to 0} \frac{\log \nu(B(x,\delta))}{\log \delta} \quad \text{and} \quad \underline{d}_\nu(x) = \liminf_{\delta\to 0} \frac{\log \nu(B(x,\delta))}{\log \delta}.$$

We say that the *local dimension* exists at x if the two values above coincide, and we denote the common value by $d_\nu(x)$. We define the *lower and upper Hausdorff dimensions* of the measure ν by

$$\underline{\dim}_H(\nu) = \inf\{\dim_H(E) : \nu(E) > 0\} = \operatorname*{ess\,inf}_{x\sim\nu} \underline{d}_\nu(x)$$

and

$$\overline{\dim}_H(\nu) = \inf\{\dim_H(E) : \nu(E^c) = 0\} = \operatorname*{ess\,sup}_{x\sim\nu} \underline{d}_\nu(x).$$

For proof of the above equalities, see the book by Przytycki and Urbański [55, Section 7.6]. Let us note that the lower and upper Tricot dimensions can be defined similarly by using the upper local dimension

$$\underline{\dim}_P(\nu) = \operatorname*{ess\,inf}_{x\sim\nu} \overline{d}_\nu(x) \quad \text{and} \quad \overline{\dim}_P(\nu) = \operatorname*{ess\,sup}_{x\sim\nu} \overline{d}_\nu(x).$$

Finally, we say that the measure ν is *exact dimensional* if $\underline{\dim}_H(\nu) = \overline{\dim}_P(\nu)$. In this case, we denote the coinciding dimension values of ν by $\dim(\nu)$.

2. Quick overview of self-similar sets and measures

Let us first briefly consider the special case in which all the maps of the IFS $\Phi = \{f_i\}_{i\in\mathcal{I}}$ on $\mathbb{R}^d$ are similarities. That is, $f_i(x) = \lambda_i O_i x + t_i$, where $0 < \lambda_i < 1$ is the scaling, $O_i \in O(d,\mathbb{R})$ is an orthonormal matrix and $t_i \in \mathbb{R}^d$ is the translation parameter. Iterative construction (1.2) of the self-similar set X provides a naive cover

of X and so a natural upper bound of the dimension of X. In particular, choosing $\lambda_{\max} = \max_i \lambda_i$, one can see that

$$\mathcal{H}^s_{\lambda_{\max}^n R}(X) \le \sum_{\mathtt{i}\in\mathcal{I}^n} |f_{\mathtt{i}}(B(0,R))|^s = R^s \left(\sum_{i\in\mathcal{I}} \lambda_i^s\right)^n \quad \text{for every } n\in\mathbb{N}. \tag{2.1}$$

In the last equality, we have used that all the maps $f_{\mathtt{i}}$ are similarities, and so $f_{\mathtt{i}}(B(0,R))$ is a closed ball centred at $f_{\mathtt{i}}(0)$ with radius $R\cdot\lambda_{\mathtt{i}}$. Hence, $\dim_H(X) \le s_0(\Phi)$, where $s_0(\Phi)$ is the unique solution of the equation $\sum_{i\in\mathcal{I}} \lambda_i^s = 1$. We call the quantity $s_0(\Phi)$ the *similarity dimension* of the IFS Φ. In particular, Hutchinson [33] showed that

$$\overline{\dim}_B(X) \le \min\{d, s_0(\Phi)\}. \tag{2.2}$$

A central problem is when the lower bound $\dim_H X \ge \min\{d, s_0(\Phi)\}$ can be established. For that, we introduce some separation between the cylinder sets and the open set condition (OSC). We say that the IFS Φ satisfies the *open set condition* if there exists an open and bounded set $U\subset\mathbb{R}^d$ such that $f_i(U)\subset U$ and $f_i(U)\cap f_j(U)=\emptyset$ for every $i\ne j\in\mathcal{I}$. Many self-similar objects satisfy the OSC, for instance, the Sierpiński gasket in Figure 1. Hutchinson [33] showed the following.

Theorem 2.1. *If a self-similar IFS Φ satisfies the OSC, then*

$$\dim_H(X) = \overline{\dim}_B(X) = s_0(\Phi).$$

Falconer [19] showed that regardless of any separation condition, $\dim_H(X) = \overline{\dim}_B(X)$, and in particular, the box-counting dimension of X exists.

The idea of the proof is the following: consider the natural self-similar measure ν in (1.3) defined by the probabilities $p_i = \lambda_i^{s_0(\Phi)}$. In other words, let ν be the pushforward of the Bernoulli measure μ defined by the probability vector $(\lambda_i^{s_0(\Phi)})_{i\in\mathcal{I}}$. An easy-to-see consequence of the open set condition is that there exists a constant $C>0$ such that for every $r>0$ and every $x\in X$,

$$\#\{f_{\mathtt{i}}(U) : f_{\mathtt{i}}(U)\cap B(x,r)\ne\emptyset \text{ and } \lambda_{\mathtt{i}}\le r<\lambda_{\mathtt{i}_-}\}\le C,$$

where $\mathtt{i}_- = (i_1,\dots,i_{n-1})$ for $\mathtt{i}=(i_1,\dots,i_n)$. Since $X\subseteq\overline{U}$, we get

$$\nu(B(x,r)) \le \sum_{\substack{\mathtt{i}\in\Sigma_*: f_{\mathtt{i}}(U)\cap B(x,r)\ne\emptyset \\ \text{and } \lambda_{\mathtt{i}}\le r<\lambda_{\mathtt{i}_-}}} \lambda_{\mathtt{i}}^{s_0(\Phi)} \le C r^{s_0(\Phi)},$$

and so $\dim_H(X)\ge \underline{\dim}_H(\nu)\ge s_0(\Phi)$. In particular, we have shown the following variational principle:

$$\dim_H(X) = \max\{\dim(\pi_*\mu) : \mu \text{ is left-shift invariant and ergodic on } \Sigma\}, \tag{2.3}$$

where the dimension maximum is attained at the Bernoulli measure with probability vector $(\lambda_i^{s_0(\Phi)})_{i\in\mathcal{I}}$. Let us note that Feng and Hu [22] verified (2.3) with supremum for general self-similar sets regardless of the overlapping structure.

The dimension of self-similar measures can be studied analogously. Namely, let μ be an ergodic left-shift invariant probability measure on Σ. Let us denote by $\chi_\mu(\Phi)$ the Lyapunov exponent of μ concerning the IFS Φ. That is, $\chi_\mu(\Phi) = -\sum_{i\in\mathcal{I}} \mu([i]) \log \lambda_i$. The quantity represents the average shrinking rate of the diameter $f_{\mathtt{i}|_n}(X)$ as $n \to \infty$ for a μ-typical $\mathtt{i}$. One can show easily, relying on the Shannon–McMillan–Breiman theorem and Birkhoff's ergodic theorem, that for every ergodic measure μ on Σ,

$$\overline{\dim}_P(\pi_*\mu) \le \min\left\{d, \frac{h_\mu}{\chi_\mu(\Phi)}\right\},$$

where we recall that h_μ denotes the entropy of the measure μ. We call the quantity on the right-hand side the *Lyapunov dimension* of the measure μ. Furthermore, if the open set condition holds, then

$$\dim(\pi_*\mu) = \frac{h_\mu}{\chi_\mu(\Phi)}; \tag{2.4}$$

see Edgar [14, Theorem 5.2.5] or [8]. Feng and Hu [22] showed that regardless of any separation condition, the pushforward of any ergodic left-shift invariant measure is exact dimensional.

It is clear that $\dim_H(X) = \min\{d, s_0\}$ does not always hold. Consider, for example, the IFS $\Phi = \{f_1(x) = x/3,\ f_2(x) = x/3 + 1,\ f_3(x) = x/3 + 3\}$, which has similarity dimension 1. However, $f_1 \circ f_3 = f_2 \circ f_1$, and so the IFS $\Phi' = \{f_i \circ f_j\}_{(i,j)\in\{1,2,3\}^2\setminus\{(1,3)\}}$ has the same attractor X, and the similarity dimension of Φ' is $\log_9 8 < 1$. One of the long-standing conjectures in fractal geometry claims that having a completely coinciding map is the only way on the real line to drop the dimension of the attractor strictly below the similarity dimension. This is called the *exact overlap conjecture*. More precisely, for a self-similar IFS Φ on the line with attractor X, if $\dim_H X < \min\{1, s_0(\Phi)\}$, then there exists an exact overlap, that is, there are $\mathtt{i} \ne \mathtt{j} \in \Sigma_*$ such that $f_\mathtt{i} \equiv f_\mathtt{j}$. The higher-dimensional analogue of this conjecture is false; see, for example, [8, Example 7.6].

To handle the overlapping situation (i.e., when the open set condition fails) on the real line, Hochman [29] introduced the exponential separation condition (ESC). The IFS Φ satisfies the *exponential separation condition* if there exists an $\varepsilon > 0$ such that for infinitely many $n \in \mathbb{N}$,

$$\min_{\mathtt{i}\ne\mathtt{j}\in\mathcal{I}^n} (|\log \lambda_\mathtt{i} - \log \lambda_\mathtt{j}| + |f_\mathtt{i}(0) - f_\mathtt{j}(0)|) > \varepsilon^n.$$

Hochman [29] showed that if a self-similar IFS Φ on the line satisfies the ESC, then for every self-similar measure $\nu = \sum_i p_i (f_i)_* \nu$,

$$\dim(\nu) = \min\left\{1, \frac{-\sum_i p_i \log p_i}{-\sum_i p_i \log \lambda_i}\right\},$$

and in particular,

$$\dim_H(X) = \min\{1, s_0(\Phi)\}.$$

The proof relies on an entropy growth theorem proved by Hochman [29] using methods from additive combinatorics. For some heuristics and intuition on the proof, see the survey paper of Hochman [30].

Jordan and Rapaport [35] showed that the dimension result can be extended to a more general family of measures, that is, if the self-similar IFS Φ satisfies the ESC, then for every left-shift invariant ergodic measure μ,

$$\dim(\pi_* \mu) = \min\left\{1, \frac{h_\mu}{\chi_\mu(\Phi)}\right\}.$$

For more recent developments on the dimension theory of self-similar sets and measures, we direct the reader to the survey paper of Varjú [61].

3. Self-affine systems

3.1. The affinity and Lyapunov dimensions

Now, we turn to the main topic of this survey, the self-affine sets, namely, when the IFS $\Phi = \{f_i(x) = A_i x + t_i\}_{i \in \mathcal{I}}$ consists of invertible affine mappings. If one wants to cover the self-affine set X similarly to (2.1), one will immediately face the following problem: the sets $f_{\mathtt{i}}(B(0, R))$ are no longer balls. To overcome this difficulty, Falconer [18] introduced the singular value function.

Let A be a $d \times d$ invertible matrix. Let us denote the singular values of A by $\alpha_1(A) \geq \cdots \geq \alpha_d(A) > 0$. Note that $\alpha_1(A) = \|A\|$, where $\|\cdot\|$ is the operator norm induced by the usual Euclidean norm. Let us recall that for finite words $\mathtt{i} = (i_1, \ldots, i_n) \in \Sigma_*$, we use the notations $f_{\mathtt{i}} = f_{i_1} \circ \cdots \circ f_{i_n}$ and $A_{\mathtt{i}} = A_{i_1} \cdots A_{i_n}$. Then for every $\mathtt{i} \in \Sigma_*$, the set $f_{\mathtt{i}}(B(0, R))$ is an ellipse whose main semi-axes have lengths $\alpha_1(A_{\mathtt{i}}) \cdot R, \ldots, \alpha_d(A_{\mathtt{i}}) \cdot R$. One can cover the ellipse $f_{\mathtt{i}}(B(0, R))$ with one square of side length $\alpha_1(A_{\mathtt{i}}) \cdot R$ or with at most $\lceil \alpha_1(A_{\mathtt{i}})/\alpha_2(A_{\mathtt{i}}) \rceil$-many squares of side length $\alpha_2(A_{\mathtt{i}}) \cdot R$, and so on; see Figure 2. For a given $s \in [0, d]$ and given $d \times d$ matrix A, it is easy to see that

$$\min_{k=1,\ldots,d} \frac{\alpha_1(A) \cdots \alpha_{k-1}(A)}{\alpha_k(A)^{k-1}} \alpha_k(A)^s = \alpha_1(A) \cdots \alpha_{\lfloor s \rfloor}(A) \cdot \alpha_{\lfloor s \rfloor + 1}(A)^{s - \lfloor s \rfloor}.$$

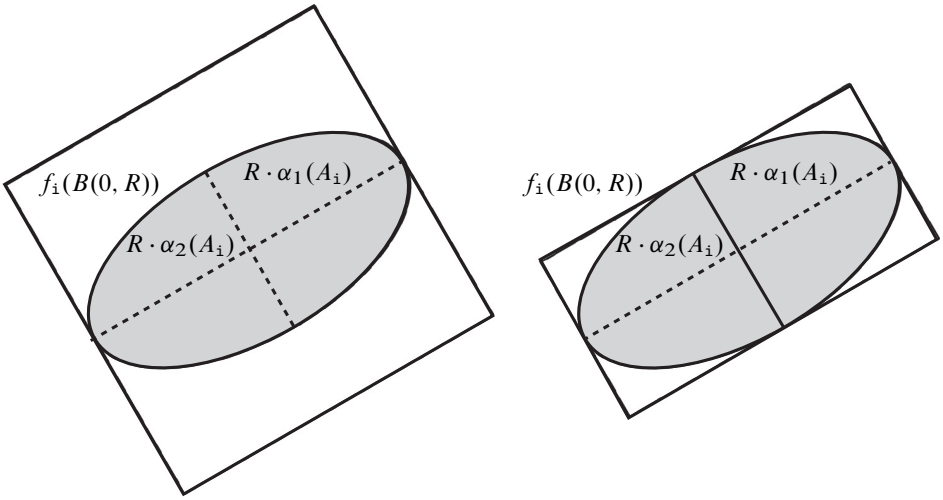


Figure 2. Two ways to cover the ellipse $f_{\mathtt{i}}(B(0,R))$ with squares of side lengths $R\cdot\alpha_1(A_{\mathtt{i}})$ and $R\cdot\alpha_2(A_{\mathtt{i}})$.

Hence, the most economic cover for a given $s>0$ using the sets $f_{\mathtt{i}}(B(0,R))$ is covering $f_{\mathtt{i}}(B(0,R))$ with squares of side length $\alpha_{\lfloor s\rfloor+1}(A_{\mathtt{i}})\cdot R$. This motivates the definition of *singular value function*. Namely, for a given $s>0$ and $d\times d$ invertible matrix A, let

$$\varphi^s(A):=\begin{cases}\alpha_1(A)^s & \text{if } s\in[0,1],\\ \alpha_1(A)\cdots\alpha_k(A)\alpha_{k+1}(A)^{s-k} & \text{if } s\in(k,k+1]\\ & \text{for some } k\in\{1,\dots,d-1\},\\ |\alpha_1(A)\cdots\alpha_d(A)|^{s/d} & \text{if } s>d.\end{cases}\tag{3.1}$$

And so taking $\alpha_{\max}=\max_i\alpha_1(A)$, we have

$$\mathcal{H}^s_{\alpha^n_{\max}R}(X)\le R^s\cdot\sum_{\mathtt{i}\in\mathcal{I}^n}\varphi^s(A_{\mathtt{i}}).$$

In particular, $\dim_H(X)\le s$ for every $s>0$ for which $\sum_{\mathtt{i}\in\mathcal{I}^n}\varphi^s(A_{\mathtt{i}})\to 0$ as $n\to\infty$. More generally, Falconer [18] showed that

$$\dim_H(X)\le\overline{\dim}_B(X)\le\min\{d,s_0(\Phi)\},\tag{3.2}$$

where

$$s_0(\Phi)=\inf\Big\{s>0:\sum_{n=1}^{\infty}\sum_{\mathtt{i}\in\mathcal{I}^n}\varphi^s(A_{\mathtt{i}})<\infty\Big\}.$$

We call the quantity $s_0(\Phi)$ the *affinity dimension* of Φ.

The singular value function is sub-multiplicative, see Falconer [18], and so by Fekete's lemma, the *sub-additive pressure function*

$$P_\Phi(s) = \lim_{n\to\infty} \frac{1}{n} \log\left(\sum_{\mathtt{i}\in\mathcal{I}^n} \varphi^s(A_\mathtt{i}) \right)$$

is well defined, strictly monotone decreasing and continuous. Furthermore, $P(0) = \log \#\mathcal{I}$ and $P(s) \to -\infty$ as $s \to \infty$. Moreover, the affinity dimension $s_0(\Phi)$ is the unique root $P_\Phi(s_0(\Phi)) = 0$. Let us note that the affinity dimension is the natural generalisation of the similarity dimension, since $\alpha_1(\lambda O) = \cdots = \alpha_d(\lambda O) = \lambda$ of every $O \in O(d, \mathbb{R})$ and $\lambda \in (0, \infty)$.

It is natural to ask whether the affinity dimension can be calculated explicitly or approximated efficiently. Unfortunately, the naive approximation $\sum_{\mathtt{i}\in\Sigma_n} \varphi^{s_n}(A_\mathtt{i}) = 1$ gives us the extremely inefficient $|s_0(\Phi) - s_n| = O(1/n)$ error rate. However, Pollicott and Vytnova [53] were able to introduce an approximation method with a super-exponential $O(e^{-cn^2})$ approximation rate, where n stands for the number of iterates required, for a certain class of planar matrices (like matrices with strictly positive entries) using zeta functions. This has been generalised recently by Morris [49] for higher dimensions.

The non-conformal structure of the IFS also necessitates the modification of the Lyapunov dimension compared to the self-similar situation following Jordan, Pollicott and Simon [34]. As we mentioned, the Lyapunov exponent measures the average shrinking rate of the cylinder sets at a typical point. However, the non-conformality of the system implies that this rate is different in different directions.

Let μ be a left-shift invariant ergodic probability measure on Σ. By Oseledets' multiplicative ergodic theorem, see Krengel's book [43], we define the ith Lyapunov exponent $\chi^i_\mu(\Phi)$ by

$$\begin{aligned}\chi^i_\mu(\Phi) &= \lim_{n\to\infty} \frac{-1}{n} \int \log \alpha_i(A_{\mathtt{i}|_n}) d\mu(\mathtt{i}) \\ &= \lim_{n\to\infty} \frac{-1}{n} \log \alpha_i(A_{\mathtt{i}|_n}) \quad \text{for } \mu\text{-almost every } \mathtt{i} \in \Sigma\end{aligned}$$

for $i = 1, \dots, d$. Then we define the *Lyapunov dimension* of the measure μ by

$$\dim_L(\mu) = \min\left\{ d, \min_{k=1,\dots,d} \left\{ k - 1 + \frac{h_\mu - \sum_{n=1}^{k-1} \chi^n_\mu(\Phi)}{\chi^k_\mu(\Phi)} \right\} \right\}.$$

Rossi [59] showed that for every ergodic left-shift invariant measure μ,

$$\underline{\dim}_H(\pi_*\mu) \le \overline{\dim}_P(\pi_*\mu) \le \dim_L(\mu). \tag{3.3}$$

A natural question arises: under which circumstances are the inequalities in (3.2) and (3.3) equalities? The first results to answer this question were considering "typical systems" in the sense of Lebesgue-typical choice of translation parameters

$\underline{t} = (t_i)_{i\in\mathcal{I}} \in \mathbb{R}^{d\cdot\#\mathcal{I}}$. More precisely, Falconer [18] and Solomyak [60] showed the following.

Theorem 3.1. *If $\|A_i\| < 1/2$ for every $i \in \mathcal{I}$, then for $d\cdot\#\mathcal{I}$-dimensional Lebesgue almost every $\underline{t} = (t_i)_{i\in\mathcal{I}}$,*

$$\dim_H(X_{\underline{t}}) = \overline{\dim}_B(X_{\underline{t}}) = \min\{d, s_0(\Phi_{\underline{t}})\},$$

where we write the IFS $\Phi_{\underline{t}} = \{x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ and the attractor $X_{\underline{t}}$ of $\Phi_{\underline{t}}$.

Solomyak [60] improved the previous bound $1/3$ of Falconer [18] to $1/2$. Let us also note that Edgar [13], based on Przytycki and Urbański [54], pointed out that the bound $1/2$ is sharp. Namely, for every $\varepsilon > 0$, there exists a planar self-affine IFS $\Phi = \{x \mapsto A_1 x + t_1,\ x \mapsto A_2 x + t_2\}$ such that $\max_i \|A_i\| = 1/2 + \varepsilon$ but $\dim_H(X) < \min\{2, s_0(\Phi)\}$ for any choice of (t_1, t_2).

A similar result to Theorem 3.1 has been shown by Jordan, Pollicott and Simon [34] for the pushforward of ergodic measures.

Theorem 3.2. *If $\|A_i\| < 1/2$ for every $i \in \mathcal{I}$ and μ is an ergodic left-shift invariant measure on Σ, then for $d\cdot\#\mathcal{I}$-dimensional Lebesgue almost every $\underline{t}$, the measure $(\pi_{\underline{t}})_*\mu$ is exact dimensional; moreover,*

$$\dim((\pi_{\underline{t}})_*\mu) = \dim_L(\mu),$$

where $\pi_{\underline{t}}\colon \Sigma \to X_{\underline{t}}$ denotes the natural projection with respect to the IFS $\Phi_{\underline{t}} = \{x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ to the attractor $X_{\underline{t}}$.

The above theorems provide us with the value of dimensions for typical systems in some sense, but they do not provide simple, verifiable, sufficient conditions for particular systems. Before the recent developments on the Ledrappier–Young formula for self-affine systems (see Section 3.5) and the entropy growth theorem introduced by Hochman [29], only very few particular examples were known; see, for instance, Hueter and Lalley [32] or Falconer [20]. In the remaining part of the paper, we will give "relatively" easily verifiable geometric and algebraic conditions which ensure that equalities hold in (3.2) and (3.3), with heuristics on the necessity of them.

3.2. Self-affine carpets

First, let us make a detour now for a class of planar self-affine systems for which we see the atypical behaviour. Let us consider the following simple example:

$$\Phi = \left\{ f_1(x,y) = \left(\frac{x}{3}, \frac{y}{4}\right), f_2(x,y) = \left(\frac{x+2}{3}, \frac{y}{4}\right), \right.$$
$$\left. f_3(x,y) = \left(\frac{x}{3}, \frac{y+3}{4}\right), f_4(x,y) = \left(\frac{x+2}{3}, \frac{y+3}{4}\right)\right\}. \tag{3.4}$$

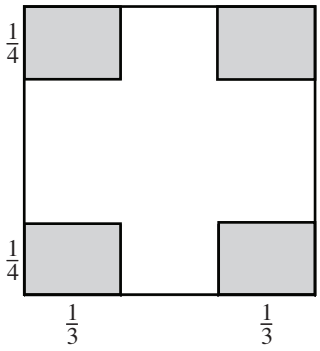


Figure 3. The maps of the IFS Φ defined in (3.4).

Every map of this IFS maps the unit square $[0,1]^2$ to itself. For a visualisation of that, see Figure 3. Let us observe that the orthogonal projections of the attractor X of Φ are self-similar sets, and X is the Cartesian product of these projections. More precisely, the orthogonal projection to the x-axis is the triadic-Cantor set $C_{1/3}$, and let us denote the projection to the y-axis by $C_{1/4}$. Then $X = C_{1/3} \times C_{1/4}$. Using dimension formula (2.2) for self-similar sets and the properties of Hausdorff dimension regarding product sets, we get that

$$\dim_H(X) = \dim_H(C_{1/3}) + \dim_H(C_{1/4}) = \frac{\log 2}{\log 3} + \frac{\log 2}{\log 4}.$$

However, it is easy to calculate the affinity dimension in this particular case. Namely, $A_i = \left(\begin{smallmatrix} 1/3 & 0 \\ 0 & 1/4 \end{smallmatrix}\right)$ for every $i \in \{1,2,3,4\}$, and so $\alpha_1(A_{\mathtt{i}}) = 3^{-|\mathtt{i}|}$ and $\alpha_2(A_{\mathtt{i}}) = 4^{-|\mathtt{i}|}$ for every $\mathtt{i} \in \Sigma_*$. Thus,

$$s_0(\Phi) = 1 + \frac{\log 4 - \log 3}{\log 4} > \frac{\log 2}{\log 3} + \frac{\log 2}{\log 4} = \dim_H(X).$$

The drop in the dimension here is due to the alignment structure. That is, since the sets $f_{\mathtt{i}}([0,1]^2)$ for high iterates are relatively long and thin rectangles positioned in columns above each other, a square of side length 3^{-n} in an optimal cover captures many cylinders at once. This leads us to a smaller dimension value than our naive cover, which covers every cylinder set individually.

Self-affine sets of this kind are collectively called *self-affine carpets* and have been intensively studied. Bedford [10] and McMullen [48] independently studied the following class: let $a > b \geq 2$ be integers, and let $A \subseteq \{0,\dots,b-1\} \times \{0,\dots,a-1\}$ be such that it contains at least two elements. Let

$$\Phi = \left\{ f_{(i,j)}(x,y) = \left(\frac{x+i}{b}, \frac{y+j}{a} \right) \right\}_{(i,j)\in A}.$$

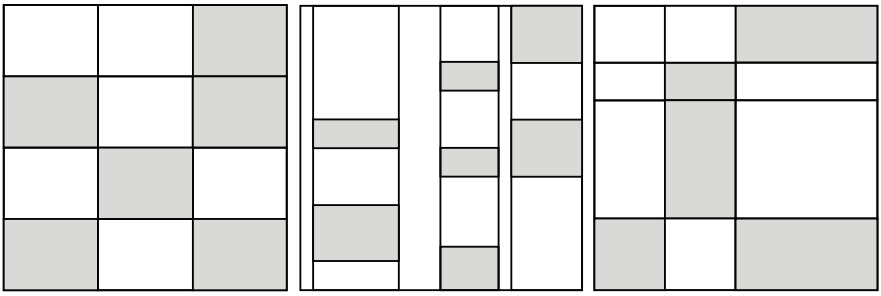


Figure 4. The images of the unit square in a Bedford–McMullen system, a Gatzouras–Lalley system and a Barański system.

For a visualisation of the images of the unit square, see the first image in Figure 4. Let A_1 be the set of $i \in \{0, \dots, b-1\}$ for which there exists j such that $(i, j) \in A$, and for every $i \in A_1$, let $T_i = \{j : (i, j) \in A\}$. Bedford [10] and McMullen [48] established the following dimension formulas:

$$\dim_B(X) = \log \#A_1 \left(\frac{1}{\log b} - \frac{1}{\log a} \right) + \frac{\log \#A}{\log a}$$

and

$$\dim_H(X) = \frac{\log\left(\sum_{i \in A_1} (\#T_i)^{\frac{\log b}{\log a}}\right)}{\log b}.$$

From the formulas above, it is clear that $\dim_H(X) = \dim_B(X)$ if and only if $\#T_i = \#T_j$ for every $i, j \in A_1$. Further simple algebraic manipulations show that

$$\dim_H(X) = s_0(\Phi) \quad \text{if and only if } \#A = \#A_1 \text{ or } \#T_i = \frac{\#A}{a} \text{ for every } i,$$

and

$$\dim_B(X) = s_0(\Phi) \quad \text{if and only if } \#A = \#A_1 \text{ or } \#A_1 = b.$$

This shows that Bedford–McMullen sets fall into the set of "exceptions" in the Falconer–Solomyak theorem for the majority of cases.

Without going into the details, we mention that several generalisations of Bedford–McMullen carpets have appeared. Lalley and Gatzouras [44] studied planar systems where they kept the alignment structure and the order of the contraction rates, but the width and height of the rectangles may vary. Barański [2] also kept the alignment structure in both x- and y-directions but not the order of the contractions. (See Figure 4 for their visualisation.) Kenyon and Peres [40] generalised the result of Bedford and McMullen for higher-dimensional systems. Feng and Wang [24] gave a formula of the box-counting dimension for general planar carpet-like systems in

which the linear parts of the maps are diagonal matrices and the images of the open unit square are disjoint, relying on the contraction ratios and the dimension of the projections to the coordinate axis, which was generalised by Kolossváry [41] for higher dimensions.

Recently, considerable attention has been paid to overlapping carpet-like systems beyond Falconer's and Solomyak's "typicality" results. Fraser and Shmerkin [27] studied overlapping Bedford–McMullen carpets, in which they allowed overlaps between the columns, relying on the result on self-similar systems by Hochman [29]. Very recently, using entropy growth theorems and additive combinatorial methods, Rapaport [58] studied overlapping carpet-like sets in arbitrary dimensions. For further discussion and dimension formulas, see, for example, [8, Section 11.5].

3.3. Self-affine systems in $\mathbb{R}^2$ and $\mathbb{R}^3$ with separation and irreducibility

It is clear from the previous section that we cannot expect easily verifiable conditions to provide equalities in (3.2) when the linear parts of the maps are diagonal. More generally, similar exceptional behaviour can be observed if the maps only preserve a common subspace corresponding to the strongest contraction rate, see, for example, Kolossváry and Simon [42], or they preserve a finite collection of subspaces; see Fraser [25]. To avoid these cases, let us introduce the concept of strong irreducibility.

Let $\mathcal{A} = \{A_i\}_{i\in\mathcal{I}}$ be a finite collection of $d \times d$ invertible matrices. We say that $\mathcal{A}$ is *strongly irreducible* if there is *no* finite collection $V_1, \dots, V_N$ of proper subspaces of $\mathbb{R}^d$ such that

$$A_i V_k \subset \bigcup_{j=1}^{N} V_j \quad \text{for every } i \in \mathcal{I} \text{ and } k \in \{1, \dots, N\}.$$

We say that a self-affine IFS $\Phi = \{x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ is strongly irreducible if the corresponding collection of matrices $\{A_i\}_{i\in\mathcal{I}}$ is strongly irreducible. It is clear that the systems discussed in Section 3.2 are not strongly irreducible. One would think naively that if we avoid alignments, and have separation, like an open set condition, we might achieve equality of the affinity dimension and the Hausdorff dimension of the attractor. Unfortunately, this is not the case. Let $\Phi = \{f_1(x) = A_1 x, f_2(x) = A_2 x\}$ be such that the matrices A_1, A_2 are invertible and have strictly positive entries and the images of the unit open square are disjoint $f_1((0,1)^2) \cap f_2((0,1)^2) = \emptyset$; see Figure 5. Clearly, the attractor $X = \{0\}$ is a single point; however, the affinity dimension of Φ is positive, since for every $\mathtt{i} \in \Sigma_*$, $\varphi^s(A_{\mathtt{i}}) \geq (\min_i \alpha_2(A_i))^{s|\mathtt{i}|}$ by the super-multiplicativity of the mini-norm. Thus, we need a stronger assumption than the open set condition.

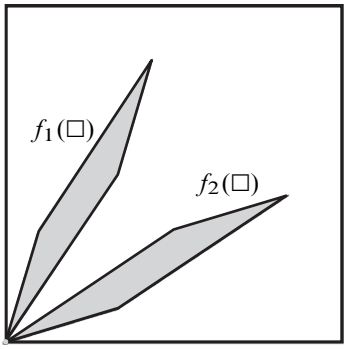


Figure 5. The IFS Φ satisfies the open set condition, but the attractor is a singleton.

We say that an IFS $\Phi = \{f_i\}_{i\in\mathcal{I}}$ satisfies the *strong open set condition (SOSC)* if there exists a non-empty, open and bounded set $U \subset \mathbb{R}^d$ such that

$$f_i(U) \subseteq U \quad \text{and} \quad f_i(U) \cap f_j(U) = \emptyset \quad \text{for every } i \neq j \in \mathcal{I} \text{ and } U \cap X \neq \emptyset,$$

where X is the attractor of Φ. Bandt and Graf [1] showed that the open set condition and the strong open set condition are equivalent for self-similar systems. This is clearly not the case for general self-affine systems.

Bárány, Hochman and Rapaport [4] showed that these two assumptions are enough for the dimension to be equal to the natural upper bounds in (3.2) and (3.3) on the plane.

Theorem 3.3. *Let $\Phi = \{f_i : x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ be a self-affine IFS on $\mathbb{R}^2$ such that it satisfies the strong open set condition and the strong irreducibility. Then for every self-affine measure $\nu = \pi_*\mu$, where μ is the corresponding Bernoulli measure,*

$$\dim(\nu) = \dim_L(\mu) \quad \textit{and} \quad \dim_H(X) = \overline{\dim}_B(X) = s_0(\Phi). \tag{3.5}$$

Note that the equation above indicates that the self-affine measure ν is exact dimensional; this and some ideas of the proof will be discussed later in Sections 3.4 and 3.5. Let us mention that in this form, the result of Bárány, Hochman and Rapaport [4] does not exclude that the IFS Φ consists only of similarities (or that the IFS Φ is conjugated to a self-similar IFS); however, that case already follows from the result of Hutchinson, Theorem 2.1. To avoid such redundancy, we introduce the concept of proximality. We say that a finite collection $\mathcal{A} = \{A_i\}_{i\in\mathcal{I}}$ of matrices is *proximal* if there is a finite product $A_{\mathtt{i}}$ for some $\mathtt{i} \in \Sigma_*$ which has maximal eigenvalue in modulus with multiplicity one. We use the same term similarly for the IFS, too.

Let us mention that the result of Bárány, Hochman and Rapaport [4] also remains valid in some not strongly irreducible cases, for example, for the graph of Takagi's function. Let us consider the parametrised family of maps

$$T_\lambda(x) := \sum_{n=0}^{\infty} \lambda^n d(2^n x, \mathbb{Z}),$$

where $d(., \mathbb{Z})$ denotes the distance from the closest integer. The graph of T_λ is a self-affine set invariant for the IFS

$$\Phi_\lambda = \left\{ f_1(x,y) = \left(\frac{x}{2}, \lambda y + \frac{x}{2}\right), f_2(x,y) = \left(\frac{x+1}{2}, \lambda y + \frac{1-x}{2}\right)\right\};$$

see the third image in Figure 1. It is clear that the IFS Φ is not strongly irreducible; however, the commonly preserved direction corresponds to the weaker contraction $\lambda > 1/2$, and the invariant directions of the strong contraction have a rich enough structure to avoid the occurrence of alignments. In particular, Bárány, Hochman and Rapaport [4, Corollary 7.6] showed the following: for every $\lambda \in (1/2, 1)$,

$$\dim_H(\text{graph}(T_\lambda)) = \overline{\dim}_B(\text{graph}(T_\lambda)) = s_0(\Phi_\lambda) = 2 - \frac{\log \lambda}{\log 2},$$

where $\text{graph}(T_\lambda)$ denotes the graph of T_λ over the interval $[0, 1]$.

The concept of exponential separation has been generalised for self-affine systems by Hochman and Rapaport [31], and they extended Theorem 3.3 for exponentially separated planar systems, omitting SOSC. Very recently, Rapaport [57] generalised the result on self-affine measures of Theorem 3.3 for systems in $\mathbb{R}^3$, and relying on that, Morris and Sert [51] established the result for the dimension of sets.

Theorem 3.4. *Let $\Phi = \{f_i : x \mapsto A_i x + t_i\}_{i \in \mathcal{I}}$ be a self-affine IFS on $\mathbb{R}^3$ such that it is proximal, strongly irreducible and it satisfies the strong open set condition. Then for every self-affine measure $\nu = \pi_* \mu$, where μ is the corresponding Bernoulli measure,*

$$\dim(\nu) = \dim_L(\mu) \quad \textit{and} \quad \dim_H(X) = \overline{\dim}_B(X) = s_0(\Phi). \tag{3.6}$$

Rapaport [57] also gained dimension results for self-affine measures in higher dimensions, but with additional, most likely technical assumptions. So, the problem is open in $\mathbb{R}^d$ for $d \geq 4$.

Let us briefly describe here the strategy of the proofs of Theorems 3.3 and 3.4 compared to the strategy of the sketch in Section 2 in the self-similar case. First, we need a sequence of self-affine measures, whose affinity dimension approximates the affinity dimension arbitrarily well, to show the results for the dimension of sets

in (3.5) and (3.6). This is needed because the methods of proving the dimension results on measures strongly rely on the structure of self-affine measures. We will discuss the possibility of this approximation in detail in Section 3.4.

Another essential tool is the so-called Ledrappier–Young formula, which allows us to study the dimension of the planar/spatial measure via its orthogonal projections. Namely, the dimension of a self-affine measure is the Lyapunov dimension if and only if the dimension of the orthogonal projections is as large as possible. We will discuss it in Section 3.5. However, the projections of ν share a lot of common properties with self-similar measures, and by proper approximations, the entropy growth theorem of Hochman [29] is applicable, which allows us to establish the maximality of the dimension of the projections in reasonably many directions.

Let us note that it is an open problem whether under the conditions of Theorems 3.3 and 3.4 we have $\dim(\pi_*\mu) = \dim_L(\mu)$ for a general left-shift invariant ergodic measure.

Another open problem is whether the dimension of self-affine sets can be achieved via a dimension-maximising ergodic left-shift invariant measure. That is, does (2.3) hold for self-affine systems? In general, the answer is negative. Das and Simmons [12] constructed a self-affine set X in $\mathbb{R}^3$ with a special alignment structure such that

$$\dim_H(X) > \sup\{\dim_H(\pi_*\mu) : \mu \text{ is ergodic and left-shift invariant}\};$$

hence, (2.3) does not hold in general. However, we do not expect this phenomenon for sets under certain irreducibility assumptions (like under the assumptions of Theorems 3.3 and 3.4). Still, we have only partial results on the existence of dimension-maximising measures; see Bárány and Rams [7].

3.4. Thermodynamic formalism

As we mentioned at the end of Section 3.3, the results on the dimension of the measures in Theorems 3.3 and 3.4 rely strongly on the structural properties of self-affine measures, so to verify the dimension results for the sets in (3.5) and (3.6), we need to find a proper way to approximate the affinity dimension $s_0(\Phi)$ via self-affine measures, possibly by taking a high iterate of the IFS Φ.

So let $\Phi = \{x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ be a self-affine IFS. As Falconer [18] pointed out, the singular value function φ^s, defined in (3.1), is sub-multiplicative, that is, for all A, B matrices, $\varphi^s(AB) \le \varphi^s(A)\varphi^s(B)$. Relying on this property, Käenmäki [36] showed that for every ergodic left-shift invariant measure μ,

$$s_0(\Phi) \ge \dim_L(\mu).$$

Moreover, there always exists a left-shift invariant ergodic measure μ such that the equality holds in the preceding formula. We call this measure *the Käenmäki measure*. Käenmäki and Vilppolainen [38] pointed out that the Käenmäki measure is not necessarily unique; however, under mild irreducibility assumptions, it is. More precisely, Feng and Käenmäki [23] showed that for planar IFS Φ, if the matrices are irreducible (there is no proper subspace V of $\mathbb{R}^2$ such that $A_i V = V$ for every $i \in \mathcal{I}$), then the Käenmäki measure is unique; furthermore, there exists a constant $C > 0$ such that for every $\mathtt{i} \in \Sigma_*$,

$$C^{-1}\varphi^{s_0(\Phi)}(A_\mathtt{i}) \le \mu([\mathtt{i}]) \le C\varphi^{s_0(\Phi)}(A_\mathtt{i}). \tag{3.7}$$

The uniqueness has been generalised for higher dimensions by Käenmäki and Morris [37] under certain irreducibility assumptions on the exterior product of matrices.

In general, it is not likely that the Käenmäki measure is a Bernoulli measure unless the linear parts of the IFS Φ have a very special structure. This has been recently verified by Morris and Sert [50].

Morris and Shmerkin [52] proved the following approximation result for planar self-affine IFS Φ with proximality, strong irreducibility: for every $\varepsilon > 0$, there exists $n \ge 1$ and a subset $\mathcal{G}_n \subset \Sigma_n$ such that the IFS $\Phi' = \{f_\mathtt{i}\}_{\mathtt{i}\in\mathcal{G}_n}$ is strongly irreducible and proximal; furthermore, the Lyapunov dimension of the self-affine measure

$$\nu' = \frac{1}{\#\mathcal{G}_n}\sum_{\mathtt{i}\in\mathcal{G}_n}(f_\mathtt{i})_*\nu'$$

is at least $s_0(\Phi) - \varepsilon$.

Property (3.7) has a crucial role in this statement. A possible natural strategy is to approximate the affinity dimension with the Lyapunov dimension of Bernoulli measures of higher order; one needs to take a sufficiently large n and a sufficiently large subset $\mathcal{G}_n \subset \Sigma_n$ such that

$$\#\mathcal{G}_n \approx e^{nh_\mu}, \quad \alpha_k(A_\mathtt{i}) \approx e^{-n\chi^k_\mu(\mu)} \quad \text{and} \quad \alpha_1(A_\mathtt{i}A_\mathtt{j}) \ge C\alpha_1(A_\mathtt{i})\alpha_1(A_\mathtt{j}) \text{ for all } \mathtt{i},\mathtt{j} \in \mathcal{G}_n,$$

where μ is the Käenmäki measure and C is a universal constant, independent of n. It is not trivial that such a constant $C > 0$ can be found, but in the proximal and strongly irreducible case, it is; see, for example, Bárány, Jordan, Käenmäki and Rams [5]. The approximation result for higher dimensions under certain irreducibility conditions was generalised recently by Morris and Sert [51].

3.5. The Ledrappier–Young formula

One of the main tools of the proofs of Theorems 3.3 and 3.4 is the exact dimensionality of self-affine measures and the so-called Ledrappier–Young formula. The problem

of the existence of local dimensions has a long history for self-affine measures and also in smooth dynamical systems. For an invariant measure in high-dimensional $C^{1+\alpha}$ systems, Ledrappier and Young [45, 46] proved the existence of the local dimensions along stable and unstable local manifolds and established a connection between conditional entropies, Lyapunov exponents and local dimensions. The adaptation of their method for self-affine systems remained an open question for almost three decades. The first special case was established by Feng and Hu [22] for self-affine systems, where the matrices were diagonal.

Let us illustrate the problem for planar self-affine systems with proximality, strong irreducibility and the strong open set condition. For a proper subspace $V \subset \mathbb{R}^2$, let us denote the orthogonal projection by $\mathrm{proj}_V \colon \mathbb{R}^2 \to V$. The projective circle, that is, the set of all proper subspaces of $\mathbb{R}^2$, is denoted by $\mathbb{P}\mathbb{R}^1$. A 2×2 invertible matrix A naturally defines an action on the projective circle $\mathbb{P}\mathbb{R}^1$, which is also denoted by A with a slight abuse of notation. Let $\nu = \pi_* \mu$ be a self-affine measure with the corresponding Bernoulli measure μ. Furthermore, let $\mathtt{i} \in \Sigma$ be a "μ-typical" point. Then take the ball $B(\pi(\mathtt{i}), \alpha_2(A_{\mathtt{i}|_n}))$ centred at $\pi(\mathtt{i})$ with radius $\alpha_2(A_{\mathtt{i}|_n})$. Heuristically, it intersects only the cylinder $f_{\mathtt{i}|_n}(X)$. Let $V \in \mathbb{P}\mathbb{R}^1$ be a subspace close to the shortest main semi-axis of $A_{\mathtt{i}|_n}(B(0,1))$. Using the invariance of the measure μ and (1.4), we get that

$$\begin{aligned} \nu(B(\pi(\mathtt{i}), \alpha_2(A_{\mathtt{i}|_n}))) &\approx \nu(f_{\mathtt{i}|_n}(X) \cap B(\pi(\mathtt{i}), \alpha_2(A_{\mathtt{i}|_n}))) \\ &\approx p_{\mathtt{i}|_n} \nu(f_{\mathtt{i}|_n}^{-1}(B(\pi(\mathtt{i}), \alpha_2(A_{\mathtt{i}|_n})))) \\ &\approx p_{\mathtt{i}|_n} \nu\left(\mathrm{proj}^{-1}_{A^T_{\mathtt{i}|_n} V} B\left(\mathrm{proj}_{A^T_{\mathtt{i}|_n} V} \pi(\sigma^n \mathtt{i}), \frac{\alpha_2(A_{\mathtt{i}|_n})}{\alpha_1(A_{\mathtt{i}|_n})}\right)\right). \end{aligned} \tag{3.8}$$

For a visualisation, see Figure 6.

Hence, studying the local dimension of ν requires understanding the local dimension of $(\mathrm{proj}_V)_* \nu$. But observe that as n goes to infinity, the direction of the orthogonal projection $A^T_{\mathtt{i}|_n} V$ takes a random walk on the projective circle similarly to the chaos game defined in Section 1.1. For a proximal and strongly irreducible collection of matrices $\mathcal{A}^T = \{A_i^T\}_{i \in \mathcal{I}}$, there exists a unique stationary Borel probability measure ν_F, called *the Furstenberg–Kesten measure*, with strictly positive Hausdorff dimension such that

$$\nu_F = \sum_{i \in \mathcal{I}} p_i (A_i^T)_* \nu_F.$$

For a proof, see, for example, the book of Bougerol and Lacroix [11]. The name of the measure originates back to the paper by Furstenberg and Kesten [28], who studied the Lyapunov exponents of random matrix products and gave a formula for it by using ν_F.

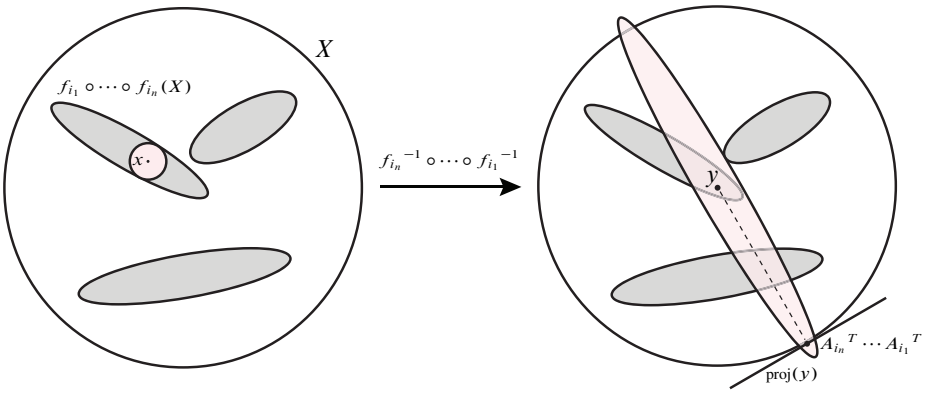


Figure 6. Applying the inverses of the maps of the IFS on a ball centred at $x = \pi(\mathtt{i})$ with radius $\alpha_2(A_{\mathtt{i}|_n})$, which is mapped to an ellipse centred at $y = \pi(\sigma^n \mathtt{i})$ comparable with $\mathrm{proj}^{-1}_{A^T_{\mathtt{i}|_n} V}\, B\big(\mathrm{proj}_{A^T_{\mathtt{i}|_n} V}\, \pi(\sigma^n \mathtt{i}), \frac{\alpha_2(A_{\mathtt{i}|_n})}{\alpha_1(A_{\mathtt{i}|_n})}\big) \cap X$.

Taking the logarithm on both sides of (3.8) and dividing by $\log \alpha_2(A_{\mathtt{i}|_n})$, the heuristic argument gives us the following, proven by Bárány and Käenmäki [6].

Theorem 3.5. *Let $\Phi = \{x \mapsto A_i x + t_i\}_{i \in \mathcal{I}}$ be a planar self-affine IFS such that it is proximal, strongly irreducible and satisfies the strong open set condition. Let $\nu = \pi_* \mu$ be a fully supported self-affine measure (i.e., $p_i > 0$ for every $i \in \mathcal{I}$) with Bernoulli measure μ. Then every fully supported self-affine measure ν is exact dimensional; moreover, for ν_F-almost every $V \in \mathbb{P}\mathbb{R}^1$, the projected measure $(\mathrm{proj}_V)_* \nu$ is exact dimensional and*

$$\dim(\nu) = \frac{h_\mu}{\chi^2_\mu(\Phi)} + \left(1 - \frac{\chi^1_\mu(\Phi)}{\chi^2_\mu(\Phi)}\right) \dim((\mathrm{proj}_V)_* \nu) \quad \textit{for } \nu_F\textit{-almost every } V.$$

For a more detailed discussion on the Furstenberg–Kesten measure and the Ledrappier–Young formula, see [8].

Theorem 3.5 has significant consequences. It is easy to see that

$$\dim(\nu) = \dim_L(\mu) \quad \text{if and only if } \dim((\mathrm{proj}_V)_* \nu) = \min\left\{1, \frac{h_\mu}{\chi^1_\mu(\Phi)}\right\} \text{ for } \nu_F\text{-a.e. } V.$$

Hence, to study the dimension of ν, one needs to understand the dimension of typical projections of ν. Several results appeared before Theorem 3.3 using the Ledrappier–Young formula, which were relying on the dimension estimates on exceptional sets of orthogonal projections of measures by Marstrand [47], Kaufman [39] and Falconer [17] and studying the dimension of ν_F; see, for example, Bárány [3], Bárány and Rams [7], Falconer and Kempton [16] and Rapaport [56]. While the proof

of Theorem 3.3 relies on the Ledrappier–Young formula and the observation that the projections of ν share a lot of common properties with self-similar measures, and by proper approximations, the entropy growth theorem of Hochman [29] is applicable.

Finally, let us mention that Feng [21] generalised the Ledrappier–Young formula in a very wide generality and established the following regularity property.

Theorem 3.6. *Let $\Phi = \{x \mapsto A_i x + t_i\}_{i\in\mathcal{I}}$ be a self-affine IFS on $\mathbb{R}^d$, and let μ be an ergodic measure on Σ. Then the measure $\pi_*\mu$ is exact dimensional.*

Acknowledgements. The author would like to thank Antti Käenmäki for his helpful and valuable suggestions. The author also acknowledges support from grants NKFI FK134251, K142169, and KKP144059 "Fractal geometry and applications" Research Group.

References

[1] C. Bandt and S. Graf, Self-similar sets 7. A characterization of self-similar fractals with positive Hausdorff measure. *Proc. Amer. Math. Soc.* **114** (1992), no. 4, 995–1001 Zbl 0823.28003 MR 1100644

[2] K. Barański, Hausdorff dimension of the limit sets of some planar geometric constructions. *Adv. Math.* **210** (2007), no. 1, 215–245 Zbl 1116.28008 MR 2298824

[3] B. Bárány, On the Ledrappier–Young formula for self-affine measures. *Math. Proc. Cambridge Philos. Soc.* **159** (2015), no. 3, 405–432 Zbl 1371.28015 MR 3413884

[4] B. Bárány, M. Hochman, and A. Rapaport, Hausdorff dimension of planar self-affine sets and measures. *Invent. Math.* **216** (2019), no. 3, 601–659 Zbl 1414.28014 MR 3955707

[5] B. Bárány, T. Jordan, A. Käenmäki, and M. Rams, Birkhoff and Lyapunov spectra on planar self-affine sets. *Int. Math. Res. Not. IMRN* **2021** (2021), no. 10, 7966–8005 Zbl 1479.37039 MR 4259165

[6] B. Bárány and A. Käenmäki, Ledrappier–Young formula and exact dimensionality of self-affine measures. *Adv. Math.* **318** (2017), 88–129 Zbl 1457.37032 MR 3689737

[7] B. Bárány and M. Rams, Dimension maximizing measures for self-affine systems. *Trans. Amer. Math. Soc.* **370** (2018), no. 1, 553–576 Zbl 1458.28003 MR 3717989

[8] B. Bárány, K. Simon, and B. Solomyak, *Self-similar and self-affine sets and measures*. Math. Surveys Monogr. 276, American Mathematical Society, Providence, RI, 2023, 451 pp. Zbl 1543.28001 MR 4661364

[9] M. F. Barnsley, *Fractals everywhere*. Academic Press, Boston, MA, 1988, 394 pp. Zbl 0691.58001 MR 0977274

[10] T. J. Bedford, *Crinkly curves, Markov partitions and dimensions*. Ph.D. thesis, University of Warwick, 1984

[11] P. Bougerol and J. Lacroix, *Products of random matrices with applications to Schrödinger operators*. Progr. Probab. 8, Birkhäuser, Boston, 1985, 284 pp. Zbl 0572.60001 MR 0886674

[12] T. Das and D. Simmons, The Hausdorff and dynamical dimensions of self-affine sponges: a dimension gap result. *Invent. Math.* **210** (2017), no. 1, 85–134 Zbl 1387.37026 MR 3698340

[13] G. A. Edgar, Fractal dimension of self-affine sets: some examples. *Rend. Circ. Mat. Palermo (2)* **28** (1992), 341–358 Zbl 0762.28005 MR 1183060

[14] G. A. Edgar, *Integral, probability, and fractal measures*. Springer, New York, 1998, 286 pp. Zbl 0893.28001 MR 1484412

[15] K. Falconer, *Fractal geometry. Mathematical foundations and applications*. John Wiley & Sons, Chichester, 1990, 288 pp. Zbl 0689.28003 MR 1102677

[16] K. Falconer and T. Kempton, Planar self-affine sets with equal Hausdorff, box and affinity dimensions. *Ergodic Theory Dynam. Systems* **38** (2018), no. 4, 1369–1388 Zbl 1388.37028 MR 3789169

[17] K. J. Falconer, Hausdorff dimension and the exceptional set of projections. *Mathematika* **29** (1982), no. 1, 109–115 Zbl 0477.28004 MR 0673510

[18] K. J. Falconer, The Hausdorff dimension of self-affine fractals. *Math. Proc. Cambridge Philos. Soc.* **103** (1988), no. 2, 339–350 Zbl 0642.28005 MR 0923687

[19] K. J. Falconer, Dimensions and measures of quasi self-similar sets. *Proc. Amer. Math. Soc.* **106** (1989), no. 2, 543–554 Zbl 0683.58034 MR 0969315

[20] K. J. Falconer, The dimension of self-affine fractals II. *Math. Proc. Cambridge Philos. Soc.* **111** (1992), no. 1, 169–179 Zbl 0797.28004 MR 1131488

[21] D.-J. Feng, Dimension of invariant measures for affine iterated function systems. *Duke Math. J.* **172** (2023), no. 4, 701–774 Zbl 1564.28002 MR 4557759

[22] D.-J. Feng and H. Hu, Dimension theory of iterated function systems. *Comm. Pure Appl. Math.* **62** (2009), no. 11, 1435–1500 Zbl 1230.37031 MR 2560042

[23] D.-J. Feng and A. Käenmäki, Equilibrium states of the pressure function for products of matrices. *Discrete Contin. Dyn. Syst.* **30** (2011), no. 3, 699–708 Zbl 1219.37027 MR 2784616

[24] D.-J. Feng and Y. Wang, A class of self-affine sets and self-affine measures. *J. Fourier Anal. Appl.* **11** (2005), no. 1, 107–124 Zbl 1091.28005 MR 2128947

[25] J. M. Fraser, On the packing dimension of box-like self-affine sets in the plane. *Nonlinearity* **25** (2012), no. 7, 2075–2092 Zbl 1247.28006 MR 2947936

[26] J. M. Fraser, *Assouad dimension and fractal geometry*. Cambridge Tracts in Math. 222, Cambridge University Press, Cambridge, 2020, 269 pp. Zbl 1467.28001 MR 4411274

[27] J. M. Fraser and P. Shmerkin, On the dimensions of a family of overlapping self-affine carpets. *Ergodic Theory Dynam. Systems* **36** (2016), no. 8, 2463–2481 Zbl 1441.37031 MR 3570020

[28] H. Furstenberg and H. Kesten, Products of random matrices. *Ann. Math. Statist.* **31** (1960), no. 2, 457–469 Zbl 0137.35501 MR 0121828

[29] M. Hochman, On self-similar sets with overlaps and inverse theorems for entropy. *Ann. of Math. (2)* **180** (2014), no. 2, 773–822 Zbl 1337.28015 MR 3224722

[30] M. Hochman, Dimension theory of self-similar sets and measures. In *Proceedings of the International Congress of Mathematicians – ICM 2018 (Rio de Janeiro 2018)*, pp. 1949–1972, Invited lectures III, World Scientific, Hackensack, NJ, 2019 Zbl 1447.28007 MR 3966837

[31] M. Hochman and A. Rapaport, Hausdorff dimension of planar self-affine sets and measures with overlaps. *J. Eur. Math. Soc. (JEMS)* **24** (2022), no. 7, 2361–2441 Zbl 1502.28005 MR 4413769

[32] I. Hueter and S. P. Lalley, Falconer's formula for the Hausdorff dimension of a self-affine set in $\mathbf{R}^2$. *Ergodic Theory Dynam. Systems* **15** (1995), no. 1, 77–97 Zbl 0867.28006 MR 1314970

[33] J. E. Hutchinson, Fractals and self similarity. *Indiana Univ. Math. J.* **30** (1981), no. 5, 713–747 Zbl 0598.28011 MR 0625600

[34] T. Jordan, M. Pollicott, and K. Simon, Hausdorff dimension for randomly perturbed self affine attractors. *Comm. Math. Phys.* **270** (2007), no. 2, 519–544 Zbl 1119.28004 MR 2276454

[35] T. Jordan and A. Rapaport, Dimension of ergodic measures projected onto self-similar sets with overlaps. *Proc. Lond. Math. Soc. (3)* **122** (2021), no. 2, 191–206 Zbl 1465.28007 MR 4218958

[36] A. Käenmäki, On natural invariant measures on generalised iterated function systems. *Ann. Acad. Sci. Fenn. Math.* **29** (2004), no. 2, 419–458 Zbl 1078.37014 MR 2097242

[37] A. Käenmäki and I. D. Morris, Structure of equilibrium states on self-affine sets and strict monotonicity of affinity dimension. *Proc. Lond. Math. Soc. (3)* **116** (2018), no. 4, 929–956 Zbl 1391.28009 MR 3789836

[38] A. Käenmäki and M. Vilppolainen, Dimension and measures on sub-self-affine sets. *Monatsh. Math.* **161** (2010), no. 3, 271–293 Zbl 1206.28013 MR 2726214

[39] R. Kaufman, On Hausdorff dimension of projections. *Mathematika* **15** (1968), no. 2, 153–155 Zbl 0165.37404 MR 0248779

[40] R. Kenyon and Y. Peres, Measures of full dimension on affine-invariant sets. *Ergodic Theory Dynam. Systems* **16** (1996), no. 2, 307–323 Zbl 0851.58028 MR 1389626

[41] I. Kolossváry, The L^q spectrum of self-affine measures on sponges. *J. Lond. Math. Soc. (2)* **108** (2023), no. 2, 666–701 Zbl 1549.37013 MR 4626722

[42] I. Kolossváry and K. Simon, Triangular Gatzouras–Lalley-type planar carpets with overlaps. *Nonlinearity* **32** (2019), no. 9, 3294–3341 Zbl 1426.28019 MR 3987800

[43] U. Krengel, *Ergodic theorems*. De Gruyter Stud. Math. 6, De Gruyter, Berlin, 1985, 357 pp. Zbl 0575.28009 MR 0797411

[44] S. P. Lalley and D. Gatzouras, Hausdorff and box dimensions of certain self-affine fractals. *Indiana Univ. Math. J.* **41** (1992), no. 2, 533–568 Zbl 0757.28011 MR 1183358

[45] F. Ledrappier and L.-S. Young, The metric entropy of diffeomorphisms. Part I: characterization of measures satisfying Pesin's entropy formula. *Ann. of Math. (2)* **122** (1985), no. 3, 509–539 Zbl 0605.58028 MR 0819556

[46] F. Ledrappier and L.-S. Young, The metric entropy of diffeomorphisms. Part II: relations between entropy, exponents and dimension. *Ann. of Math. (2)* **122** (1985), no. 3, 540–574 Zbl 1371.37012 MR 0819557

[47] J. M. Marstrand, Some fundamental geometrical properties of plane sets of fractional dimensions. *Proc. Lond. Math. Soc. (3)* **4** (1954), no. 1, 257–302 Zbl 0056.05504 MR 0063439

[48] C. McMullen, The Hausdorff dimension of general Sierpiński carpets. *Nagoya Math. J.* **96** (1984), 1–9 Zbl 0539.28003 MR 0771063

[49] I. D. Morris, Fast approximation of the affinity dimension for dominated affine iterated function systems. *Ann. Fenn. Math.* **47** (2022), no. 2, 645–694 Zbl 1505.28010 MR 4407233

[50] I. D. Morris and Ç. Sert, A converse statement to Hutchinson's theorem and a dimension gap for self-affine measures. *J. Eur. Math. Soc. (JEMS)* **25** (2023), no. 11, 4315–4367 Zbl 1550.28007 MR 4662294

[51] I. D. Morris and C. Sert, A variational principle relating self-affine measures to self-affine sets. 2023, arXiv:2303.03437v2

[52] I. D. Morris and P. Shmerkin, On equality of Hausdorff and affinity dimensions, via self-affine measures on positive subsystems. *Trans. Amer. Math. Soc.* **371** (2019), no. 3, 1547–1582 Zbl 1406.28005 MR 3894027

[53] M. Pollicott and P. Vytnova, Estimating singularity dimension. *Math. Proc. Cambridge Philos. Soc.* **158** (2015), no. 2, 223–238 Zbl 1371.28010 MR 3310242

[54] F. Przytycki and M. Urbański, On the Hausdorff dimension of some fractal sets. *Studia Math.* **93** (1989), no. 2, 155–186 Zbl 0691.58029 MR 1002918

[55] F. Przytycki and M. Urbański, *Conformal fractals: ergodic theory methods*. London Math. Soc. Lecture Note Ser. 371, Cambridge University Press, Cambridge, 2010, 354 pp. Zbl 1202.37001 MR 2656475

[56] A. Rapaport, On self-affine measures with equal Hausdorff and Lyapunov dimensions. *Trans. Amer. Math. Soc.* **370** (2018), no. 7, 4759–4783 Zbl 1386.37021 MR 3812095

[57] A. Rapaport, On self-affine measures associated to strongly irreducible and proximal systems. *Adv. Math.* **449** (2024), article no. 109734, 116 pp. Zbl 1567.28018 MR 4749342

[58] A. Rapaport, Dimension of diagonal self-affine sets and measures via nonconformal partitions. *Proc. Lond. Math. Soc. (3)* **132** (2026), no. 1, article no. e70120, 39 pp. Zbl 08159147 MR 5020586

[59] E. Rossi, Local dimensions of measures on infinitely generated self-affine sets. *J. Math. Anal. Appl.* **413** (2014), no. 2, 1030–1039 Zbl 1310.28008 MR 3159819

[60] B. Solomyak, Measure and dimension for some fractal families. *Math. Proc. Cambridge Philos. Soc.* **124** (1998), no. 3, 531–546 Zbl 0927.28006 MR 1636589

[61] P. P. Varjú, Self-similar sets and measures on the line. In *International Congress of Mathematicians (ICM 2022), Helsinki, Finland, July 6–14, 2022*, Vol. 5. Sections 9–11, pp. 3610–3634, EMS Press, Berlin, 2023 Zbl 1561.28103 MR 4680376

[62] P. Walters, *An introduction to ergodic theory*. Grad. Texts in Math. 79, Springer, New York, 1982, 250 pp. Zbl 0475.28009 MR 0648108

Balázs Bárány
Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics, Műegyetem rkp. 3., 1111 Budapest, Hungary; barany.balazs@ttk.bme.hu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/12

Applications of the renormalisation group

Roland Bauerschmidt

Abstract. The renormalisation group provides a powerful point of view to understand random systems with strong correlations. This review discusses some applications to spin systems and related lattice models obtained in recent years.

1. What is the renormalisation group?

The general idea of the renormalisation group is to study a physical system in terms of effective models for different length scales, which roughly take the same form as the original model except that various parameters are adjusted or *renormalised*. The change of length scale defines the renormalisation group dynamics. This approach is particularly relevant for systems that are interacting and have strong correlations. Strong correlations mean that many length scales contribute to the behaviour of the system in a significant way, and interacting means that these scales have a nontrivial interaction, or dynamics. An idealised goal is given by Kadanoff's block spin picture for the Ising model, illustrated in Figure 1.

Kenneth Wilson (who received the 1982 Nobel Prize in Physics for his achievements in developing the renormalisation group picture) described Kadanoff's block spin picture as follows [38]:

> *In short, the Kadanoff block picture, although absurd, will be the basis for generalizations which are not absurd, and it is helpful to understand the Kadanoff picture in differential form before studying these generalizations.*

The theoretical scope of the renormalisation group approach is rather general, but its roots and its most complete implementations to date remain in quantum and statistical field theory. The procedure of passing from a length scale to a larger one while

Mathematics Subject Classification 2020: 82B28 (primary); 60K35 (secondary).
Keywords: renormalisation, spin systems.

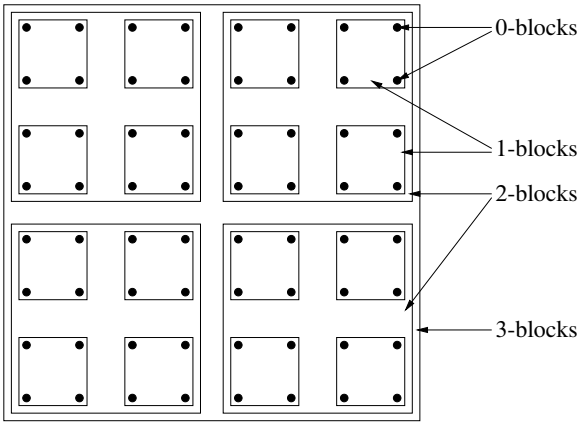


Figure 1. Block spin picture.

forgetting the microscopic details (or more precisely an implementation of this procedure) is called the *renormalisation group map*, and the point of view is to study this map as a dynamical system on the 'space of systems'.

2. Spin systems and their strongly correlated regimes

This review focuses on applications of the renormalisation group to lattice spin models. Continuum models could be included as limits $\varepsilon \to 0$ of models on rescaled lattices of mesh size ε, but we will not focus on these. In this section, we will present a high-level overview of some of the regimes in which spin systems can develop strong correlations and where the renormalisation group approach therefore becomes relevant.

From now on, $\Lambda \subset \mathbb{Z}^d$ denotes a finite set of sites that eventually tends to $\mathbb{Z}^d$. In continuum models, one would similarly take $\Lambda \subset \varepsilon\mathbb{Z}^d$ with the understanding that $\varepsilon \to 0$. Typically, we assume that Λ is a cube with periodic boundary conditions (a torus); the interesting topic of boundary conditions is mostly undeveloped in the renormalisation group context.

To each site $x \in \Lambda$ we associate a spin φ_x. The spins can have internal degrees of freedom (i.e., they could be vector valued), and their components can be 'bosonic' or 'fermionic'. The bosonic components are normally ordinary real numbers, while fermionic spins are odd elements of a Grassmann algebra.

The expectations of a spin system are roughly of the form

$$\langle F \rangle = \frac{1}{Z} e^{\frac{1}{2}(\varphi, \Delta\varphi) - \sum_{x \in \Lambda} V(\varphi_x)} \prod_{x \in \Lambda} d\varphi_x, \tag{2.1}$$

where Δ is the (negative definite) lattice Laplace operator:

$$\Delta f(x) = \sum_{y \sim x} (f(y) - f(x)), \tag{2.2}$$

and V is a local potential, and $d\varphi_x$ is an appropriate a priori measure (such as the Lebesgue measure). In several examples, we consider that the local potential is replaced by a nonlinear constraint on the spins. Constraints can be seen as degenerate potentials.

Some examples and regimes where strong correlation occur are given as follows. Precise results for representative models in all regimes are given in Section 3.

2.1. Critical temperature

Spin systems develop strong correlations over long distances near critical temperatures of continuous phase transitions.

The prototypical example is the Ising model, given at inverse temperature $\beta > 0$ (and assuming there is no external field) by

$$\langle F \rangle_\beta \propto \sum_{\sigma \in \{\pm 1\}^\Lambda} F(\sigma)\, e^{\frac{1}{2}\beta(\sigma, \Delta\sigma)}. \tag{2.3}$$

The closely related Ginzburg–Landau φ^4 model is defined for $g > 0$ and $\nu \in \mathbb{R}$ by

$$\langle F \rangle_{g,\nu} \propto \int_{\mathbb{R}^\Lambda} F(\varphi)\, e^{\frac{1}{2}(\varphi, \Delta\varphi) - \sum_{x\in\Lambda} \frac{1}{4} g \varphi_x^4 - \sum_{x \in \Lambda} \frac{1}{2} \nu \varphi_x^2} \prod_{x \in \Lambda} d\varphi_x. \tag{2.4}$$

For the Ginzburg–Landau φ^4 model, we consider $g > 0$ fixed and the parameter ν then takes the role of an effective temperature. Both models are expected to have essentially identical critical behaviour. In particular, both models have critical temperatures, denoted β_c and $\nu_c(g)$, respectively, which separate an ordered low-temperature phase and a disordered high-temperature phase, both of which have exponential truncated correlation decay. At and near the critical temperature, strong correlations develop. These can be quantified through various correlation functions and critical exponents. For example, in the Ising model, it is known that for $d \neq 3$ (and expected also in $d = 3$),

$$\langle \sigma_0 \sigma_x \rangle_{\beta_c} \approx |x|^{-(d-2+\eta)}. \tag{2.5}$$

The critical exponent η is 0 for $d > 4$; more delicately, it is also 0 in $d = 4$ (without logarithmic correlation, which, however, appear in closely related quantities), and it takes (or is expected to take) a nonzero value in $d = 2, 3$. The exact same exponents are expected for the φ^4 model.

The rigorous renormalisation group approach has, in particular, been used to study such models in the upper critical dimension $d = 4$. For long-range models

in which the nearest-neighbour Laplacian $-\Delta$ is replaced by a fractional power of $-\Delta$ (which then has power-law decay), models slightly below the upper critical dimension (which one may think of as a version of dimension '$d = 4 - \varepsilon$' but with important features missing) can also be studied; this is a version of the Wilson–Fisher 'ε-expansion' [39].

The critical φ^4 type problem has been a prototypical problem considered in particular in [31, 35, 36] and more closely related to this review in [9–11, 22–25].

2.2. Low temperature

In spin systems with continuous symmetries, strong correlations are expected not only at their critical temperatures, but also in their low-temperature phases, where the continuous symmetry is spontaneously broken.

The prototypical example of a spin system with continuous symmetry is the classical Heisenberg model (or $O(n)$ model), defined by

$$\langle F \rangle_\beta \propto \int_{\sigma \in (\mathbb{S}^{n-1})^\Lambda} F(\sigma)\, e^{\frac{1}{2}\beta(\sigma, \Delta\sigma)} \prod_{x \in \Lambda} d\sigma_x. \tag{2.6}$$

Parallel to the situation of the Ising and φ^4 models, one can alternatively consider the n-component Ginzburg–Landau version of the $O(n)$ model, and not much is expected to change in the long-distance behaviour. At high temperatures, standard high temperature expansion methods show that correlations decay exponentially. Different from the situation of discrete spins (or discrete symmetry), the low-temperature phase is now strongly correlated as well. At low temperature, spins typically align in one direction (spontaneous symmetry breaking), but fluctuations orthogonal to the direction of broken symmetry are energetically inexpensive. These strong fluctuations orthogonal to the direction of broken symmetry are known as the Goldstone mode; see [32] for a review. For example, in $d \geq 3$ and $\beta > \beta_0$, it is the case that

$$\langle \sigma_0 \cdot \sigma_x \rangle_\beta \approx m(\beta) + \frac{c(\beta)}{\beta |x|^{d-2}}. \tag{2.7}$$

This was proven in [1–7], but a more tractable renormalisation group analysis of these models with nonabelian symmetry remains an outstanding problem. In some situation, reflection positivity and correlation inequalities can provide softer results [18–20, 33].

2.3. High temperature

In spin systems with noncompact symmetry groups, strong correlations can also occur in their high-temperature phases corresponding to unbroken symmetry.

The prototypical example is the two-dimensional Discrete Gaussian model, which can formally be regarded as the lattice Gaussian free field conditioned to be integer valued and is a model for crystal interface in $2+1$ dimensions. The expectation of the Discrete Gaussian model is defined for $F : \mathbb{Z}^\Lambda \to \mathbb{R}$ and inverse temperature $\beta > 0$ by

$$\langle F \rangle_\beta \propto \sum_{\sigma \in \mathbb{Z}^\Lambda : \sigma_0 = 0} F(\varphi)\, e^{\frac{1}{2}\beta(\sigma, \Delta\sigma)}. \tag{2.8}$$

As for the previous models, there is a closely related Ginzburg–Landau version of the Discrete Gaussian model, known as the lattice sine-Gordon model. It is again expected to have essentially identical long-distance behaviour. It is defined for $\beta > 0$ and $z \in \mathbb{R}$ by the expectation

$$\langle F \rangle_{\beta,z} \propto \int_{\mathbb{R}^\Lambda : \varphi_0 = 0} F(\varphi)\, e^{\frac{1}{2}(\varphi, \Delta\varphi) - \sum_{x \in \Lambda} 2z \cos(\sqrt{\beta}\varphi)}\, d\varphi. \tag{2.9}$$

In our conventions, the parameter β of the Discrete Gaussian model corresponds to $1/\beta$ in the sine-Gordon model. The reason for this convention is that sine-Gordon model has another interpretation as a Coulomb gas at inverse temperature β, and one can recover the Discrete Gaussian model in an appropriate $z \to \infty$ limit (which is analogous to the $g \to \infty$ limit in which the Ising model can be recovered from the φ^4 model). In two dimensions, these models have strong fluctuations at high temperature where they are close to an ordinary Gaussian free field, and the renormalisation group approach is particularly relevant in this regime. For example, for the Discrete Gaussian model on $\mathbb{Z}^2$, $\beta \le \beta_0$, variances behave as

$$\langle (\varphi_x - \varphi_0)^2 \rangle_\beta \approx c(\beta) \log|x|, \tag{2.10}$$

whereas $\langle (\varphi_x - \varphi_0)^2 \rangle_\beta$ is bounded for large β. This transition is the roughening transition of the crystal interface, or Kosterlitz–Thouless transition. The mathematical understanding of this transition was pioneered in [29].

2.4. Continuum limits and other remarks

Continuum limits. In continuum limits of the above systems (when they exist), strong correlations also occur at short distances. As a prototypical example, we consider the continuum φ^4_d measure which exists if $d < 4$ as the $\varepsilon \to 0$ limit (on a suitable space of generalised functions or in the sense of correlation functions) of the regularised measure with expectation

$$\langle F \rangle^\varepsilon_{\lambda,\mu} = \int_{\mathbb{R}^{\Lambda_\varepsilon}} F(\varphi) e^{-\varepsilon^d \sum_{x \in \Lambda_\varepsilon} \varphi_x (-\Delta^\varepsilon \varphi)_x - \varepsilon^d \sum_{x \in \Lambda_\varepsilon} (\frac{1}{4}\lambda\varphi^4 + \frac{1}{2}(\mu + a_\varepsilon(\lambda))\varphi^2)} \prod_{x \in \Lambda_\varepsilon} d\varphi_x, \tag{2.11}$$

where $a_\varepsilon(\lambda)$ is a divergent (as $\varepsilon \to 0$) counterterm needed to obtain a nontrivial continuum limit $\langle\cdot\rangle_{\lambda,\mu}$. These continuum limits satisfy, for example in $d = 3$,

$$\langle\varphi_x\varphi_y\rangle_{\lambda,\mu} \approx \frac{c}{|x-y|^{d-2}} \quad (|x-y| \to 0). \tag{2.12}$$

To connect the discussion of lattice models with continuum models, note that after rescaling the field φ by $\varepsilon^{(d-2)/2}$, the ε-regularised continuum model takes the form of a model on the unit lattice, but with coupling constants that depend on ε. In fact, the corresponding lattice coupling constant $g = \varepsilon^d\lambda$ vanishes as $\varepsilon \to 0$. This is compensated by the choice of the counterterm, which in terms of the lattice model ensures that the parameters are close to a critical point (strongly correlated), so that a macroscopic effect can survive in the limit. From the perspective of a classical continuum theory (i.e., normalising with respect to the Riemann sum $\varepsilon^d \sum_x$ instead of $\sum_x$), on the other hand, the coupling constant diverges. This is an example of the infamous infinities in quantum field theory.

Combinatorial models. Various more combinatorial models turn out to be very closely related to the above or similar models. Examples include self-avoiding walks (exactly described by a version of the lattice φ^4 model), random forests (exactly described by a particular lattice nonlinear sigma model), and interacting dimers (exactly described by a discrete version of the massless Thirring model). Other models such as percolation have completely parallel behaviour, but exact field theory formulations are not known.

3. Overview of some results

We discuss three models in which we exemplify the use of the renormalisation group to analyse the emergence of strong correlations in low-, high-, and critical-temperature regimes, as discussed in the previous section. These models are all studied in a framework for the renormalisation group going back to [21] and much further developed since then. We refer to [12] for an introduction and further references.

In all examples discussed in this section, $\Lambda = \Lambda_N$ denotes a d-dimensional discrete torus of side length L^N for integers $L > 1$ and $N \geq 1$, and L will always be chosen sufficiently large (but fixed). The infinite volume limit will be taken by taking $N \to \infty$.

3.1. High temperature: The Discrete Gaussian model

The two-dimensional *Discrete Gaussian model* is a model for a crystal interface (in $2+1$ dimensions) and has a phase transition between a smooth (localised) low-temperature phase and a rough (delocalised) high-temperature phase; a physics textbook treatment can be found in [27, Chapter 6]. The mathematical understanding of this model was pioneered by Fröhlich–Spencer [34]. Our results [15, 16] with Park and Rodriguez establish the scaling limit at high temperatures using the renormalisation group, making use of ideas previously developed by Dimock–Hurd [28] and Falco [29, 30] for the Coulomb gas.

To be precise, define the two-dimensional Discrete Gaussian model on Λ_N at inverse temperature $\beta \in (0,\infty)$ by the following expectation. For any $F : (2\pi\mathbb{Z})^{\Lambda_N} \to \mathbb{R}$ with $F(\sigma) = F(\sigma + c)$ for any constant $c \in 2\pi\mathbb{Z}$, and such that the following series converges, we define

$$\langle F\rangle_\beta^{\Lambda_N} \propto \sum_{\sigma\in\Omega^{\Lambda_N}} e^{\frac{\beta}{2}(\sigma,\Delta\sigma)}\, F(\sigma) = \sum_{\sigma\in\Omega^{\Lambda_N}} e^{-\frac{\beta}{4}\sum_{x\sim y}(\sigma_x-\sigma_y)^2}\, F(\sigma), \tag{3.1}$$

where the sum over $x \sim y$ counts every undirected edge twice and

$$\Omega^{\Lambda_N} = \{\sigma \in (2\pi\mathbb{Z})^{\Lambda_N} : \sigma_{x=0} = 0\}. \tag{3.2}$$

In [15, 16], we studied the scaling limit on the continuum torus $\mathbb{T}^2$ and on the infinite plane $\mathbb{R}^2$ for $\beta \le \beta_0$ sufficiently small (high temperature). All results hold for general finite-range step distributions, but we simplify the statement to the nearest-neighbour interaction. We also remark that in [15, 16] we used β instead of $1/\beta$.

Theorem 3.1. *There exists $\beta_0 \in (0,\infty)$ and $\beta_{\mathrm{eff}}(\beta) \in (0,\infty)$ for $\beta \le \beta_0$ (high temperature), satisfying $1/\beta_{\mathrm{eff}}(\beta) = 1/\beta + O_J(e^{-c/\beta})$ for some $c > 0$ as $\beta \to 0$, such that the following scaling limit results hold:*

(i) *The Discrete Gaussian model $\langle\cdot\rangle_\beta^{\Lambda_N}$ on the discrete torus Λ_N converges to a continuum Gaussian free field on $\mathbb{T}^2$: For any $f \in C^\infty(\mathbb{T}^2)$ with $\int f\,dx = 0$ and f_N a suitable discretisation of f on Λ_N, as $N \to \infty$,*

$$\log\langle e^{(f_N,\sigma)_{\Lambda_N}}\rangle_\beta^{\Lambda_N} \to \frac{1}{2\beta_{\mathrm{eff}}(\beta)}(f,(-\Delta_{\mathbb{T}^2})^{-1}f)_{\mathbb{T}^2}. \tag{3.3}$$

(ii) *The Discrete Gaussian model $\langle\cdot\rangle_\beta^{\mathbb{Z}^2}$ on $\mathbb{Z}^2$ converges to a continuum Gaussian free field on $\mathbb{R}^2$: For any $f \in C_c^\infty(\mathbb{R}^2)$ with $\int f\,dx = 0$ and f_ε a suitable discretisation on $\mathbb{Z}^2$, as $\varepsilon \to 0$,*

$$\log\langle e^{(f_\varepsilon,\sigma)_{\mathbb{Z}^2}}\rangle_\beta^{\mathbb{Z}^2} \to \frac{1}{2\beta_{\mathrm{eff}}(\beta)}(f,(-\Delta_{\mathbb{R}^2})^{-1}f)_{\mathbb{R}^2}. \tag{3.4}$$

3.2. Low temperature: The arboreal gas

The *arboreal gas* is the uniform measure on (unrooted) spanning forests of a weighted graph. It turns out to be the graphical representation of the nonlinear sigma model with OSp(1 | 2) symmetry. This is an exact analogue of the classical Heisenberg model (2.6) with the sphere $\mathbb{S}^{n-1}$ replaced by the fermionic hyperbolic plane $\mathbb{H}^{0|2}$ and the group $O(n)$ by the supergroup OSp(1 | 2). This point of view was pioneered in [26] and further developed with Crawford, Helmuth, and Swan in [14], to which we refer for an introduction and for results in $d = 2$.

In [13], with Crawford and Helmuth, we used the renormalisation group approach to prove the existence of a low-temperature phase in $d \geq 3$ by proving that the continuous symmetry of the nonlinear sigma model is spontaneously broken. This is an instance of the Goldstone mechanism [32].

The arboreal gas model is precisely defined as follows. For the graph with vertices $\Lambda = \Lambda_N$ and edges E (the nearest neighbours for us), a forest $F = (\Lambda, E(F))$ is an acyclic subgraph with the same vertex set Λ. Given an edge weight $\beta > 0$ (inverse temperature), the probability of a forest F under the arboreal gas measure is

$$\mathbb{P}_\beta^{\Lambda_N}[F] \equiv \frac{1}{Z_{\beta,0}^{\Lambda_N}} \beta^{|E(F)|}, \tag{3.5}$$

where $T \in F$ denotes that T is a tree in the forest, i.e., a connected component of F, and $|E(F)|$ is the number of edges in F. It is sometimes of interest (and technically useful) to add an external field $h \geq 0$ (vertex weight). The generalisation of the arboreal gas is

$$\mathbb{P}_{\beta,h}^{\Lambda_N}[F] \equiv \frac{1}{Z_{\beta,h}^{\Lambda_N}} \beta^{|E(F)|} \prod_{T \in F} (1 + h|V(T)|), \tag{3.6}$$

where $|V(T)|$ is the number of vertices in T. This generalisation with external field can also be conveniently interpreted as the standard arboreal gas on an extended graph with a ghost vertex added [13]. The arboreal gas is also known as the uniform forest model, as Bernoulli bond percolation conditioned to be acyclic, and as the $q \to 0$ limit of the q-state random-cluster model with p/q converging to β; see [37].

Theorem 3.2. *Suppose $d \geq 3$. There exists $\beta_0 \in (0, \infty)$ such that for $\beta \geq \beta_0$ there exist functions $\theta_d(\beta) = 1 - O(1/\beta)$ and $\mathsf{c}_i(\beta) = \mathsf{c}_i + O(1/\beta)$ with $\mathsf{c} > 0$, satisfying $(\mathsf{c}_2(\beta)/\mathsf{c}_1(\beta)^2)\theta_d(\beta)^2 = 1$ and $\mathsf{c}_0(\beta) = 2\mathsf{c}_1(\beta)$, and that the following hold:*

(i) *The torus two-point function has long-range order with free field correction (Goldstone mode):*

$$\mathbb{P}_\beta^{\Lambda_N}[0 \leftrightarrow x] = \theta_d(\beta)^2 + \frac{\mathsf{c}_0(\beta)}{\beta|x|^{d-2}} + O\left(\frac{1}{\beta|x|^{d-2+\kappa}}\right) + O\left(\frac{1}{\beta L^{\kappa N}}\right), \tag{3.7}$$

where $|x|$ denotes the Euclidean norm and $\kappa > 0$ a small constant.

(ii) *In any possibly subsequential infinite volume limit along Λ_N, there is an infinite tree:*

$$\mathbb{P}_\beta^{\mathbb{Z}^d}[|T_0| = \infty] = \theta_d(\beta). \tag{3.8}$$

(iii) *The truncated two-point functions $\tau_\beta(x)$ and $\sigma_\beta(x)$ satisfy*

$$\tau_\beta(x) = \frac{\mathsf{c}_1(\beta)}{\beta|x|^{d-2}} + O\left(\frac{1}{\beta|x|^{d-2+\kappa}}\right), \tag{3.9}$$

$$\sigma_\beta(x) = \frac{\mathsf{c}_2(\beta)}{\beta^2|x|^{2d-4}} + O\left(\frac{1}{\beta^2|x|^{2d-4+\kappa}}\right). \tag{3.10}$$

In the sigma-model representation of the arboreal gas, (3.7) is the exact analogue of (2.7), and the truncated correlation functions (3.9) and (3.10), see [13] for their precise definition, correspond to transversal and radial correlations of the spins, respectively. The free field decay (3.7) and (3.9) represents the Goldstone mode. The decay of radial correlations (3.10) has the explanation that the radial mode is unstable and 'decays' into two Goldstone particles as explained in [32].

3.3. Critical temperature: The $|\varphi|^4$ model

Finally, we discuss an application of the renormalisation group to critical behaviour – which is the most well-studied context. The lattice φ^4 model is defined as in (2.4), and its n-component version is defined in the same way except that all spins φ_x become vectors in $\mathbb{R}^n$. A summary of some results for the four-dimensional model at the critical point is given in the next theorem. These results were obtained with Brydges, Slade, Tomberg, and Wallace. These results extend to $n = 0$ which can be interpreted as the weakly self-avoiding walk.

Theorem 3.3. *Let $d = 4$, $n = 0, 1, \ldots$, and let $g > 0$ be small. For the n-component $|\varphi|^4$ model ($n \geq 1$), and for the weakly self-avoiding walk (corresponding to $n = 0$), there exist critical values $\nu_c = \nu_c(g, n)$ such that, as $\varepsilon \downarrow 0$ and $|x| \to \infty$, the following results hold (with constants $\mathsf{c}_* = \mathsf{c}_*(g, n)$ and C_p depending on p):*

(i) *For $n \geq 0$, the critical two-point function decays as* [10]

$$G_{g,\nu_c}(x) = \lim_{N\to\infty} \langle \varphi_0 \cdot \varphi_x \rangle_{g,\nu_c}^{\Lambda_N} \sim \mathsf{c}_G |x|^{-2}. \tag{3.11}$$

(ii) *For $n \geq 0$, the susceptibility obeys* [8, 11]

$$\chi(g, \nu_c + \varepsilon) = \lim_{N\to\infty} \sum_{x\in\Lambda_N} \langle \varphi_0 \cdot \varphi_x \rangle_{g,\nu_c+\varepsilon}^{\Lambda_N} \sim \mathsf{c}_\chi\, \varepsilon^{-1} (\log \varepsilon^{-1})^{\frac{(n+2)}{(n+8)}}. \tag{3.12}$$

(iii) *For $n \geq 0$, the correlation length of order $p > 0$ obeys* [17]

$$\frac{1}{\chi(g, \nu_c + \varepsilon)} \sum_x |x|^p G_{g,\nu_c+\varepsilon}(x) \sim C_p \mathsf{c}_\chi^{\frac{-p}{2}} \varepsilon^{\frac{-p}{2}} (\log \varepsilon^{-1})^{\frac{p(n+2)}{(2n+16)}}. \tag{3.13}$$

(iv) *For $n \geq 1$, the specific heat obeys* [8]

$$c_H(g, \nu_c + \varepsilon) \sim \mathsf{c}_H \begin{cases} (\log \varepsilon^{-1})^{\frac{(4-n)}{(n+8)}} & (n = 1, 2, 3), \\ \log \log \varepsilon^{-1} & (n = 4), \\ 1 & (n > 4). \end{cases} \tag{3.14}$$

(v) *For $n \geq 1$, the spin field on Λ converges weakly to white noise if $\nu > \nu_c$, and to a massive Gaussian free field if $\nu \downarrow \nu_c$ as $N \to \infty$ appropriately; see* [8].

Acknowledgements. I thank my collaborators on the various projects described here. These include D.C. Brydges, N. Crawford, T. Helmuth, J. Park, P.-F. Rodriguez, and G. Slade.

Funding. This work was partially supported by the ERC grant 851682 SPINRG and the NSF grant DMS-2348045.

References

[1] T. Balaban, A low temperature expansion for classical N-vector models. II. Renormalization group equations. *Comm. Math. Phys.* **182** (1996), no. 3, 675–721 Zbl 0869.60095 MR 1461947

[2] T. Balaban, Localization expansions. I. Functions of the "background" configurations. *Comm. Math. Phys.* **182** (1996), no. 1, 33–82 Zbl 0878.60072 MR 1441905

[3] T. Balaban, The variational problems for classical N-vector models. *Comm. Math. Phys.* **175** (1996), no. 3, 607–642 Zbl 0838.60092 MR 1372811

[4] T. Balaban, A low temperature expansion for classical N-vector models III. A complete inductive description, fluctuation integrals. *Comm. Math. Phys.* **196** (1998), no. 3, 485–521 Zbl 0924.60096 MR 1645208

[5] T. Balaban, Renormalization and localization expansions. II. Expectation values of the "fluctuation" measures. *Comm. Math. Phys.* **198** (1998), no. 1, 1–45 Zbl 0932.81020 MR 1657361

[6] T. Balaban, The large field renormalization operation for classical N-vector models. *Comm. Math. Phys.* **198** (1998), no. 3, 493–534 Zbl 0940.81034 MR 1670021

[7] T. Balaban and M. O'Carroll, Low temperature properties for correlation functions in classical N-vector spin models. *Comm. Math. Phys.* **199** (1999), no. 3, 493–520 Zbl 0932.81024 MR 1669693

[8] R. Bauerschmidt, D. C. Brydges, and G. Slade, Scaling limits and critical behaviour of the 4-dimensional n-component $|\varphi|^4$ spin model (Special issue in memory of Kenneth Wilson). *J. Stat. Phys.* **157** (2014), no. 4–5, 692–742 Zbl 1308.82026 MR 3269689

[9] R. Bauerschmidt, D. C. Brydges, and G. Slade, A renormalisation group method. III. Perturbative analysis. *J. Stat. Phys.* **159** (2015), no. 3, 492–529 Zbl 1319.82008 MR 3332940

[10] R. Bauerschmidt, D. C. Brydges, and G. Slade, Critical two-point function of the 4-dimensional weakly self-avoiding walk. *Comm. Math. Phys.* **338** (2015), no. 1, 169–193 Zbl 1320.82031 MR 3345374

[11] R. Bauerschmidt, D. C. Brydges, and G. Slade, Logarithmic correction for the susceptibility of the 4-dimensional weakly self-avoiding walk: a renormalisation group analysis. *Comm. Math. Phys.* **337** (2015), no. 2, 817–877 Zbl 1318.60049 MR 3339164

[12] R. Bauerschmidt, D. C. Brydges, and G. Slade, *Introduction to a renormalisation group method*. Lecture Notes in Math. 2242, Springer, Singapore, 2019, 283 pp. Zbl 1471.81001 MR 3969983

[13] R. Bauerschmidt, N. Crawford, and T. Helmuth, Percolation transition for random forests in $d \geqslant 3$. *Invent. Math.* **237** (2024), no. 2, 445–540 Zbl 07887294 MR 4768630

[14] R. Bauerschmidt, N. Crawford, T. Helmuth, and A. Swan, Random spanning forests and hyperbolic symmetry. *Comm. Math. Phys.* **381** (2021), no. 3, 1223–1261 Zbl 1470.60270 MR 4218682

[15] R. Bauerschmidt, J. Park, and P.-F. Rodriguez, The Discrete Gaussian model, I. Renormalisation group flow at high temperature. *Ann. Probab.* **52** (2024), no. 4, 1253–1359 Zbl 1546.82018 MR 4766835

[16] R. Bauerschmidt, J. Park, and P.-F. Rodriguez, The Discrete Gaussian model, II. Infinite-volume scaling limit at high temperature. *Ann. Probab.* **52** (2024), no. 4, 1360–1398 Zbl 1546.82019 MR 4766836

[17] R. Bauerschmidt, G. Slade, A. Tomberg, and B. C. Wallace, Finite-order correlation length for four-dimensional weakly self-avoiding walk and $|\varphi|^4$ spins. *Ann. Henri Poincaré* **18** (2017), no. 2, 375–402 Zbl 1357.81136 MR 3596766

[18] J. Bricmont, J.-R. Fontaine, J. L. Lebowitz, E. H. Lieb, and T. Spencer, Lattice systems with a continuous symmetry. III. Low temperature asymptotic expansion for the plane rotator model. *Comm. Math. Phys.* **78** (1981), no. 4, 545–566 MR 0606463

[19] J. Bricmont, J.-R. Fontaine, J. L. Lebowitz, and T. Spencer, Lattice systems with a continuous symmetry. I. Perturbation theory for unbounded spins. *Comm. Math. Phys.* **78** (1980), no. 2, 281–302 MR 0597750

[20] J. Bricmont, J.-R. Fontaine, J. L. Lebowitz, and T. Spencer, Lattice systems with a continuous symmetry. II. Decay of correlations. *Comm. Math. Phys.* **78** (1981), no. 3, 363–371 MR 0603499

[21] D. C. Brydges, Lectures on the renormalisation group. In S. Sheffield and T. Spencer (Eds.), *Statistical mechanics*, pp. 7–93, IAS/Park City Math. Ser. 16, American Mathematical Society, Providence, RI, 2009 Zbl 1186.82033 MR 2523458

[22] D. C. Brydges and G. Slade, A renormalisation group method. I. Gaussian integration and normed algebras. *J. Stat. Phys.* **159** (2015), no. 3, 421–460 Zbl 1317.82013 MR 3332938

[23] D. C. Brydges and G. Slade, A renormalisation group method. II. Approximation by local polynomials. *J. Stat. Phys.* **159** (2015), no. 3, 461–491 Zbl 1317.82014 MR 3332939

[24] D. C. Brydges and G. Slade, A renormalisation group method. IV. Stability analysis. *J. Stat. Phys.* **159** (2015), no. 3, 530–588 Zbl 1317.82015 MR 3332941

[25] D. C. Brydges and G. Slade, A renormalisation group method. V. A single renormalisation group step. *J. Stat. Phys.* **159** (2015), no. 3, 589–667 Zbl 1317.82016 MR 3332942

[26] S. Caracciolo, J. L. Jacobsen, H. Saleur, A. D. Sokal, and A. Sportiello, Fermionic field theory for trees and forests. *Phys. Rev. Lett.* **93** (2004), no. 8, article no. 080601, 4 pp. MR 2110547

[27] J. Cardy, *Scaling and renormalization in statistical physics*. Cambridge Lecture Notes in Phys. 5, Cambridge University Press, Cambridge, 1996, 238 pp. Zbl 0914.60002 MR 1446000

[28] J. Dimock and T. R. Hurd, Sine-Gordon revisited. *Ann. Henri Poincaré* **1** (2000), no. 3, 499–541 Zbl 1002.81035 MR 1777310

[29] P. Falco, Kosterlitz–Thouless transition line for the two dimensional Coulomb gas. *Comm. Math. Phys.* **312** (2012), no. 2, 559–609 Zbl 1254.82012 MR 2917175

[30] P. Falco, Critical exponents of the two dimensional Coulomb gas at the Berezinskii–Kosterlitz–Thouless transition. 2013, arXiv:1311.2237v2

[31] J. Feldman, J. Magnen, V. Rivasseau, and R. Sénéor, Construction and Borel summability of infrared Φ^4_4 by a phase space expansion. *Comm. Math. Phys.* **109** (1987), no. 3, 437–480 MR 0882810

[32] J. Fröhlich, Phase transitions, spontaneous symmetry breaking, and Goldstone's theorem. In T. Chakraborty (Ed.), *Encyclopedia of condensed matter physics*. 2nd edn., pp. 158–173, Academic Press, 2024

[33] J. Fröhlich, B. Simon, and T. Spencer, Infrared bounds, phase transitions and continuous symmetry breaking. *Comm. Math. Phys.* **50** (1976), no. 1, 79–95 MR 0421531

[34] J. Fröhlich and T. Spencer, The Kosterlitz–Thouless transition in two-dimensional Abelian spin systems and the Coulomb gas. *Comm. Math. Phys.* **81** (1981), no. 4, 527–602 MR 0634447

[35] K. Gawędzki and A. Kupiainen, Massless lattice φ^4_4 theory: rigorous control of a renormalizable asymptotically free model. *Comm. Math. Phys.* **99** (1985), no. 2, 197–252 MR 0790736

[36] K. Gawędzki and A. Kupiainen, Asymptotic freedom beyond perturbation theory. In *Phénomènes critiques, systèmes aléatoires, théories de jauge, Part I, II (Les Houches, 1984)*, pp. 185–292, North-Holland, Amsterdam, 1986 Zbl 0706.47039 MR 0880526

[37] G. Grimmett, *The random-cluster model*. Grundlehren Math. Wiss. 333, Springer, Berlin, 2006, 377 pp. Zbl 1122.60087 MR 2243761

[38] K. G. Wilson, Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture. *Phys. Rev. B* **4** (1971), 3174–3183 Zbl 1236.82017

[39] K. G. Wilson and J. Kogut, The renormalization group and the ε expansion. *Phys. Rep.* **12** (1974), no. 2, 75–199

Roland Bauerschmidt
Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA; bauerschmidt@cims.nyu.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/13

A free boundary approach to spectral optimization problems: Extremal Neumann eigenvalues

Dorin Bucur

Abstract. This is an extended version of the invited talk given at the 9th European Congress of Mathematics. It is concerned with the geometric optimization of Neumann eigenvalues of the Laplace operator on domains of $\mathbb{R}^N$ and on spheres. The main question is to maximize the k-th eigenvalue of the Neumann Laplacian under a volume constraint; we approach it in the context of free boundary problems with no energy interface. Together with a discussion about existence/non-existence of optimal geometries and possible relaxation on densities, we point out how numerical approximations play a key role in understanding the problem. A particular attention is given to the lowest two non-trivial eigenvalues for which a full answer is available in any dimension of the Euclidean space and in the hyperbolic space, but not on spheres, where a surprising phenomenon occurs: while a complete answer can be given for the second eigenvalue, a series of open problems remain unsolved for the first one. The results presented in the talk have been obtained in collaborations with A. Henrot, R. Laugesen, E. Martinet, M. Nahon and E. Oudet.

1. Introduction and statement of the result

To start, let us fix $N \geq 1$ and denote by $\mathbb{R}^N$ the Euclidean space of dimension N. Let $\rho : \mathbb{R}^N \to [0, 1]$ be a measurable function (throughout the paper called density), $\rho \not\equiv 0$. We set $\mu_0(\rho) = 0$, and for $k \geq 1$ we define

$$\mu_k(\rho) = \inf_{S \in \mathcal{S}_k} \max_{u \in S} \frac{\int_{\mathbb{R}^N} |\nabla u|^2 \rho dx}{\int_{\mathbb{R}^N} u^2 \rho dx}, \tag{1.1}$$

where $\mathcal{S}_k$ is the family of all subspaces of dimension k in

$$\left\{ u \cdot 1_{\{\rho(x)>0\}} : u \in C_c^\infty(\mathbb{R}^N), \int_{\mathbb{R}^N} \rho u dx = 0 \right\}.$$

Mathematics Subject Classification 2020: 49Q10.
Keywords: spectral geometry, Neumann eigenvalues, shape optimization.

We note that, in general, $\forall k \in \mathbb{N}$, $0 \le \mu_k(\rho) \le \mu_{k+1}(\rho)$. In the absence of any further information on ρ, the variational quantities $\mu_k(\rho)$ are well defined (and called, by abuse of language, eigenvalues), but they are not necessarily related to a well-posed eigenvalue problem; it is even possible that all of them vanish.

Assume now that $\Omega \subseteq \mathbb{R}^N$ is a bounded, open, Lipschitz set, not necessarily connected. Taking $\rho = 1_\Omega$, the characteristic function of the set Ω, the values $\mu_k(\rho)$ are precisely the eigenvalues of the Laplace operator with Neumann boundary conditions on the set Ω. Following (1.1), we denote these eigenvalues, counting the multiplicity

$$0 = \mu_0(\Omega) \le \mu_1(\Omega) \le \mu_2(\Omega) \le \cdots \to +\infty.$$

If Ω is connected, then $\mu_1(\Omega) > 0$; otherwise, $\mu_1(\Omega) = 0$.

Moreover, there exists $u_k \in H^1(\Omega) \setminus \{0\}$ such that the equation

$$\begin{cases} -\Delta u_k = \mu_k(\Omega) u_k & \text{in } \Omega, \\ \dfrac{\partial u_k}{\partial n} = 0 & \text{on } \partial\Omega, \end{cases}$$

is satisfied in a weak sense, n being the outward normal at $\partial\Omega$.

The generic problem we are dealing with is the following: given $m > 0$, find $\Omega \subseteq \mathbb{R}^N$ solution of the following shape optimization problem

$$\max\{\mu_k(\Omega) : \Omega \subseteq \mathbb{R}^N, |\Omega| = m\}. \tag{1.2}$$

Equivalently, taking advantage of the homogeneity of the eigenvalues to rescalings in $\mathbb{R}^N$, this is equivalent to

$$\max_{\Omega \subseteq \mathbb{R}^N} |\Omega|^{\frac{2}{N}} \mu_k(\Omega).$$

The main questions can be phrased as follows:

(1) Does problem (1.2) have a solution?

(2a) If so, what are the qualitative properties of an optimal domain: is the boundary smooth, does the domain have symmetries, can the exact solution be identified?

(2b) If no existence, describe the behavior of a maximizing sequence and understand possible relaxed limits (Γ-convergence, homogenization, etc.).

(3) If an analytic identification of the solution is not possible, then approximate it numerically.

The maximization problem (1.2) is directly related to the Pólya conjecture, stating that the k-th eigenvalue is bounded from above by the main term of the Weyl

asymptotic, namely that the following inequality occurs

$$\forall k \in \mathbb{N}, \quad \mu_k(\Omega) \leq \frac{4\pi^2 k^{\frac{2}{N}}}{(\omega_N |\Omega|)^{\frac{2}{N}}}, \tag{1.3}$$

where ω_N is the volume of the unit ball of $\mathbb{R}^N$. The Pólya conjecture has been intensively studied in the last 50 years; for now, it is only known to hold for some particular classes of domains. The connection between the Pólya conjecture and problem (1.2) stems from the idea that inequality (1.3) is enough to be proved *only* for the maximizing set of (1.2), possibly taking advantage of some properties coming out from its optimality.

An incomplete collection of results concerning problem (1.2) is the following:

- Szegö 1954, [22]: the disc minimizes $\frac{1}{|\Omega|}\left(\frac{1}{\mu_1(\Omega)} + \frac{1}{\mu_2(\Omega)}\right)$, in $\mathbb{R}^2$ among smooth simply connected sets.
- Weinberger 1956, [24]: the ball maximizes $|\Omega|^{\frac{2}{N}}\mu_1(\Omega)$, in $\mathbb{R}^N$ among Lipschitz sets.
- Girouard, Nadirashvili, Polterovich 2009, [13]: the union of two equal disjoint discs maximizes $|\Omega|\mu_2(\Omega)$, in $\mathbb{R}^2$ among smooth simply connected sets.
- Poliquin and Roy-Fortin 2012, [21], μ_{22} is not maximized by a ball or union of balls. This result excludes that all maximizers in (1.2) are union of balls. The order 22 could possibly be decreased using a certified numerical approach. It is likely to be equal to 3 in $\mathbb{R}^2$.
- Bucur, Henrot 2019, [7]: the union of two equal disjoint balls maximizes $|\Omega|^{\frac{2}{N}}$ $\mu_2(\Omega)$, in $\mathbb{R}^N$ among Lipschitz sets. Moreover, the result extends to densities $\rho : \mathbb{R}^N \to [0, 1]$, provided the volume constraint is replaced by a mass constraint (details are given in the next section).
- Freitas, Laugesen 2022, [12]: extend the maximization result for μ_2 (in the class of domains) to the hyperbolic space and give a different topological argument.

2. Main results

A recent series of results, presented below, stems from the following theorem from [7]. Although its main purpose was to find the maximum of μ_2 in the class of Lipschitz domains, the extension to densities plays a fundamental role in understanding problem (1.2).

Theorem 2.1. *Let $\rho : \mathbb{R}^N \to [0, 1]$, $\rho \in L^1(\mathbb{R}^N)$. Then, denoting $B_1 \subseteq \mathbb{R}^N$ the ball of radius 1, the following inequalities hold:*

(1) $\mu_1(\rho)\big(\int_{\mathbb{R}^N}\rho dx\big)^{\frac{2}{N}} \le |B_1|^{\frac{2}{N}}\mu_1(B_1)$.
(2) $\mu_2(\rho)\big(\int_{\mathbb{R}^N}\rho dx\big)^{\frac{2}{N}} \le 2^{\frac{2}{N}}|B_1|^{\frac{2}{N}}\mu_1(B_1)$.

If equality occurs in (1) *or* (2)*, then* ρ *equals a.e. the characteristic function of a ball or two equal disjoint balls, respectively.*

For the proof of the theorem, we refer to [7] (see also [9, 12]). It is obtained by direct comparison between $\mu_1(\rho)$, $\mu_2(\rho)$ and the corresponding eigenvalues on a ball, two balls, respectively. This is done by building a suitable pack of test functions. For μ_1, they are precisely the Weinberger functions, while for μ_2, they are obtained by suitable gluing couples of two Weinberger functions associated to two (not necessarily disjoint) equal balls. A topological argument is necessary in order to certify the existence of a pack of valid test functions.

The main observation coming out from the previous theorem is that instead of working with the shape optimization problem (1.2), one could work with the density optimization problem

$$\max\left\{\left(\int_{\mathbb{R}^N}\rho dx\right)^{\frac{2}{N}}\mu_k(\rho) : \rho : \mathbb{R}^N \to [0,1]\right\} := \mu_k^*. \tag{2.1}$$

The main question is to understand whether for every $k \in \mathbb{N}$ the two problems have the same solution. Note that the competitors in problem (2.1) include the competitors of problem (1.2), so that the maximum in (2.1) cannot be lower than in (1.2). The advantage to work with (2.1) comes from the flexibility offered by the space of competitors: $L^\infty(\mathbb{R}^N)\cap L^1(\mathbb{R}^N)$ has a much better functional structure than the class of bounded, open, Lipschitz sets.

Coming back to (2.1), the following existence result has been proved in [10].

Theorem 2.2. *Problem* (2.1) *admits a solution as a collection of at most* k *densities. Precisely, there exists* $1 \le l \le k$ *and* l *densities* $\rho_1, \dots, \rho_l$ *such that* $\sum_{j=1}^l \int_{\mathbb{R}^N}\rho_j dx = 1$*, together with integers* $i_1, \dots, i_l \ge 0$ *satisfying* $k+1 = \sum_{j=1}^l (i_j+1)$*, and such that*

$$\mu_k^* = \mu_{i_1}(\rho_1) = \cdots = \mu_{i_l}(\rho_l).$$

Above, the concept of *collection* of densities plays the same role as that of connected components for an open set, with the global "spectrum" of the collection being the union of "spectra" of the individual densities. Roughly speaking, we compute the eigenvalues of each density separately and order them together, counting the multiplicities. There is no information on the display of those densities in the space, but, of course, if the supports of all densities were bounded, then one can build a unique density in $\mathbb{R}^N$ as a solution to problem (2.1).

For the proof, we refer to [10]. Here, I will only give a series of observations.

Localized version: Equibounded supports. In order to better understand the local behavior, we add an extra hypothesis and modify the problem in order to localize it. The purpose is to get some a priori compactness. Assume that all competing densities have a prescribed mass (say equal to 1) and their support is contained in a given bounded, measurable set $D \subseteq \mathbb{R}^N$. We replace (2.1) with

$$\max\left\{\mu_k(\rho) : \rho : \mathbb{R}^N \to [0,1],\ \int_{\mathbb{R}^N} \rho dx = 1,\ \rho|_{\mathbb{R}^N \setminus D} = 0\right\}. \tag{2.2}$$

Of course, if $D = \mathbb{R}^N$, problems (2.2) and (2.1) are equivalent. If D is bounded, problem (2.2) is, a priori, weaker than (2.1). However, in this case, the proof of existence of a solution is immediate by the direct method. Indeed, since D is bounded, taking a maximizing sequence $(\rho_n)_n$, one subtracts a sequence (still denoted using the same index) such that

$$\rho_n \rightharpoonup \rho \text{ weakly} - * \text{ in } L^\infty(D).$$

It is not complicated to prove that

$$\mu_k(\rho) \geq \limsup_{n \to +\infty} \mu_k(\rho_n),$$

and, since $1 \in L^1(D)$, the constraint $\int_D \rho dx = 1$ is obtained by passage to the limit. The solution may depend on D. In order to recover some global information from the local one, i.e., to pass from bounded boxes D to the full space $\mathbb{R}^N$, the standard way would be to use optimality and to identify some useful properties of the best density. However, for now, there are not many results in this direction, apart from the concentration results described below.

Global optimization: Densities with support in $\mathbb{R}^N$. If one removes the bounded box, the support of ρ_n may fill the full space $\mathbb{R}^N$. A weak limit of the sequence $(\rho_n)_n$ could possibly be identified, but, for instance, if the supports of ρ_n blow away from the origin, the weak limit would be the zero density, which is not admissible since it does not satisfy the mass constraint. The key to proving the existence is to get control of the concentration of the support of ρ_n.

This is done in the next two lemmas. The fundamental tool, due to Grigor'yan, Netrusov and Yau [14], describes the concentration phenomenon of a general measure on annuli. In turn, we relate this concentration phenomenon to the magnitude of the k-th Neumann eigenvalue. The following two results are proved in [10].

Lemma 2.3. *Let $\rho \in L^1(\mathbb{R}^N, [0,1])$, $\int_{\mathbb{R}^N} \rho dx = 1$ such that $\mu_k(\rho) > 0$. There exist a dimensional constant c_N and a ball B_{x,R^*} with*

$$R^* = \sqrt{\frac{4(k+1)}{c_N \mu_k(\rho)}}$$

such that

$$\int_{B_{x,R^*}} \rho dx \geq \frac{c_N}{k+1}.$$

Roughly speaking, the result asserts that if $\mu_k(\rho)$ is "large" (above 0), then for some R^*, which is "low", there is a ball of radius R^* containing a fixed proportion of the total mass of ρ. On the contrary, if $\mu_k(\rho)$ was close to 0, then R^* would be very large. For a maximizing sequence in (2.1), this gives a uniform control of the concentration of the mass on balls of fixed radius.

Since only a fraction of mass is concentrated in the balls found above, in order to explain what happens with the missing mass, we give the following.

Lemma 2.4. *Suppose* $\rho \in L^1(\mathbb{R}^N, [0,1])$, $\int_{\mathbb{R}^N} \rho dx = 1$. *Assume that* $\rho = \rho_0 + \rho_1 + \cdots + \rho_j$ *and*

$$\min_{i \neq j} \operatorname{dist}(\{\rho_i > 0\}, \{\rho_j > 0\}) = R > 0.$$

Then, for every $l \in \{1, \ldots, j\}$, *there exist* $R_l^* > 0$ *and* $x_l \in \mathbb{R}^N$ *satisfying*

$$\frac{1}{R_l^*} \geq \left[\frac{1}{2(k+1)} \left(\frac{\mu_k(\rho) c_N \int_{\mathbb{R}^N} \rho_l dx}{\int_{\mathbb{R}^N} \rho_l dx + \int_{\mathbb{R}^N} \rho_0 dx} \right)^{\frac{1}{2}} - \frac{3}{2R} \right]^+ \tag{2.3}$$

and

$$\int_{B_{x_l,R_l^*}} \rho_l \geq \frac{c_N}{k+1} \int_{\mathbb{R}^N} \rho_l dx,$$

where c_N *is a dimensional constant.*

As a consequence of these two concentration lemmas, one can prove the existence result of Theorem 2.2 by a kind of concentration-compactness argument. At this point, one may suspect that when k grows, the maximum achieved on densities in Theorem 2.2 is (much) higher than the one corresponding to classical domains (1.2). The following result proved in [10] shows that the growth is controlled by the same constants found by Kröger in [17], for domains. In other words, even if the maximal values in (2.1) are higher than the similar ones for domains in (1.2), their order of magnitude is the same.

Theorem 2.5. *Let* $N \geq 2$, $\rho \in L^1(\mathbb{R}^N, [0,1])$, $\rho \not\equiv 0$. *Then*

$$\forall k \in \mathbb{N}, \quad \mu_k(\rho) \leq 4\pi^2 \left(\frac{(N+2)k}{2\omega_N} \frac{1}{\|\rho\|_1} \right)^{\frac{2}{N}}.$$

This inequality supports the possible validity of a general form of the Pólya conjecture, for densities. Previous results, obtained by Colbois, El Soufi and Savo [11], justify this inequality with a generic constant, on a Riemannian manifold.

Explicit answer to Theorem 2.2 and Pólya conjecture in dimension 1. The original problem (1.2) has no interest in dimension 1, since any open set is union of intervals, so that its Neumann eigenvalues are described exactly, ordering the eigenvalues of all intervals. However, problem (2.1) has a deep meaning on densities and is far from being trivial, even in dimension 1. The following result, from [10], not only proves the validity of the Pólya conjecture for densities in dimension 1, but also gives precisely the structure of the maximizing densities. They are the characteristic functions of unions of k equal intervals. In particular, problem (2.1) has the same solution as problem (1.2).

Theorem 2.6 (One space dimension). *In $\mathbb{R}$, for every $\rho : \mathbb{R} \to [0,1]$, $\rho \in L^1(\mathbb{R})$, we have*

$$\forall k \in \mathbb{N}, \quad \mu_k(\rho) \leq \frac{\pi^2 k^2}{\left(\int_{\mathbb{R}} \rho dx\right)^2}.$$

For every k, equality is attained when ρ equals a.e. the characteristic function of a union of at most k open, pairwise disjoint segments each having length equal to an entire multiple of $\frac{1}{k}\int_{\mathbb{R}} \rho dx$.

As a consequence, one gets the following sharp inequality for Sturm–Liouville eigenvalues.

Corollary 2.7. *Let $\rho_1, \rho_2 \in C^1([0,1],(0,1])$, $\|\rho_1\|_\infty = \|\rho_2\|_\infty = 1$. We consider the (classical) eigenvalue problem*

$$\begin{cases} -(\rho_1 u')' = \mu_k \rho_2 u & \text{on } (0,1), \\ u'(0) = u'(1) = 0. \end{cases}$$

Then

$$\forall k \geq 0, \quad \mu_k \leq \frac{\pi^2 k^2}{\min\left(\int_0^1 \rho_1, \int_0^1 \rho_2\right)^2}.$$

Numerical approximations. In dimension 2, as we do not know what the explicit solutions to problem (2.1) for $k \geq 3$ are, it is of interest to perform numerical computations in order to get an intuition on whether the optimal density corresponds to a characteristic function, or not. We refer the reader to [10] and [19] for such numerical computations in $\mathbb{R}^2$ and on the Euclidean sphere of $\mathbb{R}^3$, respectively.

In [10, Table 2], the numerics suggest that the maximum of μ_3 in $\mathbb{R}^2$ for problem (2.1) is a density, not a domain. However, the associated value is only 0.5% higher than the corresponding one on domains, obtained by Antunes and Freitas in [2].

Below, we briefly describe the numerical approach. First, one has to work with densities $\rho : D \to [0,1]$ having support in a bounded set $D \subseteq \mathbb{R}^2$ (say a rectangle).

Second, since for an arbitrary density the eigenvalue problem is not well posed, one has to regularize it as follows. Let $\varepsilon > 0$, and consider the eigenvalue problem

$$\begin{cases} -\mathrm{div}[(\rho+\varepsilon)\nabla u] = \mu_k^\varepsilon(\rho)(\rho+\varepsilon^2)u & \text{on } D, \\ \partial_n u = 0 & \text{on } \partial D. \end{cases} \tag{2.4}$$

Clearly, this problem is well posed and, numerically, one can approximate its eigenvalues and eigenfunctions by finite elements in $H^1(D)$. Even if $\mu_k(\rho)$ has only a variational meaning and does not correspond to a classical eigenvalue, following [10] we have

$$\forall k \in \mathbb{N}, \quad \text{for } \varepsilon \to 0 : \mu_k^\varepsilon(\rho) \to \mu_k(\rho)$$

and

$$\max_{\substack{\rho\in L^1(D,[0,1]) \\ \int\rho=m}} \mu_k^\varepsilon(\rho) \xrightarrow[\varepsilon\to 0]{} \max_{\substack{\rho\in L^1(D,[0,1]) \\ \int\rho=m}} \mu_k(\rho).$$

These arguments justify the numerical approach of the original problem (2.1) via its ε-regularization.

We point out that the presence of ε^2 (instead of the more expected ε^1) on the right-hand side of (2.4) is necessary to avoid the appearance of spurious eigenvalues.

We refer the reader to [10] for numerical approximations of the solutions of (2.1) for $k = 3, \dots, 8$ in $\mathbb{R}^2$ and for comparisons with the numerical solutions of (1.2) (see also [2]).

3. Maximization of Neumann eigenvalues on Euclidean spheres

Let us denote by $\mathbb{S}^N \subseteq \mathbb{R}^{N+1}$ the unit sphere of $\mathbb{R}^{N+1}$. We denote by $\Delta_{\mathbb{S}^N}$ the Laplace–Beltrami operator on $\mathbb{S}^N$, and for $\Omega \subseteq \mathbb{S}^N$ smooth, we denote by $\mu_k(\Omega)$ the k-th eigenvalue of the Laplace–Beltrami operator with Neumann boundary conditions on Ω

$$\begin{cases} -\Delta_{\mathbb{S}^N} u = \mu_k(\Omega)u & \text{in } \Omega, \\ \dfrac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases}$$

Similarly to the Euclidean case, the eigenvalues are obtained by the min-max formula. If $\rho : \mathbb{S}^N \to [0,1]$ is measurable, we can define the relaxed eigenvalues associated to ρ, denoted $\mu_k(\rho)$, in the same spirit as (1.1) replacing the usual gradient with the tangential one ($\mu_0(\rho) = 0$ and multiplicities are counted).

We recall the following existence result from [19].

Theorem 3.1. *For every $m \in (0, |\mathbb{S}^N|)$, for every $k \geq 1$, the problem*

$$\max\left\{\mu_k(\rho) : \rho : \mathbb{S}^N \to [0,1], \int_{\mathbb{S}^N} \rho d\sigma = m\right\}$$

has at least a solution.

Note that the proof of Theorem 3.1 is a straightforward consequence of the upper semi-continuity result

$$\rho_n \to \rho \text{ weakly in } * - L^\infty(\mathbb{S}^N) \Longrightarrow \mu_k(\rho) \geq \limsup_{n\to+\infty} \mu_k(\rho_n).$$

With respect to the Euclidean case, the main difference which makes this existence result direct is the compactness of $\mathbb{S}^N$, which implies that $1 \in L^1(\mathbb{S}^N)$ and so the mass constraint on the density is preserved by weak $* - L^\infty$ convergence.

A result on spheres, similar to Szegö–Weinberger's one, has been obtained by Ashbaugh and Benguria in [4]. However, it requires an extra hypothesis, namely the inclusion of the competing domains in a hemisphere.

Theorem 3.2 (Ashbaugh–Benguria). *Let $m \leq \frac{|\mathbb{S}^N|}{2}$. If $\Omega \subseteq \mathbb{S}^N \cap \{x_1 \geq 0\}$ and $|\Omega| = m$, then*

$$\mu_1(\Omega) \leq \mu_1(B^m),$$

where $B^m \subseteq \mathbb{S}^N$ is a geodesic ball of volume m.

It is a long-standing open problem to understand whether the hemisphere inclusion hypothesis is only technical or whether it is really unavoidable.

The following result from [9] is quite surprising. It is the spherical counterpart of Theorem 2.1, point (2), which completely answers the maximization question for μ_2. A full understanding of the similar counterpart of point (1), concerning the maximization of μ_1, is missing.

Theorem 3.3. *Let $0 < m < |\mathbb{S}^N|$. If $\rho : \mathbb{S}^N \to [0,1]$ is measurable such that $\int_{\mathbb{S}^N} \rho d\sigma = m$. Then*

$$\mu_2(\rho) \leq \mu_1(B^{\frac{m}{2}}).$$

The original statement in [9, Theorem 2] is given for domains: if $\Omega \subseteq \mathbb{S}^N$, $|\Omega| = m$, then $\mu_2(\Omega) \leq \mu_1(B^{\frac{m}{2}})$. The proof follows the footsteps of Theorem 2.1 from [7], the construction of the test functions being adapted specifically to spheres. They are obtained by gluing two test functions of Ashbaugh–Benguria type. A topological argument justifies the existence of suitable test functions satisfying the orthogonality conditions. We point out that the zeros are counted modulo 4, correcting a sign

error in the original computation from [7]. The passage to densities is quite direct since both the topological argument and the mass transplantation procedure are not sensitive to the difference between a characteristic function and a density.

An important consequence of Theorem 3.3 is the following extension of the Ashbaugh–Benguria result (Theorem 3.2, above). This extension, even if it does not completely solve the problem, is a step forward towards better understanding it. It is a consequence of the substantial enlargement of the family of test functions (albeit for μ_2 and not for μ_1), by the gluing procedure. For the proof, we refer to [9].

Theorem 3.4. *Let $m \leq \frac{|\mathbb{S}^N|}{2}$ and B^m be a geodesic ball of mass m. If $\rho : \mathbb{S}^N \to [0,1]$ is measurable such that $\int_{\mathbb{S}^N} \rho d\sigma = m$, $\rho \cdot 1_{B^m} = 0$ (i.e., its support lays in the complement of B^m), then*

$$\mu_1(\rho) \leq \mu_1(B^m).$$

In particular, if $\Omega \subseteq \mathbb{S}^N \setminus B^m$ is such that $|\Omega| = m$, then

$$\mu_1(\Omega) \leq \mu_1(B^m).$$

Note that in Theorem 3.4 the room allowed for Ω is larger than a hemisphere. A domain $\Omega \subseteq \mathbb{S}^N$ which does not satisfy the inclusion constraint should intersect every geodesic ball of mass m of the sphere. If m is asymptotically vanishing, then such a domain would approximately fill the full sphere.

Some remarks are in order, all of them being suggested by the numerical computations from [9] and [19].

Remark 3.5. Numerical experiments suggest that on the unit sphere $\mathbb{S}^2 \subseteq \mathbb{R}^3$ there are values of $m \in (4, 2\pi)$ such that if the exclusion constraint above $\rho \cdot 1_{B^m} = 0$ is dropped (for every spherical cap of mass m), a relaxation phenomenon occurs. This leads to an effective construction of a density of mass m such that

$$\mu_1(\rho) > \mu_1(B^m).$$

This is only a numerical observation, not certified theoretically.

Remark 3.6. Numerical experiments based on the same strategy as in Remark 3.5, now under the exclusion constraint $\rho \cdot 1_{B^m} = 0$ for some spherical cap B^m of mass m, turn out to be very stable and produce clear evidence of the expected theoretical result from Theorem 3.4. This gives credit to the validity of the numerical approach and of the relaxation phenomenon observed in Remark 3.5.

Remark 3.7. If m is small, $m < 3$, numerics from [19] do not suggest that the relaxation phenomenon occurs in the absence of exclusion constraint. This observation

could support the thesis that the spherical cap is optimal if m is small, without any further constraint. However, these computations need to be carefully interpreted in order to understand the strength of this suggestion.

Remark 3.8. There do exist values $4\pi > m > 2\pi$ for which the spherical cap is not the optimal geometry. In other words, there exists an open set (which is not simply connected and presents several holes) satisfying the area constraint, but with a higher first eigenvalue than the corresponding geodesic ball. This is numerically observed in [19] and theoretically justified in [8].

4. Further remarks

4.1. More general functionals of eigenvalues

Ashbaugh and Benguria conjectured in [3] that for smooth domains $\Omega \subseteq \mathbb{R}^N$ of volume m, the following inequality holds

$$\sum_{k=1}^{N} \frac{1}{\mu_k(\Omega)} \geq \frac{N}{\mu_1(B^m)}.$$

A step towards the proof of the inequality was done by in [25], where Xia and Wang prove that

$$\sum_{k=1}^{N-1} \frac{1}{\mu_k(\Omega)} \geq \frac{N-1}{\mu_1(B^m)},$$

with equality if and only if $\Omega = B^m$. The result also holds in the hyperbolic space [25] and on Euclidean spheres [5]. On spheres, however, one required as further hypothesis that Ω is contained in a hemisphere!

By comparison to Theorem 3.3, in [9], for $\Omega \subseteq \mathbb{S}^N$, $|\Omega| = m$, without any inclusion constraint, the following is proved

$$\sum_{k=2}^{N} \frac{1}{\mu_k(\Omega)} \geq \frac{N-1}{\mu_1(B^{\frac{m}{2}})},$$

with equality if and only if Ω is the union of two disjoint geodesic balls of volume $\frac{m}{2}$.

4.2. Quantitative versions

In recent years, a series of isoperimetric and spectral inequalities have been proved to hold in a quantitative form, i.e., with a geometric reminder measuring a distance to

the optimal set. In particular, in [6], Brasco and Pratelli proved a sharp quantitative form of the Szegö–Weinberger inequality: for every $\Omega \subseteq \mathbb{R}^N$ smooth,

$$|B|^{\frac{2}{N}}\mu_1(B) - |\Omega|^{\frac{2}{N}}\mu_1(\Omega) \geq C_N \mathcal{A}(\Omega)^2,$$

where by $\mathcal{A}(\Omega)$ one denotes the Fraenkel asymmetry of Ω

$$\mathcal{A}(\Omega) = \inf\left\{\frac{|\Omega \Delta B|}{|\Omega|} : B \subseteq \mathbb{R}^N \text{ ball such that } |B| = |\Omega|\right\},$$

and C_N is a dimensional constant. The power 2 over $\mathcal{A}(\Omega)$ is sharp.

Concerning the second inequality of Theorem 2.1, in the class of domains, a quantitative version was proved in [23], reading

$$2^{\frac{2}{N}}|B|^{\frac{2}{N}}\mu_1(B) - |\Omega|^{\frac{2}{N}}\mu_2(\Omega) \geq C_N \mathcal{A}_2(\Omega)^{N+1},$$

where the asymmetry term $A_2(\Omega)$ is of Fraenkel type and quantifies a distance of Ω to a union of two disjoint balls of half measure

$$\mathcal{A}_2(\Omega) = \inf\left\{\frac{|\Omega \Delta (B_1 \cup B_2)}{|\Omega|} : B_1, B_2 \subseteq \mathbb{R}^N \text{ disjoint balls such that } |B_1| = |B_2| = \frac{1}{2}|\Omega|\right\}.$$

4.3. Other constraints

Several other constraints are naturally imposed on the competing domains.

Simply connectedness. In a recent paper [18], Langford and Laugesen prove that the maximum of μ_1 is attained on a geodesic ball among simply connected sets of $\mathbb{S}^2$ with prescribed area m, provided $m \leq 4\pi\frac{16}{17}$. Note that this is not contradictory with the numerical observations of [19] and the theoretical justifications from [8] (see Remark 3.8 above), since the domains having a higher first eigenvalue than the corresponding spherical caps have multiple holes. In the class of simply connected sets, an interesting question is to understand whether the limitation of the area to $4\pi\frac{16}{17}$ is only technical, or not.

Convexity. Payne and Weinberger have proved in [20] that if $\Omega \subseteq \mathbb{R}^N$ is bounded, open and convex, then

$$\mu_1(\Omega) \geq \frac{\pi^2}{D_\Omega^2},$$

where D_Ω is the diameter of Ω. This inequality is sharp, the lower bound being asymptotically attained by a sequence of parallelepipeds of the form $\Omega_\varepsilon = (0,1) \times$

$(0,\varepsilon)^{N-1}$. In [1], this inequality is improved by a quantitative term

$$\mu_1(\Omega) \geq \frac{\pi^2}{D_\Omega^2} + C_N \frac{a_{2,\Omega}^2}{D_\Omega^4},$$

where $a_{2,\Omega}$ is the second largest semiaxis of the John ellipsoid of Ω. This result answers, in a stronger version, a conjecture of Hang and Wang [15].

In the same class of convex subsets of $\mathbb{R}^N$, Kröger [17] found sharp upper bounds for μ_k; precisely, he proved that

- for $N = 2$,
$$\mu_k(\Omega) \leq \frac{(2j_{0,1} + (k-1)\pi)^2}{D_\Omega^2}$$
- for $N = 3$,
$$\mu_k(\Omega) \leq \frac{(k+1)^2\pi^2}{D_\Omega^2}$$
- for $N \geq 4$, then
$$\mu_k(\Omega) \leq \frac{\left(j_{\frac{d-2}{2},\lfloor\frac{k+1}{2}\rfloor} + j_{\frac{d-2}{2},\lfloor\frac{k+2}{2}\rfloor}\right)^2}{D_\Omega^2}.$$

Above, $j_{\nu,m}$ is the m-th zero of the Bessel function J_ν, and $\lfloor a \rfloor$ denotes the integer part of a. These inequalities have been revisited with a different point of view in [16]. The upper bounds are sharp and are attained asymptotically on a sequence of domains collapsing to a segment. The geometry of the collapsing domains is described in [16]. All these inequalities have a quantitative version.

References

[1] V. Amato, D. Bucur, and I. Fragalà, The geometric size of the fundamental gap. [v1] 2024, [v2] 2025, arXiv:2407.01341v2

[2] P. R. S. Antunes and P. Freitas, Numerical optimization of low eigenvalues of the Dirichlet and Neumann Laplacians. *J. Optim. Theory Appl.* **154** (2012), no. 1, 235–257 Zbl 1252.90076 MR 2931377

[3] M. S. Ashbaugh and R. D. Benguria, Universal bounds for the low eigenvalues of Neumann Laplacians in n dimensions. *SIAM J. Math. Anal.* **24** (1993), no. 3, 557–570 Zbl 0796.35122 MR 1215424

[4] M. S. Ashbaugh and R. D. Benguria, Sharp upper bound to the first nonzero Neumann eigenvalue for bounded domains in spaces of constant curvature. *J. Lond. Math. Soc. (2)* **52** (1995), no. 2, 402–416 Zbl 0835.58037 MR 1356151

[5] R. D. Benguria, B. Brandolini, and F. Chiacchio, A sharp estimate for Neumann eigenvalues of the Laplace–Beltrami operator for domains in a hemisphere. *Commun. Contemp. Math.* **22** (2020), no. 3, article no. 1950018, 9 pp. Zbl 1440.58003 MR 4082221

[6] L. Brasco and A. Pratelli, Sharp stability of some spectral inequalities. *Geom. Funct. Anal.* **22** (2012), no. 1, 107–135 Zbl 1241.47015 MR 2899684

[7] D. Bucur and A. Henrot, Maximization of the second non-trivial Neumann eigenvalue. *Acta Math.* **222** (2019), no. 2, 337–361 Zbl 1423.35271 MR 3974477

[8] D. Bucur, R. S. Laugesen, E. Martinet, and M. Nahon, Spherical caps do not always maximize Neumann eigenvalues on the sphere. *Geom. Funct. Anal.* **35** (2025), no. 5, 1313–1345 Zbl 08137324 MR 4990469

[9] D. Bucur, E. Martinet, and M. Nahon, Sharp inequalities for Neumann eigenvalues on the sphere. *J. Differential Geom.* **131** (2025), no. 1, 43–63 Zbl 08097545 MR 4947549

[10] D. Bucur, E. Martinet, and E. Oudet, Maximization of Neumann eigenvalues. *Arch. Ration. Mech. Anal.* **247** (2023), no. 2, article no. 19, 36 pp. Zbl 1517.35144 MR 4558988

[11] B. Colbois, A. El Soufi, and A. Savo, Eigenvalues of the Laplacian on a compact manifold with density. *Comm. Anal. Geom.* **23** (2015), no. 3, 639–670 Zbl 1317.58031 MR 3310527

[12] P. Freitas and R. S. Laugesen, Two balls maximize the third Neumann eigenvalue in hyperbolic space. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **23** (2022), no. 3, 1325–1355 Zbl 1500.35216 MR 4497746

[13] A. Girouard, N. Nadirashvili, and I. Polterovich, Maximization of the second positive Neumann eigenvalue for planar domains. *J. Differential Geom.* **83** (2009), no. 3, 637–661 Zbl 1186.35120 MR 2581359

[14] A. Grigor'yan, Y. Netrusov, and S.-T. Yau, *Eigenvalues of elliptic operators and geometric applications*, pp. 147–217, Surv. Differ. Geom. 9, International Press, Somerville, MA, 2004 Zbl 1061.58027 MR 2195408

[15] F. Hang and X. Wang, A remark on Zhong–Yang's eigenvalue estimate. *Int. Math. Res. Not. IMRN* **2007** (2007), no. 18, article no. rnm064, 9 pp. Zbl 1129.53021 MR 2358887

[16] A. Henrot and M. Michetti, Optimal bounds for Neumann eigenvalues in terms of the diameter. *Ann. Math. Qué.* **48** (2024), no. 2, 277–308 Zbl 1555.35208 MR 4798677

[17] P. Kröger, On upper bounds for high order Neumann eigenvalues of convex domains in Euclidean space. *Proc. Amer. Math. Soc.* **127** (1999), no. 6, 1665–1669 Zbl 0911.35079 MR 1486739

[18] J. J. Langford and R. S. Laugesen, Maximizers beyond the hemisphere for the second Neumann eigenvalue. *Math. Ann.* **386** (2023), no. 3–4, 2255–2281 Zbl 1518.35516 MR 4612417

[19] E. Martinet, Numerical optimization of Neumann eigenvalues of domains in the sphere. *J. Comput. Phys.* **508** (2024), article no. 113002, 26 pp. Zbl 1561.65456 MR 4733245

[20] L. E. Payne and H. F. Weinberger, An optimal Poincaré inequality for convex domains. *Arch. Ration. Mech. Anal.* **5** (1960), no. 1, 286–292 Zbl 0099.08402 MR 0117419

[21] G. Poliquin and G. Roy-Fortin, Wolf–Keller theorem for Neumann eigenvalues. *Ann. Sci. Math. Qué.* **36** (2012), no. 1, 169–178 Zbl 1292.35180 MR 3113300

[22] G. Szegö, Inequalities for certain eigenvalues of a membrane of given area. *J. Ration. Mech. Anal.* **3** (1954), no. 3, 343–356 Zbl 0055.08802 MR 0061749

[23] K. Wang and H. Wu, A quantitative Bucur–Henrot inequality. *Math. Nachr.* **295** (2022), no. 12, 2436–2451 Zbl 1534.35272 MR 4553586

[24] H. F. Weinberger, An isoperimetric inequality for the N-dimensional free membrane problem. *J. Ration. Mech. Anal.* **5** (1956), no. 4, 633–636 Zbl 0071.09902 MR 0079286

[25] C. Xia and Q. Wang, On a conjecture of Ashbaugh and Benguria about lower eigenvalues of the Neumann Laplacian. *Math. Ann.* **385** (2023), no. 1–2, 863–879 Zbl 1512.35416 MR 4542734

Dorin Bucur
Laboratoire de Mathématiques CNRS UMR 5127, Université de Savoie Mont Blanc, Campus Scientifique, 73 376 Le Bourget-Du-Lac, France; dorin.bucur@univ-savoie.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/14

Courant, Bézout and persistence

Lev Buhovsky

Abstract. We give an overview of some recent applications of the theory of persistence modules and barcodes to spectral geometry and to the transcendental Bézout problem. The proofs of most of the results are given in previous works written with coauthors.

1. Introduction

Let (M, g) be a compact and connected Riemannian manifold, possibly with boundary, and consider the induced (positive) Laplace–Beltrami operator $\Delta_g = -\operatorname{div}_g \nabla_g$ on M. The spectral decomposition of Δ_g (in case of $\partial M \neq \emptyset$, we impose the Dirichlet boundary conditions) provides us with an infinite sequence $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \nearrow +\infty$ of eigenvalues and a corresponding sequence $f_1, f_2, \ldots$ of eigenfunctions, $\Delta_g f_k = \lambda_k f_k$, forming an orthonormal basis of $L^2(M, dV_g)$ (here dV_g is the induced Riemannian volume density on M). Since the famous Chladni experiments conducted more than two centuries ago, the study of properties of eigenfunctions has fascinated generations of mathematicians.

For a smooth (or, more generally, continuous) function on M, its nodal set is $Z_f = \{f = 0\} \subset M$, and a nodal domain of f is a connected component of the complement $M \setminus Z_f$. In the case of $M = [0, \pi]$, the k-th Dirichlet eigenfunction $f_k(x) = \sin kx$ has exactly k nodal domains. In higher dimensions, there is no such statement, but there is a fundamental upper bound.

Theorem 1.1 (R. Courant, 1923). *A Laplace–Beltrami eigenfunction f_k has at most k nodal domains.*

Note that in the above-mentioned one-dimensional example of $M = [0, \pi]$, in fact any linear combination of the first k eigenfunctions, $\sum_{\ell=1}^{k} a_\ell f_\ell(x) = \sum_{\ell=1}^{k} a_\ell \sin \ell x$, has at most $k + 1$ zeroes on $[0, \pi]$, and as a result, at most k nodal domains. Such a property of eigenfunctions is called the *extended Courant property*.

Mathematics Subject Classification 2020: 58J50 (primary); 55U99 (secondary).
Keywords: barcode, persistence module, eigenfunction, analytic map, Bézout theorem.

Definition 1.2. We say that the extended Courant property holds on (M, g) if for every k, any linear combination of first k eigenfunctions of the Laplace–Beltrami operator has at most k nodal domains.

It is not difficult to see that the extended Courant property holds in dimension 1. The higher-dimensional case is more complicated and interesting. In case of bounded Euclidean domains, it was claimed in the footnote of [19, p. 454] (second footnote in the German original [18, p. 394]) that *"Any linear combination of the first n eigenfunctions divides the domain, by means of its nodes, into no more than n subdomains. See the Göttingen dissertation of H. Herrmann, Beiträge zur Theorie der Eigenwerte und Eigenfunktionen, 1932"*. However, already in 1956, Pleijel [31, p. 550] mentioned that he could not find a proof of this statement in the literature.

In 1973, V. Arnold [3, 4] related the extended Courant property to Hilbert's 16th problem, pointing out that while the property holds for $(\mathbb{R}P^2, g_0)$, it does not hold for $(\mathbb{R}P^3, g_0)$ due to counterexamples constructed by O. Viro [36]. More recent works [5–9, 23] provided many more examples where the extended Courant property fails, both in the setting of bounded Euclidean domains and in the more general setting of Riemannian manifolds.

Denote by $\mathcal{F}_\lambda$ the space of all linear combinations of eigenfunctions with eigenvalue $\leq \lambda$. Although the extended Courant property was shown to be incorrect in many cases, one could still ask whether the number of nodal domains of any linear combination of the first k eigenfunctions is bounded by Ck, where C is a constant. Or, equivalently,[1] one could ask whether for any $\lambda \geq 0$ and any $f \in \mathcal{F}_\lambda$, the number of nodal domains of f is bounded by $C(\lambda + 1)^{n/2}$, where n is the dimension of M. In [9], we show that the answer to this question is negative, and that in fact it is impossible even to have a general asymptotic upper bound.

Theorem 1.3. *There exists a Riemannian metric g on the 2-torus, a sequence f_{k_j} of eigenfunctions of the Laplacian Δ_g, and a sequence of scalars $c_j \in \mathbb{R}$, such that for each j the function $f_{k_j} - c_j$ has infinitely many nodal domains.*

The proof of the theorem combines ideas from analysis and topology. The constructed metric is of the Liouville type: $ds^2 = Q(x)(dx^2 + dy^2)$, $(x, y) \in \mathbb{T}^2 = (\mathbb{R}/2\pi\mathbb{Z})^2$. In particular, the variables separate, and the eigenfunctions are of the form $u(x)\cos my$ and $u(x)\sin my$, where $m \in \mathbb{Z}$, and the function u satisfies the Sturm–Liouville problem

$$u''(x) + (\lambda Q(x) - m^2)u(x) = 0. \tag{1.1}$$

[1] This equivalence follows from the Weyl's law by which λ_k grows asymptotically as $k^{2/n}$ times a constant.

For such an eigenfunction f, the difference $f - c$ will have infinitely many nodal domains provided that the function u oscillates appropriately (will have "infinitely many damped oscillations") near the level c. A crucial observation is as follows: if the function $\lambda Q(x) - m^2$ vanishes at some point x_* and moreover oscillates near x_* in a suitable way, then a solution u of (1.1), $u'(x_*) = 0$, is forced to oscillate near x_* as well. Therefore, we construct $Q(x)$ such that the function $\lambda Q(x) - m^2$ suitably oscillates near the level 0, or, equivalently, the function $Q(x)$ oscillates near the level m^2/λ for infinitely many values of m. Here, the main difficulty is that the possible values of λ (the eigenvalues of the Sturm–Liouville problem (1.1)) depend on Q and m, and we somehow need to match the levels where Q oscillates with the values m^2/λ for infinitely many m's. This becomes possible by applying the Brouwer fixed point theorem in dimensions growing to infinity and by using compactness [9].

Inspired by Theorem 1.3, we ask a general question.

Question 1.4. Is it true that on any closed manifold M of dimension $n \geq 2$ there exists a Riemannian metric as in Theorem 1.3?

It is not difficult to extend Theorem 1.3 from the 2-torus case to manifolds of the type $M = S^1 \times N$ (where the manifold N is closed), by using a similar reduction to the one-dimensional Sturm–Liouville problem. But for a general M, such a reduction is not possible, and new ideas are needed.

Remark 1.5. One could also try other possible ways to generalize Theorem 1.1. First, one could ask whether the theorem extends to higher-order elliptic differential operators (e.g., clamped plate problem). Second, notice that the number of nodal domains is the 0th Betti number of the complement of the nodal set. Hence one could ask whether the bound in Theorem 1.1 extends to higher-order Betti numbers (of the complement of the nodal set of an eigenfunction). However, based on Theorem 1.3, it was shown in [10] that, in general, these two questions have negative answers.

In this survey, we will describe an approach to the extended Courant property, based on the so-called coarse count of nodal domains, which leads to positive results [10, 32, 34]. Moreover, a variant of this "coarse approach" yields applications to Bézout theorems for eigenfunctions (giving a partial answer to a question of V. Arnold and L. Polterovich) [10], as well as to the transcendental Bézout problem [11]. For more details, see Sections 2 and 4.

The coarse approach is a part of a more general picture where the main focus lies on the interplay between topology and analysis of functions [10]. The topological function theory deals with invariants of functions under diffeomorphisms and, roughly speaking, enables to study oscillations of functions by looking at the topology of their sublevel sets. This theory is based on persistence modules and barcodes,

a mathematical apparatus originating from topological data analysis. On the analysis side, one considers measurements of functions based on the Sobolev scale. For more details, see Section 3.

2. Coarse Courant and Bézout

2.1. Coarse Courant

A remarkable and fruitful idea, which goes back to [34], is to disregard small oscillations while counting nodal domains (we call it *a coarse count of nodal domains*). More precisely, we have the following definition [34].

Definition 2.1. A nodal domain Ω of a function f is called δ-deep for some $\delta > 0$ if $\max_\Omega |f| > \delta$. Denote by $m_0(f, \delta)$ the number of δ-deep nodal domains of a function f.

In [34], L. Polterovich and M. Sodin showed that on a compact and connected Riemannian surface[2] (M, g), there is an upper bound

$$m_0(f, \delta) \leq C \min(\delta^{-1}, \delta^{-2})\lambda, \tag{2.1}$$

where $f \in \mathcal{F}_\lambda$ is $L^2(M, V_g)$-normalized, and $C = C(M, g)$. This result represents a version of the extended Courant property for the coarse count of nodal domains. It was deduced from a general inequality between a *generalized Banach indicatrix* of a smooth function and a quantity involving the Sobolev $W^{2,2}$-norm of the function [34]. These and related results were later extended to the framework of the theory of *persistence modules and barcodes* [32]. In a consequent work [10], we explored this direction further and proved a version of (2.1) (*a coarse Courant theorem*) on Riemannian manifolds of any dimension.

Theorem 2.2. *Let (M, g) be a compact and connected Riemannian manifold of dimension n, and let $k > n/2$ be an integer. For any $\lambda \geq 0$ and $f \in \mathcal{F}_\lambda$ with $\|f\|_{L^2(M,V_g)} = 1$, and any $\delta > 0$, we have*

$$m_0(f, \delta) \leq C\delta^{-n/k}(\lambda + 1)^{n/2}, \tag{2.2}$$

where $C = C(M, g, k)$.

[2] By a surface, we mean a manifold of dimension 2.

The upper bound in (2.2) is sharp in terms of the power of $\lambda + 1$, as we show in the following theorem [10].

Theorem 2.3. *Suppose (M, g) is a closed Riemannian manifold. Then there exists $c = c(M, g) > 0$ such that for any $\lambda \geq 0$ and $\delta > 0$, one can find $f \in \mathcal{F}_\lambda$, $\|f\|_{L^2(M,V_g)} = 1$, for which we have*

$$m_0(f, \delta) \geq c \frac{(\lambda + 1)^{n/2}}{\max(1, \delta^2)} - 1.$$

2.2. Bézout, eigenfunctions and coarse count

As suggested in the foundational work [21] of Donnelly and Fefferman, eigenfunctions of the Laplace–Beltrami operator with eigenvalue λ are expected to share some common features with polynomials of degree $\sqrt{\lambda}$ when λ is sufficiently large. For instance, this principle can be clearly seen on the sphere S^n endowed with the standard spherical metric: harmonic polynomials of degree d on $\mathbb{R}^{n+1}$ correspond to eigenfunctions of the Laplace–Beltrami operator with the eigenvalue $d(d + n - 1)$. A real version of the Bézout theorem states that for a generic collection $p_1, \ldots, p_n$ of real polynomials on $\mathbb{R}^n$, of degrees $d_1, \ldots, d_n$, respectively, the number of common zeroes does not exceed the product $d_1 d_2 \cdots d_n$. In the 2000s, V. Arnold and L. Polterovich asked whether eigenfunctions of the Laplace–Beltrami operator satisfy a version of the Bézout theorem. The example of the round sphere S^n provides an evidence here: indeed, due to the above explicit description of the eigenfunctions, one can deduce from the Bézout theorem that for eigenfunctions $f_1, \ldots, f_n$ on S^n with eigenvalues $\Lambda_1, \ldots, \Lambda_n$, the number of common zeroes generically does not exceed $2\sqrt{\Lambda_1 \cdots \Lambda_n}$. Furthermore, it was proved in [22] that the expectation (in a natural probabilistic setting) of the number of common zeroes equals $2n^{-n/2}\sqrt{\Lambda_1 \cdots \Lambda_n}$. Similar bounds for certain homogeneous Riemannian manifolds have also been obtained in [1,2].

The "coarse approach", useful for the Courant theorem and its extensions, is also applicable to the above-mentioned question of V. Arnold and L. Polterovich. First, let us define the relevant coarse count [10].

Definition 2.4 (Coarse count of zeroes). Let $f_1, \ldots, f_n : M \to \mathbb{R}$ be a collection of continuous functions on an n-dimensional manifold M. Denote $f = (f_1, \ldots, f_n) : M \to \mathbb{R}^n$. For $\delta > 0$, define $z_0(f, \delta)$ to be the number of connected components of $f^{-1}(B_\delta)$ which contain zeroes of f (i.e., which intersect $f^{-1}(0)$), where $B_\delta = \{|x| < \delta\} \subset \mathbb{R}^n$.

The following coarse Bézout theorem for linear combinations of eigenfunctions was shown in [10].

Theorem 2.5. *Let (M, g) be a compact and connected Riemannian manifold of dimension n, and let $k > n/2$ be an integer. For every $\lambda \geq 0$, and a collection $f_1, \dots, f_n \in \mathcal{F}_\lambda$ with $\|f_j\|_{L^2(M,V_g)} = 1$ $(j = 1, \dots, n)$, for each $\delta > 0$ we have*

$$z_0(f, \delta) \leq C\delta^{-n/k}(\lambda + 1)^{n/2} + 1,$$

where $f = (f_1, \dots, f_n)$, and $C = C(M, g, k)$.

This theorem can be seen as a partial answer to the question of Arnold and Polterovich. Note, however, that in the theorem the functions f_j are assumed to lie in *the same* $\mathcal{F}_\lambda$, which corresponds to having the same degree bound in case of polynomials, while in the Bézout theorem the degrees can be different. Therefore, we ask the following question [10].

Question 2.6. Can we generalize the statement of Theorem 2.5 by assuming that $f_j \in \mathcal{F}_{\Lambda_j}$ for $j = 1, \dots, n$ and replacing $(\lambda + 1)^{n/2}$ by $\sqrt{\Lambda_1 \cdots \Lambda_n}$ in the upper bound?

Finally, it is natural to ask whether for the question of Arnold and Polterovich, the coarse approach is at all necessary. In contrast to the "Courant story" where we have the counterexample given by Theorem 1.3, an analogous counterexample for the question of Arnold and Polterovich (the "Bézout story") is missing. Despite the lack of enough evidence, we would still conjecture that such a counterexample exists. See a recent work [35] of V. Stojisavljević for a possibly related counterexample in the context of harmonic functions, motivated by the transcendental Bézout problem.

2.3. Persistent Betti numbers and the Sobolev norm

As we will see later, the coarse Courant and coarse Bézout approaches are closely related, and moreover, Theorems 2.2 and 2.5 can be naturally generalized. Let us first extend the invariants m_0, z_0 to *persistent Betti numbers* of higher degree.

Definition 2.7. Let M be a smooth manifold of dimension n, and let $f : M \to \mathbb{R}^d$ be a continuous map. For any integer $0 \leq r < n$ and any $\delta > 0$, define the persistent Betti numbers

$$m_r(f, \delta) = \dim \mathrm{Im}(H_r(\{|f| > \delta\}) \to H_r(M \setminus Z_f))$$

and

$$z_r(f, \delta) = \dim \mathrm{Im}(H_r(Z_f) \to H_r(\{|f| < \delta\})).$$

Here, $|\cdot|$ stands for the Euclidean norm on $\mathbb{R}^d$, $Z_f = \{f = 0\} \subset M$, and $H_r(X)$ stands for the r-th singular homology group of a subset $X \subset M$ with coefficients in a field $\mathbb{K}$.

It is not difficult to see that in the case of $r = 0$, this definition indeed coincides with Definitions 2.1 and 2.4. The following result generalizes Theorems 2.2 and 2.5 (cf. Remark 1.5).

Theorem 2.8. *Let (M, g) be a compact Riemannian manifold of dimension n, let $\lambda \geq 0$ and let $f_1, \ldots, f_d \in \mathcal{F}_\lambda$ be such that $\|f_i\|_{L^2(M,V_g)} = 1$ for each $1 \leq i \leq d$. Then for any integers $k > n/2$ and $0 \leq r < n$, and any $\delta > 0$, we have*

$$m_r(f, \delta) \leq C\delta^{-n/k}(\lambda + 1)^{n/2},$$
$$z_r(f, \delta) \leq C\delta^{-n/k}(\lambda + 1)^{n/2} + b_r(M),$$

where $f = (f_1, \ldots, f_d) : M \to \mathbb{R}^d$, $b_r(M) = \dim H_r(M)$ and $C = C(M, g, k, d)$.

Theorem 2.8 appears to be a spectral geometric application of a general result which compares between m_r, z_r and a Sobolev norm. Recall that given a compact Riemannian manifold M, a real number $p \geq 1$ and an integer $k \geq 1$, for a smooth map $f : M \to \mathbb{R}^d$ the Sobolev norm $\|f\|_{W^{k,p}}$ is controlled by the L^p-norms of the derivatives of f up to order k. The Sobolev space $W^{k,p}(M; \mathbb{R}^d)$ is the completion of the space $C^\infty(M, \mathbb{R}^d)$ (consisting of all smooth functions $f : M \to \mathbb{R}^d$) with respect to the Sobolev norm $\|\cdot\|_{W^{k,p}}$. Note that if $kp > n = \dim M$, then by the Sobolev embedding theorem, the norm $\|\cdot\|_{W^{k,p}}$ dominates the uniform norm, and the Sobolev space $W^{k,p}(M; \mathbb{R}^d)$ consists of continuous maps, $W^{k,p}(M; \mathbb{R}^d) \subset C(M, \mathbb{R}^d)$. We now formulate the main result of this section [10].

Theorem 2.9. *Let (M, g) be a compact Riemannian manifold of dimension n, let $p \geq 1$ and let $k > n/p$, $0 \leq r < n$ and $d \geq 1$ be integers. Then for any $f \in W^{k,p}(M; \mathbb{R}^d)$ and any $\delta > 0$, we have*

$$m_r(f, \delta) \leq C\delta^{-n/k}\|f\|_{W^{k,p}}^{n/k}$$

and

$$z_r(f, \delta) \leq C\delta^{-n/k}\|f\|_{W^{k,p}}^{n/k} + b_r(M),$$

where $b_r(M) = \dim H_r(M)$, and $C = C(M, g, p, k, d)$.

Theorem 2.8 readily follows from an application of Theorem 2.9 (with $p = 2$) and of a priori estimates for the Laplace–Beltrami operator. Therefore, it is not surprising that Theorem 2.8 generalizes to general elliptic differential operators, with Dirichlet boundary condition in the case of a non-empty boundary [10] (cf. Remark 1.5).

The statement of Theorem 2.9, involving persistent Betti numbers, is not yet an ultimate one. In Section 3.1, we discuss the language of persistence barcodes, and in Section 3.2, we explain the meaning of persistent Betti numbers in that context. This language will allow us to state Theorem 3.8 (Section 3.2), which is the main result of [10].

3. Topology and analysis of functions

3.1. Introduction to persistent modules and barcodes

The notions of a persistence module and a barcode first appeared within the framework of topological data analysis. Below, we review these notions and their relevant properties. For a detailed account on persistence theory, see [14, 30, 33].

Definition 3.1. A persistence module (V, π) over a field $\mathbb{K}$ consists of a family of vector spaces V_t, $t \in \mathbb{R}$ over $\mathbb{K}$ together with linear maps $\pi_{s,t} : V_s \to V_t$ defined for all $s \leq t$, called structure maps, which satisfy $\pi_{t,t} = \mathrm{id}_{V_t}$ for all $t \in \mathbb{R}$, as well as $\pi_{s,t} \circ \pi_{r,s} = \pi_{r,t}$ for all $r \leq s \leq t$.

Important examples of persistence modules appear as topological invariants of functions. Let $f : X \to \mathbb{R}$ be a function on a Hausdorff topological space. Define $V(f)_t = H(\{f \leq t\})$ and $\pi^f_{s,t} = (i_{s,t})_*$, where $i_{s,t} : \{f \leq s\} \to \{f \leq t\}$ are inclusions and H denotes the singular homology with coefficients in a field $\mathbb{K}$.

Persistence modules form a category, where a morphism $\phi : (V, \pi^V) \to (W, \pi^W)$ is given by a family of linear maps $\phi_t : V_t \to W_t$, $t \in \mathbb{R}$ such that for all $s \leq t$ we have $\pi^W_{s,t} \circ \phi_s = \phi_t \circ \pi^V_{s,t}$. Direct sum of a collection of persistence modules is well defined and described by a natural (pointwise for each $t \in \mathbb{R}$) construction.

Simplest, and in some sense basic, examples of persistence modules are the following.

Definition 3.2. For an interval $I \subset \mathbb{R}$ (which can be either finite or infinite), we define the interval persistence module $\mathbb{K}_I$ as

$$(\mathbb{K}_I)_t = \begin{cases} \mathbb{K}, & \text{if } t \in I, \\ 0, & \text{otherwise,} \end{cases} \qquad \pi^{\mathbb{K}_I}_{s,t} = \begin{cases} \mathrm{id}_{\mathbb{K}}, & \text{if } s, t \in I, \\ 0, & \text{otherwise.} \end{cases}$$

General persistence modules can be quite complicated. In practice, one usually poses certain natural restrictions on persistence modules, under which it becomes possible to describe them in simple terms. For example, consider the class of pointwise finite-dimensional persistence modules, namely those (V, π) such that V_t is finite dimensional for each $t \in \mathbb{R}$. Then we have a structure theorem which states that any such persistence module is a direct sum of interval persistence modules.

Definition 3.3. A barcode $\mathcal{B}$ is a multiset of intervals with finite multiplicities. The intervals in $\mathcal{B}$ are called bars.

Theorem 3.4 (Structure theorem [20]). *To every pointwise finite-dimensional persistence module (V, π), there corresponds a unique barcode $\mathcal{B}(V)$ such that*

$$(V, \pi) \cong \bigoplus_{I \in \mathcal{B}(V)} (\mathbb{K}_I, \pi^{\mathbb{K}_I}).$$

Now let $f : M \to \mathbb{R}$ be a Morse function on a compact manifold M. Recall that the persistence module $(V(f), \pi^f)$ is given by $V(f)_t = H(\{f \leq t\})$, and natural maps $\pi^f_{s,t} : V(f)_s \to V(f)_t$ induced by inclusions of sublevel sets. Then clearly $V(f)$ is pointwise finite dimensional, and hence the structure theorem gives us the corresponding barcode which we denote by $\mathcal{B}(f)$. The barcode $\mathcal{B}(f)$ consists of a finite number of bars whose endpoints are critical values of f, and it describes how the space $V(f)_t = H(\{f \leq t\})$ changes when t grows. For a finite bar $[s, t)$ in $\mathcal{B}(f)$, at the level s a new homology class is "born" and at the level t this homology class "dies". An infinite bar $[s, +\infty)$ corresponds to a homology class that is born at the level s and never dies, that is, to a genuine homology class of M. The number of infinite bars in $\mathcal{B}(f)$ is equal to the total Betti number of M. By examining examples, it becomes evident that, in a certain sense, $\mathcal{B}(f)$ reflects the oscillatory behavior of the function f. See Figure 1.

The space of barcodes admits a remarkable pseudo-metric. We will use $\langle a, b\rangle$ to denote any of the intervals (a, b), $[a, b)$, $(a, b]$, $[a, b]$. Two barcodes $\mathcal{B}, \mathcal{B}'$ are called ε-matched, $\varepsilon > 0$, if after erasing certain bars of length $< 2\varepsilon$ from each of them, there exists a bijection Φ between the remaining bars, which satisfies

$$\Phi(\langle a, b\rangle) = \langle c, d\rangle \implies |a - c|, |b - d| < \varepsilon.$$

The bottleneck distance between barcodes is defined as

$$d_{\text{bottle}}(\mathcal{B}, \mathcal{B}') = \inf\{\varepsilon \mid \mathcal{B}, \mathcal{B}' \text{ are } \varepsilon\text{-matched}\}.$$

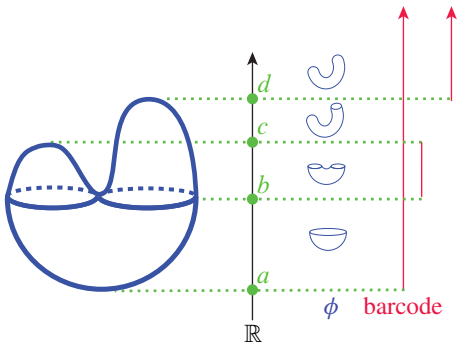


Figure 1. The barcode of the height function on a deformed sphere.

It is not difficult to see that d_{bottle} is a pseudo-metric. The counterpart of this distance is a distance between persistence modules which is called the *interleaving distance*. The direct correspondence between interleaving distance and bottleneck distance is stated in the *isometry theorem* [12, 15, 27]. The following corollary of the isometry theorem is a key feature of the theory [15].

Theorem 3.5 (Stability theorem). *Assume that M is a compact manifold and $f, g : M \to \mathbb{R}$ are Morse functions, then*

$$d_{\text{bottle}}(\mathcal{B}(f), \mathcal{B}(g)) \leq d_{C^0}(f, g).$$

3.2. The bar counting function and the Sobolev norm

It will be important for us to go beyond Morse functions and to treat general continuous functions $h : M \to \mathbb{R}$ defined on a compact manifold M. This is possible since for such h the persistence module $V(h)$ is still "good enough".

Definition 3.6. A persistence module (V, π) is called q-tame if $\pi_{s,t}$ has a finite rank for every $s < t$.

Let M be a compact manifold and $h : M \to \mathbb{R}$ be a continuous function. Approximating h by a Morse function and comparing their persistence modules, it is not difficult to deduce that the persistence module $V(h)$ is q-tame. Theorem 3.4 (the structure theorem) continues to hold also for q-tame persistence modules [13]. The barcode $\mathcal{B}(h)$ consists of a finite number of infinite bars of the form $[s, +\infty)$, and of at most countable number of finite bars of the form $[s, t)$. Moreover, by an approximation argument and by Theorem 3.5 (the stability theorem which continues to hold in the case of continuous functions), one concludes that for any $\delta > 0$ there exists at most finite number of bars of length $> \delta$ in $\mathcal{B}(h)$. This observation gives rise to an important invariant called the *bar counting function*.

Definition 3.7. For $\delta > 0$, we define $N_\delta(h)$ to be the number of bars of length $> \delta$ in the barcode $\mathcal{B}(h)$.

Let us remark that the $V(h)$ naturally decomposes into a direct sum of persistence modules $V_r(h)$, $0 \leq r \leq n$, where $V_r(h)_t = H_r(\{h \leq t\})$, and the structure maps $\pi^{h,r}_{s,t} : H_r(\{h \leq s\}) \to H_r(\{h \leq t\})$ are naturally induced by inclusions of sublevel sets ($n = \dim M$). Denote by $\mathcal{B}_r(h)$ the barcode of $V_r(h)$. As in Definition 3.6, for $\delta > 0$ we denote by $N_{r,\delta}(h)$ the number of bars of length $> \delta$ in the barcode $\mathcal{B}_r(h)$. From the isomorphism

$$(V_r(h), \pi^{h,r}) \cong \bigoplus_{I \in \mathcal{B}_r(h)} (\mathbb{K}_I, \pi^{\mathbb{K}_I}),$$

stated in the structure theorem, we get that for any $a < b$, the quantity

$$\operatorname{rk} \pi_{a,b}^{h,r} = \dim \operatorname{Im}(H_r(\{h \leq a\}) \to H_r(\{h \leq b\}))$$

equals the number of bars $[s,t)$ in $\mathcal{B}_r(h)$ such that $s \leq a < b < t$. It is also not difficult to deduce that

$$\dim \operatorname{Im}(H_r(\{h < a\}) \to H_r(\{h < b\}))$$

equals the number of bars $[s,t)$ in $\mathcal{B}_r(h)$ such that $s < a < b \leq t$, and that

$$\dim \operatorname{Im}(H_r(\{h \leq a\}) \to H_r(\{h < b\}))$$

equals the number of bars $[s,t)$ in $\mathcal{B}_r(h)$ such that $s \leq a < b \leq t$. This allows us to interpret the *persistent Betti numbers* m_r, z_r defined in Section 2.3, in the language of barcodes.

Indeed, let M be a compact manifold of dimension n, let $f : M \to \mathbb{R}^d$ be a continuous map and let $0 \leq r < n$ and $\delta > 0$. By the definition of $m_r(f,\delta)$, we have

$$m_r(f,\delta) = \dim \operatorname{Im}(H_r(\{h < -\delta\}) \to H_r(\{h < 0\})),$$

where $h = -|f|$. Hence it follows from the above that $m_r(f,\delta)$ equals the number of bars $[s,t)$ in $\mathcal{B}_r(-|f|)$ such that $s < -\delta < 0 \leq t$, that is, the number of bars in $\mathcal{B}_r(-|f|)$ whose left endpoint is less than $-\delta$ and whose right endpoint is either 0 or $+\infty$. Similarly, we conclude that $z_r(f,\delta)$ equals the number of bars in $\mathcal{B}_r(|f|)$ whose left endpoint is 0 and whose length is greater than δ. As a corollary, we get the bounds

$$m_r(f,\delta) \leq N_{r,\delta}(-|f|) \leq N_\delta(-|f|) \tag{3.1}$$

and

$$z_r(f,\delta) \leq N_{r,\delta}(|f|) \leq N_\delta(|f|). \tag{3.2}$$

Now we formulate the general theorem which compares the bar counting function and the Sobolev norm [10].

Theorem 3.8. *Let (M,g) be a compact Riemannian manifold of dimension n, let $p \geq 1$ and let $k > n/p$ and $d \geq 1$ be integers. Then for any $f \in W^{k,p}(M;\mathbb{R}^d)$ and any $\delta > 0$, we have*

$$N_\delta(|f|), N_\delta(-|f|) \leq C\delta^{-n/k}\|f\|_{W^{k,p}}^{n/k} + b(M), \tag{3.3}$$

where $b(M) = \dim H(M)$ is the total Betti number, and $C = C(M,g,p,k,d)$.

By (3.1) and (3.2), Theorem 3.8 implies[3] Theorem 2.9 from Section 2.3.

Several results akin to Theorem 3.8 were obtained in previous works. In [38], Yomdin showed an estimate on the magnitude of the oscillations of a smooth function in terms of the uniform norm of its higher derivatives (we also refer [24, 26, 37] for earlier related results). Estimates similar to (3.3), stated in the language of persistence barcodes, were known for $p = \infty$ and $k = 1$ [16], and in the case of surfaces for $p = k = 2$ [32] (see also [34] for other estimates of this kind).

4. The transcendental Bézout problem and coarse zero count

By the classical Bézout theorem, the number of common zeroes of n polynomials in n variables is generically bounded by the product of their degrees. The *transcendental Bézout problem* asks about what can be said when we have entire functions instead of polynomials. For an n-tuple $f_1, \dots, f_n \in \mathrm{Hol}(\mathbb{C}^n)$ of entire functions, the number of their isolated common zeroes can be infinite, and generally one can view f_j as polynomials of an infinite degree. In order to make the problem meaningful, introduce the zero counting function

$$\zeta(f, r) = \#(f^{-1}(0) \cap B_r),$$

where $f = (f_1, \dots, f_n) : \mathbb{C}^n \to \mathbb{C}^n$, and $B_r = \{z \in \mathbb{C}^n \mid |z| \le r\}$ is the closed Euclidean ball of radius r in $\mathbb{C}^n$ centered at the origin ($r > 0$). In addition, note that for a complex polynomial, its degree is directly related to the growth rate of the polynomial at ∞. Namely, for $p \in \mathbb{C}[z_1, \dots, z_n]$, we have $k = \deg p$ if and only if $\max_{|z| \le r} |p(z)| = \Theta(r^k)$ when $r \to +\infty$. Hence for extending the notion of the degree to entire maps, it is natural to consider the maximum modulus function

$$\mu(f, r) = \max_{|z| \le r} |f(z)|.$$

Now the transcendental Bézout problem would sound as follows: is $\zeta(f, r)$ bounded in terms of $\mu(f, r)$?

Look first at the case of dimension $n = 1$. For an entire function $f \in \mathrm{Hol}(\mathbb{C})$, by Jensen's formula, we have

$$\frac{1}{2\pi} \int_0^{2\pi} \log|f(re^{i\theta})|\, d\theta - \log|f(0)| = \int_0^r \frac{\zeta(f, t)}{t}\, dt$$

(we assume $f(0) \neq 0$), which readily implies the following corollary.

[3] An additional argument is needed for obtaining the additive constants 0 and $b_r(M)$ in the upper bounds for m_r and z_r in Theorem 2.9.

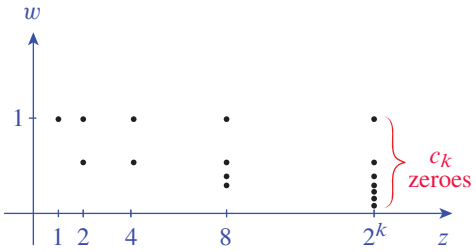


Figure 2. Zeroes of the Cornalba–Shiffman counterexample.

Corollary 4.1. *For every $a > 0$, we have*

$$\zeta(f, r) \le C \log\left(\frac{\mu(f, ar)}{|f(0)|}\right), \tag{4.1}$$

where $C = (\log a)^{-1}$.

Note that the estimate (4.1) is sharp when $f(z) = e^z + 1$. Corollary 4.1 can be viewed as an answer to the transcendental Bézout problem in dimension 1.

It is natural to ask whether the estimate (4.1) generalizes to higher dimensions. Interestingly enough, it turns out that already in dimension 2 such an estimate cannot hold because of a remarkable counterexample due to Cornalba and Shiffman [17]. Namely, they constructed an entire map $F : \mathbb{C}^2 \to \mathbb{C}^2$ such that:

- F is of order 0, that is, for each $\varepsilon > 0$, we have $\log \mu(F, r) \le C_\varepsilon r^\varepsilon$ for $r \ge 1$.
- The growth of $\zeta(F, r)$ is arbitrarily fast. More precisely, we have

$$F^{-1}(0) = \left\{\left(2^i, \frac{1}{j}\right) \mid i = 1, 2, \ldots;\ j = 1, \ldots, c_i\right\} \subset \mathbb{C}^2,$$

  where $c_1 < c_2 < \cdots$ can be any increasing sequence of positive integers. See Figure 2.

The counterexample of Cornalba and Shiffman implies the negative answer to the transcendental Bézout problem in dimension 2 and higher.[4] Nevertheless, let us show that by following the "coarse approach", we can obtain positive results. First, we define the coarse zero count in our setting which is reminiscent of Definition 2.4 from Section 2.2.

Definition 4.2. Let $f : \mathbb{C}^n \to \mathbb{C}^n$ be an analytic map. For $\delta > 0$ and $r > 0$, define $\zeta(f, r, \delta)$ to be the number of connected components of $f^{-1}(B_\delta) \cap B_r$ that contain zeroes of f. See Figure 3.

[4]For any dimension $n > 2$, define the mapping $F_n : \mathbb{C}^n \to \mathbb{C}^n$ by $F_n(z_1, z_2, \ldots, z_n) = (F(z_1, z_2), z_3, \ldots, z_n)$. Then F_n is of order 0, and the growth of $\zeta(F_n, r)$ is arbitrarily fast.

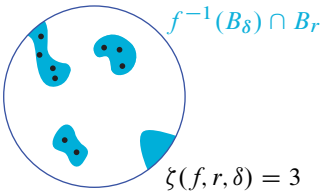


Figure 3. Coarse zero count.

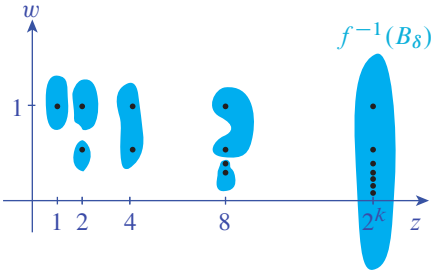


Figure 4. Coarse zero count for the Cornalba–Shiffman counterexample.

In [11], we show the following result.

Theorem 4.3. *For any analytic map $f : \mathbb{C}^n \to \mathbb{C}^n$ and any $a > 1$, for each $r > 0$ and $\delta \in (0, \mu(f, ar)/2)$, we have*

$$\zeta(f, r, \delta) \leq C \left(\log \left(\frac{\mu(f, ar)}{\delta} \right) \right)^{2n-1}, \tag{4.2}$$

where $C = C(n, a)$.

Estimates analogous to (4.2) for the usual count of zeroes had been previously proven under positive lower bounds on the Jacobian of f [25, 28, 29].

It is unclear whether the estimate (4.2) is sharp in terms of the power $2n - 1$ in the upper bound. In the examples known to us, the correct power does not exceed n. For instance, for $f(z_1, \ldots, z_n) = (e^{z_1} + 1, \ldots, e^{z_n} + 1)$, it is not difficult to see that $\zeta(f, r, \delta) \sim r^n$ and $\log \mu(f, r) \sim r$. For the Cornalba–Shiffman counterexample $F : \mathbb{C}^2 \to \mathbb{C}^2$, one verifies [11] that $\zeta(F, r, \delta) \sim \log r$ and $\log \mu(F, r) = \Theta((\log r)^2)$. Moreover, [11] contains a careful analysis of the Cornalba–Shiffman counterexample from the coarse perspective. See Figure 4.

Acknowledgments. The main results that are explicitly mentioned in this survey were obtained in [9–11]; I cordially thank all the coauthors for many interesting discussions and fruitful collaboration. I thank Leonid Polterovich for his help in the

preparation of this note. The note is a write-up of my talk at the 9th ECM, and it is my pleasure to thank the organizers for the invitation.

Funding. The author was partially supported by the Israel Science Foundation grant 1918/23.

References

[1] D. Akhiezer and B. Kazarnovskii, An estimate for the average number of common zeros of Laplacian eigenfunctions (in Russian). *Tr. Mosk. Mat. Obs.* **78** (2017), no. 1, 145–154. English translation: *Trans. Moscow Math. Soc.* **78** (2017), 123–130 Zbl 1395.53057 MR 3738081

[2] D. Akhiezer and B. Kazarnovskii, On common zeros of eigenfunctions of the Laplace operator. *Abh. Math. Semin. Univ. Hambg.* **87** (2017), no. 1, 105–111 Zbl 1367.53034 MR 3623828

[3] V. I. Arnol'd, The topology of real algebraic curves (I. G. Petrovskii's papers and developments from them) (in Russian). *Uspekhi Mat. Nauk* **28** (1973), no. 5, 260–262. English translation: [4] Zbl 0284.57001

[4] V. I. Arnold, The topology of real algebraic curves. In A. Givental, B. Khesin, A. Varchenko, V. Vassilev, and O. Viro (Eds.), *Vladimir I. Arnold – Collected works. Vol. II. Hydrodynamics, bifurcation theory, and algebraic geometry 1965–1972*. Springer, Berlin, 2014, 465 pp. MR 3185029

[5] P. Bérard, P. Charron, and B. Helffer, Non-boundedness of the number of super level domains of eigenfunctions. *J. Anal. Math.* **146** (2022), no. 1, 127–164 Zbl 1495.35129 MR 4467996

[6] P. Bérard and B. Helffer, On Courant's nodal domain property for linear combinations of eigenfunctions. Part 1. *Doc. Math.* **23** (2018), 1561–1585 Zbl 1403.35192 MR 3890963

[7] P. Bérard and B. Helffer, Level sets of certain Neumann eigenfunctions under deformation of Lipschitz domains application to the extended Courant property. *Ann. Fac. Sci. Toulouse Math. (6)* **30** (2021), no. 3, 429–462 Zbl 1479.35581 MR 4331305

[8] P. Bérard and B. Helffer, On Courant's nodal domain property for linear combinations of eigenfunctions. Part II. In S. Albeverio, A. Balslev, and R. Weder (Eds.), *Schrödinger operators, spectral analysis and number theory*, pp. 47–88, Springer Proc. Math. Stat. 348, Springer, Cham, 2021 Zbl 1472.35250 MR 4281621

[9] L. Buhovsky, A. Logunov, and M. Sodin, Eigenfunctions with infinitely many isolated critical points. *Int. Math. Res. Not. IMRN* **2020** (2020), no. 24, 10100–10113 Zbl 1459.35295 MR 4190398

[10] L. Buhovsky, J. Payette, I. Polterovich, L. Polterovich, E. Shelukhin, and V. Stojisavljević, Coarse nodal count and topological persistence. *J. Eur. Math. Soc. (JEMS)* **28** (2026), no. 7, 3131–3202 MR 5061408

[11] L. Buhovsky, I. Polterovich, L. Polterovich, E. Shelukhin, and V. Stojisavljević, Persistent transcendental Bézout theorems. *Forum Math. Sigma* **12** (2024), article no. e72, 28 pp. Zbl 1550.32008 MR 4790778

[12] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot, Proximity of persistence modules and their diagrams. In *Proceedings of the 25th annual symposium on Computational geometry, SCG '09*, pp. 237–246, ACM, New York, 2009 Zbl 1380.68387

[13] F. Chazal, W. Crawley-Boevey, and V. de Silva, The observable structure of persistence modules. *Homology Homotopy Appl.* **18** (2016), no. 2, 247–265 Zbl 1377.55015 MR 3575998

[14] F. Chazal, V. de Silva, M. Glisse, and S. Oudot, *The structure and stability of persistence modules*. SpringerBriefs Math., Springer, Cham, 2016, 120 pp. Zbl 1362.55002 MR 3524869

[15] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams. *Discrete Comput. Geom.* **37** (2007), no. 1, 103–120 Zbl 1117.54027 MR 2279866

[16] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko, Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.* **10** (2010), no. 2, 127–139 Zbl 1192.55007 MR 2594441

[17] M. Cornalba and B. Shiffman, A counterexample to the "Transcendental Bézout Problem". *Ann. of Math. (2)* **96** (1972), no. 2, 402–406 Zbl 0244.32006 MR 0311937

[18] R. Courant and D. Hilbert, *Methoden der mathematischen Physik*. Erster Band, Zweite verbesserte Auflage, Julius Springer, 1931 Zbl 0001.00501

[19] R. Courant and D. Hilbert, *Methods of mathematical physics. Vol. I*. 1st English edn., Interscience, New York, 1953, 561 pp. Zbl 0051.28802 MR 0065391

[20] W. Crawley-Boevey, Decomposition of pointwise finite-dimensional persistence modules. *J. Algebra Appl.* **14** (2015), no. 5, article no. 1550066, 8 pp. Zbl 1345.16015 MR 3323327

[21] H. Donnelly and C. Fefferman, Nodal sets of eigenfunctions on Riemannian manifolds. *Invent. Math.* **93** (1988), no. 1, 161–183 Zbl 0659.58047 MR 0943927

[22] V. M. Gichev, Some remarks on spherical harmonics (in Russian). *Algebra i Analiz* **20** (2008), no. 4, 64–86. English translation: *St. Petersburg Math. J.* **20** (2009), no. 4, 553–567 Zbl 1206.33023 MR 2473744

[23] G. M. L. Gladwell and H. Zhu, The Courant-Herrmann conjecture. *ZAMM Z. Angew. Math. Mech.* **83** (2003), no. 4, 275–281 Zbl 1221.35413 MR 1977003

[24] L. D. Ivanov, *Variations of sets and functions* (in Russian). In A. G. Vituškin (ed.), Izdat. "Nauka", Moscow, 1975, 352 pp. Zbl 0967.26500 MR 0476953

[25] S. Ji, Bézout estimate for entire holomorphic maps and their Jacobians. *Amer. J. Math.* **117** (1995), no. 2, 395–403 Zbl 0824.32001 MR 1323681

[26] A. S. Kronrod, On functions of two variables (in Russian). *Uspehi Matem. Nauk (N.S.)* **5** (1950), no. 1(35), 24–134 Zbl 0040.31603 MR 0034826

[27] M. Lesnick, The theory of the interleaving distance on multidimensional persistence modules. *Found. Comput. Math.* **15** (2015), no. 3, 613–650 Zbl 1335.55006 MR 3348168

[28] B. Q. Li, On the Bézout problem and area of interpolating varieties in $\mathbf{C}^n$, II. *Amer. J. Math.* **120** (1998), no. 6, 1191–1198 Zbl 0954.58006 MR 1657162

[29] B. Q. Li and B. A. Taylor, On the Bézout problem and area of interpolating varieties in $\mathbf{C}^n$. *Amer. J. Math.* **118** (1996), no. 5, 989–1010 Zbl 0863.32002 MR 1408496

[30] S. Y. Oudot, *Persistence theory: from quiver representations to data analysis*. Math. Surveys Monogr. 209, American Mathematical Society, Providence, RI, 2015, 218 pp. Zbl 1335.55001 MR 3408277

[31] Å. Pleijel, Remarks on Courant's nodal line theorem. *Comm. Pure Appl. Math.* **9** (1956), no. 3, 543–550 Zbl 0070.32604 MR 0080861

[32] I. Polterovich, L. Polterovich, and V. Stojisavljević, Persistence barcodes and Laplace eigenfunctions on surfaces. *Geom. Dedicata* **201** (2019), 111–138 Zbl 1420.58012 MR 3978537

[33] L. Polterovich, D. Rosen, K. Samvelyan, and J. Zhang, *Topological persistence in geometry and analysis*. Univ. Lecture Ser. 74, American Mathematical Society, Providence, RI, 2020, 128 pp. Zbl 1464.55003 MR 4249570

[34] L. Polterovich and M. Sodin, Nodal inequalities on surfaces. *Math. Proc. Cambridge Philos. Soc.* **143** (2007), no. 2, 459–467 Zbl 1127.58029 MR 2364662

[35] V. Stojisavljević, Harmonic functions with highly intersecting zero sets. 2024, arXiv: 2410.03975v1

[36] O. Ya. Viro, Construction of multicomponent real algebraic surfaces (in Russian). *Sov. Math. Dokl.* **20** (1979), no. 5, 991–995. English translation: *Dokl. Akad. Nauk SSSR* **248** (1979), no. 2, 279–282 Zbl 0435.14006 MR 0553188

[37] A. G. Vituškin, *On multi-dimensional variations* (in Russian). Gosudarstv. Izdat. Tehn.-Teor. Lit., Moscow, 1955, 220 pp. Zbl 0063.09196 MR 0075267

[38] Y. Yomdin, Global bounds for the Betti numbers of regular fibers of differentiable mappings. *Topology* **24** (1985), no. 2, 145–152 Zbl 0566.57014 MR 0793180

Lev Buhovsky
School of Mathematical Sciences, Tel Aviv University, P.O. Box 39040, 69978 Tel Aviv, Israel; levbuh@tauex.tau.ac.il

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/15

Unstable vortices and non-uniqueness for 2D Euler and α-gSQG

Ángel Castro

Abstract. In the sixties, Yudovich proved global existence and uniqueness of solutions for the 2D Euler incompressible equation in $L^1 \cap L^\infty$. This result extends to the case with a force in $L^1_t(L^1 \cap L^\infty)$. Global existence in $L^1 \cap L^p$, $1 < p < \infty$, was obtained by R. DiPerna and A. Majda in 1987, but the problem of uniqueness, in this range of p, has remained open until Vishik's work in 2018. Indeed, M. Vishik proved that there are forces in $L^1_t(L^1 \cap L^p)$, $1 < p < \infty$, such that at least two different solutions exist for 2D Euler starting in the same initial datum. A nice exposition of the proof of this result can be found in D. Albritton et al. (2024). In this paper, we will present an alternative approach to prove this theorem. More precisely, we will show a different way to construct unstable radial vortices with compact support which also give rise to non-uniqueness. This strategy also works for α-gSQG, $0 \leq \alpha \leq 1$. This family of transport equations, with velocity $\mathbf{u} = -\nabla^\perp(-\Delta)^{-1+\frac{\alpha}{2}}\theta$, for the scalar θ, interpolates from 2D Euler ($\alpha = 0$) to the surface quasi-geostrophic equation ($\alpha = 1$).

1. Introduction

This article is a written version of the talk "Unstable vortices and non-uniqueness for 2D Euler and α-gSQG" in the 9ECM. This talk consisted of a review of some results concerning non-uniqueness in 2D and 3D Euler, Navier–Stokes and generalized surface quasi-geostrophic equations and a presentation of some results by the author and his collaborators, Daniel Faraco, Francisco Mengual, and Marcos Solera-Diana. A wide part of these results is published in [19].

The main topics of this paper will be the 2D Euler equations, the surface quasi-geostrophic equation (SQG) and, an interpolation between both of them, the generalized surface quasi-geostrophic equation (α-gSQG).

The forced Navier–Stokes system models the motion of an incompressible fluid. It consists of two equations for the velocity, $\mathbf{u}(\mathbf{x}, t)$, and the pressure, $p(\mathbf{x}, t)$, and

Mathematics Subject Classification 2020: 76B03 (primary); 76D03 (secondary).
Keywords: 2D Euler, incompressible fluid, uniqueness, SQG.

reads

$$\partial_t \mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla p + \nu\Delta\mathbf{u} + \mathbf{g}, \quad (\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}^+, \tag{1.1}$$

$$\nabla\cdot\mathbf{u} = 0, \quad (\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}^+, \tag{1.2}$$

$$\mathbf{u}(\mathbf{x},0) = \mathbf{u}^\circ(\mathbf{x}), \tag{1.3}$$

where $\mathbf{u}^\circ(\mathbf{x})$ is the given initial datum, $\mathbf{g}(\mathbf{x},t)$ a divergence-free force, $\nu \geq 0$ is the viscosity and $d = 2$, or 3. The forced Euler equations arise in the case without viscosity, $\nu = 0$, and can be written in terms of the vorticity, $\omega(\mathbf{x},t)$. In dimension 2, the vorticity is given by

$$\omega = \nabla^\perp \cdot \mathbf{u},$$

and, assuming decay at infinity, one can find the following transport equation for it:

$$\partial_t \omega + \mathbf{u}\cdot\nabla\omega = f, \quad (\mathbf{x},t) \in \mathbb{R}^2 \times \mathbb{R}^+, \tag{1.4}$$

$$\mathbf{u} = \nabla^\perp \Delta^{-1}\omega, \tag{1.5}$$

$$\omega(\mathbf{x},0) = \omega^\circ(\mathbf{x}), \tag{1.6}$$

where

$$\Delta^{-1}\omega(\mathbf{x}) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(\mathbf{x}-\mathbf{y})^\perp}{|\mathbf{x}-\mathbf{y}|^2}\omega(\mathbf{y})d\mathbf{y},$$

and

$$f = \nabla^\perp \cdot \mathbf{g}.$$

The SQG equation has a geophysical origin (see [81]). Roughly speaking, it models the evolution of the temperature of the air on the Earth and comes from the general quasi-geostrophic system after some assumptions and approximations. The forced SQG has the form

$$\partial_t \theta + \mathbf{u}\cdot\nabla\theta = f, \quad (\mathbf{x},t) \in \mathbb{R}^2 \times \mathbb{R}^+, \tag{1.7}$$

$$\mathbf{u} = -\nabla^\perp(-\Delta)^{-\frac{1}{2}}\theta, \tag{1.8}$$

$$\theta(\mathbf{x},0) = \theta^\circ(\mathbf{x}), \tag{1.9}$$

where

$$(-\Delta)^{-\frac{1}{2}}\theta(\mathbf{x}) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{\theta(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|}d\mathbf{y}.$$

In [31], P. Constantin, A. Majda and E. Tabak proposed SQG as a scalar 2D model of the 3D Euler equation because of the similarity between their structures. Since then, there has been intense research on it to understand the behavior of its solutions.

The α-generalized SQG equation also has a geophysical origin and, mathematically, is an interpolation between 2D Euler and SQG. Indeed, forced α-SQG reads

$$\begin{aligned} \partial_t \theta + \mathbf{u} \cdot \nabla \theta &= f, && (\mathbf{x}, t) \in \mathbb{R}^2 \times \mathbb{R}^+, \\ \mathbf{u} &= -\nabla^\perp (-\Delta)^{-\frac{\alpha}{2}} \theta, \\ \theta(\mathbf{x}, 0) &= \theta^\circ(\mathbf{x}), \end{aligned}$$

where

$$(-\Delta)^{-\frac{\alpha}{2}} \theta(\mathbf{x}) = C_\alpha \int_{\mathbb{R}^2} \frac{\theta(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|^\alpha} d\mathbf{y}.$$

Then for $\alpha = 0$ we obtain 2D Euler, and for $\alpha = 1$ we obtain SQG. We remark that, for α-gSQG, the velocity is an operator of degree $1 - \alpha$, and this makes a big difference between the members of this family of transport equations.

In this paper, we focus on the uniqueness problem for α-gSQG, with $0 \leq \alpha \leq 1$. One could also consider the case $1 < \alpha < 2$, which has been studied in the literature, but we will skip it here. The starting point for us will be the celebrated Yudovich's theorem.

Theorem 1.1 (Yudovich). *Let $\omega_0 \in L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$ and $f \in L^1([0, T], L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2))$ for all $T > 0$. Then there exists a solution*

$$\omega(\mathbf{x}, t) \in L^\infty([0, T], L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)),$$

with

$$\mathbf{u}(\mathbf{x}, t) \in L^\infty([0, T], L^2(\mathbb{R}^2)),$$

to (1.4), (1.5) and (1.6) for all $T < \infty$. In addition, this solution is unique in the class $L^\infty([0, T), L^1(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2))$.

Actually, Yudovich proved this theorem in [91] with $f = 0$, but his proof can be extended to the case with $f \in L^1_t(L^1 \cap L^\infty)$ as shown in [2].

The question of existence in $L^1 \cap L^p$ with $1 < p < \infty$ was addressed by R. DiPerna and A. Majda [48], who gave a positive answer. However, the question of whether there can be more than one solution in the class $L^\infty_t(L^1 \cap L^p)$, with $1 < p < \infty$, for the same initial datum remains open today in most of the cases. However, very recently, there have been some remarkable advances in this question. In [86, 87], M. Vishik proved the following theorem.

Theorem 1.2 (Vishik). *For any $2 < p < \infty$, there exists a force*

$$f \in L^1_t(L^1 \cap L^p) \quad \text{with } g \in L^1_t L^2, \tag{1.10}$$

where $g = \nabla^{\perp}\Delta^{-1} f$*, with the property that there are two different solutions*

$$\omega_0, \omega_1 \in L^{\infty}_t(L^1 \cap L^p) \quad \textit{with } v_0, v_1 \in L^{\infty}_t L^2, \tag{1.11}$$

where $v_j = \nabla^{\perp}\Delta^{-1}\omega_j$*, to the Euler equation starting from* $\omega^{\circ} = 0$*.*

Thus, this result states that there exists at least a force such that uniqueness fails for 2D Euler in the class $L^{\infty}_t(L^1 \cap L^p)$ for any $1 < p < \infty$. A nice exposition of Vishik's proof of Theorem 1.2 can be found in [2], where the authors, apart from elaborating on some details, improve it in several aspects. In addition, in [19], the author, together with D. Faraco, F. Mengual and M. Solera-Diana, gave a proof of Theorem 1.2 which is different in several points from Vishik's proof: mainly, in the construction of the unstable vortex which gives rise to non-uniqueness. The main purpose of this paper is the presentation of this construction and its extension to α-gSQG. After Vishik's result, the main open question in this context is whether Theorem 1.2 can be proven with $f = 0$. In [9], this problem has been partially solved. Indeed, E. Bruè, M. Colombo and A. Kumar claim that there exists a small ε such that uniqueness fails in $L^1 \cap L^{1+\varepsilon}$. Despite the great interest of this last paper, its analysis exceeds the scope of this review, and we will not go into details about it.

This paper is organized as follows. In Section 2, we will review some results concerning local and global existence for 2D Euler, SQG and α-SQG. In Section 3, we will present some literature dealing with the problem of non-uniqueness in fluid mechanics without forcing. This section will consist of two parts: in the first one, we shall consider papers that approach the problem using convex integration, and in the second one, those dealing with self-similar solutions. In Section 4, we will give some more details about what is known about uniqueness in fluid mechanics when we introduce a force. Finally, in Section 5, we will sketch our proof of Theorem 1.2.

2. Local and global existence of solutions

Global existence and uniqueness of classical solutions to the 2D Euler equations was proven by W. Wolibner [89] and E. Hölder [63] with velocity $C^{k,\gamma}$ for $k \geq 1$ and $\gamma > 0$. Local well-posedness in $C^{k\gamma}$ had been previously established by L. Lichtenstein [77] and N. Gunther [61]. Local well-posedness in H^s, $s > \frac{d}{2} + 1$, was proven by D. Ebin and J. Marsden [49] in a compact domain and by T. Kato [72] in $\mathbb{R}^d$. Local well-posedness in $W^{s,p}$, $s > \frac{d}{p} + 1$, was provided by J.P. Bourguignon and H. Brezis [6] in a compact domain and by T. Kato and G. Ponce [73] in $\mathbb{R}^d$.

The global well-posedness in 2D sharply contrasts with the 3D case due to vortex stretching, as demonstrated by T. Elgindi [50], who showed finite-time singularities with $\omega^{\circ} \in C^{\gamma}$ for some $0 < \gamma \ll 1$ in the unforced case (see also [24,39]). D. Córdoba

and L. Martínez-Zoroa [35] upgrade the regularity of the singular solutions in the forced case with $f \in L_t^\infty C^{\frac{1}{2}-\varepsilon}$. For numerical studies about singularities in $\mathbb{R}^3$, one can consult [64] and [88]. In the case with boundaries, we refer to the works [25] and [26] where J. Chen and T. Hou claimed finite-time singularities for smooth initial data and finite energy. Let us also mention the work [76] by Kiselev and Šverák for the related issue on small-scale creation in two dimensions.

J. Bourgain and D. Li [5] and, independently, T. Elgindi and N. Masmoudi [51] showed strong ill-posedness in C^k, $k \geq 1$. See [4] for ill-posedness in $H^{\frac{d}{2}+1}$ and [74] for ill-posedness in $W^{\frac{d}{p}+1,p}$. Instantaneous gap loss of Sobolev regularity with initial vorticity in H^β, $0 < \beta < 1$, has been proven in [37] by D. Córdoba, L. Martínez-Zoroa and W. Ożański.

For SQG, global existence of solutions in L^2 was provided by S.G. Resnick [82]. F. Marchand [78] extended this result to L^p for $p > \frac{4}{3}$. Local well-posedness in H^s, $s > 2$, can be found in [20] and [31] (see also [32] for well-posedness in a bounded domain). J. Wu [90] showed existence in $C^r \cap L^q$, $r > 1$ and $q > 1$. The problem of whether SQG presents finite-time singularities is one of the most outstanding problems in mathematical fluid mechanics. A. Kiselev and F. Nazarov [75] showed norm inflation and S. Friedlander and R. Shvydkoy [53] unstable eigenvalues in the spectrum. In addition, exponential growth of the C^2-norm has been shown by S. He and A. Kiselev [62]. There are also examples of global solutions with non-trivial dynamics. See, for example, [3, 18, 57, 58]. Ill-posedness in C^k, $k \geq 2$, and in H^s, $s \in (3/2, 2)$, was proven by D. Córdoba and L. Martínez-Zoroa [34] (see also [51, 69]). Instantaneous loss of regularity in H^s, with $s \in (3/2, 2)$, has been recently proven in [38].

In the case of α-gSQG, in [21], D. Chae, P. Constantin, D. Córdoba and F. Gancedo extended Resnick's result, showing also global existence in L^2. In [22], a global existence criterion is shown: solutions in C^σ, $\sigma > 1$, exist as far as the $L^1([0,T), C^\alpha)$-norm is under control. The possible formation of singularities is still unknown. Finally, let us remark that existence and uniqueness in the spaces $H^{(1+\alpha)+}$ was proven in [23] (see Theorem 1.2 and take $\beta = 2 - \alpha - \varepsilon$ and $P(\xi) = |\xi|^{-\varepsilon}$). For the more singular case $\alpha > 1$, the interested reader can consult [21] for well-posedness and [36] for ill-posedness.

3. Uniqueness without forcing

In this section, we will take the force equal to zero and approach the problem of uniqueness into two different ways: convex integration and self-similar solutions.

3.1. Convex integration and non-uniqueness

In the last decades, convex integration has been applied to the uniqueness problem in incompressible fluid mechanics with great success. We can highlight, for example, the proof of Onsager's conjecture for the Euler equations. This conjecture states that there exist solutions $\mathbf{u} \in C^\alpha$, with $\alpha < \frac{1}{3}$, which dissipate kinetic energy. It was solved in 3D in [12] and [65] after a series of works which used convex integration. The fact that solutions of the Euler equations in C^α, with $\alpha > \frac{1}{3}$, do conserve energy had been previously proven in [33]. See also the work [28] by A. Cheskidov, P. Constantin, S. Friedlander and R. Shvydkoy on critical regularity.

The first proofs of non-uniqueness for the Euler equations were provided independently by V. Scheffer [83] and A. Shnirelman [85]. They proved the existence of solutions with compact support in time whose velocity $\mathbf{u} \in L^2_{t,x}$. After that, C. De Lellis and L. Székelyhidi [42] introduced a systematic way to run convex integration in the problem which produced solutions with $\mathbf{u} \in L^\infty_t L^2_x$ and compact support in space and time. In a series of papers, the same authors improved the regularity and properties of these non-unique solutions (see [43–46]). Regularity $C^{\frac{1}{5}}$ was achieved in [10] by T. Buckmaster, C. De Lellis, P. Isett and L. Székelyhidi and regularity $L^1_t C^{\frac{1}{3}-}_x$ in [11]. Finally, P. Isett proved, in [65], non-uniqueness for 3D Euler with velocity in $C^{\frac{1}{3}-}$ and T. Buckmaster, C. De Lellis, V. Vicol and L. Székelyhidi dissipation of kinetic energy also in $C^{\frac{1}{3}-}$. See also the work [80] by M. Novack and V. Vicol on an intermittent Onsager theorem, based on their joint work [13] with T. Buckmaster and N. Masmoudi, as well as the work [55] on the L^3-based strong Onsager conjecture by V. Giri, H. Kwon and M. Novack.

The dissipative part of the Onsager conjecture in dimension 2 is harder to deal with and was solved recently in [56] by V. Giri and R.-O. Radu by means of convex integration. We refer to the recent work [47] by De Rosa and Park for the related issue on anomalous dissipation in two dimensions.

To the best of our knowledge, the problem of non-uniqueness for the unforced Euler equation in the velocity class $L^2 \cap C^\gamma$ remains open in the regime $\frac{1}{3} \leq \gamma < 1$.

In [79], F. Mengual proved the existence, for any $2 < p < \infty$, of initial data $v^\circ \in L^2$ with $\omega^\circ \in L^1 \cap L^p$ for which there are infinitely many admissible solutions $v \in C_t L^2$. This result shows the sharpness of Yudovich's proof of uniqueness, but with the drawback that the vorticity information is lost for positive times.

We emphasize that the convex integration method allowed for proving non-uniqueness of Navier–Stokes solutions in [15] by T. Buckmaster and V. Vicol, with $\mathbf{u} \in C_t H^\beta$ for small β, and sharpness of the Ladyzhenskaya–Prodi–Serrin criteria [29] by A. Cheskidov and X. Luo.

Convex integration has also been applied to study the lack of uniqueness for SQG. The first result in that direction was provided by T. Buckmaster, S. Shkoller and

V. Vicol [14]. They proved non-uniqueness with $\Lambda^{-1}\theta \in C_t^\sigma C_x^\beta$, with $\frac{1}{2} < \beta < \frac{4}{5}$ and $\sigma < \frac{\beta}{(2-\beta)}$. P. Isett and A. Ma improved this regularity in [67] and obtained an analogous result for α-gSQG in [68] (see also [27]). Very recently, what is considered the dissipative part of Onsager's conjecture for SQG has been proven: independently, M. Dai, V. Giri and R.-O. Radu [41] and S. Looi and P. Isett [66] have proven energy dissipation and non-uniqueness of solutions in $C_{t,x}^{0-}$.

3.2. Self-similar solutions and non-uniqueness

In this section, we will explain the relation between non-uniqueness and the stability of self-similar solutions. For the sake of clarity, we will take as a model the 2D Euler equation, although similar arguments can be applied to other equations including, of course, α-gSQG. We have borrowed the following argument from [71].

We will look for self-similar solutions, i.e., solutions with the form

$$\omega(\mathbf{x}, t) = t^\alpha \Omega(t^\beta \mathbf{x}),$$

where $\alpha, \beta \in \mathbb{R}$. Introducing this ansatz in (1.4), we get

$$\alpha t^{\alpha-1}(\alpha\Omega(\xi) + \beta\xi \cdot \nabla\Omega(\xi)) + t^{2\alpha}\mathbf{u}[\Omega](\xi) \cdot \nabla\Omega(\xi) = 0,$$

where

$$\mathbf{u}[\Omega](\xi) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{(\xi - \bar{\xi})^\perp}{|\xi - \bar{\xi}|^2} d\bar{\xi}, \quad \text{and} \quad \xi = t^\beta \mathbf{x}.$$

Thus, taking $\alpha = -1$, we obtain the time-independent equation

$$\Omega(\xi) + (\beta\xi + \mathbf{u}[\Omega](\xi)) \cdot \nabla\Omega(\xi) = 0. \tag{3.1}$$

Let us assume that Ω satisfies (3.1) and look for solutions of the form

$$\omega(\mathbf{x}, t) = t^{-1}\Omega(t^\beta \mathbf{x}) + t^{-1}\bar{\Omega}(t^\beta \mathbf{x}, t).$$

Thus, $\bar{\Omega}$ needs to satisfy

$$\begin{aligned} t\partial_t\bar{\Omega}(\xi, t) + \bar{\Omega}(\xi, t) + (\beta\xi + \mathbf{u}[\Omega](\xi, t)) \cdot \nabla\bar{\Omega}(\xi, t) \\ + \mathbf{u}[\bar{\Omega}](\xi, t) \cdot \nabla\bar{\Omega}(\xi, t) = 0. \end{aligned} \tag{3.2}$$

At the linear level, we find that (3.2) reads

$$t\partial_t\Omega_{\text{lin}}(\xi, t) - \mathcal{L}[\Omega]\Omega_{\text{lin}} = 0, \tag{3.3}$$

where

$$\mathcal{L}_{\text{lin}}\Omega = -1 - (\beta\xi + \mathbf{u}[\Omega]) \cdot \nabla.$$

Next, let us assume that there exists W (nice) such that

$$\mathcal{L}[\Omega]W = \lambda W,$$

with $\lambda \in \mathbb{C}$ and $\mathrm{Re}(\lambda) > 0$. These properties give us the following solution for (3.3)

$$\Omega_{\mathrm{lin}}(\xi, t) = t^{\lambda} W(\xi),$$

which starts from zero at time $t = 0$, i.e., $\Omega_{\mathrm{lin}}(\xi, 0) = 0$. Thus, we have two different solutions starting from zero for (3.3), namely, Ω_{lin} and 0.

Going further, we now write

$$\omega(\mathbf{x}, t) = t^{-1}\Omega(\xi) + t^{-1}\Omega_{\mathrm{lin}}(\xi, t) + t^{-1}\Omega_{\mathrm{cor}}(\xi, t)$$

to try to include the nonlinear terms in (3.2) that we have neglected in (3.3). Thus,

$$\begin{aligned} t\,\partial_t \Omega_{\mathrm{cor}}(\xi, t) - \mathcal{L}[\Omega]\Omega_{\mathrm{cor}} &= \mathcal{F}[\Omega_{\mathrm{lin}}, \Omega_{\mathrm{cor}}](\xi, t), \\ \Omega_{\mathrm{cor}}(\xi, 0) &= 0, \end{aligned}$$

where

$$\mathcal{F}[\Omega_{\mathrm{lin}}, \Omega_{\mathrm{cor}}] = -\mathbf{u}[\Omega_{\mathrm{lin}} + \Omega_{\mathrm{cor}}] \cdot \nabla(\Omega_{\mathrm{lin}} + \Omega_{\mathrm{cor}}).$$

Notice that the initial condition $\Omega_{\mathrm{cor}}(\xi, 0) = 0$ gives us some hope to find a solution $\Omega_{\mathrm{cor}}(\xi, t)$ of a smaller order than $\Omega_{\mathrm{lin}}(\xi, t)$ at $t = 0$. Let us assume that we have such a solution. We have then constructed two of them, namely,

$$\begin{aligned} \omega_0(\mathbf{x}, t) &= t^{-1}\Omega(t^{\beta}\mathbf{x}), \\ \omega_1(\mathbf{x}, t) &= t^{-1}\Omega(t^{\beta}\mathbf{x}) + t^{-1}\Omega_{\mathrm{lin}}(t^{\beta}\mathbf{x}, t) + t^{-1}\Omega_{\mathrm{cor}}(t^{\beta}\mathbf{x}, t). \end{aligned}$$

Finally, if Ω, Ω_{lin} and $\Omega_{\mathrm{cor}} \in L^{\infty}_t L^p$, a straightforward computation shows that

$$\|\omega_i(\cdot, t)\|_{L^p} \leq C t^{-1-\frac{2\beta}{p}}$$

for both $i = 0, 1$. If $p = \infty$, we fail; however, for any $p < \infty$, we can take β such that

$$\omega_i \to 0 \quad \text{in } L^p$$

for both $i = 0, 1$.

In dimension 2, it is well known that Leray–Hopf solutions of the Navier–Stokes systems are unique. However, the same question in 3D is an open problem. Let us remind the reader of the definition of a Leray–Hopf solution. We will include in it a force, since it will be needed in the following sections.

Definition 3.1. Let $T > 0$, $\mathbf{u}^\circ \in L^2(\mathbb{R}^3)$ be a divergence-free vector field and $\mathbf{g} \in L^1([0,T); L^2(\mathbb{R}^3))$. A Leray–Hopf solution on $\mathbb{R}^3 \times (0,T)$ with initial datum $\mathbf{u}^\circ$ and force $\mathbf{g}$ is a divergence-free vector field

$$\mathbf{u} \in L^\infty([0,T), L^2(\mathbb{R}^3)) \cap L^2([0,T), H^1(\mathbb{R}^3))$$

which

(1) belongs to $C_w([0,T], L^2(\mathbb{R}^3))$ and attains the initial data: $\mathbf{u}(x,0) = \mathbf{u}^\circ$,

(2) solves the forced Navier–Stokes equations in the sense of distributions on $\mathbb{R}^3 \times (0,T)$ for some pressure $p \in L^1_{\text{loc}}(\mathbb{R}^3 \times (0,T))$ and

(3) satisfies the energy inequality for all $t \in (0,T]$:

$$\frac{1}{2}\int_{\mathbb{R}^3} |\mathbf{u}(\mathbf{x},t)|^2 d\mathbf{x} + \int_0^T |\nabla \mathbf{u}(\mathbf{x},s)|^2 d\mathbf{x} ds \leq \frac{1}{2}\int_{\mathbb{R}^2} |\mathbf{u}^\circ(\mathbf{x})|^2 d\mathbf{x} + \int_0^T \mathbf{g}(\mathbf{x},s)\cdot \mathbf{u}(\mathbf{x},s) d\mathbf{x} ds.$$

In a couple of papers, [70, 71], H. Jia and V. Sverák explored the ideas at the beginning of this section to study the uniqueness of Leray–Hopf solutions of the 3D Navier–Stokes equations (without forcing). Indeed, in [70], they proved the existence of self-similar solutions.

Theorem 3.2 (Jia, Sverák). *Assume $\mathbf{u}^\circ$ is scale invariant and locally Hölder continuous in $\mathbb{R}^3 \setminus \{0\}$ with $\nabla \cdot \mathbf{u}^\circ = 0$ in $\mathbb{R}^3$. Then the Navier–Stokes equations have at least one scale-invariant solution $\mathbf{u}(\mathbf{x},t)$ which is smooth in $\mathbb{R}^3 \times (0,\infty)$ and locally Hölder continuous in $\mathbb{R}^3 \times [0,\infty) \setminus \{(0,0)\}$.*

Here, a scale-invariant solution satisfies $\lambda \mathbf{u}(\lambda x, \lambda^2 t) = \mathbf{u}(\mathbf{x},t)$, and the theorem is proven by looking at self-similar solutions of the form $\mathbf{u}(\mathbf{x},t) = t^{-\frac{1}{2}}\mathbf{U}t'(\mathbf{x}t^{-\frac{1}{2}})$.

After this theorem, they also proved that these self-similar solutions could yield non-uniqueness of Leray–Hopf solutions in [71]. In order to do it, they need to assume that, roughly speaking, there exist unstable self-similar solutions satisfying several conditions. We shall make this statement more precise. The linearization of the Navier–Stokes equations around a family of self-similar solutions $\mathbf{U}_\sigma$, parameterized by σ, reads

$$t\partial_t \phi = \mathcal{L}_\sigma \phi,$$

where

$$\mathcal{L}_\sigma \phi \equiv \Delta \Phi + \frac{\mathbf{x}}{2}\cdot\nabla\phi + \frac{1}{2}\phi - \mathbf{U}_\sigma \nabla\phi - \phi\nabla\mathbf{U}_\sigma. \tag{3.4}$$

We define the domain of $\mathcal{L}_\sigma$ as

$$\mathcal{D} = \{\phi \in X : \partial_j\phi,\ \partial^2_{ij}\phi \in X, \text{ and } \mathbf{x}\cdot\nabla\phi \in X\},$$

with

$$X = \{\phi \in L^4(\mathbb{R}^3) \cap L^2(\mathbb{R}^3) : \nabla \cdot \phi = 0\}.$$

Then two different Leray–Hopf solutions exist for Navier–Stokes starting in the same initial data if the following condition (B) is satisfied:

(B) For $\sigma < \sigma_0$, the eigenvalues of $\mathcal{L}_\sigma \subset \{\lambda \in \mathbb{C} : \mathrm{Re} < -\delta < 0\} \cup \{\lambda_1(\sigma)\}$ for some $\delta > 0$ and $\lambda_1(\sigma) < 0$. The eigenvalue $\lambda_1(\sigma)$ is simple. Moreover, $\lim_{\sigma \to \sigma_0^-} \lambda_1(\sigma) = 0$.

In [71], an additional condition (A) is given, which also yields non-uniqueness. Remarkably, in [60], J. Guillod and V. Šverak found numerical evidence of the existence of self-similar solutions satisfying condition (B).

There have been several authors who have studied the existence of self-similar solutions for 2D Euler. In [52], V. Elling showed the existence of self-similar solutions satisfying

$$\lim_{t \to 0^+} \omega(\mathbf{x}, t) = r^{-\frac{1}{\mu}} \widetilde{\omega}(\theta). \tag{3.5}$$

More precisely, the result can be formulated as follows.

Theorem 3.3 (Elling). *Given $\varepsilon > 0$ and $\mu \in \left(\frac{2}{3}, \infty\right)$, there is an $N_0 \in \mathbb{N}$ so that a weak solution of 2D Euler satisfying (3.5) exists for all initial data $\widetilde{\omega}$ satisfying the following conditions:*

(1) *Periodicity: $\widetilde{\omega}$ is $\frac{2\pi}{N}$-periodic for $N \geq N_0$.*

(2) *Dominant rotation: the Fourier coefficients satisfy*

$$|\widetilde{\omega}\hat{}(0)| \geq \varepsilon \sum_{n \neq 0} |\widetilde{\omega}\hat{}(n)|.$$

For α-gSQG, $0 < \alpha < 1$, introducing some new ideas, C. García and J. Gómez-Serrano [54] proved the existence of self-similar solutions.

Theorem 3.4 (García, Gómez-Serrano). *Let $\alpha \in (0, 1)$ and $\gamma \in (1, 1 + \alpha)$. Then there exists $\varepsilon > 0$ such that for any $\frac{2\pi}{N}$-periodic $\widetilde{\theta} \in B_{L^p(\mathbb{T})}(1, \varepsilon)$ and $\frac{p>1}{(1-\alpha)}$, the initial condition $\theta^\circ = r^{-(1+\gamma-\alpha)}\widetilde{\theta}$ describes a self-similar solution for α-gSQG for any $N \geq 1$.*

And, in [84], F. Shao, D. Wei and Z. Zhang improved Elling's result for 2D Euler allowing $\mu > \frac{1}{2}$, $N_0 = 2$ and $\widetilde{w} \in L^1(\mathbb{T})$.

To conclude this section, we will comment on the program by A. Bressan, R. Murray and W. Shen, collected in the papers [7, 8], to prove non-uniqueness for 2D Euler. This program is based on some numerical simulations by W. Shen which

show two different behaviors of the evolution of two different sequences of initial data which converge to the same function.

These two sequences of initial data are constructed in the following way. First, we define

$$\bar{\omega}(\mathbf{x}) = r^{-\frac{1}{\mu}}\bar{\Omega}(\theta),$$

where $\bar{\Omega} \in C^\infty$ is nonnegative periodic and such that

$$\bar{\Omega}(\theta+\pi) = \bar{\Omega}(\theta), \quad \bar{\Omega}(\theta) = 0 \quad \text{if } \frac{\pi}{4} \le \theta \le \pi.$$

And now we take

$$\omega_1^\varepsilon(\mathbf{x}) = \begin{cases} \varepsilon^{-\frac{1}{\mu}}, & |\mathbf{x}| < \varepsilon, \\ \bar{\omega}(\mathbf{x}), & |\mathbf{x}| > \varepsilon, \end{cases} \qquad \omega_2^\varepsilon(\mathbf{x}) = \begin{cases} 0, & |\mathbf{x}| < \varepsilon, \\ \bar{\omega}(\mathbf{x}), & |\mathbf{x}| > \varepsilon. \end{cases}$$

The numerical simulations by W. Shen show that, on one hand, the sequence of ω_1^ε gives rise to a spiral-shaped structure at the point $\mathbf{x} = (0,0)$. On the other hand, the sequence ω_2^ε yields two spiral-shaped structures whose origins move from the point $\mathbf{x} = (0,0)$. These two different evolutions are reminiscent of the Kaden and Pullin spirals (see [30] and references therein).

4. Uniqueness with forcing

After the seminal works [86, 87] by M. Vishik, there have been some advances in the study of the Euler and Navier–Stokes equations with forcing. Apart from the papers [86, 87] and [2] dealing with the forced 2D Euler equations, here we will highlight the work [1] where D. Albritton, E. Bruè and M. Colombo prove non-uniqueness of Leray–Hopf solutions with force in $L^1_t L^2(\mathbb{R}^3)$.

Theorem 4.1 (Albritton, Bruè, Colombo). *There exist $T > 0$, $\mathbf{g} \in L^1((0,T), L^2(\mathbb{R}^3))$ and two distinct Leray–Hopf solutions $\mathbf{u}_1$, $\mathbf{u}_2$ to the forced Navier–Stokes system on $\mathbb{R}^2 \times (0,t)$ with force $\mathbf{g}$ and initial condition $\mathbf{u}^\circ = 0$.*

In the next section, we will briefly sketch Vishik's proof in [86, 87] and [2].

To conclude this section, let us comment that, in the forced case, A. Bulut, M. Huynh and S. Palasek [16], using convex integration, proved that the regularity in 3D can be upgraded to $C^{\frac{1}{2}-}$. The same authors applied a similar strategy to SQG in [17]. For other applications of convex integration to forced equations, we refer to the work [40] by M. Dai and S. Friedlander, and the references therein.

5. Sketch of the proof

In this section, we will sketch our proof of Theorem 1.2. The strategy is similar to the one in either [86, 87] or [2], except for the construction of the unstable vortex, which is totally different. In addition, the properties of our unstable vortex, which plays a fundamental role in the proof, enable us to simplify, in [19], some arguments and computations.

The proof of Theorem 1.2 divides into three steps: Eulerian, self-similar and nonlinear instability.

5.1. Eulerian instability

We start by recalling how linear stability for the Euler equation leads to the Rayleigh stability equation, which we found convenient to express in vorticity form, contrary to the usual practice.

Given a background (steady) vorticity $\bar{\omega}$, a perturbation $\bar{\omega} + \varepsilon\omega$ is a second solution to the Euler equation if the deviation ω satisfies the equation

$$(\partial_t - \mathcal{L})\omega + \varepsilon(\mathbf{u} \cdot \nabla\omega) = 0.$$

Here, $\mathcal{L} = \mathcal{L}_{\bar{\omega}}$ is the linearization of the 2D Euler equation around $\bar{\omega}$. The linearization can be written as

$$\mathcal{L} = T + K, \tag{5.1}$$

where T is a transport operator, K is a compact operator, and they are defined by

$$T\omega = -\bar{\mathbf{u}} \cdot \nabla\omega, \quad K\omega = -\mathbf{u} \cdot \nabla\bar{\omega}.$$

We have to study the linear equation

$$(\partial_t - \mathcal{L})\omega^{\text{lin}} = 0 \tag{5.2}$$

and seek solutions

$$\omega_{\text{lin}}(t, \mathbf{x}) = e^{\lambda t} w(\mathbf{x}), \tag{5.3}$$

with $\text{Re}(\lambda) > 0$. This leads to the eigenvalue problem

$$\mathcal{L}w = \lambda w. \tag{5.4}$$

Since the operator $\mathcal{L}$ is real, the real-valued solutions to (5.2) will be recovered by replacing (5.3) with $\omega_{\text{lin}}(t, \mathbf{x}) = \text{Re}(e^{\lambda t} w(\mathbf{x}))$.

We restrict ourselves to radial profiles $\bar{\omega}(\mathbf{x}) = \bar{w}(r)$. Next, we choose the function space on which the eigenvalue problem will be solved. Given $0 > n \in \mathbb{N}$, we seek eigenfunctions in the subspace of L^2 formed by purely n-fold symmetric vorticities

$$U_n = \{w \in L^2 : w(\mathbf{x}) = w_n(r)e^{in\theta}\}. \tag{5.5}$$

Definition 5.1. We say that the vortex $\overline{\omega}$ is *unstable* if, for some $n \in \mathbb{N}$, there exists $0 \neq w \in U_n$ satisfying (5.4) with $\lambda \in \mathbb{C}_+$.

For $w \in U_n$, we arrive at the Rayleigh stability equation

$$\frac{\overline{v}_\theta}{r} w_n + \frac{\partial_r \overline{w}}{n} \int_0^\infty K_n\Big(\frac{r}{s}\Big) w_n(s)\, \mathrm{d}s = z w_n, \tag{5.6}$$

where $\overline{v}_\theta$ is the angular velocity, K_n are explicit kernels which came from (1.5), and we have replaced $\lambda = -\mathrm{i}nz$. Thus, we can write

$$w_n = h \partial_r \overline{w},$$

in terms of some profile h. For that h, equation (5.6) reads as

$$\frac{\overline{v}_\theta(r)}{r} h(r) + \frac{1}{n} \int_0^\infty K_n\Big(\frac{r}{s}\Big)(h \partial_r \overline{w})(s)\, \mathrm{d}s = z h(r), \quad r \in \operatorname{supp}(\partial_r \overline{w}). \tag{5.7}$$

We emphasize that equation (5.7) only needs to be solved in the support of $\partial_r \overline{w}$. This fact will play a very important role in our analysis. Indeed, this is the reason why we work with the Rayleigh equation in vorticity form.

Thus, Eulerian instability reduces to finding a vortex that admits nontrivial solutions (h, z) of (5.7) with $\Im z > 0$. We will show that this is indeed the case and therefore land in our first main result.

Theorem 5.2. *There exists an unstable vortex $\overline{\Omega} \in C_c^\infty(\mathbb{R}^2)$ with zero mean.*

Our proof of Theorem 5.2 is split into two steps: the construction of a piecewise constant unstable vortex and a regularization. This strategy was already proposed by E. Grenier [59] in the context of the stability of shear flows.

5.1.1. Piecewise constant unstable vortex. We now explain how to construct an unstable vortex of the form

$$\overline{w} = c 1_{[0,r_1)} - 1_{[r_1,r_2)},$$

in terms of some parameters $0 < r_1 < r_2 < \infty$ and $c > 0$, to be determined. Firstly, we choose c so that the mean of $\overline{\omega}$ equals zero

$$\int_{\mathbb{R}^2} \overline{\omega}(\mathbf{x})\, \mathrm{d}\mathbf{x} = 2\pi \int_0^\infty \overline{w}(r) r\, \mathrm{d}r = 0.$$

This condition guarantees that $\overline{v} \in L^2$. Moreover, $\overline{v}$ is Lipschitz and compactly supported with $\operatorname{supp}(\overline{v}) = \operatorname{supp}(\overline{\omega}) = B_{r_2}$. Secondly, since

$$\operatorname{supp}(\partial_r \overline{w}) = \{r_1, r_2\},$$

equation (5.7) turns into two conditions for the vector $h = (h(r_1), h(r_2))$ that can be written as a linear system

$$Ah = zh, \tag{5.8}$$

where the matrix $A \in \mathbb{R}^{2\times 2}$ depends on the parameters r_1, r_2, and the frequency n. Thus, there is an eigenvector $h \neq 0$ if and only if z is a root of the characteristic polynomial $\det(A - z)$. Finally, we can choose r_1, r_2 such that the roots z, z^* of $\det(A - z)$ satisfy $\Im z > 0$.

5.1.2. Regularization. We now explain how to prove the existence of a smooth vortex $\overline{\omega}^\varepsilon$ which is also unstable for some small $\varepsilon > 0$. Similarly to Step 5.1.1, now we need to solve the Rayleigh stability equation (5.7) in the intervals $B_\varepsilon(r_j)$ for $j = 1, 2$. We rescale variables around r_j writing $r = r_j + \varepsilon\alpha$ with $\alpha \in I = (-1, 1)$. Next, we make the asymptotic expansions for eigenfunctions and eigenvalues. Namely, we write

$$h^\varepsilon(r) = h(r_j) + \varepsilon g_j(\alpha), \quad z^\varepsilon = z + \varepsilon y,$$

for some profiles $g = (g_1, g_2) \in L^2(I)^2$ and a constant $y \in \mathbb{C}$, to be determined. We also expand

$$A^\varepsilon g = \frac{\overline{v}^\varepsilon_\theta}{r} g + \frac{1}{n}\int_0^\infty K_n\Big(\frac{r}{s}\Big)(g\partial_r \overline{w}^\varepsilon)(s)\,\mathrm{d}s = (A + \varepsilon B)g,$$

where B is another linear operator in $L^2(I)^2$. Therefore, the Rayleigh stability equation (5.7) can be rewritten in the rescaled variables as

$$(A^\varepsilon - z^\varepsilon)h^\varepsilon = (A - z)h + \varepsilon(A - z)g + \varepsilon(B - y)h + \varepsilon^2(B - y)g = 0.$$

Notice that the zeroth-order term vanishes by (5.8). Since $\varepsilon > 0$, we want to find (g, y) satisfying

$$(A - z)g = (y - B)h + \varepsilon(y - B)g. \tag{5.9}$$

The first step consists of "inverting" the operator $(A - z)$. Unfortunately, since z, z^* are the eigenvalues of A, the kernel and the image of $(A - z)$ are given by $\mathrm{Ker}(A - z) = \mathrm{span}(h)$ and $\mathrm{Im}(A - z) = \mathrm{span}(h^*)$. We bypass this obstacle by choosing y in such a way that the right-hand side of (5.9) becomes "parallel" to h^*. This allows us to write (5.9) in the form

$$(g, y) = H + \varepsilon R,$$

where H depends only on the (h, z) solving (5.8), while the remainder R also depends on (g, y). Thus, there is an explicit solution for $\varepsilon = 0$, and for small enough $\varepsilon > 0$, a fixed point argument in $L^2(I)^2 \times \mathbb{C}$ yields a solution as well. Once $\varepsilon > 0$ is fixed, Theorem 5.2 holds by taking

$$\overline{\Omega} = \overline{\omega}^\varepsilon.$$

From now on, the smooth vortex $\overline{\Omega}$ is fixed, and thus we will omit ε for the sake of simplicity. The use of capital letters will become clear in the following step.

5.2. Self-similar instability

Let us start by writing the Euler equation in self-similar coordinates

$$\tau = \frac{1}{ab}\log t, \quad \mathbf{X} = \frac{\mathbf{x}}{(abt)^{\frac{1}{a}}},$$

in terms of two parameters $0 < a, b \leq 1$ to be determined.

It is straightforward to check that (ω, f), given by the change of variables

$$\omega(t,\mathbf{x}) = \frac{1}{abt}\Omega(\tau,\mathbf{X}), \quad f(t,\mathbf{x}) = \frac{1}{(abt)^2}F(\tau,\mathbf{X}),$$

is a solution to the Euler equation if and only if (Ω, F) solves the *self-similar Euler equation*

$$\partial_\tau\Omega + \mathbf{V}\cdot\nabla\Omega - bS\Omega = F, \tag{5.10}$$

where $S = S_a$ is the linear operator defined by

$$S\Omega = (a + \mathbf{X}\cdot\nabla)\Omega. \tag{5.11}$$

Notice that (5.10) agrees with the 2D Euler equation but for the extra term $bS\Omega$, which vanishes (formally) as $b \to 0$.

We take the background solution as the time-dependent vortex

$$\omega_0(t,\mathbf{x}) = \frac{1}{abt}\overline{\Omega}(\mathbf{X}), \quad \mathbf{u}_0(t,\mathbf{x}) = (abt)^{\frac{1}{a}-1}\overline{\mathbf{V}}(\mathbf{X}), \tag{5.12}$$

where $\overline{\Omega}$ is the unstable vortex from Theorem 5.2. Given $2 < p < \infty$, we fix the parameter a in the regime

$$0 < ap < 2,$$

to guarantee that the integrability condition (1.11) is satisfied. Indeed, we have the scaling

$$\|\omega_0(t)\|_{L^p} = (abt)^{\frac{2}{ap}-1}\|\overline{\Omega}\|_{L^p}, \quad \|\mathbf{u}_0(t)\|_{L^2} = (abt)^{\frac{2}{a}-1}\|\overline{\mathbf{V}}\|_{L^2}. \tag{5.13}$$

The force is defined *ad hoc* in such a way that (5.12) becomes a solution. By the radial symmetry of the vortex (and defining the pressure properly), the quadratic term in the Euler equation vanishes, i.e., $\overline{\mathbf{V}}\cdot\nabla\overline{\Omega} = 0$. Thus, the force is defined by

$$\partial_t\omega_0 = f, \quad \partial_t\mathbf{u}_0 = \mathbf{g},$$

or, equivalently,

$$F(\mathbf{X}) = -bS\bar{\Omega}, \quad \mathbf{G}(\mathbf{X}) = b(a-S)\bar{\mathbf{V}}.$$

Observe that $(\bar{\Omega}, \bar{\mathbf{V}}, F, \mathbf{G})$ are supported on $\{|\mathbf{X}| \leq r_2 + \varepsilon\}$, and that the integrability condition (1.10) is also satisfied

$$\int_0^t \|f(s)\|_{L^p}\,\mathrm{d}s = \frac{(abt)^{\frac{2}{ap}-1}}{ab\left(\frac{2}{ap}-1\right)}\|F\|_{L^p}, \quad \int_0^t \|\mathbf{g}(s)\|_{L^2}\,\mathrm{d}s = \frac{(abt)^{\frac{2}{a}-1}}{ab\left(\frac{2}{a}-1\right)}\|G\|_{L^2}. \tag{5.14}$$

Analogously to Step 5.1, the deviation Ω of a close but different solution $\bar{\Omega} + \varepsilon\Omega$ to the self-similar Euler equation must satisfy

$$(\partial_\tau - \mathcal{L}_b)\Omega + \varepsilon(\mathbf{V}\cdot\nabla\Omega) = 0, \tag{5.15}$$

where $\mathcal{L}_b = \mathcal{L}_{b,\bar{\Omega}}$ is the linearization of the self-similar Euler equation (5.10) around $\bar{\Omega}$

$$\mathcal{L}_b = \mathcal{L} + bS \tag{5.16}$$

for S defined in (5.11). The unstable solutions in the limit $\varepsilon \to 0$ are governed by exponentially growing solutions to the linearized problem, which we denote by

$$\Omega^{\mathrm{lin}}(\tau, \mathbf{X}) = e^{\lambda\tau}W(\mathbf{X}). \tag{5.17}$$

Hence, we need to understand the corresponding eigenvalue problem

$$\mathcal{L}_b W = \lambda W, \tag{5.18}$$

this time for the linear operator $\mathcal{L}_b = \mathcal{L}_{b,\bar{\Omega}}$.

Definition 5.3. We say that the vortex $\bar{\Omega}$ is *self-similarly unstable* if, for some $n \in \mathbb{N}$, there exists $0 \neq W \in U_n$ satisfying (5.18) with $\lambda \in \mathbb{C}_+$.

Remarkably, M. Vishik showed that the spectrum $\sigma(\mathcal{L}_b) \cap \mathbb{C}_+$ converges (in a suitable sense) to $\sigma(\mathcal{L}) \cap \mathbb{C}_+$ as $b \to 0$. To this end, it is useful to decompose

$$\mathcal{L}_b = ab + T_b + K, \tag{5.19}$$

where

$$T_b\Omega = -(\bar{\mathbf{V}} - b\mathbf{X})\cdot\nabla\Omega, \quad K\Omega = -\mathbf{V}\cdot\nabla\bar{\Omega},$$

with $\mathbf{V} = \nabla^\perp\Delta^{-1}\Omega$. Notice that the first term only shifts the spectrum by ab. The transport operator T_b is invertible in $\mathbb{C}_+$. Since K is compact, the operator theory allows us to conclude that $\sigma(\mathcal{L}_b) \cap \mathbb{C}_+$ consists of isolated eigenvalues. Since the dependence on b is continuous, there is hope to find an eigenvalue $\lambda_b \in \sigma(\mathcal{L}_b) \cap \mathbb{C}_+$ close to $\lambda \in \sigma(\mathcal{L}) \cap \mathbb{C}_+$. Actually, the following result can be proven.

Theorem 5.4. *The vortex $\bar{\Omega}$ described in Theorem 5.2 is self-similarly unstable for some $b > 0$. Moreover, the corresponding eigenfunction satisfies $W \in C_c^2(\mathbb{R}^2)$.*

5.3. Nonlinear instability

The last step for proving Theorem 1.2 requires controlling the nonlinear effects. We decompose the deviation into

$$\Omega = \Omega_{\text{lin}} + \varepsilon \Omega_{\text{cor}},$$

where the correcting term Ω^{cor} must satisfy the equation

$$(\partial_\tau - \mathcal{L}_b)\Omega_{\text{cor}} + \underbrace{(\mathbf{V}_{\text{lin}} + \varepsilon \mathbf{V}_{\text{cor}}) \cdot \nabla(\Omega_{\text{lin}} + \varepsilon \Omega_{\text{cor}})}_{\mathcal{F}} = 0, \tag{5.20}$$

coupled with the initial condition

$$\Omega_{\text{cor}}|_{\tau=-\infty} = 0. \tag{5.21}$$

Recall that $\Omega_{\text{lin}} = \Re(e^{\lambda\tau} W)$, while $\mathbf{V}_{\text{lin}}$, $\mathbf{V}_{\text{cor}}$ are the velocities associated to Ω_{lin}, Ω_{cor}, respectively. If we interpret $\mathcal{F}$ as a forcing term, since the linear part decays as $e^{\Re\lambda\tau}$ and the contribution of the quadratic term is expected to be negligible for short times, one can naively expect to gain a slightly faster exponential decay by exploiting the Duhamel formula. We can prove the following result.

Theorem 5.5. *There exists a solution Ω_{cor} to (5.20) and (5.21) satisfying*

$$\|\Omega_{\text{cor}}(\tau)\|_{L^p} = o(e^{\text{Re}(\lambda)\tau}), \tag{5.22}$$

as $\tau \to -\infty$.

We conclude this section by showing how Theorems 5.4 and 5.5 allow proving Theorem 1.2. Coming back to the Eulerian coordinates, Theorems 5.4 and 5.5 imply that the vorticity

$$\omega_\varepsilon = \omega_0 + \varepsilon \omega_{\text{lin}} + \varepsilon^2 \omega_{\text{cor}}$$

satisfies that $\omega_\varepsilon|_{t=0} = 0$, and it is a solution to the Euler equation. We just need to show that for different ε's in we have different solutions. We recall that

$$\omega_0(t, \mathbf{x}) = \frac{1}{abt}\bar{\Omega}(\mathbf{X}), \quad \omega_{\text{lin}}(t, \mathbf{x}) = \frac{1}{abt}\Omega_{\text{lin}}(\tau, \mathbf{X}), \quad \omega_{\text{cor}}(t, \mathbf{x}) = \frac{1}{abt}\Omega_{\text{cor}}(\tau, \mathbf{X}),$$

where we have applied the convention that capital letters and τ stand for self-similar variables. Thus, by applying the scaling

$$\|\omega_{\text{lin}}(t)\|_{L^p} = (abt)^{\frac{2}{ap}-1}\|\Re(e^{\lambda\tau} W)\|_{L^p}, \quad \|\omega_{\text{cor}}(t)\|_{L^p} = (abt)^{\frac{2}{ap}-1} o(e^{\Re\lambda\tau}),$$

and the reverse triangle inequality, we obtain

$$(abt)^{1-\frac{2}{ap}} e^{-\Re\lambda\tau} \|(\omega_\varepsilon - \omega_{\overline{\varepsilon}})(t)\|_{L^p} \geq |\varepsilon - \overline{\varepsilon}|(\|\Re(e^{i\Im\lambda\tau} W)\|_{L^p} - o(1)).$$

If $\Im\lambda \neq 0$, the right-hand side is positive in a sequence of times $\tau_k = \tau_0 - \frac{2\pi k}{\Im z} \to -\infty$ as $k \to \infty$. If $\Im\lambda = 0$, we can assume from the beginning that W is real valued. Otherwise, it would suffice to take its imaginary part instead. This allows us to conclude that $\omega_\varepsilon \neq \omega_{\overline{\varepsilon}}$ whenever $\varepsilon \neq \overline{\varepsilon}$. Finally, it can be proven that the velocities $\mathbf{u}_\varepsilon = \nabla^\perp \Delta^{-1} \omega_\varepsilon$ belong to $L^\infty_t L^2$.

6. Extension to α-gSQG, with $0 < \alpha \leq 1$

An advantage of our construction of an unstable vortex for 2D Euler is that it can be adapted to construct unstable vortices for α-gSQG, $0 < \alpha \leq 1$. In fact, Step 5.1.1 works almost the same for the whole range $0 \leq \alpha \leq 1$. In this case, we have to consider the equation

$$\frac{1}{r}\int_0^\infty \left(I_{n,\alpha}\left(\frac{r}{s}\right)h(s) - I_{1,\alpha}\left(\frac{r}{s}\right)h(r)\right)\partial_r\overline{\theta}(s)s^{1-\alpha}\,\mathrm{d}s = zh(r), \quad r \in \operatorname{supp}(\partial_r\overline{\theta}), \tag{6.1}$$

instead of (5.7). The main difference in the analysis, when we take θ as a step function, is that the kernels $I_{n,\alpha}$ are not explicit unlike the kernel K_n in the case $\alpha = 0$. In addition, for the special case $\alpha = 1$, one has to take care of a cancellation to obtain a suitable analog of equation (5.8). In all the cases, $0 \leq \alpha \leq 1$, equation (5.8) can be analyzed to get unstable eigenvalues.

Step 5.1.2 also works in the case $0 \leq \alpha \leq 1$. Nevertheless, for $0 < \alpha \leq 1$, some extra effort is needed because of the lack of explicitness of the kernels. The main difference arises in the case $\alpha = 1$. Here, in addition to the complexity of the kernels, one has to deal with the loss of a factor $\log(\varepsilon)$ in the asymptotic expansion of equation (6.1) (with $\partial_r\overline{\theta}$ given by the regularized profile).

Step 5.2 can be adapted as well to the case $0 < \alpha < 1$. This is mainly because the operator K in (5.19) is still compact in this case. However, for Step 5.3, one needs to find new energy estimates to include the nonlinear interactions. In this issue, the degree of regularity of the velocity plays a dramatic role.

Both Steps 5.2 and 5.3 are much more involved in the case $\alpha = 1$. For example, the operator K in (5.19) is not compact anymore. In addition, the estimates we yield for $\alpha < 1$ to include the nonlinear part degenerate at $\alpha = 1$. In order to solve both problems, one needs to take into account some skew-symmetric properties of the velocity which are not needed for $\alpha < 1$.

At the end of the day, these ideas are able to produce non-uniqueness for forced α-gSQG, with $0 \leq \alpha \leq 1$, in the space H^s, $s < 1 + \alpha$. We remark again that well-posedness in H^s, with $s > 1 + \alpha$, $0 < \alpha < 1$, was proven in [23] and [31] for $\alpha = 1$. The precise statements of the theorems and proofs will be published in a forthcoming paper.

Acknowledgments. We thank D. Faraco, F. Mengual and M. Solera-Diana for their great help with the writing of this paper.

Funding. This work was partially supported by grant PID2020-114703GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by a 2023 Leonardo Grant for researchers and cultural creators, BBVA Foundation. The BBVA Foundation accepts no responsibility for the opinions, statements and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors. This work was also supported by the Severo Ochoa Programme for Centres of Excellence, grant CEX2019-000904-S, funded by MCIN/AEI/10.13039/501100011033.

The author also acknowledges support from grants RED2022-134784-T and RED2018-102650-T funded by MCIN/AEI/10.13039/501100011033.

References

[1] D. Albritton, E. Brué, and M. Colombo, Non-uniqueness of Leray solutions of the forced Navier–Stokes equations. *Ann. of Math. (2)* **196** (2022), no. 1, 415–455 Zbl 1497.35337 MR 4429263

[2] D. Albritton, E. Brué, M. Colombo, C. De Lellis, V. Giri, M. Janisch, and H. Kwon, *Instability and non-uniqueness for the 2D Euler equations, after M. Vishik*. Ann. of Math. Stud. 219, Princeton University Press, Princeton, NJ, 2024, 136 pp. Zbl 1560.35147 MR 4729618

[3] W. Ao, J. Dávila, M. del Pino, M. Musso, and J. Wei, Travelling and rotating solutions to the generalized inviscid surface quasi-geostrophic equation. *Trans. Amer. Math. Soc.* **374** (2021), no. 9, 6665–6689 Zbl 1477.35153 MR 4302173

[4] J. Bourgain and D. Li, Strong ill-posedness of the incompressible Euler equation in borderline Sobolev spaces. *Invent. Math.* **201** (2015), no. 1, 97–157 Zbl 1320.35266 MR 3359050

[5] J. Bourgain and D. Li, Strong illposedness of the incompressible Euler equation in integer C^m spaces. *Geom. Funct. Anal.* **25** (2015), no. 1, 1–86 Zbl 1480.35316 MR 3320889

[6] J. P. Bourguignon and H. Brezis, Remarks on the Euler equation. *J. Funct. Anal.* **15** (1974), no. 4, 341–363 Zbl 0279.58005 MR 0344713

[7] A. Bressan and R. Murray, On self-similar solutions to the incompressible Euler equations. *J. Differential Equations* **269** (2020), no. 6, 5142–5203 Zbl 1442.35313 MR 4104468

[8] A. Bressan and W. Shen, A posteriori error estimates for self-similar solutions to the Euler equations. *Discrete Contin. Dyn. Syst.* **41** (2021), no. 1, 113–130 Zbl 1460.35268 MR 4182316

[9] E. Bruè, M. Colombo, and A. Kumar, Flexibility of two-dimensional Euler flows with integrable vorticity. 2024, arXiv:2408.07934v1

[10] T. Buckmaster, C. De Lellis, P. Isett, and L. Székelyhidi, Jr., Anomalous dissipation for 1/5-Hölder Euler flows. *Ann. of Math. (2)* **182** (2015), no. 1, 127–172 Zbl 1330.35303 MR 3374958

[11] T. Buckmaster, C. De Lellis, and L. Székelyhidi, Jr., Dissipative Euler flows with Onsager-critical spatial regularity. *Comm. Pure Appl. Math.* **69** (2016), no. 9, 1613–1670 Zbl 1351.35109 MR 3530360

[12] T. Buckmaster, C. De Lellis, L. Székelyhidi, Jr., and V. Vicol, Onsager's conjecture for admissible weak solutions. *Comm. Pure Appl. Math.* **72** (2019), no. 2, 229–274 Zbl 1480.35317 MR 3896021

[13] T. Buckmaster, N. Masmoudi, M. Novack, and V. Vicol, *Intermittent convex integration for the 3D Euler equations*. Ann. of Math. Stud. 217, Princeton University Press, Princeton, NJ, 2023, 246 pp. Zbl 1535.35002 MR 4649134

[14] T. Buckmaster, S. Shkoller, and V. Vicol, Nonuniqueness of weak solutions to the SQG equation. *Comm. Pure Appl. Math.* **72** (2019), no. 9, 1809–1874 Zbl 1427.35200 MR 3987721

[15] T. Buckmaster and V. Vicol, Nonuniqueness of weak solutions to the Navier–Stokes equation. *Ann. of Math. (2)* **189** (2019), no. 1, 101–144 Zbl 1412.35215 MR 3898708

[16] A. Bulut, M. K. Huynh, and S. Palasek, Convex integration above the Onsager exponent for the forced Euler equations. 2023, arXiv:2301.00804v1

[17] A. Bulut, M. K. Huynh, and S. Palasek, Non-uniqueness up to the Onsager threshold for the forced SQG equation. 2023, arXiv:2310.12947v1

[18] A. Castro, D. Córdoba, and J. Gómez-Serrano, Global smooth solutions for the inviscid SQG equation. *Mem. Amer. Math. Soc.* **266** (2020), no. 1292, 89 pp. Zbl 1444.35003 MR 4126257

[19] Á. Castro, D. Faraco, F. Mengual, and M. Solera, A proof of Vishik's nonuniqueness theorem for the forced 2D Euler equation. *J. Reine Angew. Math.* **2025** (2025), no. 824, 253–288 Zbl 1569.35181 MR 4926947

[20] D. Chae, The quasi-geostrophic equation in the Triebel–Lizorkin spaces. *Nonlinearity* **16** (2003), no. 2, 479–495 Zbl 1029.35006 MR 1958612

[21] D. Chae, P. Constantin, D. Córdoba, F. Gancedo, and J. Wu, Generalized surface quasi-geostrophic equations with singular velocities. *Comm. Pure Appl. Math.* **65** (2012), no. 8, 1037–1066 Zbl 1244.35108 MR 2928091

[22] D. Chae, P. Constantin, and J. Wu, Inviscid models generalizing the two-dimensional Euler and the surface quasi-geostrophic equations. *Arch. Ration. Mech. Anal.* **202** (2011), no. 1, 35–62 Zbl 1266.76010 MR 2835862

[23] D. Chae and J. Wu, Logarithmically regularized inviscid models in borderline Sobolev spaces. *J. Math. Phys.* **53** (2012), no. 11, article no. 115601, 15 pp. Zbl 1329.76027 MR 3026546

[24] J. Chen, Remarks on the smoothness of the $C^{1,\alpha}$ asymptotically self-similar singularity in the 3D Euler and 2D Boussinesq equations. *Nonlinearity* **37** (2024), no. 6, article no. 065018, 32 pp. Zbl 1541.35368 MR 4752549

[25] J. Chen and T. Y. Hou, Finite time blowup of 2D Boussinesq and 3D Euler equations with $C^{1,\alpha}$ velocity and boundary. *Comm. Math. Phys.* **383** (2021), no. 3, 1559–1667 Zbl 1485.35071 MR 4244260

[26] J. Chen and T. Y. Hou, Stable nearly self-similar blowup of the 2D Boussinesq and 3D Euler equations with smooth data I: analysis. [v1] 2022, [v3] 2023, arXiv:2210.07191v3

[27] X. Cheng, H. Kwon, and D. Li, Non-uniqueness of steady-state weak solutions to the surface quasi-geostrophic equations. *Comm. Math. Phys.* **388** (2021), no. 3, 1281–1295 Zbl 1477.35272 MR 4340931

[28] A. Cheskidov, P. Constantin, S. Friedlander, and R. Shvydkoy, Energy conservation and Onsager's conjecture for the Euler equations. *Nonlinearity* **21** (2008), no. 6, 1233–1252 Zbl 1138.76020 MR 2422377

[29] A. Cheskidov and X. Luo, Sharp nonuniqueness for the Navier–Stokes equations. *Invent. Math.* **229** (2022), no. 3, 987–1054 Zbl 1504.35221 MR 4462623

[30] T. Cieślak, P. Kokocki, and W. S. Ożański, Well-posedness of logarithmic spiral vortex sheets. *J. Differential Equations* **389** (2024), 508–539 Zbl 1558.76028 MR 4708579

[31] P. Constantin, A. J. Majda, and E. Tabak, Formation of strong fronts in the 2-D quasigeostrophic thermal active scalar. *Nonlinearity* **7** (1994), no. 6, 1495–1533 Zbl 0809.35057 MR 1304437

[32] P. Constantin and H. Q. Nguyen, Local and global strong solutions for SQG in bounded domains. *Phys. D* **376–377** (2018), 195–203 Zbl 1398.35006 MR 3815216

[33] P. Constantin, E. Weinan and E. S. Titi, Onsager's conjecture on the energy conservation for solutions of Euler's equation. *Comm. Math. Phys.* **165** (1994), no. 1, 207–209 Zbl 0818.35085 MR 1298949

[34] D. Córdoba and L. Martínez-Zoroa, Non existence and strong ill-posedness in C^k and Sobolev spaces for SQG. *Adv. Math.* **407** (2022), article no. 108570, 74 pp. Zbl 1504.35566 MR 4452676

[35] D. Córdoba and L. Martínez-Zoroa, Blow-up for the incompressible 3D-Euler equations with uniform $C^{1,\frac{1}{2}-\varepsilon} \cap L^2$ force. 2023, arXiv:2309.08495v1

[36] D. Córdoba and L. Martínez-Zoroa, Non-existence and strong ill-posedness in $C^{k,\beta}$ for the generalized surface quasi-geostrophic equation. *Comm. Math. Phys.* **405** (2024), no. 7, article no. 170, 53 pp. Zbl 1543.35244 MR 4768540

[37] D. Córdoba, L. Martínez-Zoroa, and W. S. Ożański, Instantaneous gap loss of Sobolev regularity for the 2D incompressible Euler equations. *Duke Math. J.* **173** (2024), no. 10, 1931–1971 Zbl 1547.35518 MR 4776418

[38] D. Córdoba, L. Martínez-Zoroa, and W. S. Ożański, Instantaneous continuous loss of regularity for the SQG equation. *Adv. Math.* **481** (2025), article no. 110553, 49 pp. Zbl 08111016 MR 4963986

[39] D. Córdoba, L. Martinez-Zoroa, and F. Zheng, Finite time singularities to the 3D incompressible Euler equations for solutions in $C^\infty(\mathbb{R}^3 \setminus \{0\}) \cap C^{1,\alpha} \cap L^2$. *Ann. PDE* **11** (2025), no. 2, article no. 19, 56 pp. Zbl 1572.35171 MR 4929621

[40] M. Dai and S. Friedlander, Non-uniqueness of forced active scalar equations with even drift operators. *Phys. D* **467** (2024), article no. 134271, 13 pp. Zbl 1542.35309 MR 4769113

[41] M. Dai, V. Giri, and R.-O. Radu, An Onsager-type theorem for SQG. 2024, arXiv: 2407.02582v1

[42] C. De Lellis and L. Székelyhidi, Jr., The Euler equations as a differential inclusion. *Ann. of Math. (2)* **170** (2009), no. 3, 1417–1436 Zbl 1350.35146 MR 2600877

[43] C. De Lellis and L. Székelyhidi, Jr., On admissibility criteria for weak solutions of the Euler equations. *Arch. Ration. Mech. Anal.* **195** (2010), no. 1, 225–260 Zbl 1192.35138 MR 2564474

[44] C. De Lellis and L. Székelyhidi, Jr., The h-principle and the equations of fluid dynamics. *Bull. Amer. Math. Soc. (N.S.)* **49** (2012), no. 3, 347–375 Zbl 1254.35180 MR 2917063

[45] C. De Lellis and L. Székelyhidi, Jr., Dissipative continuous Euler flows. *Invent. Math.* **193** (2013), no. 2, 377–407 Zbl 1280.35103 MR 3090182

[46] C. De Lellis and L. Székelyhidi, Jr., Dissipative Euler flows and Onsager's conjecture. *J. Eur. Math. Soc. (JEMS)* **16** (2014), no. 7, 1467–1505 Zbl 1307.35205 MR 3254331

[47] L. De Rosa and J. Park, No anomalous dissipation in two-dimensional incompressible fluids. *SIAM J. Math. Anal.* **57** (2025), no. 5, 5771–5790 Zbl 08113981 MR 4969534

[48] R. J. Diperna and A. J. Majda, Concentrations in regularizations for 2-D incompressible flow. *Comm. Pure Appl. Math.* **40** (1987), no. 3, 301–345 Zbl 0850.76730 MR 0882068

[49] D. G. Ebin and J. Marsden, Groups of diffeomorphisms and the motion of an incompressible fluid. *Ann. of Math. (2)* **92** (1970), no. 1, 102–163 Zbl 0211.57401 MR 0271984

[50] T. M. Elgindi, Finite-time singularity formation for $C^{1,\alpha}$ solutions to the incompressible Euler equations on $\mathbb{R}^3$. *Ann. of Math. (2)* **194** (2021), no. 3, 647–727 Zbl 1492.35199 MR 4334974

[51] T. M. Elgindi and N. Masmoudi, L^∞ ill-posedness for a class of equations arising in hydrodynamics. *Arch. Ration. Mech. Anal.* **235** (2020), no. 3, 1979–2025 Zbl 1507.35153 MR 4065655

[52] V. Elling, Self-similar 2d Euler solutions with mixed-sign vorticity. *Comm. Math. Phys.* **348** (2016), no. 1, 27–68 Zbl 1351.35114 MR 3551260

[53] S. Friedlander and R. Shvydkoy, The unstable spectrum of the surface quasi-geostrophic equation. *J. Math. Fluid Mech.* **7** (2005), no. suppl. 1, S81–S93 Zbl 1064.76046 MR 2126131

[54] C. García and J. Gómez-Serrano, Self-similar spirals for the generalized surface quasi-geostrophic equations. *J. Eur. Math. Soc. (JEMS)* (2024), DOI 10.4171/JEMS/1549

[55] V. Giri, H. Kwon, and M. Novack, The L^3-based strong Onsager theorem. [v1] 2023, [v3] 2025, arXiv:2305.18509v3, to appear in *Ann. of Math. (2)*

[56] V. Giri and R.-O. Radu, The Onsager conjecture in 2D: a Newton–Nash iteration. *Invent. Math.* **238** (2024), no. 2, 691–768 Zbl 1556.35231 MR 4809443

[57] L. Godard-Cadillac, Smooth traveling-wave solutions to the inviscid surface quasi-geostrophic equations. *C. R. Math. Acad. Sci. Paris* **359** (2021), no. 1, 85–98 Zbl 1459.76021 MR 4229041

[58] P. Gravejat and D. Smets, Smooth travelling-wave solutions to the inviscid surface quasi-geostrophic equation. *Int. Math. Res. Not. IMRN* **2019** (2019), no. 6, 1744–1757 Zbl 1415.35079 MR 3932594

[59] E. Grenier, On the nonlinear instability of Euler and Prandtl equations. *Comm. Pure Appl. Math.* **53** (2000), no. 9, 1067–1091 Zbl 1048.35081 MR 1761409

[60] J. Guillod and V. Šverák, Numerical investigations of non-uniqueness for the Navier–Stokes initial value problem in borderline spaces. *J. Math. Fluid Mech.* **25** (2023), no. 3, article no. 46, 25 pp. Zbl 1522.35363 MR 4593274

[61] N. Gunther, On the motion of fluid in a moving container. *Izvestia Akad. Nauk USSR, Ser. Fiz.-Mat.* **20** (1927), 1323–1348

[62] S. He and A. Kiselev, Small-scale creation for solutions of the SQG equation. *Duke Math. J.* **170** (2021), no. 5, 1027–1041 Zbl 1473.35579 MR 4255049

[63] E. Hölder, Über die unbeschränkte Fortsetzbarkeit einer stetigen ebenen Bewegung in einer unbegrenzten inkompressiblen Flüssigkeit. *Math. Z.* **37** (1933), no. 1, 727–738 Zbl 0008.06902 MR 1545431

[64] T. Y. Hou and S. Zhang, Potential singularity of the axisymmetric Euler equations with C^α initial vorticity for a large range of α. *Multiscale Model. Simul.* **22** (2024), no. 4, 1326–1364 Zbl 1551.35344 MR 4805859

[65] P. Isett, A proof of Onsager's conjecture. *Ann. of Math. (2)* **188** (2018), no. 3, 871–963 Zbl 1416.35194 MR 3866888

[66] P. Isett and S.-Z. Looi, A proof of Onsager's conjecture for the SQG equation. [v1] 2024, [v3] 2025, arXiv:2407.02578v3

[67] P. Isett and A. Ma, A direct approach to nonuniqueness and failure of compactness for the SQG equation. *Nonlinearity* **34** (2021), no. 5, 3122–3162 Zbl 1473.35580 MR 4260790

[68] P. Isett and A. Ma, On the conservation laws and the structure of the nonlinearity for SQG and its generalizations. 2024, arXiv:2403.08279v1

[69] I.-J. Jeong and J. Kim, Strong ill-posedness for SQG in critical Sobolev spaces. *Anal. PDE* **17** (2024), no. 1, 133–170 Zbl 1533.35273 MR 4702316

[70] H. Jia and V. Šverák, Local-in-space estimates near initial time for weak solutions of the Navier–Stokes equations and forward self-similar solutions. *Invent. Math.* **196** (2014), no. 1, 233–265 Zbl 1301.35089 MR 3179576

[71] H. Jia and V. Sverak, Are the incompressible 3d Navier–Stokes equations locally ill-posed in the natural energy space? *J. Funct. Anal.* **268** (2015), no. 12, 3734–3766 Zbl 1317.35176 MR 3341963

[72] T. Kato, Nonstationary flows of viscous and ideal fluids in R^3. *J. Funct. Anal.* **9** (1972), no. 3, 296–305 Zbl 0229.76018 MR 0481652

[73] T. Kato and G. Ponce, Commutator estimates and the Euler and Navier–Stokes equations. *Comm. Pure Appl. Math.* **41** (1988), no. 7, 891–907 Zbl 0671.35066 MR 0951744

[74] J. Kim and I.-J. Jeong, A simple ill-posedness proof for incompressible Euler equations in critical Sobolev spaces. *J. Funct. Anal.* **283** (2022), no. 10, article no. 109673, 34 pp. Zbl 1496.76023 MR 4470735

[75] A. Kiselev and F. Nazarov, A simple energy pump for the surface quasi-geostrophic equation. In H. Holden and K. Karlsen (eds.), *Nonlinear partial differential equations. The Abel Symposium 2010*, pp. 175–179, Abel Symp. 7, Springer, Berlin, 2012 Zbl 1250.35034 MR 3289364

[76] A. Kiselev and V. Šverák, Small scale creation for solutions of the incompressible two-dimensional Euler equation. *Ann. of Math. (2)* **180** (2014), no. 3, 1205–1220 Zbl 1304.35521 MR 3245016

[77] L. Lichtenstein, Über einige Existenzprobleme der Hydrodynamik homogener, unzusammendrückbarer, reibungsloser Flüssigkeiten und die Helmholtzschen Wirbelsätze. *Math. Z.* **23** (1925), no. 1, 89–154 JFM 51.0658.01 MR 1544733

[78] F. Marchand, Existence and regularity of weak solutions to the quasi-geostrophic equations in the spaces L^p or $\dot{H}^{-1/2}$. *Comm. Math. Phys.* **277** (2008), no. 1, 45–67 Zbl 1134.35025 MR 2357424

[79] F. Mengual, Non-uniqueness of admissible solutions for the 2D Euler equation with L^p vortex data. *Comm. Math. Phys.* **405** (2024), no. 9, article no. 207, 28 pp. Zbl 1546.35146 MR 4788287

[80] M. Novack and V. Vicol, An intermittent Onsager theorem. *Invent. Math.* **233** (2023), no. 1, 223–323 Zbl 1530.35200 MR 4601999

[81] J. Pedlosky, *Geophysical fluid dynamics*. Springer Study Ed., Springer, New York, 1979, 626 pp. Zbl 0429.76001

[82] S. G. Resnick, *Dynamical problems in non-linear advective partial differential equations*. Ph.D. thesis, The University of Chicago, 1995 MR 2716577

[83] V. Scheffer, An inviscid flow with compact support in space-time. *J. Geom. Anal.* **3** (1993), no. 4, 343–401 Zbl 0836.76017 MR 1231007

[84] F. Shao, D. Wei, and Z. Zhang, Self-similar algebraic spiral solution of 2-D incompressible Euler equations. *Ann. PDE* **11** (2025), no. 1, article no. 13, 91 pp. Zbl 08063100 MR 4902133

[85] A. Shnirelman, On the nonuniqueness of weak solution of the Euler equation. *Comm. Pure Appl. Math.* **50** (1997), no. 12, 1261–1286 Zbl 0909.35109 MR 1476315

[86] M. Vishik, Instability and non-uniqueness in the Cauchy problem for the Euler equations of an ideal incompressible fluid. Part I. 2018, arXiv:1805.09426v1

[87] M. Vishik, Instability and non-uniqueness in the Cauchy problem for the Euler equations of an ideal incompressible fluid. Part II. 2018, arXiv:1805.09440v1

[88] Y. Wang, C.-Y. Lai, J. Gómez-Serrano, and T. Buckmaster, Asymptotic self-similar blow-up profile for three-dimensional axisymmetric Euler equations using neural networks. *Phys. Rev. Lett.* **130** (2023), no. 24, article no. 244002, 6 pp. MR 4608987

[89] W. Wolibner, Un theorème sur l'existence du mouvement plan d'un fluide parfait, homogène, incompressible, pendant un temps infiniment long. *Math. Z.* **37** (1933), no. 1, 698–726 Zbl 0008.06901 MR 1545430

[90] J. Wu, Solutions of the 2D quasi-geostrophic equation in Hölder spaces. *Nonlinear Anal.* **62** (2005), no. 4, 579–594 Zbl 1116.35348 MR 2149903

[91] V. I. Yudovich, Non-stationary flows of an ideal incompressible fluid (in Russian). *Zh. Vychisl. Mat. Mat. Fiz.* **3** (1963), no. 6, 1032–1066. Zbl 0129.19402 English translation. *USSR Comput. Math. Math. Phys.* **3** (1963), no. 6, 1407–1456 Zbl 0147.44303 MR 0158189

Ángel Castro
Instituto de Ciencias Matemáticas, Nicolás Cabrera 13-15, 28049 Madrid, Spain;
angel_castro@icmat.es

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/16

Graph product structure: Theory and applications

Vida Dujmović and David R. Wood

Abstract. Graph product structure theory describes graphs in complicated graph classes as subgraphs of products of graphs in simpler graph classes. This essay selectively surveys recent results and applications in this field.

1. Introduction

One of the aims of the field of graph structure theory is to describe the global structure of graphs in a particular class $\mathcal{G}$ in terms of simpler graphs. Graph product structure theory achieves this goal by describing graphs in $\mathcal{G}$ as subgraphs of products of graphs in simpler graph classes, typically with bounded treewidth. The *treewidth* of a graph G, denoted by $\mathrm{tw}(G)$, is the standard measure of how similar G is to a tree. Graph classes with bounded treewidth are considered to be simple; for example, numerous NP-complete problems are efficiently solvable on graphs with bounded treewidth. See [5, 39, 52] for background on treewidth, including the definition.

The following planar graph product structure theorem by Dujmović, Joret, Micek, Morin, Ueckerdt, and Wood [29] is the classical example of a graph product structure theorem. Here, a graph H is *contained* in a graph G if H is isomorphic to a subgraph of G, written $H \subsetneq G$. As illustrated in Figure 1, the *strong product* of graphs A and B, denoted by $A \boxtimes B$, is the graph with vertex set $V(A) \times V(B)$, where distinct vertices $(v, x), (w, y)$ are adjacent if

- $v = w$ and $xy \in E(B)$, or
- $x = y$ and $vw \in E(A)$, or
- $vw \in E(A)$ and $xy \in E(B)$.

Theorem 1.1 ([29]). *For every planar graph G, there is a graph H and a path P such that* $\mathrm{tw}(H) \leq 8$ *and* $G \subsetneq H \boxtimes P$.

Mathematics Subject Classification 2020: 05C10 (primary); 05C83 (secondary).
Keywords: graph, graph minor, graph product, planar graph, treewidth.

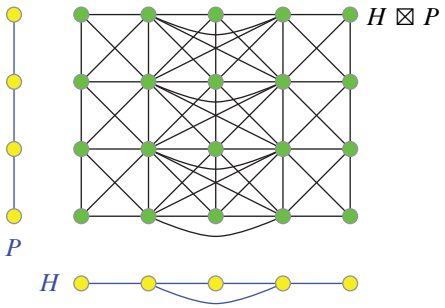


Figure 1. Example of a strong product.

Theorem 1.1 implies and generalises many results from the literature on the global structure of planar graphs, including the Lipton–Tarjan balanced separator theorem [49], the Baker decomposition and bounded local treewidth [2, 12, 37], and layered treewidth [32].

Theorem 1.1 has been the key tool for resolving several open problems regarding queue layouts [29], nonrepetitive colourings [24], centred colourings [11], adjacency labelling schemes [6, 23, 38], twin-width [7, 47], infinite graphs [45], clustered colouring [25, 26], and comparable box dimension [35]. Now we give one example.

Motivated by measuring the relative power of stacks and queues as data structures, Heath, Leighton, and Rosenberg [40, 41] defined the *stack number* $\mathrm{sn}(G)$ and *queue number* $\mathrm{qn}(G)$ of a graph G. They asked whether planar graphs have bounded queue number, which became the most important open problem in the field and a key case to consider in the comparison of the relative power of stacks and queues.

Prior to [29] the best known upper bound on the queue number of n-vertex planar graphs was $O(\log n)$ [14, 22]. With Theorem 1.1 in hand, Dujmović et al. [29] showed that an $O(1)$ bound is easily proved using already established tools. In particular, Dujmović, Morin, and Wood [34] showed in 2005 that $\mathrm{qn}(G) \leq f(\mathrm{tw}(G))$ for some function f. The best known bound is $\mathrm{qn}(G) \leq 2^{\mathrm{tw}(G)} - 1$ due to Wiechert [56]. Wood [57] showed in 2005 that $\mathrm{qn}(H \boxtimes P) \leq 3\,\mathrm{qn}(H) + 1$ for any graph H and any path P. Now consider any planar graph G. By Theorem 1.1, $G \subsetneq H \boxtimes P$, where $\mathrm{tw}(H) \leq 8$ and P is a path. Since queue number is subgraph monotone, the above results together imply that

$$\mathrm{qn}(G) \leq \mathrm{qn}(H \boxtimes P) \leq 3\,\mathrm{qn}(H) + 1 \leq 3(2^{\mathrm{tw}(H)} - 1) + 1 \leq 3(2^8 - 1) + 1 = 766.$$

This shows that planar graphs have bounded queue number, resolving the 28-year-old open problem of Heath, Leighton, and Rosenberg [40, 41].

The key tool in this proof is the planar graph product structure theorem. We emphasise that the authors of [29] were not trying to prove Theorem 1.1. They first solved the queue-number problem and then realised that their proof contained

a structural description of planar graphs that could be written as a product, as in Theorem 1.1. It turns out that, back in 2008, Wood [58] had conjectured a version of Theorem 1.1, where the subgraph relation was replaced by the weaker notion of shallow minor. Theorem 1.1 implies this conjecture.

Theorem 1.1 has been extended in various ways. The following result is the state of the art for the product structure of planar graphs.

Theorem 1.2. *Every planar graph G is contained in:*

(a) $H \boxtimes P$, *where* $\mathrm{tw}(H) \leq 6$ *and* P *is a path* [55],

(b) $H \boxtimes P \boxtimes K_2$, *where* $\mathrm{tw}(H) \leq 4$ *and* P *is a path* [55],

(c) $H \boxtimes P \boxtimes K_3$, *where* $\mathrm{tw}(H) \leq 3$ *and* P *is a path* [29].

Theorem 1.2 (c) is often the most useful version of the planar graph product structure theorem since the dependence on $\mathrm{tw}(H)$ in applications is often exponential, as in the queue-number example above. Indeed, using Theorem 1.2 (c), Dujmović et al. [29] showed that planar graphs have queue number at most 49, which was improved to 42 by Bekos, Gronemann, and Raftopoulou [3]. The treewidth 3 bound in Theorem 1.2 (c) is the best possible for any result showing that planar graphs are contained in $H \boxtimes P \boxtimes K_c$ for any constant c (see [29]). Improved bounds are known for special classes of planar graphs, such as squaregraphs [42]. Note that some extra properties hold in Theorems 1.1 and 1.2. Notably, H is a minor of the given graph G, and thus H is planar. On the other hand, Dujmović, Joret, Micek, Morin, and Wood [30] proved that in Theorems 1.1 and 1.2, H must have unbounded maximum degree, even for planar graphs G with maximum degree 5.

The proofs of Theorems 1.1 and 1.2 were inspired by an elegant result of Pilipczuk and Siebertz [51] which uses Sperner's lemma at its heart. The method is constructive and leads to an efficient algorithm for computing the mapping into the product graph. Indeed, Bose, Morin, and Odak [9] gave a linear-time algorithm for doing so. Note that Illingworth, Scott, and Wood [46] gave a purely combinatorial proof of Theorem 1.2 (c) that avoids planar embeddings and avoids the use of Sperner's lemma. Roughly speaking, their result describes how any tree decomposition of a graph excluding a fixed minor can be converted into a product-like structure.

Motivated by Theorem 1.1, Bose, Dujmović, Javarsineh, Morin, and Wood [8] introduced the following definition. The *row treewidth* of a graph G, denoted by $\mathrm{rtw}(G)$, is the minimum $c \in \mathbb{N}$ such that $G \subsetneq H \boxtimes P$ for some path P and graph H with $\mathrm{tw}(H) \leq c$. Theorem 1.2 (a) says that planar graphs have row treewidth at most 6. It is open whether planar graphs have row treewidth at most 3. Bose et al. [8] studied the relationship between row treewidth and layered treewidth $\mathrm{ltw}(G)$. While it is easily seen that $\mathrm{ltw}(G) \leq \mathrm{rtw}(G)$, Bose et al. [8] described a class of graphs G

with $\operatorname{ltw}(G) = 1$ and unbounded row treewidth. Thus, bounded row treewidth is a strictly stronger setting than bounded layered treewidth.

Other classes with bounded row treewidth include graphs embeddable on any fixed surface [29]. The best bounds are due to Distel, Hickingbotham, Huynh, and Wood [18].

Theorem 1.3 ([18]). *Every graph G of Euler genus g is contained in:*

(a) $H \boxtimes P$, *where* $\operatorname{tw}(H) \leq 2g + 6$ *and* P *is a path,*

(b) $H \boxtimes P \boxtimes K_{\max\{2g,3\}}$, *where* $\operatorname{tw}(H) \leq 3$ *and* P *is a path.*

Dujmović et al. [29] established the following characterisation of minor-closed classes with bounded row treewidth. A graph X is *apex* if $X - a$ is planar for some vertex $a \in V(X)$.

Theorem 1.4 ([29]). *A minor-closed class $\mathcal{G}$ has bounded row treewidth if and only if some apex graph is not in $\mathcal{G}$.*

The product structure of graphs excluding an apex minor has been further studied in [27, 46]. For an apex graph X, let k_X be the minimum integer such that, for some integer c, every X-minor-free graph is contained in $H \boxtimes P \boxtimes K_c$, where H is a graph with $\operatorname{tw}(H) \leq k_X$ and P is a path. Illingworth et al. [46] showed that k_X is at most the vertex cover number of X. Tightening this result, Dujmović et al. [27] showed that $\operatorname{td}(X) - 2 \leq k_X \leq 2^{\operatorname{tw}(X)-1}$, where $\operatorname{td}(X)$ is the tree depth of X. This shows that the tree depth of X determines the minimum treewidth in such a product structure theorem.

While graphs excluding an apex minor are the largest class with bounded row treewidth, graph products can be used to describe the structure of graphs excluding any fixed minor. Dujmović et al. [25] showed that graphs excluding a fixed minor and with bounded maximum degree have bounded row treewidth. Dujmović et al. [29] showed the following 'graph minor product structure theorem', where $+K_a$ means the complete join with a complete graph on a vertices.

Theorem 1.5 ([29]). *For any graph X, there exist integers k, a such that every X-minor-free graph can be obtained from clique sums of subgraphs of graphs of the form $(H \boxtimes P) + K_a$, where* $\operatorname{tw}(H) \leq k$.

This result has enabled many of the above applications, including bounded queue number, to be extended to the setting of any proper minor-closed class. As another application, Dujmović et al. [29] showed that Theorem 1.5 implies the following classical theorem of DeVos, Ding, Oporowski, Sanders, Reed, Seymour, and Vertigan [13]: for any graph X and integer $k \geq 2$, every X-minor-free graph has an

(improper) vertex k-colouring such that the union of any $k-1$ colour classes has bounded treewidth. Another application relates to Hadwiger's conjecture. Using Theorem 1.5 as a starting point, Dujmović et al. [26] showed that graphs excluding a fixed K_t-minor are $(t-1)$-colourable with monochromatic components of bounded size. This resolves the so-called clustered Hadwiger conjecture.

Graph product structure theorems have also been established for various non-minor-closed graph classes [4, 19, 33, 43]. Here, Theorems 1.1 and 1.2 are used as building blocks for establishing product structure theorems for more general classes, as illustrated by the following example. A graph is (g,k)-*planar* if it has a drawing in a surface of Euler genus g such that each edge is in at most k crossings (and no three edges cross at the same point).

Theorem 1.6. *Every (g,k)-planar graph G is contained in:*

(a) $H \boxtimes P \boxtimes K_c$, *where* $\mathrm{tw}(H) \leq O(k^3)$, *$P$ is a path, and* $c \in O(gk^2)$ [33],

(b) $H \boxtimes P \boxtimes K_c$, *where* $\mathrm{tw}(H) \leq 10^{10}$, *$P$ is a path, and* $c = c(g,k)$ [19].

We now briefly summarise some of the recent directions in which graph product structure theory is moving.

- Huynh et al. [45] extended all of the above results to the setting of countably infinite graphs. This is interesting, since in the infinite setting, the graph $H \boxtimes P$ is universal, in the sense that it contains every countable planar graph.
- Several authors have established results which demonstrate that graphs of given treewidth are contained in products of simpler graphs [10, 15, 16, 20, 21, 50]. Indeed, this direction predates Theorem 1.1 through the classical work of Ding and Oporowski [15, 16]. The characterisation of graphs with polynomial growth due to Krauthgamer and Lee [48] is another early example of a graph product structure theorem.
- Hliněný and Jedelský [44] introduced graph product structure theory for dense graph classes and the induced subgraph relation.
- A number of classical theorems about graph minors have recently been revisited in light of graph product structure theory. Dujmović, Hickingbotham, Joret, Micek, Morin, and Wood [28] showed that for every tree T of radius h, there is an integer c such that every T-minor-free graph is contained in $H \boxtimes K_c$ for some graph H with pathwidth at most $2h-1$. This is a qualitative strengthening of the excluded tree minor theorem of Robertson and Seymour [53]. They also showed that radius is the right parameter to consider in this setting, and $2h-1$ is the best possible bound. Dujmović et al. [27] proved analogous strengthenings of the grid minor theorem of Robertson and Seymour [54]. Here the right parameter of the excluded minor is tree depth.

- Several recent results show that for various graph classes $\mathcal{G}$, every n-vertex graph in $\mathcal{G}$ is contained in $H \boxtimes K_m$, where H has bounded treewidth and $m \in O(\sqrt{n})$. This direction was initiated by Illingworth et al. [46]. For planar graphs, Distel, Dujmović, Eppstein, Hickingbotham, Joret, Micek, Morin, Seweryn, and Wood [17] showed this result with $\mathrm{tw}(H) \leq 2$ which is best possible. For K_t-minor-free graphs, they showed this result with $\mathrm{tw}(H) \leq 4$. These two results are qualitative strengthenings of classical balanced separator results of Lipton and Tarjan [49] and Alon, Seymour, and Thomas [1]. Dvořák and Wood [36] studied analogous questions for any graph class admitting strongly sublinear separators. Recently, Dujmović, Joret, Micek, Morin, and Wood [31] showed that every n-vertex planar graph (and indeed any graph with bounded row treewidth) is contained in $H \boxtimes K_m$, where H is a fan graph (which has pathwidth 2) and $m \in O(\sqrt{n}\log^2 n)$.

We finish with two open problems. The first asks how far these product structure theorems can be generalised. For example, is there a product structure theorem (using other ingredients in addition to products) for any monotone graph class admitting strongly sublinear separators? For example, the Koebe disc packing theorem says every planar graph can be represented as the touching graph of discs in $\mathbb{R}^2$. What can be said about the product structure of touching graphs of spheres in $\mathbb{R}^3$? Our second open problem concerns applications. Does every monotone graph class admitting strongly sublinear separators have bounded queue number? We expect that a successful answer to this question will unlock the right structure theorem to answer the first question. The interplay between theory and applications is a feature of this field of research.

Acknowledgements. We thank all our collaborators on this topic: Édouard Bonnet, Prosenjit Bose, Rutger Campbell, Katie Clinch, Marc Distel, David Eppstein, Louis Esperet, Cyril Gavoille, Pascal Gollin, Kevin Hendrey, Jędrzej Hodor, Robert Hickingbotham, Tony Huynh, Freddie Illingworth, Mehrnoosh Javarsineh, Gwenaël Joret, Paul Jungeblut, O-joung Kwon, Hoang La, Laura Merker, Bojan Mohar, Pat Morin, Piotr Micek, Clément Rambaud, Robert Šámal, Alex Scott, Michał Seweryn, Youri Tamitegama, Jane Tan, Carsten Thomassen, Torsten Ueckerdt, Bartosz Walczak, Wendy Yi.

Funding. Research by Dujmović is supported by NSERC and a University of Ottawa Research Chair. Research by Wood is supported by the Australian Research Council and by NSERC.

References

[1] N. Alon, P. Seymour, and R. Thomas, A separator theorem for nonplanar graphs. *J. Amer. Math. Soc.* **3** (1990), no. 4, 801–808 Zbl 0747.05051 MR 1065053

[2] B. S. Baker, Approximation algorithms for NP-complete problems on planar graphs. *J. ACM* **41** (1994), no. 1, 153–180 Zbl 0807.68067 MR 1369197

[3] M. Bekos, M. Gronemann, and C. N. Raftopoulou, An improved upper bound on the queue number of planar graphs. *Algorithmica* **85** (2023), no. 2, 544–562 Zbl 1515.68223 MR 4541066

[4] M. A. Bekos, G. Da Lozzo, P. Hliněný, and M. Kaufmann, Graph product structure for h-framed graphs. In *Proceedings of the 33rd international symposium on algorithms and computation (ISAAC 2022)*, LIPIcs - Leibniz Int. Proc. Inform. 248, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2022, article no. 23, 15 pp. Zbl 1570.05113 MR 4526676

[5] H. L. Bodlaender, A partial k-arboretum of graphs with bounded treewidth. *Theoret. Comput. Sci.* **209** (1998), no. 1–2, 1–45 Zbl 0912.68148 MR 1647486

[6] M. Bonamy, C. Gavoille, and M. Pilipczuk, Shorter labeling schemes for planar graphs. *SIAM J. Discrete Math.* **36** (2022), no. 3, 2082–2099 Zbl 1548.68148 MR 4474794

[7] É. Bonnet, O.-j. Kwon, and D. R. Wood, Reduced bandwidth: a qualitative strengthening of twin-width in minor-closed classes (and beyond). *J. Combin. Theory Ser. B* **178** (2026), 27–66 MR 4998097

[8] P. Bose, V. Dujmović, M. Javarsineh, P. Morin, and D. R. Wood, Separating layered treewidth and row treewidth. *Discrete Math. Theor. Comput. Sci.* **24** (2022), no. 1, article no. 18, 10 pp. Zbl 1520.05028 MR 4425386

[9] P. Bose, P. Morin, and S. Odak, An optimal algorithm for product structure in planar graphs. In *Proceedings of the 18th Scandinavian symposium and workshops on algorithm theory (SWAT 2022)*, LIPIcs - Leibniz Int. Proc. Inform. 227, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2022, article no. 19, 14 pp. Zbl 07853691 MR 4473458

[10] R. Campbell, K. Clinch, M. Distel, J. P. Gollin, K. Hendrey, R. Hickingbotham, T. Huynh, F. Illingworth, Y. Tamitegama, J. Tan, and D. R. Wood, Product structure of graph classes with bounded treewidth. *Combin. Probab. Comput.* **33** (2024), no. 3, 351–376 Zbl 1551.05333 MR 4730906

[11] M. Dębski, S. Felsner, P. Micek, and F. Schröder, Improved bounds for centered colorings. *Adv. Comb.* (2021), article no. 8, 28 pp. Zbl 1478.05050 MR 4309118

[12] E. D. Demaine and M. T. Hajiaghayi, Diameter and treewidth in minor-closed graph families, revisited. *Algorithmica* **40** (2004), no. 3, 211–215 Zbl 1082.05086 MR 2080518

[13] M. DeVos, G. Ding, B. Oporowski, D. P. Sanders, B. Reed, P. Seymour, and D. Vertigan, Excluding any graph as a minor allows a low tree-width 2-coloring. *J. Combin. Theory Ser. B* **91** (2004), no. 1, 25–41 Zbl 1042.05036 MR 2047529

[14] G. Di Battista, F. Frati, and J. Pach, On the queue number of planar graphs. *SIAM J. Comput.* **42** (2013), no. 6, 2243–2285 Zbl 1300.05075 MR 3141759

[15] G. Ding and B. Oporowski, Some results on tree decomposition of graphs. *J. Graph Theory* **20** (1995), no. 4, 481–499 Zbl 0837.05044 MR 1358539

[16] G. Ding and B. Oporowski, On tree-partitions of graphs. *Discrete Math.* **149** (1996), no. 1–3, 45–58 Zbl 0844.05078 MR 1375097

[17] M. Distel, V. Dujmović, D. Eppstein, R. Hickingbotham, G. Joret, P. Micek, P. Morin, M. T. Seweryn, and D. R. Wood, Product structure extension of the Alon–Seymour–Thomas theorem. *SIAM J. Discrete Math.* **38** (2024), no. 3, 2095–2107 Zbl 1543.05172 MR 4770382

[18] M. Distel, R. Hickingbotham, T. Huynh, and D. R. Wood, Improved product structure for graphs on surfaces. *Discrete Math. Theor. Comput. Sci.* **24** (2022), no. 2, article no. 6, 10 pp. Zbl 1516.05145 MR 4504777

[19] M. Distel, R. Hickingbotham, M. Seweryn, and D. R. Wood, Powers of planar graphs, product structure, and blocking partitions. *Innov. Graph Theory* **1** (2024), 39–86 Zbl 1556.05032 MR 4882662

[20] M. Distel and D. R. Wood, Tree-partitions with small bounded degree trees. [v1] 2022, [v2] 2023, arXiv:2210.12577v2

[21] M. Distel and D. R. Wood, Tree-partitions with bounded degree trees. In D. R. Wood, J. de Gier, and C. E. Praeger (Eds.), *2021–2022 MATRIX annals*, pp. 203–212, MATRIX Book Ser. 5, Springer, Cham, 2024 Zbl 1548.05085 MR 4807276

[22] V. Dujmović, Graph layouts via layered separators. *J. Combin. Theory Ser. B* **110** (2015), 79–89 Zbl 1303.05036 MR 3279388

[23] V. Dujmović, L. Esperet, C. Gavoille, G. Joret, P. Micek, and P. Morin, Adjacency labelling for planar graphs (and beyond). *J. ACM* **68** (2021), no. 6, article no. 42, 33 pp. Zbl 1499.05165 MR 4402353

[24] V. Dujmović, L. Esperet, G. Joret, B. Walczak, and D. R. Wood, Planar graphs have bounded nonrepetitive chromatic number. *Adv. Comb.* (2020), article no. 5, 11 pp. Zbl 1450.05012 MR 4125346

[25] V. Dujmović, L. Esperet, P. Morin, B. Walczak, and D. R. Wood, Clustered 3-colouring graphs of bounded degree. *Combin. Probab. Comput.* **31** (2022), no. 1, 123–135 Zbl 1510.05071 MR 4356460

[26] V. Dujmović, L. Esperet, P. Morin, and D. R. Wood, Proof of the clustered Hadwiger conjecture. In *Proceedings of the 2023 IEEE 64th annual symposium on foundations of computer science (FOCS 2023)*, pp. 1921–1930, IEEE Computer Society, 2023 Zbl 08080132 MR 4720358

[27] V. Dujmović, R. Hickingbotham, J. Hodor, G. Joret, H. La, P. Micek, P. Morin, C. Rambaud, and D. R. Wood, The grid-minor theorem revisited. In *Proceedings of the 2024 annual ACM-SIAM symposium on discrete algorithms (SODA 2024)*, pp. 1241–1245, SIAM, 2024 Zbl 07951121 MR 4699293

[28] V. Dujmović, R. Hickingbotham, G. Joret, P. Micek, P. Morin, and D. R. Wood, The excluded tree minor theorem revisited. *Combin. Probab. Comput.* **33** (2024), no. 1, 85–90 Zbl 1551.05381 MR 4680491

[29] V. Dujmović, G. Joret, P. Micek, P. Morin, T. Ueckerdt, and D. R. Wood, Planar graphs have bounded queue-number. *J. ACM* **67** (2020), no. 4, article no. 22, 38 pp. Zbl 1466.05047 MR 4148600

[30] V. Dujmović, G. Joret, P. Micek, P. Morin, and D. R. Wood, Bounded-degree planar graphs do not have bounded-degree product structure. *Electron. J. Combin.* **31** (2024), no. 2, article no. P2.51, 15 pp. Zbl 1548.05281 MR 4758205

[31] V. Dujmović, G. Joret, P. Micek, P. Morin, and D. R. Wood, Planar graphs in blowups of fans. In *Proceedings of the 2025 annual ACM-SIAM symposium on discrete algorithms (SODA 2025)*, pp. 3382–3391, SIAM, 2025 MR 4863525

[32] V. Dujmović, P. Morin, and D. R. Wood, Layered separators in minor-closed graph classes with applications. *J. Combin. Theory Ser. B* **127** (2017), 111–147 Zbl 1371.05282 MR 3704658

[33] V. Dujmović, P. Morin, and D. R. Wood, Graph product structure for non-minor-closed classes. *J. Combin. Theory Ser. B* **162** (2023), 34–67 Zbl 1519.05211 MR 4586042

[34] V. Dujmovic, P. Morin, and D. R. Wood, Layout of graphs with bounded tree-width. *SIAM J. Comput.* **34** (2005), no. 3, 553–579 Zbl 1069.05055 MR 2137079

[35] Z. Dvořák, D. Gonçalves, A. Lahiri, J. Tan, and T. Ueckerdt, On comparable box dimension. In *Proceedings of the 38th international symposium on computational geometry (SoCG 2022)*, LIPIcs - Leibniz Int. Proc. Inform. 224, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2022, article no. 38, 14 pp. Zbl 1558.68235 MR 4470917

[36] Z. Dvořák and D. R. Wood, Product structure of graph classes with strongly sublinear separators. *Innov. Graph Theory* **2** (2025), 191–222 Zbl 08099014 MR 4960622

[37] D. Eppstein, Diameter and treewidth in minor-closed graph families. *Algorithmica* **27** (2000), no. 3–4, 275–291 Zbl 0963.05128 MR 1759751

[38] L. Esperet, G. Joret, and P. Morin, Sparse universal graphs for planarity. *J. Lond. Math. Soc. (2)* **108** (2023), no. 4, 1333–1357 Zbl 1527.05041 MR 4655266

[39] D. J. Harvey and D. R. Wood, Parameters tied to treewidth. *J. Graph Theory* **84** (2017), no. 4, 364–385 Zbl 1359.05030 MR 3623383

[40] L. S. Heath, F. T. Leighton, and A. L. Rosenberg, Comparing queues and stacks as mechanisms for laying out graphs. *SIAM J. Discrete Math.* **5** (1992), no. 3, 398–412 Zbl 0764.05093 MR 1172748

[41] L. S. Heath and A. L. Rosenberg, Laying out graphs using queues. *SIAM J. Comput.* **21** (1992), no. 5, 927–958 Zbl 0778.05078 MR 1181408

[42] R. Hickingbotham, P. Jungeblut, L. Merker, and D. R. Wood, The product structure of squaregraphs. *J. Graph Theory* **105** (2024), no. 2, 179–191 Zbl 1531.05057 MR 4684232

[43] R. Hickingbotham and D. R. Wood, Shallow minors, graph products, and beyond-planar graphs. *SIAM J. Discrete Math.* **38** (2024), no. 1, 1057–1089 Zbl 1535.05076 MR 4716720

[44] P. Hliněný and J. Jedelský, $\mathcal{H}$-clique-width and a hereditary analogue of product structure. In *Proceedings of the 49th international symposium on mathematical foundations of computer science (MFCS 2024)*, LIPIcs - Leibniz Int. Proc. Inform. 306, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2024, article no. 61, 16 pp. MR 4802177

[45] T. Huynh, B. Mohar, R. Šámal, C. Thomassen, and D. R. Wood, Universality in minor-closed graph classes. 2021, arXiv:2109.00327v1

[46] F. Illingworth, A. Scott, and D. R. Wood, Product structure of graphs with an excluded minor. *Trans. Amer. Math. Soc. Ser. B* **11** (2024), 1233–1248 Zbl 1554.05157 MR 4817445

[47] H. Jacob and M. Pilipczuk, Bounding twin-width for bounded-treewidth graphs, planar graphs, and bipartite graphs. In *Proceedings of the 48th international workshop on graph-theoretic concepts in computer science (WG 2022)*, pp. 287–299, Lecture Notes in Comput. Sci. 13453, Springer, 2022 Zbl 1570.05030 MR 4606560

[48] R. Krauthgamer and J. R. Lee, The intrinsic dimensionality of graphs. *Combinatorica* **27** (2007), no. 5, 551–585 Zbl 1164.05050 MR 2375717

[49] R. J. Lipton and R. E. Tarjan, A separator theorem for planar graphs. *SIAM J. Appl. Math.* **36** (1979), no. 2, 177–189 Zbl 0432.05022 MR 0524495

[50] C.-H. Liu and D. R. Wood, Quasi-tree-partitions of graphs with an excluded subgraph. [v1] 2024, [v2] 2025, arXiv:2408.00983v2

[51] M. Pilipczuk and S. Siebertz, Polynomial bounds for centered colorings on proper minor-closed graph classes. *J. Combin. Theory Ser. B* **151** (2021), 111–147 Zbl 1490.05082 MR 4278114

[52] B. A. Reed, Tree width and tangles: a new connectivity measure and some applications. In R. A. Bailey (Ed.), *Surveys in combinatorics*, pp. 87–162, London Math. Soc. Lecture Note Ser. 241, Cambridge University Press, Cambridge, 1997 Zbl 0895.05034 MR 1477746

[53] N. Robertson and P. D. Seymour, Graph minors. I. Excluding a forest. *J. Combin. Theory Ser. B* **35** (1983), no. 1, 39–61 Zbl 0521.05062 MR 0723569

[54] N. Robertson and P. D. Seymour, Graph minors. V. Excluding a planar graph. *J. Combin. Theory Ser. B* **41** (1986), no. 1, 92–114 Zbl 0598.05055 MR 0854606

[55] T. Ueckerdt, D. R. Wood, and W. Yi, An improved planar graph product structure theorem. *Electron. J. Combin.* **29** (2022), no. 2, article no. P2.51, 12 pp. Zbl 1527.05046 MR 4441087

[56] V. Wiechert, On the queue-number of graphs with bounded tree-width. *Electron. J. Combin.* **24** (2017), no. 1, article no. P1.65, 18 pp. Zbl 1358.05060 MR 3651947

[57] D. R. Wood, Queue layouts of graph products and powers. *Discrete Math. Theor. Comput. Sci.* **7** (2005), no. 1, 255–268 Zbl 1153.05325 MR 2183176

[58] D. R. Wood, The structure of Cartesian products. In *Banff workshop on graph minors*, 2008, https://www.birs.ca/workshops/2008/08w5079/report08w5079.pdf visited on 2 February 2026

Vida Dujmović
School of Computer Science and Electrical Engineering, University of Ottawa, Ottawa, Canada; vida.dujmovic@uottawa.ca

David R. Wood
School of Mathematics, Monash University, Melbourne, Australia; david.wood@monash.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/17

Weak mixing in polygonal billiards

Giovanni Forni

Abstract. We survey known and conjectural ergodic properties of billiards in rational and non-rational polygons. The directional flows of billiards in rational polygons were proved to be ergodic in almost all directions by Kerckhoff, Masur and Smillie in 1986. In the same paper, the authors derived the existence of a G_δ dense set of ergodic (non-rational) polygons. In recent joint work, we have proved analogous results for the weak mixing property. In joint work with Jon Chaika, we have proved the existence of a G_δ dense set of weakly mixing (non-rational) polygons and recently with Francisco Arana-Herrera and Jon Chaika, we have characterized the property of weak mixing in almost all directions of translation surfaces with applications to billiards in rational polygons. All these results are based on techniques of Teichmüller dynamics and their application to the weak mixing problem following the joint work with A. Avila from 2007. Recent advances rest on measure rigidity results of Eskin, Mirzakhani and Mohammadi and related work of Chaika and Eskin on the multiplicative ergodic theorem in Teichmüller dynamics and results of Filip and Wright on the relevant (Kontsevich–Zorich) renormalization cocycle.

1. Introduction

Dynamical systems have long been classified as *stable* or *random*, as in the title of the classical book [40], depending on whether nearby orbits tend to stay close to each other or on the contrary diverge exponentially fast with time. Systems in the first class, which are the domain of KAM theory and variational methods (weak KAM theory, Aubry–Mather theory), but also of exotic Liouvillean behavior, as in examples constructed by the Anosov–Katok approximation by conjugation [18], are often called *elliptic*, while systems in the second class are called *hyperbolic*. Hyperbolic systems are described by a well-developed theory, which includes uniformly hyperbolic (Anosov) systems, non-uniformly hyperbolic systems (Pesin's theory), partially hyperbolic systems, dominated splitting and so on.

Mathematics Subject Classification 2020: 37A25 (primary); 37C83, 37E35, 37F34, 30F60, 32G15 (secondary).
Keywords: billiards in polygons, rational polygonal billiards, weak mixing flows, Teichmüller flow, moduli space of Abelian differentials.

Dynamical systems with intermediate behavior, characterized, for instance, by polynomial divergence with nearby orbits, often called *parabolic*, are less understood, with the exception of homogeneous unipotent flows (the domain of Ratner's theory).

The general investigation of parabolic dynamical phenomena has come to be considered an important part of the theory of dynamical systems, especially following the vision of A. Katok (see, for instance, [29]).

Planar billiards have played an important role in the development of the theory at least since the work of G. D. Birkhoff, perhaps that of a "playground", an expression that "should be understood as 'testing area': various questions, conjectures, methods of solution, etc. in the theory of dynamical systems are 'tested' on various types of billiard problems" [34].

Thus, smooth convex (Birkhoff) billiards (Figure 1, left) are a classical example of elliptic, stable, dynamical behavior, while billiards with strictly convex scatterers (Sinai's billiards) and those with focusing-defocusing mechanism, such as the famous Bunimovich stadium, are models of hyperbolic, random, dynamics, uniform and non-uniform.

Billiards in polygons (Figure 1, right) exemplify parabolic dynamics, although the (local) behavior of orbits is elliptic with singularities [34], as splitting of nearby trajectories only occurs at vertices of the polygon.

According to A. Katok, "George D. Birkhoff was the first to consider billiards systematically as models for problems of classical mechanics. Birkhoff considered billiards only in smooth convex domains" [34].

However, R. H. Fox and R. B. Kershner took up the study of geodesics on a rational polyhedron, since "the problem may be considered as an elementary illustration of the 'billiard ball' problem considered by Birkhoff in Chapter VI of his Colloquium Publication *Dynamical Systems* and was suggested to us by Wintner" [26].

The topological dynamics of this problem was investigated several decades later by A. Zemlyakov and A. Katok [48] in paper which brought billiards in polygons fully into the developing theory of dynamical systems.

In this lecture, we review ergodic theoretic properties of billiards in polygons up to recent results by the author and collaborators.

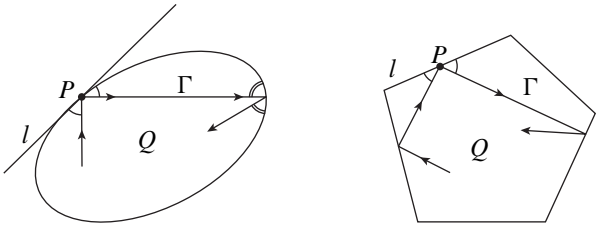


Figure 1. Convex (Birkhoff) and polygonal billiards.

2. Billiards in polygons

A *polygonal billiard* is the dynamical system given by the motion of a light ray in a polygonal (planar, Euclidean) mirror chamber P, or equivalently by the free motion of a single point mass in P with elastic collisions at the boundary. It is therefore given by straight linear motion inside the polygon and reflection at the boundary according to the law of reflection *angle of incidence = angle of reflection.*

Its (3-dimensional) phase space, after the reduction given by the (kinetic) energy integral, can be identified with the space $P \times \mathbb{T}/E$, with E expressing the reflection equivalence relation at ∂P. Indeed, the configuration space of the billiard is the polygon P, and since the kinetic energy is an integral of motion, it can be normalized. Since trajectories which hit a corner are not continued, the billiard flow is defined almost everywhere (on the complement of countably many hypersurfaces).

For background on billiards in polygons, we refer the reader to the survey by H. Masur and S. Tabachnikov [39].

2.1. Mechanical systems

The motion of two point masses m_1, m_2 on the interval $[0, 1]$ with elastic collisions between them and at the endpoints is equivalent to a billiard in a right triangle with an angle equal to $\arctan(\sqrt{m_1/m_2})$.

The motion of three point masses m_1, m_2, m_3 on a circle with elastic collisions between them is equivalent to a billiard in an acute triangle with angles

$$\arctan\left(m_i \sqrt{\frac{m_1 + m_2 + m_3}{m_1 m_2 m_3}}\right) \quad \text{for } i = 1, 2, 3.$$

This lecture reviews results and conjectures on the following fundamental question.

Question 2.1. How chaotic can a polygonal billiard be?

This question is motivated within the theory of dynamical systems by the consideration that billiards in polygons are a basic example of a parabolic system (or elliptic with singularities). The question is of interest also from the point of view of the semi-classical approach to quantum mechanics as billiards in polygons are an example of "weakly chaotic" classical dynamical system [42].

A similar, more general question concerns geodesic flow on flat compact orientable surfaces with conical singularities. In fact, the double of a Euclidean polygon is a flat sphere with conical singularities at the vertices of the polygon.

2.2. Rational versus non-rational polygons

There is a well-known fundamental dichotomy between billiards in rational and non-rational polygons.

Definition 2.2. A polygon $P \subset \mathbb{R}^2$ is called *rational* if the group $G_P \subset \mathrm{SO}(2)$ generated by reflections with respect to the edges of P is finite. Whenever G_P is infinite, the polygon P is called *non-rational*. If P is a simply connected polygon, then P is rational if and only if all of its angles are rational multiples of the round angle.

In the rational case, the map $\Theta : P \times \mathbb{T} \to \mathbb{T}/G_P$, defined as

$$\Theta : (x, v) \to [v]_{G_P} \quad \text{for all } (x, v) \in P \times \mathbb{T},$$

is a prime integral of motion, that is, a function invariant along orbits of the billiard flow, with values in the interval $\mathbb{T}/G_P$ (note that when G_P is infinite, the quotient $\mathbb{T}/G_P$ is not Hausdorff).

As a consequence, the billiard flow leaves invariant the level sets of the prime integral Θ, hence the phase space of a rational billiard is foliated by invariant surfaces. The invariant surfaces can be constructed either as an appropriate branched cover of the double of the polygon, which in the simply connected case is a flat sphere with conical points of rational angle [26] or by the so-called "unfolding construction" [48].

The invariant surfaces are *translation surfaces* since they inherit from the Euclidean plane a flat *translation* structure, that is, an atlas whose coordinate changes are given by translations of $\mathbb{R}^2$. The restrictions of the billiard flow to the invariant surfaces are linear flows called *translation flows*. Thus, the ergodic theory of rational billiards is part of that of translation flows.

The billiard flow of a rational polygon cannot even be topologically transitive, not to mention ergodic, on its (3-dimensional) phase space. The simplest meaningful ergodic theoretical questions therefore concern the restriction of the billiard flow to the random invariant surface (in the random directions) with respect to the Lebesgue measure on the angle.

The genus of the invariant surface can be computed explicitly by the Euler formula (see [39]). For any rational polygon P with $\sigma \geq 3$ edges and angles of amplitudes $(m_1\pi/n_1, \ldots, m_\sigma\pi/n_\sigma)$, the genus g_P of the invariant surfaces is given by the formula

$$g_P = 1 + \frac{\#G_P}{4}\left(\sigma - 2 - \sum_{i=1}^{\sigma} \frac{1}{n_i}\right).$$

From this formula, it follows that the invariant surfaces are tori ($g_P = 1$) only for the following *completely integrable polygons* (listed in terms of their angles):

$$\left(\frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}\right), \quad \left(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{4}\right), \quad \left(\frac{\pi}{3}, \frac{\pi}{3}, \frac{\pi}{3}\right), \quad \left(\frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{6}\right),$$

that is, for rectangles, isosceles right triangles, equilateral triangles and half-equilateral right triangles.

In recent decades the mathematical theory of generic translation flows (and of locally Hamiltonian flows on surfaces, which are singular time changes of translation flows) has developed tremendously, thanks to the discovery of *renormalization* dynamical systems for the dynamics of interval exchange transformation and translation flows: the *Rauzy–Veech–Zorich induction* algorithm [41, 43, 49] and H. Masur's *Teichmüller geodesic flow* [38]. The former can be regarded as a symbolic version of the latter, for an appropriate coding. Depending on the problem, each point of view has its own advantages; however, the connection with Teichmüller theory has brought in powerful tools from Hodge theory and algebraic geometry.

However, the theory of rational billiards has lagged behind that of generic translation flows, since translation surfaces coming from rational polygons (by unfolding) are very symmetric and therefore form a "small" exceptional set in the space of all translation surfaces. The renormalization theory cannot be directly applied, but in principle dynamical properties of rational billiards can be studied, thanks to the magic wand theorem of Eskin, Mirzakhani and Mohammadi [16, 17] and related results on the Lyapunov spectrum of the relevant renormalization cocycles, the *Rauzy–Veech cocycle* over Veech's "zippered rectangles" flow [41, 43] or its continuous time version, the *Kontsevich–Zorich cocycle* over the Teichmüller flow [13, 20–23, 36].

In contrast to the rational case, no renormalization dynamics is known for billiards in non-rational polygons. As a consequence, the mathematical theory of such billiards is much less developed. Most of the results on billiards in non-rational polygons are proved by fast approximation methods, which leverage results on translation flows and on rational polygonal billiards.

Indeed, the dichotomy between rational and non-rational billiards can be viewed as an instance of a more general dichotomy in the study of parabolic dynamical systems between "renormalizable" and "non-renormalizable": the simplest example of a renormalizable flow is a flow along the full unstable, often called "horospherical", foliation of a (partially) hyperbolic system which is a special property not verified by most parabolic flows, not even in some approximate sense.

Natural conjectures on the dynamics of non-renormalizable parabolic flows are often beyond the reach of current methods. For instance, in homogeneous dynamics, it is well known that the effective equidistribution of non-horospherical unipotent flows is much harder than that in the horospherical case (see, for instance, [37]). However, for the flow of a non-rational billiard, no renormalization approach is in sight, as there is no known hyperbolic system rescaling its orbits.

3. Dynamical properties of billiards in polygons

All billiards in polygons have *zero entropy*, as no hyperbolicity is present and the divergence of nearby orbits is subexponential [32]. In fact, nearby orbits can only diverge at corners, and the system is in fact elliptic (no-divergence) as long as the trajectories stay away from corners of the polygon.

3.1. Rational polygons

The billiard flow in a rational polygon has the following dynamical properties:

- it is *uniquely ergodic* on almost all invariant surfaces [35], with polynomial speed of ergodicity [4];
- it is *never mixing*, that is, it is not mixing on any invariant surface [31].

All of the above results are in fact proved for ALL translation surfaces. Additional results have been proved for generic translation surfaces; for instance,

- the translation flow in a fixed direction is *weakly mixing* for almost all translation surfaces with respect to the canonical absolutely continuous (Masur–Veech) measure on the relevant stratum of translation surfaces [6].

The unique ergodicity and weak mixing results can be refined with estimates on the Hausdorff dimension of the set of non-uniquely ergodic directions (see [15] and references therein for unique ergodicity and [8] for weak mixing).

In recent work with F. Arana-Herrera and J. Chaika, we have been able to completely classify rational polygons whose billiard flow is weakly mixing on almost all invariant surfaces.

Definition 3.1 ([27]). A polygon is *almost integrable* if its boundary is drawn on the planar grid obtained from a completely integrable polygon under the subgroup of the group of isometries of the plane generated by reflections with respect to its edges.

Our results prove in a strong form a conjecture of E. Gutkin [27].

Theorem 3.2 ([2]). *A billiard in a rational polygon is weakly mixing on almost all invariant surfaces (in almost all directions), unless the polygon is almost integrable or it is the rotation of a polygon with only vertical and horizontal sides such that the horizontal sides or the vertical sides have commensurable lengths.*

In light of A. Katok non-mixing result quoted above, our theorem completes the description of the weak mixing property in the random direction for rational polygonal billiards. For the reasons explained above, there were so far few results on weak mixing in rational polygons. A. Katok and E. Gutkin [28] proved the weak mixing

property in a given direction for topologically generic billiard tables in certain classes of polygons. More recently, A. Avila and V. Delecroix [5] proved a result for translation surfaces which implies weak mixing in almost all directions for regular polygons with at least five edges and other special tables (unfolding to non-arithmetic Veech surfaces).

The above theorem is derived by a standard argument from the following result for translation surfaces. We recall that a translation surface can be defined as a pair (X, ω) of a Riemann surface and a holomorphic differential ω on X. We have the following theorem.

Theorem 3.3 ([2]). *For any translation surface (X, ω), the following properties are equivalent:*

(1) *The directional flow of (X, ω) is weakly mixing in almost all directions.*

(2) *The directional flow of (X, ω) is weakly mixing in some direction.*

(3) *The surface (X, ω) does not have an affine circle factor.*

(4) *The (tautological) plane $T(\omega) = \mathbb{R} \cdot \mathrm{Re}(\omega) + \mathbb{R} \cdot \mathrm{Im}(\omega) \subset H^1(X, \mathbb{R})$ has trivial intersection with the lattice $H^1(X, \mathbb{Z})$.*

The core of the theorem is (4) $\Rightarrow$ (1).

We note that, as weak mixing is equivalent to the existence of a measurable circle factor, by the above theorem, the existence of a measurable circle factor in a positive measure set of directions implies the existence of an affine circle factor.

As mentioned above, the conclusion of the theorem was known in the special case of Veech translation surfaces [5]. D. Aulicino, A. Avila and V. Delecroix have announced a generalization to any surface in a rank one suborbifold, which includes countably many L-shaped billiards with a barrier.

3.2. Non-rational polygons

The mathematical theory of billiards in non-rational polygons is much less advanced, and the main results are derived by approximation:

- The set of ergodic polygons is a dense G_δ [35].
- The set of *topologically* weakly mixing polygons is a dense G_δ [9].

An effective version of the above ergodicity result of Kerckhoff, Masur and Smillie [35], with an estimate on the speed of approximation of the angles by rationals sufficient for ergodicity, has been proved by Y. Vorobets [45].

A few years ago, in a collaboration with J. Chaika, we proved that the weak mixing property of the billiard flow is topologically generic among polygons, a result first conjectured in the work of E. Gutkin and A. Katok [28].

Theorem 3.4 ([14]). *The set of weakly mixing polygonal billiards is a dense G_δ.*

The drawback of all of the above results is that ergodic properties are established for polygons with Liouvillean non-rational angles (that is, extremely well-approximable by rationals), which is a set of measure zero.

Conjecturally, polygons with angles of amplitudes equal to badly approximable (Diophantine) or random multiples of the round angle should exhibit stronger chaotic properties, but in the absence of a renormalization theory such conjectures remain for now out of reach.

Several conjectures can be drawn from numerical evidence in the mathematical physics literature. For instance, it is suggested that

- completely irrational acute triangles are mixing [12],
- typical right triangles may not be even ergodic [46], except in the Liouvillean case [30].

We are therefore led to formulate the following questions.

Question 3.5. The following questions are open:

- Is the non-rational billiards with random angles (with respect to the Lebesgue measure) ergodic? (see [33, Question 3])
- Does there exist a mixing polygonal billiard?
- Does there exist a relatively mixing translation surface/rational billiard?

4. Strategy and outline of the proofs

In the remainder of the paper, we outline the strategy to prove weak mixing for translation flows and interval exchange transformations (IETs) developed in [6, 8, 44] and the proofs of Theorem 3.4 from [14] and Theorem 3.3 from [2], emphasizing the novelties in the approach required in carrying out elimination of eigenvalues for an arbitrary given translation surface.

4.1. Weak mixing from eigenvalue elimination

We begin by recalling the definition of weak mixing and state some equivalent ergodic properties. A flow ϕ on (X, μ) is *weakly mixing* if

(1) for all $f, g \in L^2(X, \mu)$ of zero average,

$$\frac{1}{T}\int_0^T |\langle f \circ \phi_t, g\rangle|^2 dt \to 0;$$

(2) the Koopman operator on $L^2(X, \mu)$ has no non-constant eigenfunctions, no non-trivial atomic spectral measures; or

(3) the flow $\phi \times \psi$ on $(X \times Y, \mu \otimes \nu)$ is ergodic for any ergodic flow ψ on (Y, ν). It is enough to consider the set of all linear flows on $\mathbb{T}$; or

(4) the square flow $\phi \times \phi$ on $(X \times X, \mu \otimes \mu)$ is ergodic.

In the proof of Theorem 3.4 on the existence of weakly mixing (non-rational) polygons, we derive *square ergodicity* (the equivalent property (4)), hence weak mixing, from the following.

Theorem 4.1 ([14]). *For every translation surface (X, ω), for every $\alpha \neq 0$, the set of directions $\theta \in \mathbb{T}$ such that the directional flow ϕ^θ on (X, ω) has eigenvalue α has zero Lebesgue measure in $\mathbb{T}$.*

We note that it is crucial for the argument that the above result holds for all translation surfaces (indeed, it holds even for the genus 1 case of the flat torus).

We prove that for a dense G_δ set of polygons, the billiard flow is square ergodic by the following steps (in reverse order):

(1) (Fast approximation) Square ergodicity for a G_δ-dense set of tables follows from ergodicity for rational tables on almost all products of invariant surfaces by a Baire category argument (as in [35], from a suggestion of M. Boshernitzan and A. Katok, applied in our work to the square dynamics.)

(2) (Spectral theory of products) A product flow fails to be ergodic if and only if the factors have a common non-trivial eigenvalue. Thus, by Fubini $\phi^{\theta_1} \times \phi^{\theta_2}$ fails to be ergodic for a positive measure set of $(\theta_1, \theta_2) \in \mathbb{T}^2$ if and only if there exists α eigenvalue of ϕ^θ for a positive measure set of $\theta \in \mathbb{T}$.

In the proof of Theorem 3.2 on the weak mixing in almost all directions in translation surfaces, we prove that the Koopman operator has no non-constant eigenvalue with measurable eigenfunction (equivalent property (2)).

In summary, the proofs of Theorem 3.4 (in fact of Theorem 4.1) and of Theorem 3.3 are based on the *elimination of eigenvalues* of translation flows, more precisely of an arbitrary fixed eigenvalue in the proof of Theorem 3.4 [14] and of the full line of possible eigenvalues in the proof of Theorem 3.3.

The classical tool for the elimination of eigenvalues since W. Veech's work in the 1980s is the so-called Veech criterion [44]. This criterion has also been exploited later in [6, 8] and is sufficient for the elimination of a fixed eigenvalue as in Theorem 4.1. However, the elimination procedure presents important novelties compared with earlier works. The Veech criterion is, on the contrary, not sufficient for the elimination in the proof of Theorem 3.3. For this reason, we have introduced in [2], in the

proof of Theorem 3.3, a strengthened criterion, based on the general *partial rigidity* of translation flows [31].

4.2. The Veech criterion

We formulate the criterion in terms of the Kontsevich–Zorich (KZ) cocycle on the (real) Hodge bundle with fiber $H^1(X,\mathbb{R})$ at every translation surface (X,ω) in a stratum of the moduli space of Abelian differentials. Indeed, the moduli space of translation surfaces is stratified by fixing multiplicities of zeros of the holomorphic 1-form ω. Strata are determined by a vector $\kappa=(k_1,\dots,k_\sigma)\in\mathbb{N}^\sigma$ of multiplicities of the zeros, which satisfies the topological constraint $\sum_{i=1}^{\sigma}k_i=2g-2$ for surfaces of genus $g\geq 2$.

Let $\widehat{\mathcal{H}}(\kappa)$ and $\mathcal{H}(\kappa):=\widehat{\mathcal{H}}(\kappa)/\Gamma_g$ denote a stratum in the Teichmüller space, respectively, in the moduli space, of Abelian differentials on Riemann surfaces of genus $g\geq 2$ (the moduli space $\mathcal{H}(\kappa)$ is the quotient of the Teichmüller space $\widehat{\mathcal{H}}(\kappa)$ by the action of the mapping class group Γ_g). There is a natural action of the group $\mathrm{SL}(2,\mathbb{R})$ on $\widehat{\mathcal{H}}(\kappa)$ which commutes with that of Γ_g, hence a natural action of $\mathrm{SL}(2,\mathbb{R})$ on $\mathcal{H}(\kappa)$. This action can be defined by post-composition of charts of a translation atlas with elements of $\mathrm{SL}(2,\mathbb{R})$ as linear maps of $\mathbb{R}^2$ or equivalently for all $(X,\omega)\in\mathcal{H}_g^{(1)}$,

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}\cdot\begin{pmatrix}\mathrm{Re}(\omega)\\ \mathrm{Im}(\omega)\end{pmatrix}=\begin{pmatrix}a\,\mathrm{Re}(\omega)+b\,\mathrm{Im}(\omega)\\ c\,\mathrm{Re}(\omega)+d\,\mathrm{Im}(\omega)\end{pmatrix}\quad\text{for}\quad\begin{pmatrix}a & b\\ c & d\end{pmatrix}\in\mathrm{SL}(2,\mathbb{R}).$$

Definition 4.2. The *KZ cocycle* $\{g_t^{\mathrm{KZ}}\}$ is the restriction to the *Teichmüller flow* $\{g_t\}=\mathrm{diag}(e^t,e^{-t})<\mathrm{SL}(2,\mathbb{R})$ of the projection to the bundle $(\widehat{\mathcal{H}}(\kappa)\times H^1(\mathbb{R}))/\Gamma_g$ of the cocycle

$$\mathrm{SL}(2,\mathbb{R})\times\mathrm{id}:\widehat{\mathcal{H}}(\kappa)\times H^1(\mathbb{R})\to\widehat{\mathcal{H}}(\kappa)\times H^1(\mathbb{R}).$$

We can now state a weak mixing criterion for translation flows from [44].

Theorem 4.3 (Veech criterion). *If the real number α is an eigenvalue for the vertical flow ϕ_t on (X,ω), then the following holds: for every compact set $K\subset\mathcal{H}(\kappa)$, there is $\ell_K>0$ such that, if for some $\ell\geq\ell_K$ and a diverging sequence (t_n), for all $n\in\mathbb{N}$, we have $g_{t_n-\ell}(\omega),g_{t_n}(\omega),g_{t_n+\ell}(\omega)\in K$, then*

$$\lim_{n\to+\infty}g_{t_n}^{\mathrm{KZ}}(\omega,\alpha\cdot\mathrm{Re}(\omega))=0\quad\textit{in}\quad H^1(\mathbb{R})/H^1(\mathbb{Z}).$$

Recent developments concerning the Veech criterion include the proof, based on Hodge theory, of a slightly stronger version [25] and an effective Veech criterion to prove *effective weak mixing* (polynomial speed of convergence of Cesaro

averages of correlations) for translation flows [10, 11, 24] and interval exchange transformations [7].

Proof of the Veech criterion (outline). Let K be a fixed compact subset of the moduli space of Abelian differentials. It is enough to prove that for a given canonical basis $\mathcal{B}(t) := \{a_1(t), \dots, a_g(t), b_1(t), \dots, b_g(t)\}$ of the homology of the surface, composed of curves of bounded length for $g_t(\omega) \in K$, the periods

$$\int_{a_1(t)} \alpha \operatorname{Im}(\omega), \dots, \int_{a_g(t)} \alpha \operatorname{Im}(\omega), \int_{b_1(t)} \alpha \operatorname{Im}(\omega), \dots, \int_{b_g(t)} \alpha \operatorname{Im}(\omega)$$

are close to integers. By definition of the Teichmüller flow, there exists a constant $C_K > 0$ such that any loop $\gamma(t) \in \mathcal{B}(t)$ has horizontal length at least $C_K^{-1} e^t$ and vertical length at most $C_K e^{-t}$ on (X, ω). Let us assume that all homology classes of loops in $\mathcal{B}(t)$ can be realized by paths composed of exactly one (long) vertical and one (short) horizontal segment. Let f denote an eigenfunction of eigenvalue $\alpha \in \mathbb{R} \setminus \{0\}$ for the vertical flow ϕ_s on (X, ω). We can integrate the eigenvalue equation for f along the vertical segment $\gamma_v(t) = \{\phi_s(x) \mid 0 \le t \le L_v(t)\}$ of any path $\gamma(t) \in \mathcal{B}(t)$ to get

$$f \circ \phi_{L_v(t)}(x) = e^{2\pi i \alpha L_v(t)} f(x).$$

Since the distance $d(x, \phi_{L_v(t)}(x))$ on (X, ω) equals the length of vertical component of γ, and since f can be chosen with constant unit modulus, under the hypothesis that f is continuous, we have that, as $t \to +\infty$,

$$\left| \frac{f \circ \phi_{L_v(t)}(x)}{f(x)} - 1 \right| \to 0,$$

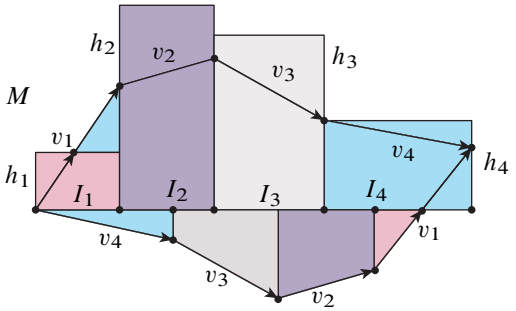
hence

$$d(\alpha L_v(t), \mathbb{Z}) = d\left(\int_{\gamma(t)} \alpha \operatorname{Im}(\omega), \mathbb{Z} \right) \to 0.$$

The argument is therefore complete in case there is a continuous eigenfunction of eigenvalue $\alpha \neq 0$. In general, the eigenfunction may be only measurable. In this case, in order to apply Luzin's theorem, which states that restrictions of the eigenfunction to sets with complements of arbitrarily small measure are continuous, we assume that the vertical segment of every loop $\gamma(t) \in \mathcal{B}(t)$ (in fact of a subset of $\mathcal{B}(t)$ which spans the homology of the surface) *belongs to a vertical rectangle $R_{\gamma(t)}$ of area uniformly bounded below*. The lower bound on the area of the rectangles ensures that Luzin sets with complement of sufficiently small measure intersect all of them, and therefore that it is possible to choose loops with vertical components starting at points of continuity of the eigenfunction. In the statement of the Veech criterion, the condition that the Teichmüller orbit has three consecutive returns (at times $t_n - \ell$, t_n and

$t_n + \ell$) to a compact set has precisely the role of enforcing a uniform equivalence of the area of all vertical rectangles $R_{\gamma(t)}$ in a decomposition of the surfaces, hence a uniform lower bound on the area of each of them.

In Figure 2, which illustrates the Veech criterion, the symbol $\mathcal{RV}$ denotes the Rauzy–Veech map, a discrete time version of the Teichmüller flow, and the sequence



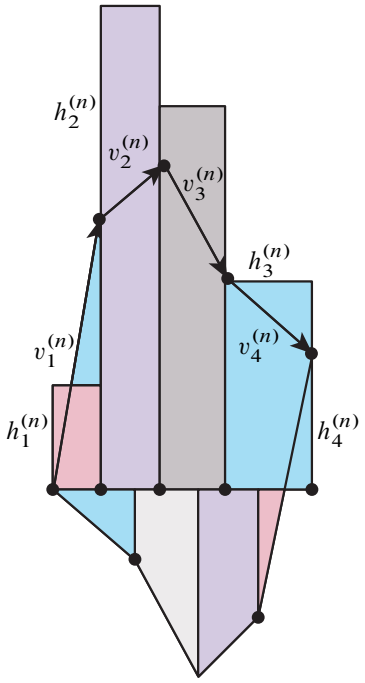
Eigenfunction equation

$$\frac{\partial u}{\partial y} = 2\pi i \alpha u$$

$$u \in L^2(M)$$

$$u(x, h_k) = e^{2\pi i \alpha h_k} u(x, 0) \quad \text{for } x \in I_k$$

Permutation: $(4\ 3\ 2\ 1) = \sigma$
Horizontal data: $(\ell_1, \ell_2, \ell_3, \ell_4) = \ell$
Vertical data: $(h_1, h_2, h_3, h_4) = h \in H(\sigma) = \mathbb{R}^4$



Horizontal data: $(\ell_1^{(n)}, \ell_2^{(n)}, \ell_3^{(n)}, \ell_4^{(n)}) = \ell^{(n)}$
Vertical data: $(h_1^{(n)}, h_2^{(n)}, h_3^{(n)}, h_4^{(n)}) = h^{(n)}$
$(\ell^{(n)}, h^{(n)}) = (\mathcal{RV})^n(\ell, h)$

$$u(x, h_k^{(n)}) = e^{2\pi i \alpha h_k^{(n)}} u(x, 0) \quad \text{for } x \in I_k$$

$u(x, h_k^{(n)}) \approx u(x, 0)$ (Luzin's theorem)
$\Rightarrow \quad \alpha h^{(n)} = \alpha A_n(\ell)(h)$
close to $\mathbb{Z}^4 \subset \mathbb{R}^4$

Figure 2. The Veech criterion.

of matrices $(A_n(\ell)(h))$ represents the Rauzy–Veech cocycle, a discrete version of the KZ cocyle, along the orbit $(\mathcal{RV}^n(\ell, h))$. We are neglecting to keep track of the permutation along the orbit of the Rauzy–Veech map. ∎

4.3. Elimination of eigenvalues

The Veech criterion leads to the notion of a *weak stable space* $W_\omega \subset H^1(X, \mathbb{R})$ at almost all (X, ω), as the set of all cohomology classes such that the conclusion of the Veech criterion holds, that is, the set of all $c \in H^1(X, \mathbb{R})$ such that $g_t^{\mathrm{KZ}}(\omega, c)$ "converges" to $H^1(\mathbb{Z})$. The Veech criterion implies that the vertical flow on a translation surface $(X, \omega) \in \mathcal{H}$ is weakly mixing if $\mathbb{R}[\mathrm{Re}(\omega)] \cap W_\omega = \{0\}$.

We list below the basic elementary properties of the weak stable space:

(1) W_ω is saturated by subspaces parallel to the stable space of the Kontsevich–Zorich cocycle on $H^1(\mathbb{R})$.

(2) W_ω coincides with the union of integer translates of the stable space if the forward Teichmüller orbit is bounded.

(3) W_ω only depends on $\mathrm{Im}(\omega)$.

A couple of elimination schemes were introduced in [6]: the first, called *linear elimination*, is carried out on the heights parameters (first return times) of translation flows and is sufficient to prove that almost all translation flows are weakly mixing (but not for the proof of weak mixing for almost all IETs); the second, called *probabilistic elimination*, is carried out on the transverse parameters of a translation flows for a fixed vector of heights parameters, and it is adapted to the case of IETs. In fact, for IETs on d intervals, the vector of heights can be taken to be $(1, \ldots, 1) \in \mathbb{R}^d$.

4.3.1. Linear elimination. The linear elimination is based on a Hausdorff dimension bound from [6], which can be stated as follows.

Lemma 4.4. *For almost all forward Oseledets regular points (X, ω) with respect to the Kontsevich–Zorich cocycle, we have the following inequality for the Hausdorff codimension of the weak stable space:*

$$\text{H-codim}(W_\omega) \geq \#\{\text{positive KZ exponents}\}.$$

The above result also holds for the effective weak stable space [11, 24].

For *Veech surfaces*, the weak stable space W_ω is the union of integer translates of the Oseledets stable space [5]. Avila, Aulicino and Delecroix have announced a generalization of this result to all rank one $\mathrm{SL}(2, \mathbb{R})$-invariant suborbifolds.

For any $\mathrm{SL}(2, \mathbb{R})$-invariant suborbifold $\mathcal{M}$, the number of positive KZ exponents in $T\mathcal{M} \subset H^1(X, \Sigma, \mathbb{R})$ is equal to the rank of $\mathcal{M}$ [19], hence when $\mathrm{rank}(\mathcal{M}) \geq 2$ weak mixing of the vertical flow holds for almost all translation surfaces in $\mathcal{M}$ (with respect

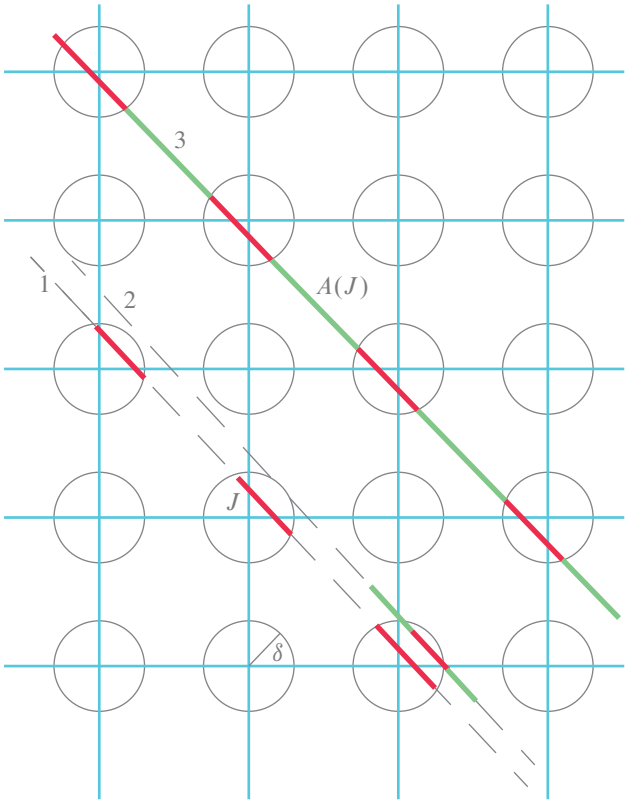


Figure 3. Repulsion and spreading.

to the canonical $\mathrm{SL}(2,\mathbb{R})$-invariant probability measure) since W_ω only depends on $\mathrm{Re}(\omega)$ and the vector of heights (return times) of the vertical flow is $\mathrm{Im}(\omega)$ (and both vector belong to the projection of $T\mathcal{M}$ in $H^1(X,\mathbb{C})$). In fact, by dimensional reasons, we have that $\mathbb{R}[\mathrm{Im}(\omega)] \cap W_\omega = \{0\}$ for almost all $\mathrm{Im}(\omega)$.

4.3.2. Probabilistic elimination. In this elimination scheme, introduced in [6], the vector $\mathbb{R}[\mathrm{Re}(\omega)]$ is fixed. The scheme analyzes the dynamics on the bundle $H^1(\mathbb{R})$ modulo $H^1(\mathbb{Z})$ with toral fibers over strata of the moduli space of line segments approximately aligned with the top unstable Oseledets subspace.

There are two main effects (see Figure 3):

- *Repulsion*: under the hypothesis that the second Lyapunov exponent of the KZ cocycle is strictly positive ($\lambda_2 > 0$), line segments are repelled away from the nearest integer point (under the action of the KZ cocycle).

- *Spreading*: for sufficiently long excursions of the Teichmüller orbit, a line segment can get dilated (under the action of the KZ cocycle) and spread to several line segments near multiple nearby integer points.

We note that the phenomenon of "spreading" is a priori compatible with the Veech criterion, since the latter does not constrain the orbit of the cohomology class under the Kontsevich–Zorich cocycle when the Teichmüller orbit is outside of a given compact subset of a stratum.

With sufficiently high probability, a line segment is repelled "far enough" from integer points and eliminated, since by the Veech criterion its points cannot be eigenvalues (for the vertical flow). It is possible to estimate that the probability of escaping elimination by spreading goes to zero with time [6]. A version of the probabilistic elimination for effective weak mixing of IETs is carried out in [7].

The probabilistic elimination is well adapted to the proof of (effective) weak mixing for directional flows of almost all translation surfaces with respect to the Masur–Veech measure on strata [6, 7].

4.4. Weakly mixing polygonal billiards

We recall that proof of density of weakly mixing (non-rational) polygons is based on the elimination *for a given translation surface* of a single (fixed) eigenvalue, that is, on the proof of Theorem 4.1. The elimination is carried out separately for each irreducible subbundle of the Hodge (real cohomology) bundle:

- For the "tautological" plane $T_\omega = \mathbb{R}[\mathrm{Re}(\omega)] + \mathbb{R}[\mathrm{Im}(\omega)]$: an elementary transversality argument to eliminate intersections with the stable subspace $\mathbb{R}[\mathrm{Im}(\omega)]$.
- For isometric subbundles: Veech's isometric elimination from [44].
- For "hyperbolic" (at least one positive exponent) subbundles: repulsion without spreading for a positive measure set of directions (for any given surface).

We note that, for the KZ cocycle, irreducible subbundles are either hyperbolic or isometric, never "parabolic" [19] (in the sense that either there is exponential growth of typical vectors or there is no growth at all).

4.4.1. Controlled hyperbolic elimination. The crucial novelty in the above elimination scheme is a controlled repulsion which achieves elimination (again based on the Veech criterion) before spreading occurs. The essential technical ingredient is the following large deviation result.

Theorem 4.5 ([14]). *For any translation surface (X, ω), there exist $\lambda > 0$ and $\delta > 0$ such that (along the Teichmüller horocycle orbit segment $\{h_s\omega \mid s \in [-1, 1]\}$) for large $T > 0$ and for $c \in H^1(X, \mathbb{R})$, we have*

$$\mathrm{Leb}\{s \in [-1, 1] | g_T^{\mathrm{KZ}}(h_s\omega, c)| < e^{\lambda T}\} < e^{-\delta T}.$$

This result both improves upon and complements large deviation estimates of [1, 5, 8]. It is based on the results of Eskin and Mirzakhani [16] and Eskin, Mirzakhani and Mohammadi [17], on the results of Chaika and Eskin on the Oseledets theorem in almost all directions for arbitrary surfaces [13].

By the recurrence result of J. Athreya [3], for a fixed surface, the probability of an excursion outside a given compact set $\mathcal{K}$ of a stratum is exponentially small with the length of the excursion: there exists $\eta > 0$ such that, for any $S > 0$,

$$\mathrm{Leb}\{s \in [-1,1] \mid U_{t\in[S,S+T]} g_t(h_s\omega) \cap \mathcal{K} = \emptyset\} < e^{-\eta T}.$$

From the above estimates, we derive that if at some time we have a cohomology class $c \in H^1(X,\mathbb{R})$ such that $d(c, H^1(\mathbb{Z})) \approx \gamma$ (sufficiently small), then there exist an integer class $z \in H^1(X,\mathbb{Z})$ and a time $t := t(s)$ within $[-\log\gamma/2, -3\log\gamma/4]$ such that, outside a set of s of measure γ^σ ($\sigma > 0$),

$$g_t(h_s\omega) \in \mathcal{K} \quad \text{and} \quad d(g_t^{\mathrm{KZ}}(h_s\omega, c), H^1(\mathbb{Z})) \approx \gamma^{-\lambda}\|c - z\| = d(c, H^1(\mathbb{Z}))^{1-\lambda}.$$

Iteration of this argument produces controlled repulsion away from the integer lattice outside a set of "small" measure (for γ small). Indeed, the norm of the cohomology class $c - z$ (hence the distance of the cohomology class from the integer lattice) grows under the Kontsevich–Zorich cocycle like $\gamma^{(1-\lambda)^k} \to 1$ as $k \to \infty$, while the measure of the discarded set is at most

$$\sum_{j=0}^{\infty} \gamma^{\sigma(1-\lambda)^j},$$

a quantity which converges to zero as $\gamma \to 0$.

A density type argument then completes the proof of Theorem 3.4.

4.5. Weak mixing in rational polygonal billiards

The challenge in this work is to go from elimination of a single eigenvalue to the elimination of the whole line of potential eigenvalues. By the Veech criterion, it is not restrictive to focus on a single segment $J \subset \mathbb{R}\cdot\mathrm{Im}(\omega)$ contained in a small neighborhood of a lattice point $z \in H^1(X,\mathbb{Z})$. Let

$$H^1(X,\mathbb{R}) = T_\omega \oplus T_\omega^\perp$$

denote the decomposition of the cohomology space as a direct sum of the so-called tautological subbundle $T_\omega = \mathbb{R}\,\mathrm{Re}(\omega) \oplus \mathbb{R}\,\mathrm{Im}(\omega)$ and of its symplectic orthogonal. Let π_ω denote the projection onto the subspace $T_\omega^\perp$. An important remark is that, since $J \subset \mathbb{R}[\mathrm{Im}(\omega)] \subset T_\omega$, we have

$$\pi_\omega(J - z) = -\pi_\omega(z) \in T_\omega^\perp,$$

which is a single vector (for all points of J) to which the controlled repulsion argument outlined in Section 4.4.1 from [14] can be applied. However, since the projection of J belongs to the unstable subspace of the Kontsevich–Zorich cocycle, the Veech criterion cannot avoid dilation and spreading of the interval J to a collection of intervals near finitely many lattice points whenever the Teichmüller orbit makes an excursion outside a given compact set.

This difficulty is overcome in [2] by introducing a refined weak mixing criterion, which implies that, as long as the projection $\pi_\omega(z)$ remains sufficiently small under the Kontsevich–Zorich cocycle, the tautological projection of the interval $J - z$ remains confined to a neighborhood of the origin. Hence, there is no spreading in the tautological plane, and the controlled repulsion on $T_\omega^\perp$, by increasing the norm of $\pi_\omega(z)$, leads to a contradiction with the Veech criterion.

In the argument, it is important that we can restrict to subbundle on which the Kontsevich–Zorich cocycle is fully hyperbolic. Indeed, the line $\mathbb{R} \cdot \mathrm{Im}(\omega)$ is contained in the tautological plane $T_\omega \subset H^1(X, \mathbb{R})$, which in turn is contained in the subbundle given by the sum of all of its Galois conjugate of the projection of the tangent space of the $\mathrm{SL}(2, \mathbb{R})$-orbit closure of (X, ω) onto the absolute cohomology. It was proved by A. Wright [47] that this Galois subbundle is defined over $\mathbb{Q}$, hence it has a well-defined integral lattice. The Veech criterion still holds for the restriction of the cocycle to the Galois subbundle, hence the argument can be carried out in this restricted set up. In addition, by a result of S. Filip [19], all the Lyapunov exponents of the Kontsevich–Zorich cocycle on the Galois subbundle are all non-zero.

4.5.1. A weak mixing criterion. The refined weak mixing criterion is based on the notion of a rigidity configuration. Roughly speaking, a rigidity configuration for a translation surface is an immersed rectangle of large vertical size and small horizontal size which contains an embedded sub-rectangle whose area is uniformly bounded from below. A *rigidity configuration* is a pair (J, R) with J a horizontal interval and R an embedded rectangle with top and bottom horizontal sides in J. It is characterized by the following parameters: the vertical height of the maximal immersed rectangle with base J, the vertical size $V > 0$ of R, the horizontal length $H > 0$ of J, the area $\sigma > 0$ of the maximal embedded sub-rectangle $R^* \subset R$ (see [2, Definition 3.2]).

We then have the following crucial existence statement.

Proposition 4.6 ([2, Proposition 4.1]). *Let $\mathcal{H}^1$ be a stratum of unit area Abelian differentials. Then, for every compact set $\mathcal{K} \subseteq \mathcal{H}^1$, there exist constants $L_0 = L_0(\mathcal{K}) \geq 1$ and $C = C(\mathcal{K}) \geq 1$ such that, for every $(X, \omega) \in \mathcal{K}$ and every $L \geq L_0$, there exists a (V, H, σ, L)-rigidity configuration (J, R) of (X, ω) with the following properties:*

(1) $L/C \leq V \leq C \cdot L$.

(2) $\sigma \geq 1/C$.

(3) *If* $0 < t < \log L$ *is such that* $g_t(X,\omega) \in \mathcal{K}$, *then* $H < C \cdot e^{-t}$.

An argument reminiscent of the proof of Veech's criterion (see [44] or [14]) then establishes the following (see [2, Definition 3.3 and Lemma 3.4]).

Theorem 4.7 (Weak mixing criterion). *Let* $(X,\omega) \in \mathcal{H}$ *be a translation surface. If the vertical flow is ergodic and* α *is an eigenvalue for the vertical flow on* (X,ω), *then for every* $\varepsilon > 0$ *and compact* $\mathcal{K} \subset \mathcal{H}^1$, *there exists* $t_0 := t_0(\varepsilon, \mathcal{K}, \alpha) > 0$ *such that for all* $t > t_0$, *both of the following two conditions hold:*

(1) *For all* (V, H, σ, L)*-rigidity configuration such that* $V \geq e^t$, *at least one of the following holds:*

$$\sigma < \varepsilon \quad or \quad d_{\mathbb{R}}(\alpha V, \mathbb{Z}) < \varepsilon.$$

(2) *(Veech criterion) for all* $L > t$ *such that* $(X_L, \omega_L) \in \mathcal{K}$,

$$d(g_L \cdot \alpha \operatorname{Im}(\omega), H^1(X_L;\mathbb{Z})) < \varepsilon.$$

Let us recall the splitting $H^1(X,\mathbb{R}) = T_\omega \oplus T_\omega^\perp$ as a direct sum of the tautological plane and of its symplectic orthogonal, and let, as above, π_ω be the projection onto $T_\omega^\perp$ and $(I - \pi_\omega)$ the projection onto the tautological plane. Let then $(I - \pi_\omega)^+$ denote the composition of the projection onto T_ω and the projection onto the unstable (1-dimensional) subspace of T_ω along the (1-dimensional) stable space.

Proposition 4.8. *Let* $\mathcal{H}^1$ *be a stratum of unit area Abelian differentials over a moduli space of Riemann surfaces. Then, for every compact set* $\mathcal{K} \subseteq \mathcal{H}^1$, *there exist constants* $C = C(\mathcal{K}) \geq 1$ *and* $t_0 := t_0(\mathcal{K}) > 0$ *with the following property. Let* $(X,\omega) \in \mathcal{H}^1$ *be any translation surface. Consider* $t \geq t_0$ *such that the following conditions hold:*

(1) $g_t(X,\omega) = (X_t, \omega_t) \in \mathcal{K}$.

(2) *There exists* $z \in H^1(X_t,\mathbb{Z})$ *such that* $d(g_t^{\mathrm{KZ}} \cdot \alpha \operatorname{Im}(\omega), z) < 1/(2C^2)$.

Suppose $\alpha \in \mathbb{R} \setminus \{0\}$ *has the property that* $d_{\mathbb{R}}(V\alpha,\mathbb{Z}) < 1/C$ *for all* $V > e^t/C''$ *whenever* (X,ω) *has a* $(V, H, 1/C, L)$*-rigidity configuration. Let* $\beta \in \mathbb{R}$ *be the unique real number such that* $(I - \pi_\omega)^+(z) = g_t^{\mathrm{KZ}} \cdot \beta \operatorname{Im}(\omega) \in H^1(X_t,\mathbb{R})$. *Then,*

$$d_{\mathbb{R}}(\alpha,\beta) \leq C \cdot \|\pi_\omega(z)\| \cdot e^{-t}.$$

The significance of the above proposition stands in the fact that as long as the norm $\|\pi_\omega(z)\|$ stays small, since the vector $(I - \pi_\omega)(z - \beta \operatorname{Im}(\omega))$ belongs to the stable Oseledets space in the tautological plane, the norm of the vector $z - \alpha \operatorname{Im}(\omega)$ remains small for a time of the order $\log\|\pi_\omega(z)\|$. It then follows from the large deviation Theorem 4.5 as in Section 4.4.1 that if $\gamma := \|\pi_\omega(z)\|$ is sufficiently small, there

exists a time $t := t(s) \in [-\log \gamma/2, -3\log \gamma/4]$ such that, outside a set of s of measure γ^σ ($\sigma > 0$),

$$g_t(h_s\omega) \in \mathcal{K} \quad \text{and} \quad \|g_t^{\mathrm{KZ}}(\pi_\omega(z))\| \approx \gamma^{-\lambda}\|\pi_\omega(z)\| = \|\pi_\omega(z)\|^{1-\lambda}.$$

Iteration of this argument produces controlled repulsion away from the integer lattice outside a set of "small" measure (for γ small). Indeed, by Proposition 4.8, there exists $z \in H^1(\mathbb{Z}) \setminus \{0\}$ such that the norm of the vector $z - \alpha\,\mathrm{Im}(\omega)$ remains under control (and less than $1/2$) up to the time $t \in [-\log \gamma/2, -3\log \gamma/4]$, while the norm of its projection $\pi_\omega(z - \alpha\,\mathrm{Im}(\omega)) = \pi_\omega(z)$ grows, under the assumption that $\pi_\omega(z) \neq 0$. The latter assumption is ensured by the hypotheses that the tautological plane T_ω *contains no non-zero integer points*. By a finite iteration of the procedure just described, we can prove that, outside of a set of parameters of small measure, the norm $\|g_t^{\mathrm{KZ}}(z - \alpha\,\mathrm{Im}(\omega))\| = d(g_t^{\mathrm{KZ}}(\alpha\,\mathrm{Im}(\omega)), \mathbb{Z})$ grows larger than a preassigned constant, thereby contradicting the Veech criterion for weak mixing and completing the argument.

Acknowledgments. The author is grateful to F. Arana-Herrera and J. Chaika for the privilege of collaborating with them.

Funding. This work was partially supported by NSF Grant DMS-2154208.

References

[1] H. Al-Saqban, P. Apisa, A. Erchenko, O. Khalil, S. Mirzadeh, and C. Uyanik, Exceptional directions for the Teichmüller geodesic flow and Hausdorff dimension. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 5, 1423–1476 Zbl 1475.37038 MR 4244511

[2] F. Arana-Herrera, J. Chaika, and G. Forni, Weak mixing in rational billiards. 2024, arXiv: 2410.11117v1

[3] J. S. Athreya, Quantitative recurrence and large deviations for Teichmuller geodesic flow. *Geom. Dedicata* **119** (2006), no. 1, 121–140 Zbl 1108.32007 MR 2247652

[4] J. S. Athreya and G. Forni, Deviation of ergodic averages for rational polygonal billiards. *Duke Math. J.* **144** (2008), no. 2, 285–319 Zbl 1153.37001 MR 2437681

[5] A. Avila and V. Delecroix, Weak mixing directions in non-arithmetic Veech surfaces. *J. Amer. Math. Soc.* **29** (2016), no. 4, 1167–1208 Zbl 1373.37005 MR 3522612

[6] A. Avila and G. Forni, Weak mixing for interval exchange transformations and translation flows. *Ann. of Math. (2)* **165** (2007), no. 2, 637–664 Zbl 1136.37003 MR 2299743

[7] A. Avila, G. Forni, and P. Safaee, Quantitative weak mixing for interval exchange transformations. *Geom. Funct. Anal.* **33** (2023), no. 1, 1–56 Zbl 1514.37012 MR 4561147

[8] A. Avila and M. Leguil, Weak mixing properties of interval exchange transformations & translation flows. *Bull. Soc. Math. France* **146** (2018), no. 2, 391–426 Zbl 1402.37002 MR 3933880

[9] J. Bobok and S. Troubetzkoy, Topologically weakly mixing polygonal billiards. *Proc. Amer. Math. Soc.* **151** (2023), no. 8, 3391–3400 Zbl 1523.37035 MR 4591774

[10] A. I. Bufetov and B. Solomyak, A spectral cocycle for substitution systems and translation flows. *J. Anal. Math.* **141** (2020), no. 1, 165–205 Zbl 1462.37025 MR 4174040

[11] A. I. Bufetov and B. Solomyak, Hölder regularity for the spectrum of translation flows. *J. Éc. polytech. Math.* **8** (2021), 279–310 Zbl 1470.37051 MR 4218159

[12] G. Casati and T. Prosen, Mixing property of triangular billiards. *Phys. Rev. Lett.* **83** (1999), no. 23, 4729–4732

[13] J. Chaika and A. Eskin, Every flat surface is Birkhoff and Oseledets generic in almost every direction. *J. Mod. Dyn.* **9** (2015), no. 1, 1–23 Zbl 1358.37008 MR 3395258

[14] J. Chaika and G. Forni, Weakly mixing polygonal billiards. [v1] 2020, [v2] 2025, arXiv: 2003.00890v2, to appear in *Ann. of Math. (2)*

[15] J. Chaika and H. Masur, The set of non-uniquely ergodic d-IETs has Hausdorff codimension 1/2. *Invent. Math.* **222** (2020), no. 3, 749–832 Zbl 1462.37026 MR 4169051

[16] A. Eskin and M. Mirzakhani, Invariant and stationary measures for the $\mathrm{SL}(2,\mathbb{R})$ action on moduli space. *Publ. Math. Inst. Hautes Études Sci.* **127** (2018), 95–324 Zbl 1478.37002 MR 3814652

[17] A. Eskin, M. Mirzakhani, and A. Mohammadi, Isolation, equidistribution, and orbit closures for the $\mathrm{SL}(2,\mathbb{R})$ action on moduli space. *Ann. of Math. (2)* **182** (2015), no. 2, 673–721 Zbl 1357.37040 MR 3418528

[18] B. Fayad and A. Katok, Constructions in elliptic dynamics. *Ergodic Theory Dynam. Systems* **24** (2004), no. 5, 1477–1520 Zbl 1089.37012 MR 2104594

[19] S. Filip, Semisimplicity and rigidity of the Kontsevich–Zorich cocycle. *Invent. Math.* **205** (2016), no. 3, 617–670 Zbl 1368.14013 MR 3539923

[20] S. Filip, Zero Lyapunov exponents and monodromy of the Kontsevich–Zorich cocycle. *Duke Math. J.* **166** (2017), no. 4, 657–706 Zbl 1370.37066 MR 3619303

[21] G. Forni, Deviation of ergodic averages for area-preserving flows on surfaces of higher genus. *Ann. of Math. (2)* **155** (2002), no. 1, 1–103 Zbl 1034.37003 MR 1888794

[22] G. Forni, On the Lyapunov exponents of the Kontsevich–Zorich cocycle. In B. Hasselblatt and A. Katok (eds.), *Handbook of dynamical systems*, Vol. 1B, pp. 549–580, Elsevier, Amsterdam, 2006 Zbl 1130.37302 MR 2186248

[23] G. Forni, A geometric criterion for the nonuniform hyperbolicity of the Kontsevich–Zorich cocycle. *J. Mod. Dyn.* **5** (2011), no. 2, 355–395 Zbl 1238.30031 MR 2820565

[24] G. Forni, Twisted translation flows and effective weak mixing. *J. Eur. Math. Soc. (JEMS)* **24** (2022), no. 12, 4225–4276 Zbl 1505.37009 MR 4493624

[25] G. Forni, Effective unique ergodicity and weak mixing of translation flows. In C. Bonanno, A. Sorrentino, and C. Ulcigrai (eds.), *Modern aspects of dynamical systems*, pp. 161–221, Lecture Notes in Math. 2347 – CIME Found. Subser., Springer, Cham, 2024 Zbl 08107407 MR 4823777

[26] R. H. Fox and R. B. Kershner, Concerning the transitive properties of geodesics on a rational polyhedron. *Duke Math. J.* **2** (1936), no. 1, 147–150 Zbl 0014.03401 MR 1545913

[27] E. Gutkin, Billiards in polygons. *Phys. D* **19** (1986), no. 3, 311–333 Zbl 0593.58016 MR 0844706

[28] E. Gutkin and A. Katok, Weakly mixing billiards. In X. Gómez-Mont, J. A. Seade, and A. Verjovsky (eds.), *Holomorphic dynamics (Mexico, 1986)*, pp. 163–176, Lecture Notes in Math. 1345, Springer, Berlin, 1988 Zbl 0673.58028 MR 0980958

[29] B. Hasselblatt and A. Katok, Principal structures. In B. Hasselblatt and A. Katok (eds.), *Handbook of dynamical systems*, Vol. 1A, pp. 1–203, North-Holland, Amsterdam, 2002 Zbl 1047.37001 MR 1928518

[30] Z. Hwang, C. A. Marx, J. J. Seaward, S. Jitomirskaya, and M. Olshanii, Growth of the Wang–Casati–Prosen counter in an integrable billiard. *SciPost Phys.* **14** (2023), no. 2, article no. 017, 16 pp. Zbl 1551.37043 MR 4546215

[31] A. Katok, Interval exchange transformations and some special flows are not mixing. *Israel J. Math.* **35** (1980), no. 4, 301–310 Zbl 0437.28009 MR 0594335

[32] A. Katok, The growth rate for the number of singular and periodic orbits for a polygonal billiard. *Comm. Math. Phys.* **111** (1987), no. 1, 151–160 Zbl 0631.58020 MR 0896765

[33] A. Katok, Five most resistant problems in dynamics. Clay Mathematics Institute and MSRI Conference on Recent Progress in Dynamics, 2004, http://akatok.s3-website-us-east-1.amazonaws.com/pub/5problems-expanded.pdf visited on 1 February 2026

[34] A. B. Katok, Billiard table as a playground for a mathematician. In V. Prasolov and Y. Ilyashenko (eds.), *Surveys in modern mathematics*, pp. 216–242, London Math. Soc. Lecture Note Ser. 321, Cambridge University Press, Cambridge, 2005 Zbl 1147.37329 MR 2166929

[35] S. Kerckhoff, H. Masur, and J. Smillie, Ergodicity of billiard flows and quadratic differentials. *Ann. of Math. (2)* **124** (1986), no. 2, 293–311 Zbl 0637.58010 MR 0855297

[36] M. Kontsevich and A. Zorich, Lyapunov exponents and Hodge theory. In J. M. Drouffe and J. B. Zuber (eds.), *Proceedings of the conference on the mathematical beauty of physics (Saclay, 1996)*, pp. 318–332, Adv. Ser. Math. Phys. 24, World Scientific, River Edge, NJ, 1997 Zbl 1058.37508 MR 1490861

[37] E. Lindenstrauss, A. Mohammadi, and Z. Wang, Polynomial effective equidistribution. *C. R. Math. Acad. Sci. Paris* **361** (2023), 507–520 Zbl 1528.37006 MR 4548163

[38] H. Masur, Interval exchange transformations and measured foliations. *Ann. of Math. (2)* **115** (1982), no. 1, 169–200 Zbl 0497.28012 MR 0644018

[39] H. Masur and S. Tabachnikov, Rational billiards and flat structures. In B. Hasselblatt and A. Katok (eds.), *Handbook of dynamical systems*, Vol. 1A, pp. 1015–1089, North-Holland, Amsterdam, 2002 Zbl 1057.37034 MR 1928530

[40] J. Moser, *Stable and Random Motions in Dynamical Systems: With Special Emphasis on Celestial Mechanics (AM-77)*. REV-revised, Princeton Landmarks Math. Phys., Princeton University Press, Princeton, 1973, 216 pp. Zbl 0271.70009 MR 0442980

[41] G. Rauzy, Échanges d'intervalles et transformations induites. *Acta Arith.* **34** (1979), no. 4, 315–328 Zbl 0414.28018 MR 0543205

[42] P. J. Richens and M. V. Berry, Pseudointegrable systems in classical and quantum mechanics. *Phys. D* **2** (1981), no. 3, 495–512 Zbl 1194.37150 MR 0625449

[43] W. A. Veech, Gauss measures for transformations on the space of interval exchange maps. *Ann. of Math. (2)* **115** (1982), no. 2, 201–242 Zbl 0486.28014 MR 0644019

[44] W. A. Veech, The metric theory of interval exchange transformations. I. Generic spectral properties. *Amer. J. Math.* **106** (1984), no. 6, 1331–1359 Zbl 0631.28006 MR 0765582

[45] Ya. B. Vorobets, Ergodicity of billiards in polygons (in Russian). *Mat. Sb.* **188** (1997), no. 3, 65–112. English translation. *Sb. Math.* **188** (1997), no. 3, 389–434 Zbl 0886.58065 MR 1462024

[46] J. Wang, G. Casati, and T. Prosen, Nonergodicity and localization of invariant measure for two colliding masses. *Phys. Rev. E* **89** (2014), no. 4, article no. 042918, 4 pp.

[47] A. Wright, The field of definition of affine invariant submanifolds of the moduli space of abelian differentials. *Geom. Topol.* **18** (2014), no. 3, 1323–1341 Zbl 1320.32019 MR 3254934

[48] A. N. Zemlyakov and A. B. Katok, Topological transitivity of billiards in polygons (in Russian). *Mat. Zametki* **18** (1975), no. 2, 291–300 Zbl 0315.58014. English translation. *Math. Notes* **18** (1975), no. 2, 760–764 Zbl 0323.58012 MR 0399423

[49] A. Zorich, Finite Gauss measure on the space of interval exchange transformations. Lyapunov exponents. *Ann. Inst. Fourier (Grenoble)* **46** (1996), no. 2, 325–370 Zbl 0853.28007 MR 1393518

Giovanni Forni
Department of Mathematics, University of Maryland at College Park, 4176 Campus Drive - William E. Kirwan Hall, College Park, MD 20742-4015, USA; gforni@umd.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/18

Partition regularity of homogeneous quadratics: Current trends and challenges

Nikos Frantzikinakis

Abstract. Suppose we partition the integers into finitely many cells. Can we always find a solution of the equation $x^2 + y^2 = z^2$ with x, y, z on the same cell? What about more general homogeneous quadratic equations in three variables? These are basic questions in arithmetic Ramsey theory, which have recently been partially answered using ideas inspired by ergodic theory and tools such as Gowers uniformity properties and concentration estimates of bounded multiplicative functions. The aim of this article is to provide an introduction to this exciting research area, explaining the main ideas behind the recent progress and some of the important challenges that lie ahead.

1. Introduction

A general principle in arithmetic Ramsey theory is that if you finitely partition $\mathbb{N} := \{1, 2, \dots\}$, then a cell of the partition will have a lot of structure. By structure, we mean solutions to algebraic equations or systems of algebraic equations. This is nicely illustrated by a theorem of Schur [53] and van der Waerden [59] and its extension by Rado [51], which covers systems of linear equations. In the case of a single linear equation, it takes a particularly nice form that we will explain next.

Definition 1.1. We say that the equation $P(x, y, z) = 0$ is *partition regular* if for every finite partition of $\mathbb{N}$ there exist distinct x, y, z on the same cell that satisfy the equation.

Rado's theorem states that for $a, b, c \in \mathbb{N}$, the equation

$$ax + by = cz$$

Mathematics Subject Classification 2020: 05D10 (primary); 11N37, 11B30, 37A44 (secondary).
Keywords: partition regularity, Pythagorean triples, multiplicative functions, concentration inequalities, Gowers uniformity.

is partition regular if and only if either a, b, or $a + b$ equals c; we call such a triple of positive integers (a, b, c) a *Rado triple*. To prove the necessity of the conditions, take a prime $p > a + b + c$ and partition the positive integers into $p - 1$ cells according to the last non-zero digit in their base p expansion. Such colorings arise from level sets of some modified Dirichlet characters (see Section 5.2) and will be our main source of counterexamples in this article.

Higher-degree polynomial equations have proved much harder to tackle. Perhaps the most well-known problem in this area is that of Erdős and Graham [26,27], which asks whether Pythagorean triples are partition regular, i.e., whether the equation

$$x^2 + y^2 = z^2 \tag{1.1}$$

is partition regular. Graham in [26] places the origin of the problem in the late 1970s and offered $250 for its solution, noting that "there is actually very little data (in either direction) to know which way to guess". While this may have been true a decade ago, there have been some positive developments in recent years. The case where one allows only partitions of $\mathbb{N}$ into two sets was verified in 2016 with the help of a computer search [39]; this endeavor was hailed as the "longest mathematical proof" at the time, consuming 200 terabytes of data [45]. Also, partition regularity of Pythagorean triples for level sets of finite-valued multiplicative functions was recently established [23].

The seminal work of Furstenberg [24] and Sárközy [52], culminating in the influential polynomial Szemerédi theorem of Bergelson and Leibman [7], allows us to deal with shift-invariant systems of polynomial equations. Regarding equations that resemble (1.1), here are some examples that are known to be or not to be partition regular:

(i) $x^2 + y = z$ is partition regular (Bergelson [4]).

(ii) $x + y = z^2$ is not partition regular (Csikvári–Gyarmati–Sárközy [14]).

(iii) $x^2 + y^2 = z$ is not partition regular (variant of Green–Lindqvist [36, Section 2]).

(iv) $x^2 + y = z^2$ is partition regular (Moreira [48]).

The second equation is partition regular if we consider partitions with only two cells [36, 49] or if we only require x, y to belong to the same cell [41]. There are a variety of tools used to prove these results, such as elementary combinatorial techniques, Fourier analysis, topological dynamics, and ergodic theory. Other partition regularity results of similar flavor can be found in [2,3,6,16–18,46].

However, in the fully non-linear setting, the only known results involve equations in more than three variables. For example, a result of Chow–Lindqvist–Prendiville [13] establishes that the equation

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = x_5^2$$

is partition regular (see also [9–11, 50] for related results).

Here, we focus on homogeneous quadratic equations in three variables, for which the methodology developed to deal with previous results has so far proved inadequate. A problem that drives much recent research in this area is the following.

Conjecture 1. *If $a, b, c \in \mathbb{N}$, then the equation*

$$ax^2 + by^2 = cz^2 \tag{1.2}$$

is partition regular if and only if (a, b, c) is a Rado triple.

The necessity of the conditions follows by combining Rado's theorem [51] with the fact that partition regularity of the equation $P(x^2, y^2, z^2) = 0$ implies partition regularity of the equation $P(x, y, z) = 0$. Regarding sufficiency, unfortunately there is no instance of a Rado triple for which the partition regularity of (1.2) has been proved. Three typical cases to keep in mind are the equations

$$x^2 + y^2 = z^2, \quad x^2 + 2y^2 = z^2, \quad x^2 + y^2 = 2z^2,$$

which, as we will see, have particular difficulties that are representative of the general case. The last equation was conjectured to be partition regular by Gyarmati–Ruzsa [37] and may even be density regular with respect to the natural density, i.e., it may have non-trivial solutions on any subset of the integers with positive upper density.

Although we are not in a position to address the question of partition regularity for triples, we will describe a systematic method, based on recent work by the author, Klurman, and Moreira [22, 23] and older work by the author and Host [21], which allows us to address the problem for pairs in several cases.

Definition 1.2. Given non-zero $a, b, c \in \mathbb{Z}$, we say that equation (1.2) is *partition regular with respect to x, y* if for every finite coloring of $\mathbb{N}$ there exist distinct $x, y \in \mathbb{N}$, with the same color, and $z \in \mathbb{N}$ that satisfy (1.2). Similarly, we define the partition regularity of (1.2) with respect to the variables x, z and y, z.

One of our main results in [23] is that Pythagorean pairs are partition regular, which answers a question that appeared in [18, 19, 21, 31].

Theorem 1.3. *The Pythagorean equation* (1.1) *is partition regular with respect to all pairs of variables.*

Our starting point is to link our problem with one concerning the asymptotic properties of bounded multiplicative functions via a representation result of Bochner–Herglotz; the details of this link are explained in Section 3.3. Let us just say that multiplicative functions play the role of characters in multiplicative Fourier analysis, which is an alternative to additive Fourier analysis that is worth exploring for the Pythagorean equation, or any other equation that is dilation invariant but not translation invariant. As we explain in Section 7.3 below, there is no known analogous "Bochner–Herglotz representation result" in the context of the partition regularity of Pythagorean triples, and an approach based on ergodic theory may be better suited to this problem. This is similar to the situation with shift-invariant patterns, where patterns of length two can be handled by a variety of tools, including Fourier analysis, but ergodic theory tools are better suited for patterns of length three and above, especially when these patterns are sparse.

In [22], we extended our methodology to cover partition regularity with respect to the variables x, y of more general equations of the form $ax^2 + by^2 = cz^2$, where a, b, c are non-zero integers such that either ac or bc is a square. We will discuss why equations like $x^2 \pm y^2 = 2z^2$ are harder to handle. In all of these problems, our approach depends on an in-depth study of the asymptotic behavior of the averages

$$\frac{1}{N^2} \sum_{1\leq m,n\leq N} f(P_1(m,n)) \cdot \overline{f(P_2(m,n))},$$

where f is an arbitrary completely multiplicative function taking values on the unit circle, and P_1, P_2 are binary quadratic forms used to parametrize the solutions of (1.2). We use Gowers uniformity properties to deal with "minor arc" multiplicative functions, which is the aperiodic case discussed in Section 4, and we use concentration estimates to deal with "major arc" multiplicative functions, which is the pretentious case discussed in Section 5.

Before these works, partition regularity of pairs was proved by the author and Host in [21], which covered, for example, partition regularity of the equation $16x^2 + 9y^2 = z^2$ with respect to x, y, but missed the case of Pythagorean pairs for reasons we will explain later. Extending these ideas, Sun in [54, 55] established partition regularity with respect to x, y for the equation $x^2 - y^2 = z^2$ when $\mathbb{N}$ is replaced by the ring of integers of a larger number field, such as the Gaussian integers. However, the methods used there do not apply to $\mathbb{N}$.

One may also pursue similar partition regularity questions on the rationals, but as it turns out, in the case of homogeneous equations (or dilation-invariant patterns), partition regularity over $\mathbb{Q}$ and $\mathbb{Z}$ is an equivalent problem [18, Proposition 2.13].

This is not the case for non-homogeneous patterns, and there are several interesting patterns that are known to be partition regular over the rationals, but it is not known whether they are partition regular over the integers; see, for example, some recent results in [1, 8].

Notation. We let $\mathbb{N} := \{1, 2, \ldots\}$, $\mathbb{Z}_+ := \{0, 1, 2, \ldots\}$, $\mathbb{R}_+ := [0, +\infty)$, $\mathbb{S}^1 \subset \mathbb{C}$ be the unit circle, and $\mathbb{U}$ be the closed complex unit disk. With $\mathbb{P}$ we denote the set of primes, and throughout we use the letter p to denote prime numbers.

For $t \in \mathbb{R}$, we let $e(t) := e^{2\pi i t}$. For $z \in \mathbb{C}$, with $\Re(z)$ and $\mathrm{Im}(z)$ we denote the real and imaginary parts of z, respectively, and we also let $\exp(z) := e^z$.

For $N \in \mathbb{N}$, we let $[N] := \{1, \ldots, N\}$. We often denote sequences $a\colon \mathbb{N} \to \mathbb{U}$ by $(a(n))$ instead of $(a(n))_{n\in\mathbb{N}}$.

If A is a finite non-empty subset of the integers and $a\colon A \to \mathbb{C}$, we let

$$\mathbb{E}_{n\in A}\, a(n) := \frac{1}{|A|}\sum_{n\in A} a(n).$$

We write $a(n) \ll b(n)$ if for some $C > 0$ we have $a(n) \le C\, b(n)$ for every $n \in \mathbb{N}$.

Throughout this article, the letter f is typically used for multiplicative functions and the letter χ for Dirichlet characters. With $\mathcal{M}$ we denote the set of completely multiplicative functions with modulus 1.

2. Partition regularity of generalized Pythagorean pairs

We state here the main results concerning the partition regularity of pairs for homogeneous quadratic equations and also make a crucial reduction to a density regularity statement involving parametrizations of the solution sets of the respective pairs.

2.1. Main results

A variant of Conjecture 1 that covers partition regularity of pairs for (1.2) is the following one.

Conjecture 2. *Suppose $a, b, c \in \mathbb{Z}$ are non-zero and at least one of the integers $ac, bc, (a+b)c$ is a square, then (1.2) is partition regular with respect to x, y.*

If (1.2) is partition regular with respect to all three variables, then it is obviously partition regular for any pair of variables. However, the converse is not true: [23, Theorem 1.1] implies that the equation $x^2 + y^2 = 4z^2$ is partition regular with respect to any pair of variables, but this equation is not partition regular with respect to all three variables since $(1, 1, 4)$ is not a Rado triple.

In [22], we prove the following result, which partially answers Conjecture 2.

Theorem 2.1. *If $a, b, c \in \mathbb{Z}$ are non-zero and ac or bc is a square, then (1.2) is partition regular with respect to x, y.*

Taking $a = b = c = 1$ and $a = -b = c = 1$, we get that the equation $x^2 + y^2 = z^2$ is partition regular with respect to all pairs of variables x, y, z. Taking $a = c = 1, b = 2$, and $a = c = 1, b = -2$, we get that the equation $x^2 + 2y^2 = z^2$ is partition regular with respect to the pairs x, y and y, z. We will see later why the pair x, z is harder to handle for this equation, and why a similar problem arises with the equation $x^2 + y^2 = 2z^2$, this time with respect to all pairs of variables. In this regard, we can only give a conditional result, assuming a difficult Chowla–Elliott-type conjecture (see Section 6.5.2) for a class of "sufficiently aperiodic" completely multiplicative functions.

Theorem 2.2. *Suppose that Conjecture 3 below holds. If $a, b, c \in \mathbb{Z}$ are non-zero and $(a + b)c$ is a square (perhaps zero), then (1.2) is partition regular with respect to x, y.*

Although the conditionality of this result is an unpleasant feature, we can cover, unconditionally, partitions generated by pretentious multiplicative functions (see [22, Theorem 1.4] for the precise statement), which are the only known sources of counterexamples to the partition regularity of pairs of homogeneous quadratic equations.

An immediate consequence of Theorems 2.1 and 2.2 is that conditionally on the truth of Conjecture 3 in Section 6.5.2, we have that if (a, b, c) is a Rado triple, then equation (1.2) is partition regular with respect to all pairs of variables x, y, z. We get this result unconditionally if $a = c$ or $b = c$.

2.2. Parametric reformulation

To prove our main results, we will work with the parametric form of the solutions of equation (1.2). We give three representative examples of such parametrizations (below, k, m, n are arbitrary integers):

(i) $x^2 + y^2 = z^2$: $x = k\,(m^2 - n^2)$, $y = k\,(2mn)$, $z = k\,(m^2 + n^2)$.

(ii) $x^2 + 2y^2 = z^2$: $x = k\,(m^2 - 2n^2)$, $y = k\,(2mn)$, $z = k\,(m^2 + 2n^2)$.

(iii) $x^2 + y^2 = 2z^2$: $x = k\,(m^2 - n^2 + 2mn)$, $y = k\,(m^2 - n^2 - 2mn)$, $z = k\,(m^2 + n^2)$.

To give the reader a sense of how certain features of these parametrizations affect our analysis, we make a few remarks that will be properly justified when we go deeper into the proof outline of our main results.

In the case of (i), in order to prove partition regularity with respect to x, y, it suffices to show that for every finite partition of $\mathbb{N}$, there exist $k, m, n \in \mathbb{N}$ such that the integers

$$k(m^2 - n^2) \quad \text{and} \quad k(2mn)$$

are distinct and belong to the same partition cell. The fact that both polynomials $m^2 - n^2$ and $2mn$ are products of linear forms simplifies the argument substantially. On the other hand, the parameter z uses the irreducible quadratic form $m^2 + n^2$, which takes values on a zero density subset of $\mathbb{N}$, and this makes the corresponding analysis of partition regularity with respect to the variables x, z (or y, z) much more complicated. On the positive side, the polynomial $m^2 + n^2$ is a norm form (higher-degree irreducible polynomials that do not have this property are out of reach at the moment), and the corresponding ring of integers one needs to work with in part of our analysis is the Gaussian integers, which is a principal ideal domain with finitely many units. All these things reduce the complexity of our argument.

In the case of (ii), handling the parameters y, z is similar to the case of (i). However, handling the parameters x, y is much harder, since the ring of integers associated with the polynomial $m^2 - 2n^2$ is $\mathbb{Z}[\sqrt{2}]$, which has infinitely many units; this adds another layer of problems to overcome. Finally, the variables x, z depend on two irreducible quadratic forms, which is the most difficult case to handle, and we can offer only conditional results in this case. The same problem appears in equation (iii); in this case, all pairs depend on two irreducible quadratic forms, so we can only prove conditional results in these cases.

2.3. Density regularity

It will be more convenient for us to prove stronger versions of Theorems 2.1 and 2.2, which deal with density regularity. Roughly, the stronger statements will state that every set with positive multiplicative density contains the pairs we are interested in, written in parametric form.

We recall some standard notions. A *multiplicative Følner sequence in* $\mathbb{N}$ is a sequence $\Phi = (\Phi_K)_{K=1}^{\infty}$ of finite subsets of $\mathbb{N}$ that is asymptotically invariant under dilation, in the sense that

$$\lim_{K\to\infty} \frac{|\Phi_K \cap (x \cdot \Phi_K)|}{|\Phi_K|} = 1 \quad \text{for all } x \in \mathbb{N}.$$

An example of a multiplicative Følner sequence is given by

$$\Phi_K := \left\{ \prod_{p \le K} p^{a_p} : 0 \le a_p \le K \right\}, \quad K \in \mathbb{N}.$$

The *upper multiplicative density* of a set $\Lambda \subset \mathbb{N}$ with respect to a multiplicative Følner sequence $\Phi = (\Phi_K)_{K=1}^{\infty}$ is the quantity

$$\bar{d}_{\Phi}(\Lambda) := \limsup_{K\to\infty} \frac{|\Phi_K \cap \Lambda|}{|\Phi_K|}.$$

For example, for every $r \in \mathbb{N}$, the set $r\mathbb{Z} + j$ has multiplicative density 1 if $j = 0$ and multiplicative density 0 if $j = 1, \dots, r-1$, so a set with positive multiplicative density should contain many elements that are highly divisible. The set of squares has multiplicative density 0, while the set of integers $n \in \mathbb{N}$ such that $2^{2k} \mid n$ but $2^{2k+1} \nmid n$ for some $k \in \mathbb{Z}_+$ has multiplicative density $1/2$.

Definition 2.3. We say that the pair $P_1, P_2 \in \mathbb{Z}[m,n]$ is *good for density regularity* if whenever $\Lambda \subset \mathbb{N}$ satisfies $\bar{d}_{\Phi}(\Lambda) > 0$ for some multiplicative Følner sequence Φ, there exist $m, n \in \mathbb{N}$ such that $P_1(m,n)$ and $P_2(m,n)$ are distinct positive integers and

$$\bar{d}_{\Phi}((P_1(m,n))^{-1}\Lambda \cap (P_2(m,n))^{-1}\Lambda) > 0, \tag{2.1}$$

where $r^{-1}\Lambda := \{n \in \mathbb{N} : rn \in \Lambda\}$.

If the pair $P_1, P_2 \in \mathbb{Z}[m,n]$ is good for density regularity, then for any $\Lambda \subset \mathbb{N}$ that satisfies $\bar{d}_{\Phi}(\Lambda) > 0$ for some multiplicative Følner sequence Φ, the intersection $(P_1(m,n))^{-1}\Lambda \cap (P_2(m,n))^{-1}\Lambda$ has positive multiplicative density and in particular is non-empty for some $m, n \in \mathbb{N}$. Taking any $k \in (P_1(m,n))^{-1}\Lambda \cap (P_2(m,n))^{-1}\Lambda$, it follows that $kP_1(m,n)$, $kP_2(m,n)$ are distinct and they both belong to Λ.

By considering appropriate parametrizations of solutions of (1.2) when ac or bc is a square, it turns out that in order to prove Theorem 2.1 it suffices to establish the following result.

Theorem 2.4. *Suppose that for some $\alpha, \gamma \in \mathbb{N}$ and non-zero $\beta \in \mathbb{Z}$, we have*

$$P_1(m,n) = \alpha m^2 + \beta n^2, \quad P_2(m,n) = \gamma mn.$$

Then the pair of polynomials P_1, P_2 is good for density regularity.

More generally, our argument shows density regularity if at least one of the two polynomials is reducible and not a square, and the two polynomials are not multiples of each other.

On the other hand, appropriate parametrizations of solutions of (1.2) when $(a+b)c$ is a square show that Theorem 2.2 is an immediate consequence of the following result.

Theorem 2.5. *Suppose that for some $\alpha, \beta, \alpha', \beta' \in \mathbb{Z}$, the polynomials*

$$P_1(m,n) = m^2 + \alpha mn + \beta n^2, \quad P_2(m,n) = m^2 + \alpha' mn + \beta' n^2$$

are distinct and irreducible, and Conjecture 3 holds for P_1, P_2 (see Section 6.5.2). Then the pair of polynomials P_1, P_2 is good for density regularity.

More generally, our argument shows density regularity if P_1, P_2 are irreducible quadratic forms, not multiples of each other, such that either $P_1(1,0) = P_2(1,0)$ or $P_1(0,1) = P_2(0,1)$. If neither of the last two conditions on the coefficients of m^2 and n^2 is met, then the corresponding pair of polynomials is often not partition regular.

We note that our arguments also give that the positivity property (2.1) occurs quite often; in fact, it holds for a set of $m, n \in \mathbb{N}$ with positive lower density (taken over the squares $[N] \times [N]$).

2.4. Other results

Our methodology is quite flexible and can be applied to other partition regularity problems of pairs of dilation-invariant patterns and related results. Here is a list of some additional examples:

(i) (More general homogeneous quadratic equations) The methods of this article can be used to prove partition and density regularity of more general equations $P(x, y, z) = 0$, where P is a homogeneous quadratic form. For example, take $a, b \in \mathbb{N}$, $d \in \mathbb{Z}$, and consider the equation

$$ax^2 + by^2 + dxy = az^2. \tag{2.2}$$

Direct computation shows that (2.2) is satisfied when

$$x = k(bm^2 - an^2), \quad y = km(-dm + 2an), \quad z = k(bm^2 - dmn + an^2),$$

where $k, m, n \in \mathbb{Z}$. Using the remark following Theorem 2.4, we get that (2.2) is density regular with respect to x, y and y, z. We also get density regularity with respect to x, z if the polynomial $bm^2 - an^2$ is reducible, i.e., if ab is a square. The last assumption can be removed if we assume that Conjecture 3 in Section 6.5.2 holds; to see this, use the remark following Theorem 2.5.

(ii) (Pythagorean triples on level sets of multiplicative functions) In [23, Theorem 1.7] we showed that for any completely multiplicative function $f: \mathbb{N} \to \mathbb{S}^1$ taking finitely many values, there exist $x, y, z \in \mathbb{N}$ such that

$$x^2 + y^2 = z^2 \quad \text{and} \quad f(x) = f(y) = f(z) = 1.$$

(iii) (A question from [18]) Answering a question of Donoso–Le–Moreira–Sun from [18], we showed in [23, Theorem 1.8] (our argument also uses the main result from [56]) that the patterns $(km^2, k(n^2+n))$ are density regular.

(iv) (An extension of a result of Klurman–Mangerel) Recently, Charamaras–Mountakis–Tsinas [12] used a variant of the Q-trick, which we explain in Section 6, to show, among other things, that for every $a \in \mathbb{N}$ and every pair of completely multiplicative functions $f, g: \mathbb{N} \to \mathbb{S}^1$, we have

$$\liminf_{n\to\infty} |f(an+1) - g(an)| = 0, \tag{2.3}$$

and we cannot replace 1 by 2 even when $a = 1$. When $f = g$, equation (2.3) was previously established by Klurman–Mangerel [43, Theorem 1.1] for $a = 1$ and for general $a \in \mathbb{N}$ by Donoso–Le–Moreira–Sun [18, Corollary 1.7].

3. A model problem and connection with multiplicative functions

3.1. An additive model problem

Our approach to solving the density regularity problems of Theorems 2.4 and 2.5 tries to model Furstenberg's ergodic-theoretic proof of Sárközy's theorem. Let us briefly describe this argument. The goal is to show that if Λ is a positive density subset of the integers, i.e.,

$$\overline{d}(\Lambda) := \limsup_{N\to\infty} \frac{|\Lambda \cap [N]|}{N} > 0,$$

then there exists $n \in \mathbb{N}$ such that

$$\overline{d}(\Lambda \cap (\Lambda - n^2)) > 0. \tag{3.1}$$

Using a correspondence principle of Furstenberg [24], it suffices to show that for every "additive" measure-preserving system (X, μ, T), i.e., probability space $(X, \mathcal{X}, \mu)$, measure-preserving transformation $T: X \to X$, and measurable set A with $\mu(A) > 0$, there exists $n \in \mathbb{N}$ such that $\mu(A \cap T^{-n^2}A) > 0$. In fact, we show the stronger property

$$\liminf_{N\to\infty} \mathbb{E}_{n\in[N]}\, \mu(A \cap T^{-n^2}A) > 0. \tag{3.2}$$

To prove (3.2), we have two ways to proceed. One is to first show that the rational Kronecker factor is characteristic (by appealing to the Hilbert space version of van der Corput's lemma) and then to deal with the rational Kronecker factor of the system using the good divisibility properties of n^2. Unfortunately, this approach cannot be modeled very well in our multiplicative setting, since our patterns use a

mixture of addition and multiplication which does not allow cancellation when using a multiplicative version of van der Corput's lemma.

The second way to prove (3.2) is to use a representation theorem for sequences of the form $n \mapsto \mu(A \cap T^{-n}A)$ or more generally positive density sequences in $\mathbb{Z}$. It follows from Herglotz's theorem that there exists a (positive) bounded Borel measure σ on $[0, 1)$ (depending on A) such that

$$\mu(A \cap T^{-n}A) = \int e(nt)\, d\sigma(t) \tag{3.3}$$

for every $n \in \mathbb{Z}$. Averaging over $n \in [N]$, taking the limit as $N \to \infty$, and using the ergodic theorem to lower bound the left side and the bounded convergence theorem to compute the right side, we deduce that $\sigma(\{0\}) \geq (\mu(A))^2 > 0$.

It follows that in order to establish (3.2), it suffices to show that if σ is a bounded (positive) measure on $[0, 1)$ such that $\sigma(\{0\}) = \delta > 0$ and $\widehat{\sigma}(n) \geq 0$ for every $n \in \mathbb{N}$, then

$$\liminf_{N\to\infty} \mathbb{E}_{n\in[N]} \int e(n^2 t)\, d\sigma(t) > 0. \tag{3.4}$$

The advantage of this new setting is that exponential sums along squares can be estimated efficiently, and this helps us analyze the left-hand side in (3.4). Indeed, it is a well-known consequence of Weyl's theorem that if t is irrational, we have

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, e(n^2 t) = 0. \tag{3.5}$$

Using the bounded convergence theorem, this essentially allows us to reduce matters to handling the "rational" part of the measure σ, which can be done using the divisibility properties of n^2 as follows: For $A \subset [0, 1)$, let

$$\sigma_A := \sigma|_A$$

and choose $Q \in \mathbb{N}$ so that

$$\sigma_{\mathbb{Q}}([0, 1) \setminus A_Q) \leq \frac{\delta^2}{2}, \quad \text{where } A_Q := \left\{0, \frac{1}{Q}, \ldots, \frac{(Q-1)}{Q}\right\}. \tag{3.6}$$

Using the positivity property $\widehat{\sigma}(n) \geq 0, n \in \mathbb{N}$, we get that (3.4) would follow from

$$\liminf_{N\to\infty} \mathbb{E}_{n\in[N]} \int e(Q^2 n^2 t)\, d\sigma(t) > 0; \tag{3.7}$$

indeed, the limit in (3.4) is at least $1/Q$ times the limit in (3.7). Using (3.5) and the bounded convergence theorem, we get that it suffices to show that

$$\liminf_{N\to\infty} \mathbb{E}_{n\in[N]} \int e(Q^2 n^2 t)\, d\sigma_{\mathbb{Q}}(t) > 0. \tag{3.8}$$

Since $Q^2t^2 = 0 \pmod 1$ for every $t \in A_Q$, we have

$$\int e(Q^2n^2t)\, d\sigma_{A_Q}(t) = \int 1\, d\sigma_{A_Q}(t) = \sigma(A_Q) \geq \sigma(\{0\}) = \delta^2 \quad \text{for every } n \in \mathbb{N}. \tag{3.9}$$

Combining (3.6) and (3.9), it follows that the limit in (3.8) is at least $\delta^2/2$, and the proof is complete.

We also note that the use of ergodic theory in this argument is rather superficial, since it was only employed to reduce the density positivity property (3.1) to the analytic positivity property (3.4). This step could also be done directly by showing that the sequence $n \mapsto d^*(\Lambda \cap (\Lambda - n))$ is positive definite, where d^* is a density that agrees with $\overline{d}$ on Λ and is defined along a subsequence of the integers so that all relevant limits exist.

3.2. Bird's eye view of the modified strategy

We plan to use the general principles of the strategy explained in the previous subsection to establish the density statements of Theorems 2.4 and 2.5, with some necessary modifications that we will briefly summarize now. In the following sections, we will give a much more detailed account of how to implement these modifications, and it will become clear why the exact form of the homogeneous polynomials we have to work with greatly affects the difficulty of the problem.

(i) (Representation result) We can replace the "additive" measure-preserving system (X, μ, T) by a measure-preserving system (X, μ, T_n) that has multiplicative structure, but we opt for a more direct approach and represent directly some multiplicative densities that are naturally linked to our problem. The fact that these densities are dilation invariant and not translation invariant leads to a representation result analogous to (3.3), where the place of the linear exponential sequences $n \mapsto e(nt)$ is taken by sequences with multiplicative structure, i.e., sequences $f\colon \mathbb{N} \to \mathbb{S}^1$ that satisfy $f(mn) = f(m) \cdot f(n)$ for every $m, n \in \mathbb{N}$. See (3.10) for the exact identity.

(ii) (Reduced analytic statement) Using the previous representation result, we reduce the proof of Theorems 2.4 and 2.5 to showing a positivity property analogous to (3.4), which involves averages of integrals of multiplicative functions evaluated over homogeneous polynomials in two variables. See Theorems 3.2 and 3.3 below for the exact statements. It is the case, however, that we have a much larger variety of multiplicative functions than we have linear exponential sequences, and studying their asymptotic behavior turns out to be a much more delicate and interesting task. Similar to the additive case, we consider two complementary cases, the "structured" and the

"random" cases, to handle the "major arcs" and "minor arcs" of the spectral measure.

(iii) (Minor arcs) In the additive setting, we used Weyl's theorem to show that for irrational numbers, we have the vanishing property (3.5). In the multiplicative setting, the place of irrational numbers is taken by the "aperiodic" multiplicative functions, for which we prove, using a rather involved argument, several vanishing statements about their correlations; see Section 4 for details.

(iv) (Major arcs) In the additive setting, we were able to handle the part of the spectral measure σ that was supported on rational numbers by passing to a subprogression and then using identity (3.9). In the multiplicative setting, the place of rational numbers is taken by the "pretentious" multiplicative functions for which we are able to prove concentration estimates as a substitute for identity (3.9). The exact statements appear in Section 5.

(v) (The Q-trick) In the additive setting, it was easy to combine the information about the major and minor arcs to simplify expression (3.7) to a point where the positivity property is obvious. In the multiplicative setting, the simpler expression we arrive at has no obvious positivity property, so it will be much less clear how to conclude. A key manipulation here is what we call the Q-trick, which allows us to obtain a property analogous to (3.8) for some highly divisible Q. We explain in Section 6 how this key maneuver is used to complete the proof.

3.3. From partition regularity to multiplicative functions

The starting point for the proofs of Theorems 2.4 and 2.5 is to reformulate them as positivity properties involving integrals of averages of completely multiplicative functions. As we also mentioned before, this key step can only be performed for partition and density regularity results of pairs and has no useful analog we know of for triples (though see Section 7.3 for an ergodic reformulation).

3.3.1. Multiplicative functions. A function $f\colon \mathbb{N} \to \mathbb{U}$, where $\mathbb{U}$ is the complex unit disk, is called *multiplicative* if

$$f(mn) = f(m) \cdot f(n) \quad \text{whenever } (m,n) = 1.$$

If necessary, we extend f to an even function in $\mathbb{Z}$ by letting $f(0) := 0$ and $f(-n) := f(n)$ for $n \in \mathbb{N}$. It is called *completely multiplicative* if the previous identity holds for all $m, n \in \mathbb{N}$. We let

$$\mathcal{M} := \{f\colon \mathbb{N} \to \mathbb{S}^1 \text{ is a completely multiplicative function}\},$$

where $\mathbb{S}^1$ is the unit circle. Throughout, we assume that $\mathcal{M}$ is equipped with the topology of pointwise convergence; then $\mathcal{M}$ is a metrizable compact space with this topology.

3.3.2. Reduction to a statement about multiplicative functions. We will use a representation result of Bochner–Herglotz for positive definite functions on locally compact Abelian groups. We need it only for the discrete group $\mathbb{Q}_+$ with multiplication, in which case we can identify $\mathcal{M}$ with the Pontryagin dual of $(\mathbb{Q}_+, \times)$. We say that $A\colon \mathbb{Q}_+ \to \mathbb{C}$ is positive definite on $(\mathbb{Q}_+, \times)$ if for every $n \in \mathbb{N}$, all $r_1, \dots, r_n \in \mathbb{Q}^+$, and all $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ we have

$$\sum_{i,j=1}^{n} \lambda_i \overline{\lambda_j}\, A(r_i\, r_j^{-1}) \geq 0.$$

Theorem 3.1 (Bochner–Herglotz). *Let $A\colon \mathbb{Q}_+ \to \mathbb{C}$ be positive definite on $(\mathbb{Q}_+, \times)$. Then there exists a finite Borel measure σ on $\mathcal{M}$ such that*

$$A\Big(\frac{r}{s}\Big) = \int_{\mathcal{M}} f(r) \cdot \overline{f(s)}\, d\sigma(f) \quad \textit{for all } r, s \in \mathbb{N}.$$

To use this representation result in our setting, we can argue as follows (an alternative way is via ergodic theory). Given a Følner sequence $\Phi = (\Phi_N)$ of subsets of $\mathbb{N}$, we can pass to a subsequence $\Phi' = (\Phi'_N)$ along which $d_{\Phi'}(\Lambda) = d_{\Phi}(\Lambda)$, and the limits

$$d_{\Phi'}((r^{-1}\Lambda) \cap (s^{-1}\Lambda)) := \lim_{N\to\infty} \frac{|(r^{-1}\Lambda) \cap (s^{-1}\Lambda) \cap \Phi'_N|}{|\Phi'_N|}$$

exist for all $r, s \in \mathbb{Z}$. Then the sequence $r/s \mapsto d_{\Phi'}((r^{-1}\Lambda) \cap (s^{-1}\Lambda))$ from $(\mathbb{Q}_+, \times)$ to $[0, 1]$ is well defined (we use that $d_{\Phi'}$ is invariant under dilations), and a direct computation shows that it is positive definite. Using Theorem 3.1, we get that there exists a finite Borel measure σ on $\mathcal{M}$ such that

$$d_{\Phi'}((r^{-1}\Lambda) \cap (s^{-1}\Lambda)) = \int_{\mathcal{M}} f(r) \cdot \overline{f(s)}\, d\sigma(f) \quad \text{for all } r, s \in \mathbb{N}. \tag{3.10}$$

Furthermore, it is not hard to show (see [21, Section 10.2]) that

$$\sigma(\{1\}) \geq (d_{\Phi'}(\Lambda))^2.$$

Using the previous two properties, it is easy to see that in order to prove Theorem 2.4, it suffices to prove the following result.

Theorem 3.2. *Let σ be a bounded Borel measure on $\mathcal{M}$ such that $\sigma(\{1\}) > 0$ and*

$$\int_{\mathcal{M}} f(r) \cdot \overline{f(s)}\, d\sigma(f) \geq 0 \quad \text{for every } r, s \in \mathbb{N}. \tag{3.11}$$

Let also $\alpha, \gamma \in \mathbb{N}$ and non-zero $\beta \in \mathbb{Z}$. Then

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]} \int_{\mathcal{M}} f(\alpha m^2 + \beta n^2) \cdot \overline{f(\gamma mn)}\, d\sigma(f) > 0. \tag{3.12}$$

In order to prove Theorem 2.5, it suffices to establish the following result.

Theorem 3.3. *Let σ be as in Theorem 3.2 and $\alpha, \beta, \alpha', \beta' \in \mathbb{Z}$ be such that*

$$P_1(m,n) = m^2 + \alpha mn + \beta n^2, \quad P_2(m,n) = m^2 + \alpha' mn + \beta' n^2$$

are distinct and irreducible quadratic forms. Suppose that either Conjecture 3 in Section 6.5.2 holds for P_1, P_2 or the measure σ is supported on the class of pretentious multiplicative functions (see Section 5.1). Then

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]} \int_{\mathcal{M}} f(P_1(m,n)) \cdot \overline{f(P_2(m,n))}\, d\sigma(f) > 0. \tag{3.13}$$

Both results were established in [22], with special cases previously obtained in [21, 23].

One may wonder about our choice of averaging over $m, n \in \mathbb{N}$ in (3.12) and (3.13). Why not take iterated limits $\liminf_{N\to\infty} \mathbb{E}_{n\in[N]}$, $\liminf_{M\to\infty} \mathbb{E}_{m\in[M]}$ or perhaps take multiplicative averages over m, n? Taking iterated limits would lead to much harder problems for "minor arc" multiplicative functions. Although in some cases the required vanishing property can be obtained for logarithmic averages using the work of Tao–Teräväinen [57], the exceptional class of "minor arc" multiplicative functions for which this method is inapplicable would render this argument not very useful. Finally, taking multiplicative averages over the parameters m, n leads to problems that we are not able to handle, both for "major arc" and "minor arc" multiplicative functions.

3.4. Proof strategy

In order to establish the necessary positivity properties of Theorems 3.2 and 3.3, we use the theory of completely multiplicative functions and proceed in two independent steps, each using a distinctly different methodology.

(i) (Minor arcs) We first study the part of σ that is supported on aperiodic multiplicative functions. The key step here is to establish Gowers uniformity properties of arbitrary aperiodic multiplicative functions and then use

them show that the averages in (3.12) vanish (this last step is trickier when $P(m,n) := \alpha m^2 + \beta n^2$ is irreducible). To prove the uniformity properties, we combine the inverse theorem of Green–Tao-Ziegler [35] with some quite delicate equidistribution results of variable nilsequences. Our approach is described in more detail in Section 4.

(ii) (Major arcs) Next, we study the complementary case, covering the part of σ that is supported on pretentious multiplicative functions (see Definition 5.1). In this case, we can establish concentration estimates, a feature of pretentious multiplicative functions that is not shared by the aperiodic ones. The details are given in Section 5. We use these concentration estimates to simplify substantially the averages in (3.12) when the range of m, n is restricted to a suitable positive density subset of $\mathbb{N} \times \mathbb{N}$ depending on some parameter Q. We then show that a suitable multiplicative average, over the parameter Q, of these simplified expressions is zero for all but an exceptional set of multiplicative functions and it is positive for the rest. We call this the Q-trick, and it easily implies the required positivity. The details are given in Section 6.

4. Aperiodic multiplicative functions: Vanishing of correlations

4.1. Definition and characterization

The first class of multiplicative functions that will play a crucial role in our arguments is the aperiodic ones.

Definition 4.1. The multiplicative function $f: \mathbb{N} \to \mathbb{U}$ is *aperiodic* if for every $a, b \in \mathbb{N}$ we have

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, f(an+b) = 0. \tag{4.1}$$

Sometimes in the literature, the term *non-pretentious* is used instead of the term aperiodic. It can be shown that $f: \mathbb{N} \to \mathbb{U}$ is aperiodic if and only if

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, f(n)\cdot\chi(n) = 0 \quad \text{for every Dirichlet character } \chi. \tag{4.2}$$

Examples of multiplicative functions that are aperiodic are the Liouville and the Möbius function; this follows from a variant of the prime number theorem on arithmetic progressions.

A rather amazing and very useful fact for our purposes is that correlations of aperiodic multiplicative functions have very strong vanishing properties, which we will describe in the following subsections.

4.2. Gowers uniformity and inverse theorem

We will define some seminorms on $\ell^\infty(\mathbb{N})$ that will be crucial for our study. Roughly, they are given by the limit as $N \to \infty$ of the Gowers norms of the sequence, restricted to the interval $[N]$.

Definition 4.2 (Gowers norms on $\mathbb{Z}_N$ [25]). Let $N \in \mathbb{N}$ and $a: \mathbb{Z}_N \to \mathbb{C}$. For $s \in \mathbb{N}$, the *Gowers $U^s(\mathbb{Z}_N)$-norm* $\|a\|_{U^s(\mathbb{Z}_N)}$ is defined inductively as follows: For every $h \in \mathbb{Z}_N$, we write $a_h(n) := a(n+h)$. We let

$$\|a\|_{U^1(\mathbb{Z}_N)} := |\mathbb{E}_{n\in\mathbb{Z}_N} a(n)|,$$

and for every $s \geq 1$, we let

$$\|a\|_{U^{s+1}(\mathbb{Z}_N)} := (\mathbb{E}_{h\in\mathbb{Z}_N} \|a \cdot \overline{a}_h\|_{U^s(\mathbb{Z}_N)}^{2^s})^{\frac{1}{2^{s+1}}}.$$

For example,

$$\|a\|_{U^2(\mathbb{Z}_N)}^4 = \mathbb{E}_{n,h_1,h_2\in\mathbb{Z}_N} a(n)\,\overline{a(n+h_1)}\,\overline{a(n+h_2)}\,a(n+h_1+h_2),$$

and for the $U^s(\mathbb{Z}_N)$ norms for $s \geq 3$, a similar but more complicated closed formula can be given.

Definition 4.3 (Gowers seminorms on $\mathbb{N}$). If $a: \mathbb{N} \to \mathbb{C}$ is a bounded sequence, we let

$$\|a\|_{U^s(\mathbb{N})} = \limsup_{N\to\infty} \|a_N\|_{U^s(\mathbb{Z}_N)},$$

where $a_N: \mathbb{Z}_N \to \mathbb{C}$ is the periodic extension of $a \cdot \mathbf{1}_{[N]}$ to $\mathbb{Z}_N$.

It can be shown that $\|\cdot\|_{U^s(\mathbb{N})}$ satisfies the triangle inequality, thus defining a seminorm on $\ell^\infty(\mathbb{N})$, and for every $s \in \mathbb{N}$, we have

$$\|a\|_{U^s(\mathbb{N})} \leq \|a\|_{U^{s+1}(\mathbb{N})}.$$

We say that the bounded sequence $a: \mathbb{N} \to \mathbb{C}$ is U^s-uniform if

$$\|a\|_{U^s(\mathbb{N})} = 0.$$

For example, if α is an irrational number and $s \in \mathbb{N}$, it can be shown that the sequence $(e(n^s\alpha))$ is U^s-uniform but not U^{s+1}-uniform, while if α is a rational non-integer, it is U^1-uniform but not U^2-uniform. It can also be shown that if c is a non-integer positive real number, then the sequence $(e(n^c))$ is U^s-uniform for all $s \in \mathbb{N}$.

Using Fourier inversion and the Parseval identity in $\mathbb{Z}_N$, for $N \in \mathbb{N}$, it can be shown that if $a\colon \mathbb{N} \to \mathbb{U}$ is a sequence, then

$$\|a\|_{U^2(\mathbb{N})} \le \limsup_{N\to\infty} \max_{\xi\in\mathbb{Z}_N} |\widehat{a_N}(\xi)|^{\frac{1}{2}},$$

where the Fourier transform is taken in $\mathbb{Z}_N$. It is easy to deduce from this the equivalence

$$\|a\|_{U^2(\mathbb{N})} = 0 \iff \lim_{N\to\infty} \sup_{\alpha\in\mathbb{R}} |\mathbb{E}_{n\in[N]}\, a(n)\cdot e(n\alpha)| = 0. \tag{4.3}$$

In order to give a generalization of this result that covers U^s-uniformity for $s \ge 3$, we have to first account for additional obstructions to uniformity. For example, when $s = 3$, it can be seen that

$$\|a\|_{U^3(\mathbb{N})} > 0 \quad \text{when } a(n) = e(n^2\alpha + n\beta) \text{ or } a(n) = e([n\alpha]n\beta), \tag{4.4}$$

where $\alpha, \beta \in \mathbb{R}$. These turn out to be examples of nilsequences, which are defined as follows: If X is an s-step nilmanifold, i.e., $X = G/\Gamma$, where G is an s-step nilpotent Lie group and Γ is a discrete cocompact subgroup, we let

$$a(n) = F(g^n\cdot e_X), \quad n \in \mathbb{N},$$

where $g \in G$, $e_X := \Gamma$, and $F \in C(X)$. Moreover, to cover the second example in (4.4), we must allow a function F that is not continuous but is Riemann integrable in the orbit closure of g. It is an amazing fact that this class of sequences suffices to give a generalization of (4.3) that covers U^s-uniformity. This is a consequence of the inverse theorem of Green–Tao–Ziegler [35], first proved in [32] for $s = 3$, which is the only case we will need for applications to partition regularity. We state it in the following convenient form.

Theorem 4.4. *Let $a\colon \mathbb{N} \to \mathbb{U}$ be a sequence and $s \in \mathbb{N}$. Then the following two properties are equivalent:*

(i) $\|a\|_{U^{s+1}(\mathbb{N})} = 0$.

(ii) *For every s-step nilmanifold $X = G/\Gamma$ and $F \in C(X)$, we have*

$$\lim_{N\to\infty} \sup_{g\in G} |\mathbb{E}_{n\in[N]}\, a(n)\cdot F(g^n\cdot e_X)| = 0.$$

Moreover, in checking the second condition, one can assume that F is smooth, and using what is known as the "vertical Fourier decomposition", one can further assume that F is a non-trivial vertical nilcharacter, i.e., assuming that G is s-step

nilpotent but not $(s-1)$-step nilpotent, we have that there exists $\chi\colon X_s \to \mathbb{S}^1$, a non-trivial character of the compact Abelian group $X_s := G_s/(G_s \cap \Gamma)$, such that

$$F(ux) = \chi(u) \cdot F(x)$$

for every $x \in X$ and $u \in G_s$. It easily follows from this that $\int F\, dm_X = 0$. We note that in the case where $X = \mathbb{T}$, non-trivial vertical nilcharacters are functions of the form $F(x) = e(kx)$ for some non-zero $k \in \mathbb{Z}$.

We will see in the next subsection how one can use this result to establish the U^s-uniformity of aperiodic multiplicative functions.

4.3. Gowers uniformity of aperiodic multiplicative functions

A very strong and perhaps a priori unexpected property is that aperiodic multiplicative functions are Gowers uniform of all orders.

To verify that they are U^2-uniform, we will use the following orthogonality criterion due to Daboussi–Kátai [15,40].

Theorem 4.5. *Let $a\colon \mathbb{N} \to \mathbb{U}$ be a sequence such that*

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, a(pn) \cdot \overline{a(qn)} = 0 \tag{4.5}$$

for all distinct primes p, q. Then for every multiplicative function $f\colon \mathbb{N} \to \mathbb{U}$, we have

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, f(n) \cdot a(n) = 0.$$

Applying Theorem 4.5 for $a(n) := e(n\alpha)$, it follows that

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, f(n) \cdot e(n\alpha) = 0 \tag{4.6}$$

for every irrational α. Since aperiodic multiplicative functions, by definition, do not correlate with $e(n\alpha)$ for α rational, it follows that for those multiplicative functions, (4.6) holds for every $\alpha \in \mathbb{R}$. It is only slightly more difficult to show that a similar property holds even if we allow the phase α to depend on N [21, Lemma 9.1], namely

$$\lim_{N\to\infty} \sup_{\alpha\in\mathbb{R}} |\mathbb{E}_{n\in[N]}\, f(n) \cdot e(n\alpha)| = 0. \tag{4.7}$$

A direct consequence of this identity and the characterization of U^2-uniformity (see (4.3)) is that

$$\|f\|_{U^2(\mathbb{N})} = 0.$$

It is also true, but much harder to prove, that a similar property holds for higher-order uniformity. This was first done for the Liouville and the Möbius functions for $s = 3$

by Green–Tao [33] and then by Green–Tao–Ziegler [34, 35] for general $s \in \mathbb{N}$. Their approach used a Type I and Type II sums analysis to verify property (ii) of Theorem 4.4, and this required explicit quantitative control for averages of multiplicative functions along arithmetic progressions, not shared by general aperiodic multiplicative functions (and the exceptional class is too large to be ignored in our applications). Using a somewhat different argument, the previous results were extended in a joint paper by the author and Host in [21, Theorem 2.5] to general aperiodic multiplicative functions (see also Matthiesen [47] for an alternative approach to a related result).

Theorem 4.6. *Let $f : \mathbb{N} \to \mathbb{U}$ be an aperiodic multiplicative function. Then for every $s \in \mathbb{N}$, we have*

$$\|f\|_{U^s(\mathbb{N})} = 0.$$

Let us briefly sketch some key ideas of the proof of this result. We start by using Theorem 4.4. It reduces the problem to establishing a strengthening of (4.7) where the place of $e(n\alpha)$ is taken by nilsequences. Namely, we need to show that if $X = G/\Gamma$ is a fixed s-step nilmanifold and $F \in C(X)$ is a non-trivial vertical nilcharacter, then

$$\lim_{N\to\infty} \sup_{b\in G} |\mathbb{E}_{n\in[N]}\, f(n) \cdot F(b^n \cdot e_X)| = 0. \tag{4.8}$$

Verifying this property is a rather laborious task, so let us just explain how it works in the much simpler setting where we do not take the supremum over b, the nilmanifold X is connected, and the sequence $(b^n \cdot e_X)$ is equidistributed on X. So in this case, we use the above-mentioned orthogonality criterion of Daboussi–Kátai for $a(n) := F(b^n \cdot e_X)$. We get that it suffices to verify the vanishing property (4.5) for all distinct primes p, q or, equivalently, that

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, F(b^{pn} \cdot e_X) \cdot \overline{F(b^{qn} \cdot e_X)} = 0. \tag{4.9}$$

It can be shown that the sequence

$$((b^{pn} \cdot e_X, b^{qn} \cdot e_X))_{n\in\mathbb{N}}$$

is equidistributed in a connected subnilmanifold $Y = H/\Delta$ of $X \times X$. Equation (4.9) would thus follow if we showed that

$$\int F(x) \cdot \overline{F(x')}\, dm_Y(x, x') = 0. \tag{4.10}$$

This is not a simple task, since the structure of the possible nilmanifolds Y depends on b in a way that is not easy to determine explicitly. The key property that allows us

to prove (4.10) is that no matter how complicated Y may be, it always satisfies the following:

$$\{(g^p, g^q)\colon g \in G\} \subset H \cdot (G_2 \times G_2).$$

This follows by using that the sequences $(b^{pn} \cdot e_X)$ and $(b^{qn} \cdot e_X)$ are both equidistributed in X. By taking iterated commutators of elements on the left, we deduce the following key property:

$$(u^{p^s}, u^{q^s}) \in H \quad \text{for every } u \in G_s.$$

This invariance property easily implies that $(x, x') \mapsto F(x) \cdot \overline{F(x')}$ is a vertical nilcharacter of Y with non-zero frequency, so (4.10) holds.

If the place of the sequence b^n takes a more general polynomial sequence, say $a^n b^{n^2}$ for some $a, b \in G$, we first establish that there exist normal subgroups G^1, G^2 of G such that $G^1 \cdot G^2 = G$ and

$$\{(g_1^p g_2^{p^2}, g_1^q g_2^{q^2})\colon g_1 \in G^1,\ g_2 \in G^2\} \subset H \cdot (G_2 \times G_2),$$

and by taking iterated commutators of elements on the left, we deduce that

$$\text{the set } \{u \in G_s\colon (u^{p^j}, u^{q^j}) \in H \text{ for some } j \in \mathbb{N}\} \text{ generates } G_s. \tag{4.11}$$

From this, we can again prove (4.10).

The actual argument is substantially more complicated because in order to verify (4.8), we have to prove a variant of (4.9) involving variable polynomial sequences, in which case the required invariance property, analogous to (4.11), is much harder to identify and prove. However, the previous summary accurately summarizes the skeleton of the proof of Theorem 4.6.

4.4. Vanishing of correlations

An important component in the proof of Theorem 3.2, which covers the case of Pythagorean pairs $(m^2 \pm n^2, 2mn)$, is that for aperiodic multiplicative functions, several correlations vanish, a result obtained in [21, Theorem 2.6].

Theorem 4.7. *Let $f\colon \mathbb{N} \to \mathbb{U}$ be an aperiodic multiplicative function. Then*

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f(m^2 - n^2) \cdot \overline{f(mn)} = 0. \tag{4.12}$$

To prove this, one first uses some standard maneuvering to bound the left side by a multiple of $\|f\|_{U^3(\mathbb{Z})}$, which vanishes by Theorem 4.6. A similar argument leads to the following more general result: If $f_1\colon \mathbb{N} \to \mathbb{U}$ is an aperiodic multiplicative function and $f_2, \dots, f_\ell\colon \mathbb{N} \to \mathbb{U}$ are arbitrary sequences, then

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f_1(L_1(m,n)) \cdots f_\ell(L_\ell(m,n)) = 0$$

for all linear forms $L_1, \dots, L_\ell \colon \mathbb{N}^2 \to \mathbb{N}$ such that L_1 is not a multiple of $L_2, \dots, L_\ell$.

To handle Pythagorean pairs of the form $(2mn, m^2+n^2)$, we need the following result.

Theorem 4.8. *Let $f \colon \mathbb{N} \to \mathbb{U}$ be an aperiodic completely multiplicative function. Then*

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f(mn)\cdot \overline{f(m^2+n^2)} = 0. \tag{4.13}$$

To prove this, we first represent the sequence $(m,n) \mapsto f(m^2+n^2)$ as a multiplicative function on the Gaussian integers. To do this, we let $g \colon \mathbb{Z}[i] \to \mathbb{U}$ by $g(z) := \overline{f(|z|^2)}$, and note that for $z = m+in$, we have $g(z) = \overline{f(m^2+n^2)}$. Then (4.13) takes the following form:

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f(\Re(z))\cdot f(\mathrm{Im}(z))\cdot g(z) = 0.$$

Using a variant of Theorem 4.5 for bounded sequences on the Gaussian integers, it turns out that the last identity follows if we show that

$$\lim_{N\to\infty} \mathbb{E}_{z\in R_N}\, f(\Re(pz))\cdot f(\mathrm{Im}(pz))\cdot \overline{f(\Re(qz))}\cdot \overline{f(\mathrm{Im}(qz))} = 0$$

holds for all but finitely many distinct non-real Gaussian primes p, q, where

$$R_N := \{z = m+in \colon m,n \in [N]\}.$$

Using some elementary reasoning on the ring $\mathbb{Z}[i]$, it is not hard to deduce that this follows from

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f(L_1(m,n))\cdot f(L_2(m,n))\cdot \overline{f(L_3(m,n))}\cdot \overline{f(L_4(m,n))} = 0,$$

where f is aperiodic and $L_1, L_2, L_3, L_4 \colon \mathbb{N}^2 \to \mathbb{N}$ are pairwise independent linear forms (recall that we extend f to the negative integers as an even function if needed). Using standard arguments, this follows from the U^3-uniformity of f, i.e., the $s=3$ case of Theorem 4.6.

4.5. More general binary quadratic forms

The results of the previous subsection can be extended to a more general setting, in which one considers the correlations of an aperiodic completely multiplicative function involving several linear forms and at most one irreducible quadratic form. A good representative example is the identity

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f(mn)\cdot \overline{f(m^2+dn^2)} = 0,$$

where $d \in \mathbb{Z}$ is not a square of an integer. The case in which the expression $m^2 + dn^2$ is replaced by an arbitrary irreducible binary quadratic form $P(m, n)$ is only marginally more complicated.

Suppose first that $d > 0$. If $d \equiv 1, 2 \pmod 4$, then $m^2 + dn^2$ is a norm form on the ring of integers $\mathbb{Z}[\sqrt{-d}]$, which is a unique factorization domain. The argument employed in the case of $d = 1$ can be applied without significant alterations to this case, even in the context of a non-principal ideal domain. If $d \equiv 3 \pmod 4$, then $\mathbb{Z}[\sqrt{-d}]$ is not a unique factorization domain, and this causes problems for our argument. When, say $d = 3$, this issue can be circumvented by working with the binary form $m^2 + mn + n^2$ instead of the form $m^2 + 3n^2$. The advantage now is that this is a norm form on the ring $\mathbb{Z}[(1 + \sqrt{-3})/2]$, which is a unique factorization domain. The identity

$$m^2 + 3n^2 = (m - n)^2 + (m - n)(2n) + (2n)^2$$

and a "change of variables" argument are then employed to transfer results from one binary form to the other.

The case $d < 0$ presents a more substantial challenge. Let us assume that $d = -2$ and explain what the issue is in this case. An element $z \in \mathbb{Z}[\sqrt{2}]$ is said to be *C-regular* if, for all $N \in \mathbb{N}$, the following condition is satisfied:

$$z^{-1} R_N \subset R_{C|\mathcal{N}(z)|^{-\frac{1}{2}} N},$$

where $R_N := \{m + n\sqrt{2} : m, n \in [N]\}$. In order to prove a useful variant of Theorem 4.5, we would like to know that for some fixed $C > 0$, all elements of the ring $\mathbb{Z}[\sqrt{2}]$ are C-regular. Unfortunately, this property does not hold, for the reason that $\mathbb{Z}[\sqrt{2}]$ has infinitely many units, so every element has infinitely many associates, and not all of them have good regularity properties. What allows us to overcome this obstacle is a result proved by Sun in [55], namely that for a fixed $C > 0$, every element has a C-regular associate. This enables us to prove a variant of Theorem 4.5 for $\mathbb{Z}[\sqrt{2}]$ and more general quadratic integer rings, which is applicable for our purposes.

By combining the previous ideas, we prove the following result in [22, Theorem 3.1].

Theorem 4.9. *Let $s \in \mathbb{N}$ and $f_1, \dots, f_s, g : \mathbb{N} \to \mathbb{S}^1$ be completely multiplicative functions that are extended to even sequences in $\mathbb{Z}$ and suppose that f_1 is aperiodic. Let also $L_1, \dots, L_s$ be linear forms in two variables with integer coefficients and suppose that either $s = 1$ and L_1 is non-trivial or $s \geq 2$ and the forms L_1, L_j are linearly independent for $j = 2, \dots, s$. Finally, suppose that the form*

$$P(m, n) := \alpha m^2 + \beta mn + \gamma n^2,$$

where $\alpha, \beta, \gamma \in \mathbb{Z}$, is irreducible. Then

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]} \prod_{j=1}^{s} f_j(L_j(m,n)) \cdot g(P(m,n)) = 0.$$

For the partition regularity applications we have in mind, we only use the case $s = 2$.

5. Pretentious multiplicative functions: Concentration estimates

Now that we have at our disposal the vanishing property of the correlations of aperiodic multiplicative functions given by Theorem 4.9, we turn our attention to the study of those that are not aperiodic. In this case, we prove an approximate periodicity property that is analytically expressed in the form of the concentration estimates given in Theorems 5.5, 5.6, and 5.8. In the non-aperiodic case, this will allow us to substantially simplify the averages that appear in Theorems 3.2 and 3.3.

5.1. Pretentious distance and characterization of aperiodicity

Our first goal is to give a useful characterization of non-aperiodic multiplicative functions. The following are two potential obstructions to aperiodicity:

(i) (Dirichlet characters) A *Dirichlet character* is a periodic completely multiplicative function. Then $\chi(Qn+1) = 1$ whenever Q is a multiple of a period of χ and (4.1) fails for $a = Q$ and $b = 1$.

(ii) (Archimedean characters) An *Archimedean character* is a completely multiplicative function of the form $n \mapsto n^{it}$, $n \in \mathbb{N}$, where $t \in \mathbb{N}$. Then $\mathbb{E}_{n\in[N]} n^{it}$ is asymptotically equal to $N^{it}/(1+it)$, and so (4.1) fails for $a = 1$ and $b = 0$.

It is a surprising and very useful fact that these are the only obstructions to aperiodicity. A completely multiplicative function that is not aperiodic is, in a sense that will be made precise in Theorem 5.2 below, close to a product of a Dirichlet character and an Archimedean character.

Following Granville and Soundararajan [28–30], we define the notion of pretentiousness and a related distance between multiplicative functions.

Definition 5.1. If $f, g\colon \mathbb{N} \to \mathbb{U}$ are multiplicative functions, we define their distance as

$$\mathbb{D}^2(f,g) := \sum_{p\in\mathbb{P}} \frac{1}{p}(1 - \Re(f(p)\cdot\overline{g(p)})).$$

(Note that for $f, g: \mathbb{N} \to \mathbb{S}^1$, we have $\mathbb{D}^2(f, g) = 1/2 \sum_{p \in \mathbb{P}} 1/p |f(p) - g(p)|^2$.)

It can be shown (see [29] or [30, Section 2.1.1]) that $\mathbb{D}$ satisfies the triangle inequality

$$\mathbb{D}(f, g) \leq \mathbb{D}(f, h) + \mathbb{D}(h, g)$$

for all $f, g, h: \mathbb{N} \to \mathbb{U}$. Also, for all $f_1, f_2, g_1, g_2: \mathbb{N} \to \mathbb{U}$, we have (see [28, Lemma 3.1])

$$\mathbb{D}(f_1 f_2, g_1 g_2) \leq \mathbb{D}(f_1, g_1) + \mathbb{D}(f_2, g_2). \tag{5.1}$$

We say that f *pretends to be* g and write $f \sim g$ if $\mathbb{D}(f, g) < +\infty$. It follows from (5.1) that if $f_1 \sim g_1$ and $f_2 \sim g_2$, then $f_1 f_2 \sim g_1 g_2$.

The next result is a direct consequence of the characterization of aperiodicity given by (4.2) and the Wirsing–Halász mean value theorem [38], which gives necessary and sufficient conditions for a multiplicative function to have mean value 0.

Theorem 5.2. *A completely multiplicative function $f: \mathbb{N} \to \mathbb{U}$ is not aperiodic if and only if there exist $t \in \mathbb{R}$ and a Dirichlet character χ such that $f \sim \chi \cdot n^{it}$, i.e., $\mathbb{D}(f, \chi \cdot n^{it}) < +\infty$.*

This motivates the following definition.

Definition 5.3. We say that f is *pretentious* if $f \sim \chi \cdot n^{it}$ for some $t \in \mathbb{R}$ and Dirichlet character χ.

The value of t is uniquely determined; this follows from (5.1) and the fact that $n^{it} \not\sim \chi$ for every non-zero $t \in \mathbb{R}$ and Dirichlet character χ (see, for example, [30, Corollary 11.4] or [29, Proposition 7]).

It follows from Theorem 5.2 that whenever a completely multiplicative function $f: \mathbb{N} \to \mathbb{U}$ is not aperiodic, then it is pretentious.

5.2. Examples of pretentious multiplicative functions

We give several illustrative examples of pretentious multiplicative functions in order to cover all the possible varying behaviors that are relevant to our study. In the examples that follow, we describe the multiplicative function only on prime numbers and assume that it is extended to a completely multiplicative function on the positive integers.

(i) (Modified Dirichlet characters) If χ is a Dirichlet character, then $\chi(p) = 0$ if and only if p divides the minimal period of χ and $\chi(p)$ is a root of

unity otherwise. We let $\widetilde{\chi}\colon \mathbb{N} \to \mathbb{S}^1$ be the completely multiplicative function satisfying

$$\widetilde{\chi}(p) := \begin{cases} \chi(p) & \text{if } \chi(p) \neq 0, \\ 1 & \text{if } \chi(p) = 0. \end{cases}$$

Then $\widetilde{\chi}$ is not necessarily periodic and $\widetilde{\chi} \sim \chi$.

(ii) (Finite support) Let $f(p_1) = \cdots = f(p_\ell) := -1$ and $f(p) = 1$ for $p \notin \{p_1, \dots, p_\ell\}$. Then $f \sim 1$.

(iii) (Infinite support 1-pretentious) Let $f(p) := -1$ for $p \in \mathbb{P}_0$ and $f(p) = 1$ for $p \notin \mathbb{P}_0$, where $\mathbb{P}_0 \subset \mathbb{P}$ is infinite and $\sum_{p\in\mathbb{P}_0} 1/p < +\infty$. Then $f \sim 1$.

(iv) (1-Pretentious oscillatory) $f(p) := e(1/\log\log p)$, $p \in \mathbb{P}$. Then $f \sim 1$, but it turns out that f does not have a mean value, in fact, $\mathbb{E}_{n\in[N]}\, f(n)$ is asymptotically equal to $e(w(N))$, where $w(N) = c \log\log\log N$ for some $c > 0$.

(v) (Archimedean factor) $f(n) := n^{it} \cdot g(n)$ for some $t \in \mathbb{R}$ and g as in the previous examples. Then $f \sim \chi \cdot n^{it}$ if g is as in (i) and $f \sim n^{it}$ if g is as in (ii)–(iv).

We mention some features of the previous examples that will motivate the concentration results to be proved in the next subsections.

In examples (i) and (ii), there exists Q_0 such that if $Q \mid Q_0$, then

$$f(Qn + 1) = 1 \quad \text{for every } n \in \mathbb{N}.$$

Indeed, take Q_0 be the period of χ in the first case and $Q_0 := p_1 \cdots p_\ell$ in the second.

In example (iii), let $\varepsilon > 0$ and $K = K(\varepsilon)$ be such that $\sum_{p\in\mathbb{P}_0,\, p>K} 1/p < \varepsilon$. For $Q_0 := \prod_{p\le K} p$, we have that if $Q_0 \mid Q$, then $\overline{d}(n \in \mathbb{N}\colon f(Qn + 1) \neq 1) \le \varepsilon$, and hence

$$\limsup_{N\to\infty} \mathbb{E}_{n\in[N]} |f(Qn + 1) - 1| \ll \varepsilon.$$

In example (iv), let $\varepsilon > 0$ and $K := K(\varepsilon)$ be such that $\sum_{p\in\mathbb{P},\, p>K} 1/p(1 - \Re(f(p))) < \varepsilon$. For $Q_0 = \prod_{p\le K} p$, we have that if $Q_0 \mid Q$, then

$$\limsup_{N\to\infty} \mathbb{E}_{n\in[N]} |f(Qn + 1) - e(w(N))| \ll \varepsilon,$$

where $w(N)$ increases to infinity slowly, in fact, like $c \log\log\log N$ for some $c > 0$.

In example (v), we have

$$\lim_{N\to\infty} \mathbb{E}_{n\in[N]} |f(Qn + 1) - (Qn)^{it} \cdot g(Qn + 1)| = 0.$$

Combining this with the previous estimates for g, we get analogous results for f.

The general principle that can be derived from these examples is that if $f \sim \chi$, then for highly divisible values of Q, the values $f(Qn+1)$ concentrate around 1 or a slowly growing oscillatory factor. The next subsection makes this principle explicit. Later, we also give variants of it that apply to arbitrary binary quadratic forms in place of n.

5.3. Linear concentration estimates: Statements

To get a sense of the concentration estimates we will be using in a simpler setting, let us first consider the case where a pretentious multiplicative function takes on finitely many values on the primes.

Theorem 5.4. *Let $f: \mathbb{N} \to \mathbb{U}$ be a pretentious multiplicative function and suppose that the set $\{f(p): p \in \mathbb{P}\}$ is finite. Then for every $\varepsilon > 0$, there exists $Q_0 = Q_0(\varepsilon)$ such that if $Q \in \mathbb{N}$ is such that $Q_0 \mid Q$, then*

$$\limsup_{N\to\infty} \mathbb{E}_{n\in[N]} |f(Qn+1) - 1| \le \varepsilon. \tag{5.2}$$

Such concentration results are distinctive properties of pretentious multiplicative functions. Indeed, for every $Q \in \mathbb{N}$, an aperiodic multiplicative function has average 0 on the arithmetic progression $Q\mathbb{N}+1$, hence (5.2) fails if $\varepsilon < 1$.

We also note that a related property of "Besicovitch rational almost periodicity", obtained in [15, Theorem 6] for pretentious multiplicative functions $f: \mathbb{N} \to \mathbb{U}$ that take finitely many values on primes, states that for every $\varepsilon > 0$, there exists a periodic sequence a_ε such that

$$\limsup_{N\to\infty} \mathbb{E}_{n\in[N]} |f(n) - a_\varepsilon(n)| \le \varepsilon.$$

Unfortunately, for large values of Q, this estimate alone cannot give us useful information about the values of $f(Qn+1)$, and so it is not helpful for our applications.

For our purposes, we need an extension of Theorem 5.4 that covers all pretentious multiplicative functions and also allows for some uniformity in our choice of Q. To state it, we have to introduce some notation. For $K \in \mathbb{N}$, let

$$\Phi_K := \Big\{ \prod_{p\le K} p^{a_p} : K < a_p \le 2K \Big\}. \tag{5.3}$$

If $f \sim \chi \cdot n^{it}$ for some Dirichlet character χ and $t \in \mathbb{R}$ and $K, N \in \mathbb{N}$, we let

$$F_N(f, K) := \sum_{K<p\le N} \frac{1}{p} (f(p) \cdot \overline{\chi(p)} \cdot p^{-it} - 1). \tag{5.4}$$

Since $f \sim \chi \cdot n^{it}$, we have that the limit $\lim_{N\to\infty} \Re(F_N(F,1))$ exists. But there are cases where $\lim_{N\to\infty} |\mathrm{Im}(F_N(f,1))| = +\infty$; this holds in example (iv) of Section 5.2.

Theorem 5.5 ([44, Lemma 2.5]). *Let $f\colon \mathbb{N} \to \mathbb{U}$ be a multiplicative function and suppose that $f \sim \chi \cdot n^{it}$ for some Dirichlet character χ and $t \in \mathbb{R}$. Then*

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{n\in[N]} |f(Qn+1) - (Qn)^{it} \cdot \exp(F_N(f,K))| = 0, \quad (5.5)$$

where Φ_K and $F_N(f,K)$ are as in (5.3) and (5.4), respectively.

Note that in the case of multiplicative functions that take finitely many values on primes, we have $t = 0$, and the series defining $F_N(f,K)$ can be shown to converge, so in our statement we can replace the term $F_N(f,K)$ with 0. This gives a statement similar to the one in Theorem 5.4.

5.4. Linear concentration estimates: Proof

There are two main steps in the proof of Theorem 5.5. In the first step, we deal with additive functions on the complex numbers and prove the concentration estimate (5.8) under assumption (5.7). We describe the key steps of this argument; the reader will find the missing details, for example, in [58, Chapter III.3.2]. In the second step, given a pretentious multiplicative function f, we apply the previous results for the additive function h defined in (5.12) to derive the required concentration estimate for f.

Step 1 (Turán–Kubilius inequality). The function $h\colon \mathbb{N} \to \mathbb{C}$ is *additive* if it satisfies

$$h(mn) = h(m) + h(n) \quad \text{whenever } (m,n) = 1.$$

For simplicity, we also assume that

$$h(p^k) = 0 \quad \text{for every } p \in \mathbb{P},\ k \geq 2, \quad (5.6)$$

and remark that this simplifying assumption can be easily removed. The Turán–Kubilius inequality implies that if

$$\sum_{p\in\mathbb{P}} \frac{|h(p)|^2}{p} < \infty, \quad (5.7)$$

then

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{n\in[N]} |h(Qn+1) - H_N(h,K)|^2 = 0, \quad (5.8)$$

where Φ_K is as in (5.3), and

$$H_N(h,K) := \sum_{K<p\le N} \frac{h(p)}{p}.$$

For $Q \in \Phi_K$, $H_N(h,K)$ is approximately equal to $\mathbb{E}_{n\in[N]}h(Qn+1)$.

We mention here the important ingredients of the proof of (5.8). In the next subsections, we list the essential changes that need to be made to get analogous extensions for binary quadratic forms.

Let

$$w_{Q,N}(p,q) := \frac{1}{N} \sum_{n\in[N],\, p,q\,||\,Qn+1} 1,$$

where $p \,||\, Qn+1$ means that $p \mid Qn+1$ but $p^2 \nmid Qn+1$. To prove (5.8), we expand the square and compute it using the identity

$$h(Qn+1) = \sum_{p||Qn+1} h(p) = \sum_{p>K,\, p||Qn+1} h(p),$$

which follows from the additivity of h, our simplifying assumption $h(p^k)=0$ for $k\ge 2$, and the fact that if $Q\in\Phi_K$, we have $p\nmid Qn+1$ for all primes $p\le K$. We deduce that

$$\mathbb{E}_{n\in[N]}\, h(Qn+1) = \sum_{K<p\le N} w_{Q,N}(p,p)\cdot h(p) \tag{5.9}$$

and

$$\mathbb{E}_{n\in[N]}|h(Qn+1)|^2 = \sum_{K<p,q\le N} w_{Q,N}(p,q)\cdot h(p)\cdot \overline{h(q)}. \tag{5.10}$$

By direct computation, we get the approximate identities

$$w_{Q,N}(p,p) = \frac{1}{p}\Big(1-\frac{1}{p}\Big) + O\Big(\frac{1}{N}\Big)$$

and

$$w_{Q,N}(p,q) = \frac{1}{pq}\Big(1-\frac{1}{p}\Big)\Big(1-\frac{1}{q}\Big) + O\Big(\frac{1}{N}\Big),$$

which hold for all primes p,q with $p\ne q$. Importantly, for $p\ne q$, we have

$$w_{Q,N}(p,q) = w_{Q,N}(p,p)\cdot w_{Q,N}(q,q) + O\Big(\frac{1}{N}\Big). \tag{5.11}$$

(The fact that the error term in (5.11) does not depend on Q is responsible for the uniformity with respect to Q in (5.5).) Using (5.9), (5.10), (5.11), it is possible to show that (5.8) holds. To do this, we use that the terms of the form $\sum_{K<p\le N}|h(p)|^2/p$

have negligible contribution in (5.8) because of (5.7). Also, error terms of the form $O(1/N)$ contribute a total error $O\big(\sum_{p,q:pq\leq QN+1} 1/N\big) = O(Q\log\log N/\log N)$, which is also negligible for (5.8). This last fact is no longer true for polynomials in two variables and respective averages, because there are too many error terms, which causes significant problems in our treatment of analogous concentration estimates for sums of two squares and more general irreducible binary quadratic forms.

Step 2 (From additive to multiplicative). The second step is to pass the previous result from "pretentious" additive functions to pretentious multiplicative functions. To do this, we first reduce to the case where $f\sim 1$ and then define the additive function $h\colon \mathbb{N}\to\mathbb{C}$ on prime powers as follows:

$$h(p^k) := f(p^k) - 1, \quad p\in\mathbb{P},\ k\in\mathbb{N}. \tag{5.12}$$

We also make the simplifying assumption $f(p^k) = 1$ for $k\geq 2$, hence (5.6) holds. We can remove this assumption using that $f\sim 1$, which implies that the contribution of prime powers greater than 1 is negligible in (5.5).

Since $f\sim 1$, we get that (5.7) holds, and hence (5.8) holds from Step 1. Our goal is to relate the values of $f(Qn+1)$ with those of $h(Qn+1)$ and derive (5.5) from (5.8). We will use the approximate identity

$$z = e^{z-1} + O(|z-1|^2), \tag{5.13}$$

which holds for $|z|\leq 1$. Note that since $f\sim 1$, we have that $f(p)-1$ is close to 0 for "most" primes p, so for $z_p := f(p)$, we have that (5.13) gives a sensible approximation. We deduce that

$$f(Qn+1) = \prod_{p\mid\mid Qn+1,\, p>K} f(p) = \prod_{p\mid\mid Qn+1,\, p>K} (\exp(h(p)) + O(|h(p)|^2)),$$

where we used that for $Q\in\Phi_K$, only primes greater than K divide $Qn+1$. If we ignore the error terms on the right side (this can be done using (5.7)), we get that it is equal to

$$\prod_{p\mid\mid Qn+1} \exp(h(p)) = \exp(h(Qn+1)),$$

where we used (5.6). Summarizing, we have just shown that in establishing (5.5), we can replace $f(Qn+1)$ with $\exp(h(Qn+1))$. So to prove (5.5), it remains to show that (note that since $f\sim 1$, we have $t=0$)

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{n\in[N]} |\exp(h(Qn+1)) - \exp(F_N(f,K))| = 0.$$

Using the estimate $|e^z - e^w| \ll |z - w|$, which holds for $\Re(z), \Re(w) \le 0$, and the Cauchy–Schwarz inequality, it suffices to show that

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{n\in[N]} |h(Qn+1) - F_N(f,K)|^2 = 0.$$

Since $F_N(f,K) = H_N(h,K)$, this follows from (5.8).

5.5. Concentration estimates along sums of two squares

To prove that the pair of polynomials $2mn, m^2+n^2$ is density regular (which implies that $x^2+y^2=z^2$ is partition regular with respect to y,z), we need to verify Theorem 3.2 when $\alpha=\beta=1, \gamma=2$. To do this, we need a variant of Theorem 5.5 that deals with concentration estimates along sums of two squares. We give the precise statement next.

If $f \sim \chi \cdot n^{it}$ for some Dirichlet character χ and $t \in \mathbb{R}$, and $K, N \in \mathbb{N}$, we let

$$G_N(f,K) := \sum_{K<p\le N,\, p\equiv 1 \,(\mathrm{mod}\ 4)} \frac{1}{p}\,(f(p)\cdot\overline{\chi(p)}\cdot p^{-it} - 1). \tag{5.14}$$

This expression is similar to $F_N(f,K)$ in (5.4), but we omit primes that are congruent to 3 (mod 4) in the summation. This is due to the fact that no prime of this form is a sum of two squares, which ultimately makes the contribution of such primes negligible for our purposes.

Theorem 5.6 ([23, Proposition 2.11]). *Let $f:\mathbb{N}\to\mathbb{U}$ be a multiplicative function and suppose that $f \sim \chi\cdot n^{it}$ for some Dirichlet character χ and $t\in\mathbb{R}$. Then*

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{m,n\in[N]} \big| f((Qm+1)^2+(Qn)^2) - Q^{2it} \cdot (m^2+n^2)^{it} \cdot \exp(G_N(f,K))\big| = 0, \tag{5.15}$$

where Φ_K and $G_N(f,K)$ are as in (5.3) and (5.14), respectively.

If f is pretentious and takes finitely many values in the primes, then it can be shown that $f\sim\chi$ for some Dirichlet character χ, and the series defining $G_N(f,K)$ converges as $N\to\infty$, so the term $Q^{2it}\cdot(m^2+n^2)^{it}\cdot\exp(G_N(f,K))$ in (5.15) can be replaced by 1.

Our strategy is similar to the one used to prove the linear concentration estimates in Section 5.3. However, to carry out the first step, which is the Turán–Kubilius estimate for additive functions taken along sums of two squares, there are some additional important difficulties that we must resolve.

Our goal is to show that if

$$\sum_{p\in\mathbb{P}} \frac{|h(p)|^2}{p} < \infty, \tag{5.16}$$

and h_N is the restriction of h on primes less than N, i.e.,

$$h_N(p^k) := \begin{cases} h(p^k) & \text{if } p \le N, \\ 0 & \text{otherwise,} \end{cases}$$

then we have the following variant of the Turán–Kubilius inequality (5.8) that deals with sums of two squares

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{m,n\in[N]} |h_N((Qm+1)^2+(Qn)^2) - H_N(h,K)|^2 = 0, \tag{5.17}$$

where

$$H_N(h,K) := \sum_{K<p\le N,\ p\equiv 1 \pmod 4} \frac{h(p)}{p}.$$

We could replace h_N with h in (5.17), but because of some error terms that appear in the next step of the argument, such a result would not be more useful. For convenience, we also assume that $h(p) = 0$ for all primes that are congruent to 3 (mod 4) and $h(p^k) = 0$ for all primes p and $k \ge 2$; both conditions can be easily removed.

Following the plan used to prove the linear Turán–Kubilius inequality (5.8), we must compute

$$w_{Q,N}(p,q) := \frac{1}{N^2} \sum_{m,n\in[N],\ p,q\,|\,(Qm+1)^2+(Qn)^2} 1$$

whenever $p, q \equiv 1 \pmod 4$. Using that -1 is a quadratic residue (mod p) whenever $p \equiv 1 \pmod 4$ is a quadratic residue and applying the Chinese remainder theorem, we find that

$$w_{Q,N}(p,p) = \frac{2}{p}\Big(1-\frac{1}{p}\Big)^2 + O\Big(\frac{1}{N}\Big),$$

and

$$w_{Q,N}(p,q) = \frac{4}{pq}\Big(1-\frac{1}{p}\Big)^2\Big(1-\frac{1}{q}\Big)^2 + O\Big(\frac{1}{N}\Big), \tag{5.18}$$

which holds for all primes p, q with $p \ne q$ that are congruent to 1 (mod 4). Importantly, for primes $p \ne q$, congruent to 1 (mod 4), we again have the approximate identity

$$w_{Q,N}(p,q) = w_{Q,N}(p,p)\cdot w_{Q,N}(q,q) + O\Big(\frac{1}{N}\Big).$$

In contrast to the linear case, in our case, because we are averaging over two variables m, n, the errors incurred by terms of the form $O(1/N)$ are too many and cannot be easily neglected. To deal with this problem, we decompose h as a sum $h_1 + h_2$ of two additive functions, "supported" on primes less than $\sqrt{N}$ and primes greater than $\sqrt{N}$, respectively. The $O(1/N)$ errors turn out to be negligible for h_1. For h_2, they can be handled using the following simple but important fact.

Lemma 5.7. *If p, q are distinct primes congruent to* 1 (mod 4)*, then*

$$w_{Q,N}(p,p) \ll \frac{Q^2}{p}, \quad w_{Q,N}(p,q) \ll \frac{Q^2}{pq},$$

where the implicit constant is absolute.

The proof of this fact is simple and is based on the elementary estimate

$$\sum_{k \le n} r_2(k) \ll n,$$

where $r_2(n)$ is the number of representations of k as a sum of two squares. Lemma 5.7 gives better upper bounds than (5.18) when pq is much larger than N, bounds that turn out to be sufficient to show that the contribution of h_2 in (5.15) is negligible ((5.16) with h_2 instead of h is also used here).

The second step of the proof is the transition from additive to multiplicative functions. We follow the same approach as in the linear case described above. We first reduce to the case $f \sim 1$, in which case we get that (5.15) holds with f_N instead of f, where f_N is the restriction of f on primes less than N, i.e.,

$$f_N(p^k) := \begin{cases} f(p^k) & \text{if } p \le N, \\ 1 & \text{otherwise.} \end{cases}$$

The error terms incurred in this transition allow us to deal only with additive and multiplicative functions supported on primes $p \le N$; this is the reason why in (5.17) a larger range of primes in the support of h_N would not have been helpful.

Finally, to replace f_N with f in (5.15), we use the factorization $f = f_N \cdot g_N$, where g_N is the restriction of f on primes greater than N, defined in a similar way to f_N. We then use a slight variant of Lemma 5.7 together with our hypothesis $f \sim 1$ to show that the contribution of g_N in (5.15) is negligible and complete the proof of Theorem 5.6.

5.6. Concentration estimates for general binary quadratic forms

To prove Theorems 3.2 and 3.3, which are needed to cover partition and density regularity results of the more general quadratic equations $ax^2 + by^2 = cz^2$, we need

to establish variants of Theorem 5.6 in which the place of m^2+n^2 is taken by the more general irreducible binary quadratic form $P(m,n)=\alpha m^2+\beta mn+\gamma n^2$, where $\alpha,\beta,\gamma\in\mathbb{Z}$. To simplify the notation a bit, we give the exact statement when $P(m,n)=m^2+dn^2$ and $d\in\mathbb{Z}$ is a square-free integer, the general case is similar, and the exact statements can be found in [22, Proposition 4.9]. We first have to modify the definition of $G_N(f,K)$ from (5.14). If (a/p) denotes the Legendre symbol and $d\in\mathbb{Z}$ is square-free, we let

$$\mathcal{P}_d:=\Big\{p\in\mathbb{P}\colon \Big(\frac{-d}{p}\Big)=1\Big\}. \tag{5.19}$$

For example, $\mathcal{P}_1:=\{p\equiv 1 \pmod 4\}$, $\mathcal{P}_2:=\{p\equiv 1,3 \pmod 8\}$, $\mathcal{P}_{-2}:=\{p\equiv 1,7 \pmod 8\}$. It can be shown that for square-free d, the set $\mathcal{P}_d$ has relative density $1/2$ on the primes, and it always consists of the set of primes that belong to finitely many congruence classes. For example, using quadratic reciprocity, we get that if q is an odd prime, then

$$\mathcal{P}_q=\{p\equiv \pm k^2 \pmod{4q}\colon k\in[4q] \text{ is odd and coprime to } 4q\}.$$

If $f\sim\chi\cdot n^{it}$ for some Dirichlet character χ and $t\in\mathbb{R}$, and $K,N\in\mathbb{N}$, we let

$$G_{d,N}(f,K):=2\sum_{K<p\le N,\ p\in\mathcal{P}_d}\frac{1}{p}\,(f(p)\cdot\overline{\chi(p)}\cdot n^{-it}-1). \tag{5.20}$$

Theorem 5.8 ([22, Proposition 4.9]). *Let $f\colon\mathbb{N}\to\mathbb{U}$ be a multiplicative function and suppose that $f\sim\chi\cdot n^{it}$ for some Dirichlet character χ and $t\in\mathbb{R}$. Then for every square-free $d\in\mathbb{Z}$ such that the form m^2+dn^2 is irreducible, we have*

$$\lim_{K\to\infty}\limsup_{N\to\infty}\max_{Q\in\Phi_K}\mathbb{E}_{m,n\in[N]}\big|f((Qm+1)^2+d\,(Qn)^2)\\ -Q^{2it}\cdot(m^2+dn^2)^{it}\cdot\exp(G_{d,N}(f,K))\big|=0,$$

where Φ_K and $G_{d,N}(f,K)$ are as in (5.3) and (5.20), respectively.

We also prove variants of concentration estimates along more general lattices of the form $\{(Qm+a,Qn+b)\colon m,n\in\mathbb{N}\}$ for suitable $a,b\in\mathbb{Z}_+$; such results are not needed for the proof of Theorem 3.2 but turn out to be essential for the proof of Theorem 3.3.

To prove Theorem 5.8, we follow the approach used to prove Theorem 5.6, but there are quite a few additional difficulties. The most serious problem is to find a variant of Lemma 5.7 that allows us to discount the contribution of large primes in the various steps of the proof. For $d>0$, the most problematic case is when the corresponding quadratic integer ring is not a principal ideal domain, and additional,

more substantial problems arise when $d < 0$. For example, suppose that $d < 0$ is a square-free integer such that $d \equiv 1, 2 \pmod 4$, and we want to establish for all primes $p \in \mathcal{P}_d$ an estimate of the form

$$|\{m, n \in [N]: p \mid m^2 + dn^2\}| \ll_d \frac{N^2}{p}, \tag{5.21}$$

where the implicit constant depends only on d. A priori, it is not clear that such an estimate holds, because the counting is affected by the number of the solutions of Pell's equation $m^2 + dn^2 = 1$ in the box $[1, N] \times [1, N]$, which grows as $N \to \infty$, roughly like $\log N$ (unlike the case $d > 0$, where the corresponding number is bounded by 4).

Fortunately, we can perform a careful count and use classical but non-trivial facts from algebraic number theory to show that (5.21) indeed holds. We first note that the left-hand side in (5.21) is bounded by

$$\sum_{r \le \frac{2N^2}{p}} C_{d,N}(pr),$$

where $C_{d,N}(k) := |\{m, n \in [N]: |m^2 + dn^2| = k\}|$. Using the fact that there exists a unit $u \in \mathbb{Z}[\sqrt{-d}] \subset \mathbb{R}$ such that $u > 1$ and every other unit has the form $\pm u^t$ for some $t \in \mathbb{Z}$ (this is the Dirichlet unit theorem for the ring $\mathbb{Z}[\sqrt{-d}]$), it is possible to show that there exists $c_d > 0$ such that

$$C_{d,N}(k) \ll \log\left(\frac{c_d N}{\sqrt{k}}\right) \cdot \mathcal{C}_d(k), \tag{5.22}$$

where

$$\mathcal{C}_d(k) := |\{\text{ideals in } \mathbb{Z}[\sqrt{-d}] \text{ with norm } k\}|.$$

It is important for our argument that the term $\sqrt{k}$ appears in the denominator on the right-hand side of (5.22), our argument would not work without it. So (5.21) follows if we establish the bound

$$\sum_{r \le \frac{2N^2}{p}} \log\left(\frac{c_d N}{\sqrt{pr}}\right) \cdot \mathcal{C}_d(pr) \ll_d \frac{N^2}{p}. \tag{5.23}$$

To prove this, we crucially use the estimate

$$\mathcal{C}_d(pr) \le 2\,\mathcal{C}_d(p), \tag{5.24}$$

which holds for all $p \in \mathcal{P}_d$. This seemingly simple fact is not easy to prove from first principles, especially in the case where the corresponding ring of integers is

not a principal ideal domain. However, its proof follows easily from well-known but non-trivial facts in algebraic number theory: the multiplicativity of $\mathcal{C}_d(k)$, the estimate $\mathcal{C}_d(p^{k+1}) \le \mathcal{C}_d(p^k) + 1$, and the fact that $\mathcal{C}_d(p^k) \ge 1$ for all $p \in \mathcal{P}_d$ and $k \in \mathbb{N}$. The proof of (5.23) follows easily by using (5.24) and then applying partial summation combined with the estimate

$$\sum_{r=1}^{N} \mathcal{C}_d(r) \ll_d N,$$

which follows from the ideal counting theorem.

6. Connecting the pieces (the Q-trick)

In this section, we show how we can combine the ingredients of the previous two sections to prove Theorems 3.2 and 3.3. We start with the simplest case and then explain the necessary changes needed to cover the increasingly more complicated cases.

6.1. The pair $(2mn, m^2 - n^2)$

We start with the proof of Theorem 3.2 when $\alpha = 1$, $\beta = -1$, $\gamma = 2$, which suffices to prove partition and density regularity of $x^2 + y^2 = z^2$ with respect to the variables x, y. Using the positivity property (3.11), it suffices to show that there exists $Q \in \mathbb{N}$ for which

$$\int_{\mathcal{M}} L(f, Q)\, d\sigma(f) > 0, \tag{6.1}$$

where

$$L(f, Q) := \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, f((Qm+1)^2 - (Qn)^2) \cdot \overline{f(2(Qm+1)Qn)}. \tag{6.2}$$

Note that in establishing the above reduction, we used the bounded convergence theorem and the non-trivial fact that the limit defining $L(f, Q)$ exists, which follows from [20, Theorem 1.4]. The proof of the existence of the limit is non-trivial and can be avoided at the cost of making the argument a bit more complex.

We also note that we could have used in our averaging any other lattice of the form $(Qm + a, Qn)$, where a is some fixed number independent of Q; what is important is that on one of the two coordinates we use Qn (or Qm) in order to later on factor out a term $f(Q)$, which is an essential element of a key maneuver that we call the "Q-trick".

It easily follows from Theorem 4.7 that if the multiplicative function f is aperiodic, then

$$L(f, Q) = 0 \quad \text{for every } Q \in \mathbb{N}.$$

Hence, (6.1) would follow if we show that

$$\int_{\mathcal{M}_p} L(f, Q)\, d\sigma(f) > 0 \quad \text{holds for some } Q \in \mathbb{N}, \tag{6.3}$$

where

$$\mathcal{M}_p := \{f \in \mathcal{M} \text{ is pretentious}\}.$$

Note that the integrals in (6.3) can be shown to be real. To prove (6.3), we will use the linear concentration estimates of Theorem 5.5 to substantially simplify the value of the limit in (6.2) and then use a trick, called the Q-trick, to show that for some particular choice of Q, we get positivity.

6.1.1. Finitely generated case. To better illustrate the argument, we first assume that the measure σ is supported on the subset of pretentious multiplicative functions that are finitely generated, i.e., the set $\{f(p)\colon p \in \mathbb{P}\}$ is a finite subset of the unit circle. The advantage in this case is that $f \sim \chi$ for some Dirichlet character χ (so n^{it} is no longer needed), and also the oscillatory factor $\exp(F_N(f, K))$ in (5.5) can be taken to be 1. So in this case, the concentration estimate in (5.5) takes the much simpler form

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} \mathbb{E}_{n\in[N]} |f(Qn + 1) - 1| = 0, \tag{6.4}$$

where Φ_K is as in (5.3). Since

$$Qm + 1 \pm Qn = Q(m \pm n) + 1,$$

with a bit of fiddling (see [23, Lemma 3.1]), (6.4) implies that for all practical purposes, the limit $L(f, Q)$ in (6.2) can be approximated by $f(Q) \cdot A(f)$, where

$$A(f) := \lim_{N\to\infty} \mathbb{E}_{n\in[N]}\, \overline{f(2n)}. \tag{6.5}$$

More precisely, we have

$$\lim_{K\to\infty} \max_{Q\in\Phi_K} |L(f, Q) - f(Q) \cdot A(f)| = 0. \tag{6.6}$$

It follows from this that in order to establish (6.3), it suffices to study the integrals

$$\int_{\mathcal{M}_p} f(Q) \cdot A(f)\, d\sigma(f).$$

A priori, however, these expressions do not satisfy any obvious positivity property, and it seems that we are at a dead end. The simple but crucial observation is that the multiplicative average of this last expression, taken over $Q \in \Phi_K$ and letting $K \to \infty$, does satisfy a nice positivity property, namely,

$$\lim_{K\to\infty} \mathbb{E}_{Q\in\Phi_K} \int_{\mathcal{M}_p} f(Q)\cdot A(f)\, d\sigma(f) = \sigma(\{1\}) > 0. \tag{6.7}$$

This is because for every completely multiplicative function $f \neq 1$, we have

$$\lim_{K\to\infty} \mathbb{E}_{Q\in\Phi_K} f(Q) = 0,$$

and so the bounded convergence theorem gives that (6.7) holds. Combining (6.6) and (6.7), we get that

$$\lim_{K\to\infty} \mathbb{E}_{Q\in\Phi_K} \int_{\mathcal{M}_p} L(f,Q)\, d\sigma(f) = \sigma(\{1\}) > 0,$$

hence (6.3) holds.

6.1.2. General case. We now consider the general case of measures supported by multiplicative functions that are not necessarily finitely generated. In this case, suppose that $f \sim \chi \cdot n^{it}$ for some Dirichlet character χ and $t \in \mathbb{R}$. Then the concentration estimates of Theorem 4.7 give that

$$\lim_{K\to\infty} \max_{Q\in\Phi_K} |L(f,Q) - f(Q)\cdot Q^{it}\cdot B(f,K)| = 0,$$

where

$$B(f,K) := \exp(2F_{2N}(f,K))\cdot\overline{\exp(F_N(f,K))}\cdot \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\,(m^2-n^2)^{it} \cdot m^{-it}\cdot\overline{f(2n)}.$$

Although the expressions $B(f,K)$ are much more complicated than the expressions $A(f)$ in (6.5), this is of little concern to us; what matters is that for $Q \in \Phi_K$, they are constant in Q. This allows us to show that as long as $f \neq n^{-it}$, we have

$$\lim_{K\to\infty} \mathbb{E}_{Q\in\Phi_K} f(Q)\cdot Q^{it}\cdot B(f,K) = 0.$$

As before, using the bounded convergence theorem, we conclude that

$$\lim_{K\to\infty} \mathbb{E}_{Q\in\Phi_K} \int_{\mathcal{M}_p} L(f,Q)\, d\sigma(f) = \int_{\mathcal{A}} B(f)\, d\sigma(f),$$

where

$$\mathcal{A} := \{(n^{it})_{n\in\mathbb{N}} : t \in \mathbb{R}\}$$

and

$$B(f) := \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, (m^2-n^2)^{it}\cdot(2mn)^{-it}$$

when $f = n^{it}$. In order to complete the proof, it suffices to prove the positivity of $\int_{\mathcal{A}} B(f)\, d\sigma(f)$. We do this in two different ways below, the second being the more flexible.

(Endgame I: Miraculous positivity). The limit defining $B(f)$ is easily shown to be

$$\int_0^1 \int_0^1 (x-y)^{it}\cdot(2xy)^{it}\, dx\, dy.$$

It can be shown that this double integral has a positive real part for every $t\in\mathbb{R}$ by explicitly evaluating the integral (we did this computation using some computer software). Since $\sigma(\{1\}) > 0$, this implies that $\int_{\mathcal{A}} B(f)\, d\sigma(f) > 0$ and completes the proof.

However, this miraculous positivity property fails when we replace $2mn$ with mn or $3mn$ and probably only works for kmn when $k = 2$, i.e., only for the case of Pythagorean pairs! Next, we will give an alternative, more complicated, but robust method that works in much higher generality (this is needed for equations other than the Pythagorean).

(Endgame II: A more robust method). We introduce some weights that effectively allow us to replace all non-trivial Archimedean characters with 1. For every $\delta\in(0,1/2)$, we let

$$w_\delta(m,n) := \mathbf{1}_{I_\delta}((m^2-n^2)^i\cdot(mn)^{-i})\cdot\mathbf{1}_{m>n},\quad m,n\in\mathbb{N}, \tag{6.8}$$

where I_δ is an arc in $\mathbb{S}^1$ around 1 with length δ. It is not difficult to show that for all $\delta > 0$, we have

$$\mu_\delta := \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, w_\delta(m,n) > 0 \tag{6.9}$$

and also

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]} \sup_{Q\in\mathbb{N}} |w_\delta(Qm+1,Qn) - w_\delta(m,n)| = 0,$$

so for our purposes, the weights $w_\delta(m,n)$ and $w_\delta(Qm+1,Qn)$ can be used interchangeably. It is also not hard to show that for aperiodic multiplicative functions f, we have the vanishing property (4.12) even if we insert the weights $w_\delta(m,n)$ in the average. Using these properties, we can work out the previous argument, replacing the expression $L(f,Q)$ in (6.2) by

$$L_\delta(f,Q) := \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, w_\delta(m,n)\cdot f((Qm+1)^2-(Qn)^2) \cdot \overline{f(2(Qm+1)Qn)}.$$

In this way, we reduce matters to showing that for some sufficiently small $\delta > 0$, we have

$$\int_{\mathcal{A}} B_\delta(f)\, d\sigma(f) > 0,$$

where

$$B_\delta(n^{it}) := \lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, w_\delta(m,n)\cdot (m^2-n^2)^{it}\cdot (2mn)^{-it}$$

when $f = n^{it}$. The advantage of this weighted average is that for any fixed $t \in \mathbb{R}$, because of the definition of the weight in (6.8), if δ is small enough, then $B_\delta(f)$ is close to 1. This allows us to show with some fiddling that if δ is small enough, then we have the lower bound

$$\int_{\mathcal{A}} B_\delta(f)\, d\sigma(f) \geq \frac{1}{2}\cdot \mu_\delta \cdot \sigma(\mathcal{A}) \geq \frac{1}{2}\cdot \mu_\delta \cdot \sigma(\{1\}) > 0,$$

where μ_δ is as in (6.9). This completes the proof in a robust way that allows us to deal with positivity properties of more general pairs such as $(\ell mn, \ell'(m^2-n^2))$, where $\ell, \ell' \in \mathbb{N}$ are arbitrary. The reader will find more details in [23, Section 4].

6.2. Comparison with the strategy from [21]

In [21], we were able to show partition regularity for the pairs $(m^2-n^2, m(m+2n))$ but not for the pairs $(m^2-n^2, 2mn)$. Let us briefly explain why. The main technical tool used in [21] was a structural result for general multiplicative functions which roughly states the following (see [21, Theorem 2.1] for the detailed statement): For every $f \in \mathcal{M}$, $\varepsilon > 0$, and $s \in \mathbb{N}$, there exist $Q \in \mathbb{N}$ and $C > 0$ such that for every $N \in \mathbb{N}$, we have a decomposition

$$f(n) = f_{\mathrm{st}}(n) + f_{\mathrm{un}}(n) + f_{\mathrm{er}}(n), \quad n \in [N],$$

such that

(i) $|f_{\mathrm{st}}(n+Q) - f_{\mathrm{st}}(n)| \leq C/N$ for every $n \in [N-Q]$;

(ii) $\|f_{\mathrm{un}}\|_{U^s[N]}$ is "very small" in a sense that will not be made precise here;

(iii) $\mathbb{E}_{n\in[N]}\, |f_{\mathrm{er}}(n)| \leq \varepsilon$.

In fact, the applications to partition regularity require some uniformity in the choice of Q and C with respect to $f \in \mathcal{M}$, but that will not concern us here.

Let us see how we can use these properties to get some positiveness for fixed $f \in \mathcal{M}$ and averages of the form

$$\mathbb{E}_{m,n\in[N]}\, f(m^2-(Qn)^2)\cdot \overline{f(m(m+2Qn))}.$$

Using properties (ii) and (iii), it is not hard to see that these averages are $O(\varepsilon)$-close to the averages

$$\mathbb{E}_{m,n\in[N]}\, f_{\mathrm{st}}(m-Qn)\cdot f_{\mathrm{st}}(m+Qn)\cdot\overline{f_{\mathrm{st}}(m)}\cdot\overline{f_{\mathrm{st}}(m+2Qn)}.$$

Assuming for simplicity that $f_{\mathrm{st}}(n+Q)=f_{\mathrm{st}}(n)$ for every $n\in\mathbb{N}$, the last average is equal to

$$\mathbb{E}_{m,n\in[N]}\,|f_{\mathrm{st}}(m)|^4.$$

This positive lower bound is crucial for the argument in [21] and with some maneuvering can be used to prove partition regularity of the patterns $(m^2-n^2, m(m+2n))$. However, it does not appear to be possible to modify this argument in a simple way in order to deal with the pairs $(m^2-n^2, 2mn)$, the reason being that if $m=0$, or $n=0$, the two coordinates of this pair do not coincide.

Instead, one can try to restrict both coordinates m and n to suitable progressions (just as we did in Section 6.1) and work with the averages

$$\mathbb{E}_{m,n\in[N]}\, f((Qm+1)^2-(Qn)^2)\cdot\overline{f((Qm+1)(Qn))}. \tag{6.10}$$

Unfortunately, the previous decomposition result seems inadequate for analyzing these expressions because we have no useful control over the error term f_{er} along the progression $Q\,\mathbb{N}$. The important advantage of the concentration estimates described in Section 5 is that for pretentious multiplicative functions, we obtain an error term that is small even when we restrict both variables m,n to arithmetic progressions. This is crucial for the purposes of analyzing the asymptotic behavior of the averages in (6.10). Moreover, we will see in the next subsection that the concentration estimates described in Section 5.5 allow us to effectively analyze averages of the form

$$\mathbb{E}_{m,n\in[N]}\, f((Qm+1)^2+(Qn)^2)\cdot\overline{f((Qm+1)(Qn))}$$

for pretentious multiplicative functions. Such a task was beyond the limits of the methodology used in the article [21].

6.3. The pair $(2mn, m^2+n^2)$

To establish Theorem 3.2 in this case ($\alpha=\beta=1$, $\gamma=2$), and thus deduce partition and density regularity of $x^2+y^2=z^2$ with respect to the variables x,z, we follow the argument of the previous subsection with some necessary and rather laborious changes. We introduce the averages

$$\begin{aligned}L'_{\delta,N}(f,Q) := \mathbb{E}_{m,n\in[N]}\,\widetilde{w}_\delta(m,n)\cdot f((Qm+1)^2\\ +(Qn)^2)\cdot\overline{f(2(Qm+1)Qn)},\end{aligned} \tag{6.11}$$

where the weights $\widetilde{w}_\delta$ are defined similarly to (6.8) but with $m^2 - n^2$ replaced by $m^2 + n^2$. In order to study their asymptotic behavior, as $N \to \infty$, we first show using Theorem 4.8 that in the aperiodic case, they vanish for every $Q \in \mathbb{N}$.

We then study their asymptotic behavior when f is pretentious. To simplify their form, we use the non-linear concentration estimates of Theorem 5.6. This is an important difference, since the non-linear concentration estimates were not previously known and turned out to be much harder to obtain than the linear ones. Another technical annoyance is that, contrary to what happened in the previous subsection, the limit $\lim_{N\to\infty} L'_{\delta,N}(f,Q)$ may not exist, the reason being that the oscillatory factors resulting from the concentration estimates for the quadratic and linear terms do not conveniently cancel. In any case, we can show that if $f \sim \chi \cdot n^{it}$ for some Dirichlet character χ and $t \in \mathbb{R}$, then

$$\lim_{K\to\infty} \limsup_{N\to\infty} \max_{Q\in\Phi_K} |L'_{\delta,N}(f,Q) - f(Q)\cdot Q^{it}\cdot B_{\delta,N}(f,K)| = 0,$$

where

$$B_{\delta,N}(f,K) := \exp(G_N(f,K))\cdot\overline{\exp(F_N(f,K))}\cdot \mathbb{E}_{m,n\in[N]}\,\widetilde{w}_\delta(m,n)\cdot(m^2+n^2)^{it} \cdot m^{-it}\cdot\overline{f(2n)},$$

and $F_N(f,K)$ and $G_N(f,K)$ are given by (5.4) and (5.14), respectively. Again, the complexity of the expression $B_{\delta,N}(f,K)$ will not bother us; the only important fact is that for $Q \in \Phi_K$, it is constant in Q. This allows us to show that as long as $f \neq n^{-it}$, we have

$$\lim_{K\to\infty} \limsup_{N\to\infty} |\mathbb{E}_{Q\in\Phi_K} f(Q)\cdot Q^{it}\cdot B_{\delta,N}(f,K)| = 0.$$

After this point, arguing as in the previous subsection, we get that for some $\delta_0 > 0$, we have the following positivity property:

$$\liminf_{K\to\infty} \liminf_{N\to\infty} \mathbb{E}_{Q\in\Phi_K} \int_{\mathcal{M}} L'_{\delta_0,N}(f,Q)\,d\sigma(f) > 0, \tag{6.12}$$

where $L'_{\delta_0,N}(f,Q)$ is given by (6.11). It is slightly annoying that $\mathbb{E}_{Q\in\Phi_K}$ and $\liminf_{N\to\infty}$ are not interchangeable (in the previous subsection, the existence of the limit as $N \to \infty$ allowed us to do this). To overcome this and finish the argument, we argue as follows. We note that using (6.12) and the fact that $\widetilde{w}_\delta(m,n)$ is approximately equal to $\widetilde{w}_\delta(Qm+1,Qn)$, we get that there exist $K_0 \in \mathbb{N}$ and $Q_N \in \Phi_{K_0}$, $N \in \mathbb{N}$, such that

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, I(Q_N m + 1, Q_N n) =: a > 0, \tag{6.13}$$

where

$$I(m,n) := \int_{\mathcal{M}} \widetilde{w}_\delta(m,n) \cdot f(m^2+n^2) \cdot \overline{f(2mn)}\, d\sigma(f).$$

Since $I(m,n) \geq 0$ for every $m,n \in \mathbb{N}$ and Q_N belongs to a finite set bounded by, say, M, we deduce from (6.13) that

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, I(m,n) = \frac{a}{M^2} > 0.$$

The reader will find the details of this argument in [23, Section 6].

6.4. The pair (mn, m^2+dn^2)

The two main new ingredients needed to cover this more general case are Theorem 4.9, which establishes the vanishing property of correlations in the case of aperiodic multiplicative functions, and Theorem 5.8, which establishes concentration estimates for pretentious multiplicative functions for the binary quadratic form m^2+dn^2 when it is irreducible. Modulo these two results, which require considerable additional effort to prove, the required positivity property goes along the lines described in the previous subsections. The details appear in [22, Section 5].

We remark that employing an analogous approach, we can cover all pairs $(P_1(m,n), P_2(m,n))$, where at least one of the P_1, P_2 is a reducible binary quadratic form that is not a square.

6.5. The pair (m^2+2n^2, m^2-2n^2)

We need to cover this pair to prove partition and density regularity of $x^2+2y^2=z^2$ with respect to the variables x,z, (see the parametrization in Section 2.2). The situation is similar for the equation $x^2+y^2=2z^2$ with respect to any pair of variables. In order to prove the positivity property of Theorem 3.3 for $\alpha=\alpha'=0$, $\beta=2$, $\beta'=-2$, we need to study the correlations

$$\mathbb{E}_{m,n\in[N]}\, f(m^2+2n^2) \cdot \overline{f(m^2-2n^2)} \tag{6.14}$$

for arbitrary $f \in \mathcal{M}$. It is expected that asymptotic properties of f along m^2+2n^2 and $m-2n^2$ are only affected by its values on primes in $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, where

$$\mathcal{P}_1 := \{p \equiv 1,3 \pmod 8\}, \quad \mathcal{P}_2 := \{p \equiv 1,7 \pmod 8\}. \tag{6.15}$$

(Using the notation in (5.19), we should have denoted the two sets by $\mathcal{P}_2$ and $\mathcal{P}_{-2}$, respectively; we do not do this to simplify the notation later.)

Analogously to Definition 5.1, if $\mathcal{P}$ is a subset of the primes, we let

$$\mathbb{D}^2_{\mathcal{P}}(f,g) := \sum_{p\in\mathcal{P}} \frac{1}{p}(1-\Re(f(p)\cdot\overline{g(p)})),$$

and write $f \sim_{\mathcal{P}} g$ if $\mathbb{D}_{\mathcal{P}}(f,g) < +\infty$. The multiplicative function f is $\mathcal{P}$-pretentious if $\mathbb{D}_{\mathcal{P}}(f, \chi\cdot n^{it}) < +\infty$ for some Dirichlet character χ and $t\in\mathbb{R}$ and *$\mathcal{P}$-aperiodic* if it is not $\mathcal{P}$-pretentious. Note that for $\mathcal{P}=\mathbb{P}$, we get the same definition as in Section 5.1.

6.5.1. The pretentious case. Let $\mathcal{P}_1, \mathcal{P}_2$ be as in (6.15) and suppose that f is $\mathcal{P}_1$-pretentious and $\mathcal{P}_2$-pretentious. Furthermore, for convenience, suppose that

$$f \sim_{\mathcal{P}_1} \chi_1 \quad \text{and} \quad f \sim_{\mathcal{P}_2} \chi_2 \tag{6.16}$$

for some Dirichlet characters χ_1, χ_2. We will ignore Archimedean characters in this discussion, since they can be covered by introducing weights and using a straightforward variant of the argument given in Section 6.3.

Our aim is to use again the strategy described in the previous subsections. However, substantial new complications arise since we no longer have a reducible polynomial like mn to work with, and this makes the Q-trick harder to implement. Moreover, Theorem 5.8 gives that there exist multiplicative functions which concentrate on different oscillatory factors along m^2+2n^2 and m^2-2n^2 that do not cancel each other. This casts doubt on whether it is even possible in this case to establish positivity via concentration estimates. Fortunately, we will be able to do this by adapting the Q-trick, this time averaging over appropriate "translates" rather than "dilates" of an appropriate choice of lattice. Our goal is to ensure that for those multiplicative functions that concentrate along m^2+2n^2 and m^2-2n^2 around different oscillatory factors, a certain multiplicative average over Q will vanish, and for the remaining multiplicative functions, we can guarantee that they concentrate on the same oscillatory factor, which leads to a favorable cancellation.

Let us explain in more detail how the previous plan is implemented. For a given $K\in\mathbb{N}$ and $l_p, l'_p \in [1, 3K/2]$, let

$$Q_1 := \prod_{p\equiv 3 \,(\mathrm{mod}\ 8),\, p\le K,\, p\nmid q} p^{l_p}, \quad Q_2 := \prod_{p\equiv 7 \,(\mathrm{mod}\ 8),\, p\le K,\, p\nmid q} p^{l'_p}. \tag{6.17}$$

Let q be a common period of the multiplicative functions χ_1 and χ_2. We plan to apply concentration estimates on the lattice

$$\{(Q_K m + a, Q_K n + b)\colon m,n\in\mathbb{N}\},$$

where

$$Q_K := \prod_{p \le K} p^{2K}$$

and $a, b \in [Q]$ depend on Q_1, Q_2 and satisfy the following two properties:

$$a^2 + 2b^2 \equiv a^2 - 2b^2 \equiv 1 \pmod{q} \tag{6.18}$$

and

$$(a^2 + 2b^2, Q) = Q_1, \quad (a^2 - 2b^2, Q) = Q_2. \tag{6.19}$$

To show that a choice of a, b that satisfy (6.18) and (6.19) is possible, we use the Chinese remainder theorem and the fact that for $p \equiv 3 \pmod 8$ the congruences $p^{l_p} \mid\mid a^2 + 2b^2$ and $p \nmid a^2 - 2b^2$ have a solution, and similarly, for $p \equiv 7 \pmod 8$ the congruences $p^{l_p} \mid\mid a^2 - 2b^2$ and $p \nmid a^2 + 2b^2$ have a solution.

Suppose now that K is sufficiently large, Q_1, Q_2 are as in (6.17), and $a = a_{Q_1,Q_2}, b = b_{Q_1,Q_2}$ satisfy (6.18) and (6.19). For convenience, we let

$$P_1(m,n) := m^2 + 2n^2, \quad P_2(m,n) := m^2 - 2n^2.$$

Using a variant of Theorem 5.8 that allows us to work with lattices of the form $Q \cdot \mathbb{N} \times \mathbb{N} + (a,b)$, we get that for $j = 1,2$ and $Q = Q_K$, the values of $f(P_j(Qm + a, Qn + b))$ are concentrated at

$$f(Q_j) \cdot \chi_j\left(\frac{P_j(a,b)}{Q_j}\right) \cdot \exp(G_{j,N}(f,K)), \tag{6.20}$$

(we used that $Q_j = (P_j(a,b), Q)$, $j = 1,2$ and $q \prod_{p \le K} p$ divides Q/Q_j) where

$$G_{j,N}(f,K) := 2 \sum_{K < p \le N,\, p \in \mathcal{P}_j} \frac{1}{p} (f(p) \cdot \overline{\chi_j(p)} \cdot -1), \quad j = 1,2.$$

Note that for $j = 1,2$, we have $(Q_j, q) = 1$, hence $|\chi_j(Q_j)| = 1$, and also since (6.18) holds, we have $\chi_j(P_j(a,b)) = 1$. It follows that the expression in (6.20) is equal to

$$f(Q_j) \cdot \overline{\chi_j(Q_j)} \cdot \exp(G_{j,N}(f,K)).$$

Hence, $f(P_1(Qm + a, Qn + b)) \cdot \overline{f(P_2(Qm + a, Qn + b))}$ concentrates at

$$f(Q_1) \cdot \overline{\chi_1(Q_1)} \cdot \exp(G_{1,N}(f,K)) \cdot \overline{f(Q_2)} \cdot \chi_2(Q_2) \cdot \overline{\exp(G_{2,N}(f,K))}. \tag{6.21}$$

We consider two cases.

Case 1. Suppose that for primes p that satisfy $p \nmid q$, we have

$$f(p) \neq \chi_1(p) \quad \text{for some } p \equiv 3 \pmod 8$$

or

$$f(p) \neq \chi_2(p) \quad \text{for some } p \equiv 7 \pmod 8.$$

Then the previous facts easily imply that

$$\lim_{K\to\infty} \limsup_{N\to\infty} |\mathbb{E}_{Q_1\in\Phi_{1,K}, Q_2\in\Phi_{2,K}} L_N(f, K, Q_1, Q_2)| = 0, \tag{6.22}$$

where

$$L_N(f, K, Q_1, Q_2) := \mathbb{E}_{m,n\in[N]} f(P_j(Q_K m + a_{Q_1,Q_2}, Q_K n + b_{Q_1,Q_2})) \cdot \overline{f(P_j(Q_K m + a_{Q_1,Q_2}, Q_K n + b_{Q_1,Q_2}))}, \tag{6.23}$$

and for $j = 1, 2$,

$$\Phi_{j,K} := \Big\{ \prod_{p\in\mathcal{P}_j\cap[K]} p^{l_p} : K < l_p \leq \frac{3K}{2} \Big\}.$$

Case 2. Suppose that for primes p that satisfy $p \nmid q$, we have

$$f(p) = \chi_1(p) \quad \text{for all } p \equiv 3 \pmod 8 \tag{6.24a}$$

and

$$f(p) = \chi_3(p) \quad \text{for all } p \equiv 7 \pmod 8, \tag{6.24b}$$

then the expression in (6.21) equals

$$\exp(G_{1,N}(f, K)) \cdot \overline{\exp(G_{2,N}(f, K))}. \tag{6.25}$$

Using (6.24a) and (6.24b), we get that for K large enough and $N > K$, we have

$$G_{1,N}(f, K) = G_{2,N}(f, K).$$

Thus, the expression in (6.25) equals

$$\exp(2\Re(G_{1,N}(f, K))).$$

Because our standing assumption (6.16) is that $f \sim_{P_1} \chi_1$, this last expression approaches 1 as $N \to \infty$ uniformly on $N > K$. Combining the above, we get

$$\lim_{K\to\infty} \limsup_{N\to\infty} \sup_{Q_1\in\Phi_{1,K}, Q_2\in\Phi_{2,K}} |L_N(f, K, Q_1, Q_2) - 1| = 0, \tag{6.26}$$

where $L_N(f, K, Q_1, Q_2)$ is as in (6.23).

Combining (6.22) and (6.26), we get the positiveness properties needed to prove Theorem 3.3.

6.5.2. The aperiodic case. In the case where the multiplicative function $f \in \mathcal{M}$ is $\mathcal{P}_1$-aperiodic or $\mathcal{P}_2$-aperiodic, we expect the averages in (6.14) to vanish as $N \to \infty$. Unfortunately, this property turns out to be very difficult to establish because, unlike the cases treated in the previous subsections, we have no way of bounding these averages by a usable substitute for the Gowers uniformity seminorms of the multiplicative function f. In fact, the needed vanishing property is not even known for the Liouville function, i.e., it is not known that

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \lambda(m^2+2n^2)\cdot\lambda(m^2-2n^2) = 0.$$

For the applications we have in mind, we would be equally happy to show this for logarithmic averages in place of Cesàro.

Definition 6.1. If $P \in \mathbb{Z}[m,n]$ is an irreducible binary quadratic form with discriminant d_P, we let

$$\mathcal{P}_P := \Big\{p \in \mathbb{P}\colon \Big(\frac{d_P}{p}\Big) = 1\Big\}.$$

We say that the irreducible binary quadratic forms P_1, P_2 are *good for vanishing of correlations of aperiodic multiplicative functions* whenever the following property holds: If $f_1, f_2\colon \mathbb{N} \to \mathbb{S}^1$ are completely multiplicative functions such that either f_1 is $\mathcal{P}_{P_1}$-aperiodic or f_2 is $\mathcal{P}_{P_2}$-aperiodic, then for every $Q \in \mathbb{N}$ and $a, b \in \mathbb{Z}_+$, we have

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \mathbf{1}_S(Qm+a, Qn+b)\cdot f_1(P_1(Qm+a, Qn+b)) \\ \cdot f_2(P_2(Qm+a, Qn+b)) = 0,$$

where $S := \{m,n \in \mathbb{N}\colon P_1(m,n) > 0, P_2(m,n) > 0\}$.

Conjecture 3. *If $P_1, P_2 \in \mathbb{Z}[m,n]$ are irreducible binary quadratic forms that are not multiples of each other, then they are good for vanishing of correlations of aperiodic multiplicative functions.*

If this conjecture is true, we show in [22, Section 6] how Theorem 3.3 follows using the ideas summarized in the previous subsection.

7. More challenges

In this section, we present some open questions linked to the topics already discussed.

7.1. Vanishing of correlations along two irreducible quadratics

The following problem covers a special case of Conjecture 3 that seems very challenging.

Problem 1. If λ is the Liouville function, show that

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \lambda(m^2+n^2)\cdot\lambda(m^2+2n^2) = 0.$$

For the applications of this article, we could replace the Cesàro averages $\mathbb{E}_{m,n\in[N]}$ with the logarithmic averages $1/(\log N)^2 \sum_{m,n\in[N]} 1/mn$. In fact, it appears that no explicit example of two irreducible binary quadratic forms P_1, P_2 is known for which

$$\lim_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \lambda(P_1(m,n))\cdot\lambda(P_2(m,n)) = 0$$

or a similar property with logarithmic averages.

7.2. Partition regularity of "most" generalized Pythagorean triples

Conjecture 3 asks to show that for all Rado triples (a,b,c), equation (1.2) is partition regular. At the moment, this seems very hard to prove, so even showing that we have partition regularity for a positive proportion, or even better, for most Rado triples, would be desirable.

Problem 2. Show that for "most" choices of Rado triples $(a,b,c)\in\mathbb{N}^3$, the equation

$$ax^2+by^2=cz^2$$

is partition regular.

We do not make explicit what "most" means, but let us say that a subset of Rado's triples with relative density 1 with respect to the standard density would qualify. At the moment, it is not even known that there is a single Rado triple for which we have partition regularity.

7.3. Multiple recurrence results for multiplicative actions

Although the methods of this article are particularly well suited to partition regularity problems of pairs, they seem inadequate on their own to deal with problems of triples. The reason is that we lack a representation result for sequences such as

$$(x,y,z)\mapsto d(x^{-1}\Lambda\cap y^{-1}\Lambda\cap z^{-1}\Lambda),$$

which was our starting point in our approach to partition regularity of pairs. It is understood that a useful representation similar to the Bochner–Herglotz theorem

should be very hard to prove and that new sequences, not just multiplicative functions, should be used.

An alternative approach, which has been very successful for translation-invariant patterns of arbitrary length, is to use the formalism of ergodic theory to provide sufficient statements for partition regularity and then to develop the necessary toolbox in ergodic theory to prove them. The sufficient statements in ergodic theory are described next.

We say that (X, μ, T_n) is a *multiplicative action* if $(X, \mathcal{X}, \mu)$ is a probability space, and for $n \in \mathbb{Z}$, the transformations $T_n : X \to X$ are invertible, measure preserving, and satisfy the relation $T_0 = T_1 := \mathrm{id}$ and $T_m \circ T_n = T_{mn}$ for all $m, n \in \mathbb{Z}$. Using a variant of the correspondence principle of Furstenberg [24] that appears in [5], it turns out that in order to show that the equation $x^2 + y^2 = z^2$ is partition regular, it suffices to prove the following multiple recurrence result.

Problem 3. Let (X, μ, T_n) be a multiplicative action an A a measurable set such that $X = \bigcup_{j=1}^{k} T_j^{-1} A$ for some $k \in \mathbb{N}$. Show that

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \mu(T^{-1}_{m^2-n^2} A \cap T^{-1}_{2mn} A \cap T^{-1}_{m^2+n^2} A) > 0. \tag{7.1}$$

Unlike the partition regularity results for pairs, such as those in Theorem 2.4, we cannot expect to have density regularity results for the equation $ax^2 + by^2 = cz^2$ when $a + b \neq c$. If, say, $a = b = c = 1$, an example by Bergelson shows that for suitably chosen $\alpha \in \mathbb{R}$, the set $\Lambda := \{n \in \mathbb{N} : \{n^2\alpha\} \in [1/5, 2/5)\}$ has positive multiplicative density but does not contain a Pythagorean triple (this problem already appears for the equation $x + y = z$). This is the reason why in Problem 3 we cannot expect to have a multiple recurrence property for any set A with positive measure. On the other hand, if the action is finitely generated, i.e., the set $\{T_p, p \in \mathbb{P}\}$ is finite, we expect that (7.1) holds for all sets A with positive measure.

In order to show that the equation $x^2 + y^2 = 2z^2$ is partition regular (in fact, density regular with respect to multiplicative density), it suffices to prove the following.

Problem 4. Let (X, μ, T_n) be a multiplicative action, and let $A \in \mathcal{X}$ with $\mu(A) > 0$. Show that

$$\liminf_{N\to\infty} \mathbb{E}_{m,n\in[N]}\, \mu(T^{-1}_{m^2-n^2+2mn} A \cap T^{-1}_{m^2-n^2-2mn} A \cap T^{-1}_{m^2+n^2} A) > 0.$$

The problem is open even if the action is finitely generated, and we omit one of the three terms in the intersection.

7.4. Higher-degree homogeneous polynomials

As we have seen in Section 5.6, for every irreducible binary quadratic form $P(m,n)$, all pretentious multiplicative functions enjoy strong concentration properties. Does a similar fact hold for not necessarily homogeneous polynomials or higher-degree homogeneous polynomials? By slightly modifying an example of Klurman [42, Lemma 2.1], one can construct pretentious (and aperiodic) multiplicative functions $f:\mathbb{N}\to\{-1,1\}$ so that the averages

$$\mathbb{E}_{n\in[N]}\, f(n^2+1)$$

do not converge. It follows that in this case, we cannot have concentration estimates similar to those given in Theorem 5.4. Although, as we have shown in Theorem 5.6, such counterexamples do not work for, say, $P(m,n)=m^2+n^2$, it seems likely that they can be found for higher-degree irreducible homogeneous polynomials, but this is not so easy to achieve.

Problem 5. Let $a\in\mathbb{Z}$ and $d\geq 3$ be such that the polynomial m^d+an^d is irreducible. Show that there exists a pretentious completely multiplicative function $f:\mathbb{N}\to\{-1,1\}$ such that the averages

$$\mathbb{E}_{m,n\in[N]}\, f(m^d+an^d)$$

do not converge as $N\to\infty$.

We say that a sequence $a:\mathbb{N}^2\to\mathbb{Z}$ is *good for single recurrence of multiplicative actions* if the following holds: For every multiplicative action (X,μ,T_n) and $A\in\mathcal{X}$ with $\mu(A)>0$, we have

$$\mu(A\cap T^{-1}_{a(m,n)}A)>0$$

for some $m,n\in\mathbb{N}$ such that $a(m,n)\neq 0$. Is the polynomial $P(m,n)=m^3+2n^3$ good for single recurrence of multiplicative actions? The following is a more general problem.

Problem 6. Let $P\in\mathbb{Z}[m,n]$ be a homogeneous polynomial with $P(1,0)=1$. Is it true that for every multiplicative action (X,μ,T_n) and $A\in\mathcal{X}$ with $\mu(A)>0$, we have

$$\liminf_{N\to\infty}\mathbb{E}_{m,n\in[N]}\,\mu(A\cap T^{-1}_{P(m,n)}A)>0\,?$$

We cannot remove the assumption $P(1,0)=1$ since the set $\{2n^2:n\in\mathbb{N}\}$ can easily be seen to be bad for measurable recurrence [18, Example 3.11].

It follows from [18, Theorem 2.1 and Example 5.15] that the answer is positive when $\deg(P)\leq 2$. It seems very likely that, using the methodology developed

in [22], we can also get a positive answer for reducible polynomials of degree 3. But for $P(m,n) = m^3 + 2n^3$, it is not even clear how to show that for every completely multiplicative function $f:\mathbb{N}\to\{-1,1\}$, we have $f(m^3+2n^3)=1$ for a set of $m,n\in\mathbb{N}$ with positive lower density.

Acknowledgments. I am grateful to my co-authors on articles [21–23], B. Host, O. Klurman, and J. Moreira; without them, these works would not have been possible. I would also like to thank A. Mountakis and D. Sclosa for a number of useful comments.

Funding. The author was supported by the research grant ELIDEK HFRI-Next-GenerationEU-15689.

References

[1] R. Alweiss, Monochromatic sums and products over $\mathbb{Q}$. 2023, arXiv:2307.08901v5

[2] R. Alweiss, Monochromatic sums and products of polynomials. *Discrete Anal.* (2024), article no. 5, 7 pp. Zbl 08092730 MR 4793285

[3] J. M. Barrett, M. Lupini, and J. Moreira, On Rado conditions for nonlinear Diophantine equations. *European J. Combin.* **94** (2021), article no. 103277, 20 pp. Zbl 1468.11076 MR 4192111

[4] V. Bergelson, Ergodic Ramsey theory. In *Logic and combinatorics (Proc. AMS-IMS-SIAM Conf., Arcata, Calif., 1985)*, pp. 63–87, Contemp. Math. 65, American Mathematical Society, Providence, RI, 1987 Zbl 0642.10052 MR 0891243

[5] V. Bergelson, Multiplicatively large sets and ergodic Ramsey theory. *Israel J. Math.* **148** (2005), no. 1, 23–40 Zbl 1093.11003 MR 2191223

[6] V. Bergelson, J. H. Johnson, Jr., and J. Moreira, New polynomial and multidimensional extensions of classical partition results. *J. Combin. Theory Ser. A* **147** (2017), 119–154 Zbl 1352.05023 MR 3589892

[7] V. Bergelson and A. Leibman, Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J. Amer. Math. Soc.* **9** (1996), no. 3, 725–753 Zbl 0870.11015 MR 1325795

[8] M. Bowen and M. Sabok, Monochromatic products and sums in the rationals. *Forum Math. Pi* **12** (2024), article no. e17, 12 pp. Zbl 1551.05400 MR 4812188

[9] T. D. Browning and S. M. Prendiville, A transference approach to a Roth-type theorem in the squares. *Int. Math. Res. Not. IMRN* **2017** (2017), no. 7, 2219–2248 Zbl 1405.11131 MR 3658196

[10] J. Chapman, Partition regularity for systems of diagonal equations. *Int. Math. Res. Not. IMRN* **2022** (2022), no. 17, 13272–13316 Zbl 1507.11016 MR 4475278

[11] J. Chapman and S. Chow, Generalised Rado and Roth criteria. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **26** (2025), no. 3, 1263–1312 Zbl 08129930 MR 4991571

[12] D. Charamaras, A. Mountakis, and K. Tsinas, On multiplicative recurrence along linear patterns. *J. Lond. Math. Soc. (2)* **112** (2025), no. 3, article no. e70292, 53 pp. Zbl 08104426 MR 4960398

[13] S. Chow, S. Lindqvist, and S. Prendiville, Rado's criterion over squares and higher powers. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 6, 1925–1997 Zbl 1497.11028 MR 4244520

[14] P. Csikvári, K. Gyarmati, and A. Sárközy, Density and Ramsey type results on algebraic equations with restricted solution sets. *Combinatorica* **32** (2012), no. 4, 425–449 Zbl 1286.11026 MR 2965285

[15] H. Daboussi and H. Delange, On multiplicative arithmetical functions whose modulus does not exceed one. *J. Lond. Math. Soc. (2)* **26** (1982), no. 2, 245–264 Zbl 0499.10052 MR 0675168

[16] M. Di Nasso and L. Luperi Baglini, Ramsey properties of nonlinear Diophantine equations. *Adv. Math.* **324** (2018), 84–117 Zbl 1376.05159 MR 3733882

[17] M. Di Nasso and M. Riggio, Fermat-like equations that are not partition regular. *Combinatorica* **38** (2018), no. 5, 1067–1078 Zbl 1438.11075 MR 3884778

[18] S. Donoso, A. N. Le, J. Moreira, and W. Sun, Additive averages of multiplicative correlation sequences and applications. *J. Anal. Math.* **149** (2023), no. 2, 719–761 Zbl 1522.37065 MR 4594405

[19] N. Frantzikinakis, Some open problems on multiple ergodic averages. *Bull. Hellenic Math. Soc.* **60** (2016), 41–90 Zbl 1425.37004 MR 3613710

[20] N. Frantzikinakis and B. Host, Asymptotics for multilinear averages of multiplicative functions. *Math. Proc. Cambridge Philos. Soc.* **161** (2016), no. 1, 87–101 Zbl 1371.11134 MR 3505671

[21] N. Frantzikinakis and B. Host, Higher order Fourier analysis of multiplicative functions and applications. *J. Amer. Math. Soc.* **30** (2017), no. 1, 67–157 Zbl 1355.11094 MR 3556289

[22] N. Frantzikinakis, O. Klurman, and J. Moreira, Partition regularity of generalized Pythagorean pairs. 2024, arXiv:2407.08360v2

[23] N. Frantzikinakis, O. Klurman, and J. Moreira, Partition regularity of Pythagorean pairs. *Forum Math. Pi* **13** (2025), article no. e5, 52 pp. Zbl 1558.05134 MR 4862834

[24] H. Furstenberg, Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J. Anal. Math.* **31** (1977), no. 1, 204–256 Zbl 0347.28016 MR 0498471

[25] W. T. Gowers, A new proof of Szemerédi's theorem. *Geom. Funct. Anal.* **11** (2001), no. 3, 465–588 Zbl 1028.11005 MR 1844079

[26] R. Graham, Some of my favorite problems in Ramsey theory. In B. Landman, M. B. Nathanson, J. Nesetril, R. J. Nowakowski and C. Pomerance (Eds.), *Combinatorial number theory*, pp. 229–236, De Gruyter, Berlin, 2007 Zbl 1124.05087 MR 2337049

[27] R. Graham, Old and new problems and results in Ramsey theory. In E. Győri, G. O. H. Katona, L. Lovász, and G. Sági (Eds.), *Horizons of combinatorics*, pp. 105–118, Bolyai Soc. Math. Stud. 17, Springer, Berlin, 2008 Zbl 1178.05094 MR 2432529

[28] A. Granville and K. Soundararajan, Large character sums: pretentious characters and the Pólya–Vinogradov theorem. *J. Amer. Math. Soc.* **20** (2007), no. 2, 357–384 Zbl 1210.11090 MR 2276774

[29] A. Granville and K. Soundararajan, Pretentious multiplicative functions and an inequality for the zeta-function. In J.-M. De Koninck, A. Granville, and F. Luca (Eds.), *Anatomy of integers*, pp. 191–197, CRM Proc. Lecture Notes 46, American Mathematical Society, Providence, RI, 2008 Zbl 1187.11032 MR 2437976

[30] A. Granville and K. Soundararajan, Multiplicative number theory: the pretentious approach. Book manuscript in preparation

[31] B. Green, 100 open problems. Manuscript

[32] B. Green and T. Tao, An inverse theorem for the Gowers $U^3(G)$ norm. *Proc. Edinb. Math. Soc. (2)* **51** (2008), no. 1, 73–153 Zbl 1202.11013 MR 2391635

[33] B. Green and T. Tao, Quadratic uniformity of the Möbius function. *Ann. Inst. Fourier (Grenoble)* **58** (2008), no. 6, 1863–1935 Zbl 1160.11017 MR 2473624

[34] B. Green and T. Tao, The Möbius function is strongly orthogonal to nilsequences. *Ann. of Math. (2)* **175** (2012), no. 2, 541–566 Zbl 1347.37019 MR 2877066

[35] B. Green, T. Tao, and T. Ziegler, An inverse theorem for the Gowers $U^{s+1}[N]$-norm. *Ann. of Math. (2)* **176** (2012), no. 2, 1231–1372 Zbl 1282.11007 MR 2950773

[36] B. J. Green and S. Lindqvist, Monochromatic solutions to $x + y = z^2$. *Canad. J. Math.* **71** (2019), no. 3, 579–605 Zbl 1439.11076 MR 3952491

[37] K. Gyarmati and I. Z. Ruzsa, A set of squares without arithmetic progressions. *Acta Arith.* **155** (2012), no. 1, 109–115 Zbl 1293.11015 MR 2982433

[38] G. Halász, Über die Mittelwerte multiplikativer zahlentheoretischer Funktionen. *Acta Math. Acad. Sci. Hungar.* **19** (1968), 365–403 Zbl 0165.05804 MR 0230694

[39] M. J. H. Heule, O. Kullmann, and V. W. Marek, Solving and verifying the Boolean Pythagorean triples problem via cube-and-conquer. In *Proceedings of the 19th International Conference on Theory and Applications of Satisfiability Testing—SAT 2016*, pp. 228–245, Lecture Notes in Comput. Sci. 9710, Springer, Cham, 2016 Zbl 1403.68226 MR 3534782

[40] I. Kátai, A remark on a theorem of H. Daboussi. *Acta Math. Hungar.* **47** (1986), no. 1–2, 223–225 Zbl 0607.10034 MR 0836415

[41] A. Khalfalah and E. Szemerédi, On the number of monochromatic solutions of $x + y = z^2$. *Combin. Probab. Comput.* **15** (2006), no. 1–2, 213–227 Zbl 1093.11017 MR 2195583

[42] O. Klurman, Correlations of multiplicative functions and applications. *Compos. Math.* **153** (2017), no. 8, 1622–1657 Zbl 1434.11202 MR 3705270

[43] O. Klurman and A. P. Mangerel, Rigidity theorems for multiplicative functions. *Math. Ann.* **372** (2018), no. 1–2, 651–697 Zbl 1457.11131 MR 3856825

[44] O. Klurman, A. P. Mangerel, C. Pohoata, and J. Teräväinen, Multiplicative functions that are close to their mean. *Trans. Amer. Math. Soc.* **374** (2021), no. 11, 7967–7990 Zbl 1487.11086 MR 4328688

[45] E. Lamb, Two-hundred-terabyte maths proof is largest ever. *Nature* **534** (2016), no. 7605, 17–18

[46] H. Lefmann, On partition regular systems of equations. *J. Combin. Theory Ser. A* **58** (1991), no. 1, 35–53 Zbl 0746.05076 MR 1119700

[47] L. Matthiesen, Generalized Fourier coefficients of multiplicative functions. *Algebra Number Theory* **12** (2018), no. 6, 1311–1400 Zbl 1444.11201 MR 3864201

[48] J. Moreira, Monochromatic sums and products in $\mathbb{N}$. *Ann. of Math. (2)* **185** (2017), no. 3, 1069–1090 Zbl 1430.05128 MR 3664819

[49] P. P. Pach, Monochromatic solutions to $x + y = z^2$ in the interval $[N, cN^4]$. *Bull. Lond. Math. Soc.* **50** (2018), no. 6, 1113–1116 Zbl 1439.11077 MR 3891947

[50] S. Prendiville, Counting monochromatic solutions to diagonal Diophantine equations. *Discrete Anal.* (2021), article no. 14, 47 pp. Zbl 1511.11035 MR 4317277

[51] R. Rado, Studien zur Kombinatorik. *Math. Z.* **36** (1933), no. 1, 424–470 Zbl 0006.14603 MR 1545354

[52] A. Sárközy, On difference sets of sequences of integers. I. *Acta Math. Acad. Sci. Hungar.* **31** (1978), no. 1–2, 125–149 Zbl 0387.10033 MR 0466059

[53] I. Schur, Über die Kongruenz $x^m + y^m \equiv z^m \pmod p$. *Jahresber. Dtsch. Math.-Ver.* **25** (1916) 114–117 Zbl 46.0193.02

[54] W. Sun, A structure theorem for multiplicative functions over the Gaussian integers and applications. *J. Anal. Math.* **134** (2018), no. 1, 55–105 Zbl 1418.11025 MR 3771476

[55] W. Sun, Sarnak's conjecture for nilsequences on arbitrary number fields and applications. *Adv. Math.* **415** (2023), article no. 108883, 91 pp. Zbl 1525.11105 MR 4543074

[56] T. Tao, The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi* **4** (2016), article no. e8, 36 pp. Zbl 1383.11116 MR 3569059

[57] T. Tao and J. Teräväinen, The structure of logarithmically averaged correlations of multiplicative functions, with applications to the Chowla and Elliott conjectures. *Duke Math. J.* **168** (2019), no. 11, 1977–2027 Zbl 1436.11115 MR 3992031

[58] G. Tenenbaum, *Introduction to analytic and probabilistic number theory*. Cambridge Stud. Adv. Math. 46, Cambridge University Press, Cambridge, 1995, 448 pp. Zbl 0831.11001 MR 1342300

[59] B. L. van der Waerden, Beweis einer Baudetschen Vermutung. *Nieuw Arch. Wiskd.* **15** (1927), 212–216 Zbl 53.0073.12

Nikos Frantzikinakis
Department of Mathematics and Applied Mathematics, University of Crete, Voutes Campus, 70013 Heraklion, Greece; frantzikinakis@uoc.gr, frantzikinakis@gmail.com

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/19

Symmetries of Grothendieck rings in representation theory

David Hernandez

Abstract. We review certain new symmetries of Grothendieck rings that have emerged in representation theory.

Categories of representations of groups, or more generally, bialgebras and quantum groups, are endowed with monoidal structures. The Grothendieck ring of such a category has a basis consisting of classes of simple objects, with structure constants defined as multiplicities of simple representations in the decomposition of tensor products. The resulting rings together with their bases can have a very intricate structure, and many of them are fundamental with important applications (for example, leading to algorithms to compute the dimension of simple representations).

In recent years, new symmetries of Grothendieck rings have emerged, which can be used to better understand these rings. We will discuss these symmetries and illustrate some of their applications.

First, cluster symmetries give powerful tools to investigate the structure of the Grothendieck rings [19]. As an application, we have obtained the proof of long-standing conjectures on character formulas for simple finite-dimensional representations of quantum affine algebras (which are infinite-dimensional quantum groups) [14].

Second, motivated by symmetries of quantum integrable models arising in physics, we established remarkable relations in associated Grothendieck rings [8], generalizing the celebrated Baxter relations; these relations led us to the discovery of a surprising new action of the Weyl group on these Grothendieck rings [9]. These Weyl group symmetries are also closely related to the cluster symmetries [15].

Mathematics Subject Classification 2020: 17B67 (primary); 13F60, 17B10, 17B37, 35J25, 82B23 (secondary).
Keywords: Grothendieck rings, monoidal categories, quantum groups, cluster algebras, categorification.

1. Grothendieck group

To start with, let us recall the general construction of Grothendieck groups.

1.1. Grothendieck group of an additive category

Let $\mathcal{C}$ be an additive category with distinguished short exact sequences

$$A \to B \to C. \tag{1.1}$$

For example, one may consider the case of an abelian category with all ordinary exact sequences.

Definition 1.1. The Grothendieck group $K(\mathcal{C})$ of $\mathcal{C}$ is the group generated by the elements $[M]$ for each isomorphism class of objects M in $\mathcal{C}$, with the relations

$$[A] + [C] = [B]$$

for each exact sequence (1.1).

We obtain the important algebraic structure $K(\mathcal{C})$ attached to the category $\mathcal{C}$. We will discuss many examples in this lecture.

The main original motivation appeared in algebraic geometry. Consider the $\mathcal{C}$ category of coherent sheaves on an algebraic variety X. The Grothendieck group

$$K(X) = K(\mathcal{C})$$

is one of the starting points of K-theory in algebraic geometry. Historically, such constructions were first studied by Grothendieck for categories of coherent and locally free sheaves on schemes, in proving the Riemann–Roch theorem.

Another basic but fundamental example is obtained for a field k: the category $\mathcal{C} = \mathcal{V}$ of finite-dimensional k-vector spaces. The isomorphism classes are parametrized by the dimension. Then $\mathcal{V}$ is semi-simple, and the direct sum descends to the sum in the Grothendieck group, that is, for V and W in $\mathcal{V}$, we have

$$[V \oplus W] = [V] + [W].$$

This implies that $K(\mathcal{V})$ is freely generated by the class $[k]$ of the one-dimensional vector space

$$K(\mathcal{V}) = \mathbb{Z}[k] \simeq \mathbb{Z}$$

as we have

$$[V] = \dim(V) \cdot [k]$$

for V in $\mathcal{V}$.

1.2. Grothendieck groups in representation theory

We now discuss the Grothendieck group of finite-dimensional representations of an algebra.

Let $\mathcal{A}$ be an algebra over a field k and $\mathcal{F}$ be the category of finite-dimensional representations of $\mathcal{A}$.

We have the following classical and fundamental result.

Theorem 1.2 (Jordan–Hölder series). *Each object M in $\mathcal{F}$ admits a series of subobjects*

$$M = M_0 \supset M_1 \supset \cdots \supset M_N = \{0\}$$

with M_i/M_{i+1} simple object.

Consider a simple object V. Then the number $n_V(M)$ of occurrences of V as a quotient M_i/M_{i+1}, depends only on M and not on the series of subobjects.

Such a series is called a composition series or a Jordan–Hölder series of the representation. A simple quotient M_i/M_{i+1} is called a composition factor of M, and its number of occurrences is called its multiplicity.

For completeness, let us briefly recall the arguments of the proof.

Proof. The existence of the series is obtained by induction on the dimension of the representation. The uniqueness of $n_V(M)$ is also proved by induction on the dimension: it is obvious for a simple representation. In general, assume that we have two series M_i and M_i' as above. If $M_1 = M_1'$, we can conclude using the induction hypothesis for this representation. Otherwise,

$$M \supset M_1 \supset M_1 \cap M_1'$$

and

$$M \supset M_1' \supset M_1 \cap M_1'$$

are the first few terms of composition series with the same composition factors. Using the induction hypothesis, they have the same composition factors. ∎

One can then define the length of a module M as the sum

$$l(M) = \sum_{V \text{ simple class in } \mathcal{F}} n_V(M).$$

As a consequence of Theorem 1.2, we obtain that

$$K(\mathcal{F}) = \bigoplus_{V \text{ simple class}} \mathbb{Z}[V]$$

has a canonical basis parametrized by simple classes. The Grothendieck group is a group with a distinguished basis.

The coefficients on the basis are the multiplicities of simple representations, that is, for M object in $\mathcal{F}$, one has

$$[M] = \sum_{V \text{ simple class}} n_V(M)[V].$$

In general, an element

$$\sum_{V \text{ simple class}} n_V[V]$$

with all $n_V \geq 0$ is called the class of an *actual* representation.

If one of the $n_V < 0$, this is the class of a *virtual* representation. The Grothendieck group allows us to treat in a uniform way actual and virtual representations.

2. Monoidal categories

We now assume that an abelian category $\mathcal{C}$ has a monoidal structure, that is, we have tensor products $M \otimes N$ for M, N objects in $\mathcal{C}$ with isomorphisms

$$(M \otimes N) \otimes P \simeq M \otimes (N \otimes P)$$

as well as a unit object $\mathbf{1}$ in $\mathcal{C}$ so that for any object M in $\mathcal{C}$,

$$M \otimes \mathbf{1} \simeq \mathbf{1} \otimes M \simeq M.$$

We have additional compatibility properties with the abelian structure (see [28]).

Remark 2.1. We do not assume that the category is necessarily a tensor category, that is, we do not assume that the category admits dualities compatible with the monoidal structure.

Let us discuss important examples of monoidal categories.

(1) The category $\mathcal{V}$ of finite-dimensional k-vector spaces has a natural monoidal structure.

(2) For a group G, the category of finite-dimensional representations of G (that is, of its group algebra $k[G]$) has a monoidal structure. The action of $g \in G$ on a pure tensor $v \otimes w \in V \otimes W$ in a tensor product of representations is defined by

$$g \cdot (v \otimes w) = (g \cdot v) \otimes (g \cdot w).$$

(3) For a Lie algebra $\mathfrak{g}$, the category of finite-dimensional representations of $\mathfrak{g}$ (that is, of its universal enveloping algebra $\mathcal{U}(\mathfrak{g})$) has a monoidal structure. The action of $g \in \mathfrak{g}$ on a pure tensor $v \otimes w \in V \otimes W$ in a tensor product of representations is defined by

$$g \cdot (v \otimes w) = (g \cdot v) \otimes w + v \otimes (g \cdot w).$$

This is an infinitesimal version of the group action above.

(4) More generally: for a Hopf algebra H, that is, an algebra with a coproduct and a counit,

$$\Delta : H \to H \otimes H \quad \text{and} \quad \varepsilon : H \to k.$$

The category of finite-dimensional representations of H has a monoidal structure. This includes all examples above and new examples, in particular the Drinfeld–Jimbo *quantum groups*

$$\mathcal{U}_q(\mathfrak{g})$$

associated to $q \in \mathbb{C}^* \setminus \{-1, 1\}$ quantum parameter and $\mathfrak{g}$ simple Lie algebra. It is a q-deformation of $\mathcal{U}(\mathfrak{g})$, originated from mathematical physics. For example, $\mathcal{U}_q(\mathfrak{sl}_2)$ is the algebra by generators E, F, $K^{\pm 1}$ and relations

$$KE = q^2 EK, \quad KF = q^{-2} FK \quad \text{and} \quad [E, F] = \frac{K - K^{-1}}{q - q^{-1}}.$$

(5) Hecke algebras, and their variations such as quiver Hecke algebras (as defined by Khovanov–Lauda and Rouquier), give rise to categories of representations which have a monoidal structure via convolutions.

(6) In various geometric contexts, convolution diagrams

$$X_1 \xleftarrow{p} X_2 \xrightarrow{q} X_3$$

of regular morphisms p, q between remarkable varieties allow us to consider convolution products and then monoidal structures on certain categories of sheaves (see, for instance, [1, 5] and references therein).

3. Grothendieck ring

3.1. Grothendieck ring of a monoidal category

Let us define the Grothendieck ring of $\mathcal{C}$ monoidal category.

Proposition 3.1. *The Grothendieck group $K(\mathcal{C})$ of $\mathcal{C}$ inherits a ring structure such that*

$$[M] \cdot [N] = [M \otimes N]$$

for M, N objects in $\mathcal{C}$ and the class $[\mathbf{1}] \in K(\mathcal{C})$ is the neutral element.

Example 3.2. For $\mathcal{C} = \mathcal{V}$ the category of finite-dimensional vector spaces on a field k, the group isomorphism

$$K(\mathcal{C}) \simeq \mathbb{Z}$$

is a ring isomorphism. This follows, for example, from the formula $\dim(V \otimes W) = \dim(V)\dim(W)$.

Remark 3.3. If the category is braided, that is,

$$M \otimes N \simeq N \otimes M$$

for all objects M, N in $\mathcal{C}$, then the Grothendieck ring is commutative. The converse statement is false: there are non-braided monoidal categories whose Grothendieck ring is commutative. We will see examples below.

3.2. Grothendieck rings in representation theory

Let $\mathcal{C}$ be the category of finite-dimensional representations of a bialgebra H. The Grothendieck ring

$$K(\mathcal{C}) = \bigoplus_{V \text{ simple}} \mathbb{Z}[V]$$

has a canonical basis $([V])_{V \text{ simple}}$, and the constant structures on this basis are positive constant structures: they are given by the multiplicities of simple modules in tensor products (sometimes called Clebsch–Gordan coefficients or fusion coefficients).

Remark 3.4. If H is a Hopf algebra, $\mathcal{C}$ is a tensor category, and dualities induce ring automorphisms of $K(\mathcal{C})$, which preserve the canonical basis.

Example 3.5. Let $\mathcal{C}$ be the category of finite-dimensional representations of the Lie algebra $\mathfrak{sl}_2(\mathbb{C})$. Then we have a ring isomorphism

$$K(\mathcal{C}) \simeq \mathbb{Z}[X]$$

with the polynomial ring in one variable X. Here,

$$X = [V], \quad \text{where } V = \mathbb{C}^2 \text{ is the natural representation.}$$

It is well known that $\mathfrak{sl}_2(\mathbb{C})$ has a simple complex representation of dimension $n+1$ for any $n \geq 0$ and that this representation is unique up to isomorphism. Let us denote by Q_n the isomorphism class of this representation in $K(\mathcal{C})$. Then the basis $(Q_n)_{n\geq 0}$ of simple classes satisfies the induction relation

$$Q_n^2 = Q_{n+1}Q_{n-1} + 1. \tag{3.1}$$

One can compute the canonical basis of simple classes

$$(Q_0, Q_1, Q_2, \ldots) = (1, X, X^2 - 1, X^3 - 2X, \ldots),$$

which is different from the standard monomial basis of $\mathbb{Z}[X]$.

Remark 3.6. Note that, independently of Grothendieck rings, it is not clear a priori that the inductive sequence (3.1) with initial conditions $Q_0 = 1$ and $Q_1 = X$, a priori in $\mathbb{Q}(X)$, defines a sequence in the ring $\mathbb{Z}[X]$.

3.3. Applications of Grothendieck rings in representation theory

The study of Grothendieck rings of categories of representations has led to interesting developments in various directions. Let us list some of them.

Representation theory. There is a close relation between the understanding of the global structure of Grothendieck rings and the description of the structure of simple representations. For example, if one knows the structure of certain simple representations, one may extract information on the simple constituents of their tensor products.

Example 3.7. This approach will be discussed below to compute the characters of simple representations of quantum affine algebras from the characters of fundamental representations, following [29] and then [14, 16].

Categorification. One can sometimes realize an algebra $\mathcal{A}$ as a Grothendieck ring:

$$\mathcal{A} \simeq K(\mathcal{C}).$$

The categorified structure $\mathcal{A}$ inherits a canonical basis from the canonical basis of the Grothendieck ring.

Example 3.8. The coordinate ring $\mathbb{C}[N]$, with N a unipotent subgroup of a Lie group G, can be realized from representations of quiver Hecke algebras [27, 32] or of quantum affine algebras [20], leading to a categorical interpretation of its dual canonical basis. A basic example is also explained in Example 3.5.

Quantum integrable models. Certain Grothendieck rings can be identified with the ring of commuting operators of a quantum integrable model via the transfer-matrix construction. Indeed, a quantum integrable model involves a space W (the quantum space) with a family of commuting operators on this space (including quantum Hamiltonians). The transfer-matrix construction defines a ring morphism

$$T : K_0(\mathcal{C}) \to \mathrm{End}(W),$$

where $\mathcal{C}$ is the category of representations of an algebra of symmetry of the system. Hence, any relation in $K_0(\mathcal{C})$ (a "universal relation") gives relations between the eigenvalues of the operators on V for any related quantum integrable model.

Example 3.9. The famous XXZ model can be realized using the representation theory of the quantum affine algebra associated to $\mathfrak{sl}_2$ (see below).

The aim of this lecture is to discuss certain symmetries of Grothendieck rings and their consequences. The main examples of symmetries we present are cluster symmetries and Weyl group symmetries.

4. Cluster algebra symmetries

We discuss the first symmetries of Grothendieck rings in this lecture in the context of cluster algebras.

4.1. A quick review of cluster algebras

The theory of cluster algebras was introduced by Fomin–Zelevinsky [7]. A cluster algebra is a commutative algebra with a distinguished set of generators grouped into overlapping subsets (the clusters, obtained by an inductive process called mutation). Each element of the cluster algebra can be expressed as a rational fraction of the elements of a cluster. From a geometric point of view, one might think about a variety with various sets of local coordinates, and each element of the algebra of functions over this variety can be expressed locally in terms of the local coordinates. Then the mutations can be understood as a regular change of coordinates.

More precisely, the cluster algebra $\mathcal{A}_Q$ is a commutative algebra associated with a quiver Q (without loops or 2-cycles). It has a distinguished set of generators (the cluster variables) defined by a combinatorial process. Cluster algebras have many incarnations in various fields, and in particular in representation theory as we will discuss below.

The cluster algebra $\mathcal{A}_Q$ attached to the quiver Q (with a set of vertices Q_0) is a subring of the field

$$\mathcal{F} = \mathbb{Q}(X_i)_{i \in Q_0},$$

with free variables X_i which are called the initial cluster variables (they form the initial cluster). The cluster algebra $\mathcal{A}_Q$ is defined as the subalgebra of $\mathcal{F}$ generated by the cluster variables, obtained inductively from the initial variables by an inductive process called mutations. For example, the first-step mutated variables X_i^* are defined by the formula

$$X_i X_i^* = \prod_{j \to i} X_j + \prod_{j \leftarrow i} X_j,$$

where the arrows $\to$ and $\leftarrow$ are those in the initial quiver Q (then the quiver also gets mutated in the process). The number of cluster variables is not necessarily finite. In addition, the cluster variables are grouped into overlapping subsets called clusters, which are all in bijection with Q_0. The cluster monomials are defined as the monomials in the cluster variables of the same cluster. In some situations, one may have additional non-mutable cluster variables: they are called frozen variables, and they belong to all clusters.

One of the fundamental properties of cluster algebras is the Laurent phenomenon: any cluster variable can be expressed as a Laurent polynomial in the cluster variables of a given seed. In the initial seed, this can be written as

$$\mathcal{A}_Q \subset \mathbb{Z}[X_i^{\pm 1}]_{i \in Q_0}.$$

Example 4.1. Consider a Q quiver of type A_2

$$\bullet \longrightarrow \bullet.$$

We have the initial cluster variables (X_1, X_2) and five cluster variables:

$$X_1, X_2, \frac{1+X_2}{X_1}, \frac{1+X_1}{X_2}, \frac{1+X_1+X_2}{X_1 X_2}.$$

There are also five clusters:

$$(X_1, X_2), \quad \left(\frac{1+X_2}{X_1}, X_2\right), \quad \left(\frac{1+X_2}{X_1}, \frac{1+X_1+X_2}{X_1 X_2}\right),$$
$$\left(\frac{1+X_1}{X_2}, \frac{1+X_1+X_2}{X_1 X_2}\right), \quad \left(\frac{1+X_1}{X_2}, X_1\right).$$

Remark 4.2. This example should not be misleading: in general, the number of clusters is not equal to the number of cluster variables.

4.2. Monoidal categorification of cluster algebras

In this lecture, we focus on the relation between cluster algebras and monoidal categories (there are also important relations with additive categories; see [26] and references therein).

Definition 4.3 ([19]). A monoidal category $\mathcal{M}$ is a *monoidal categorification* of a cluster algebra $\mathcal{A}$ if there exists a ring isomorphism

$$\phi : \mathcal{A} \to K(\mathcal{M})$$

so that the cluster monomials are sent to simple classes.

In [19], a stronger version is also discussed, where it is required that ϕ induces a bijection between the cluster monomials and the real simple classes (whose tensor square is also simple). In the definition above, the simple modules corresponding to cluster monomials are called reachable modules. They have the property to be real, as the square of a cluster monomial is a cluster monomial.

A monoidal categorification can be regarded as a cluster symmetry of the Grothendieck ring or as a categorical realization of the cluster algebra. It gives useful information on the monoidal category: for example, it points out remarkable simple representations (corresponding to cluster variables), and it gives a factorization of simple representations corresponding to cluster monomials (in the same way as cluster monomials factorize as a product of cluster variables).

The original examples [19] were obtained from the finite-dimensional representations of quantum affine algebras. Many developments in monomial categorifications in this context have followed; one can cite, for instance, [3,24,30,31] and in different contexts [4].

In all known examples, the cluster symmetry is related to the defect of symmetry of the category which produces relations in the Grothendieck ring: we have objects V, W so that

$$V \otimes W \text{ is not isomorphic to } W \otimes V$$

but

$$[V \otimes W] = [W \otimes V]$$

in the Grothendieck ring (examples will be given below).

4.3. Quantum affine algebras

Consider $\mathfrak{g}$ a complex simple Lie algebra (for example, $\mathfrak{g} = \mathfrak{sl}_2$) of rank n (that is, $n = 1$ for $\mathfrak{g} = \mathfrak{sl}_2$). Then we can construct the affine Kac–Moody algebra $\widehat{\mathfrak{g}}$, which is a central extension of the loop algebra[1]

$$\mathfrak{g} \otimes \mathbb{C}[t, t^{-1}].$$

[1] We do not consider the derivation element in this lecture.

For $q \in \mathbb{C}^*$ a quantum parameter (which is assumed not to be a root of unity), we have the corresponding quantum affine algebra $\mathcal{U}_q(\widehat{\mathfrak{g}})$. It is a Hopf algebra and a q-deformation of the universal enveloping algebra $\mathcal{U}(\widehat{\mathfrak{g}})$.

Let $\mathcal{C}$ be the category of finite-dimensional representations of $\mathcal{U}_q(\widehat{\mathfrak{g}})$: it has a very rich structure. This monoidal category is neither semi-simple nor braided. The simple objects of $\mathcal{C}$ have been parametrized in terms of Drinfeld polynomials by Chari–Pressley, that is, of n-tuples $(P_i(z))_{1\le i\le n}$ of monic polynomials. In particular, for $1 \le i \le n$ and $a \in \mathbb{C}^*$, we have the fundamental representation $V_i(a)$ which corresponds to the n-tuple $(1, \dots, 1, 1 - za, 1, \dots, 1)$ with a degree 1 polynomial in position i.

Theorem 4.4 ([12]). *The Grothendieck ring $K(\mathcal{C})$ is commutative and polynomial:*

$$K(\mathcal{C}) \simeq \mathbb{Z}[X_{i,a}]_{1\le i\le n,\, a\in\mathbb{C}^*},$$

with $X_{i,a} = [V_{i,a}]$ the class of a fundamental representation.

4.4. Monoidal categorification – Examples

As a first example, consider the complex Lie algebra $\mathfrak{g} = \mathfrak{sl}_2$. We define $\mathcal{M}$ as the monoidal Serre subcategory of the category $\mathcal{C}$ of finite-dimensional representations of $\mathcal{U}_q(\widehat{\mathfrak{sl}_2})$-representations generated by two fundamental representations

$$V_1(1), V_1(q^2).$$

Then we obtain a monoidal categorification

$$K(\mathcal{M}) \simeq \mathcal{A}_Q$$

with Q of type A_2 (as above) with one frozen variable. The total number of cluster variables is 3. Indeed, there is a simple representation W in $\mathcal{M}$ of dimension 3 so that we have exact sequences

$$0 \to \mathbf{1} \to V_1(q^2) \otimes V_1(1) \to W \to 0,$$
$$0 \to W \to V_1(1) \otimes V_1(q^2) \to \mathbf{1} \to 0,$$

which come from degenerated braidings (which are not isomorphisms as the two tensor products are not isomorphic)

$$V_1(q^2) \otimes V_1(1) \to V_1(1) \otimes V_1(q^2) \quad \text{and} \quad V_1(1) \otimes V_1(q^2) \to V_1(q^2) \otimes V_1(1).$$

These exact sequences correspond to the unique mutation relation in the cluster algebra $\mathcal{A}_Q$:

$$[V_1(q^2)][V_1(1)] = 1 + [W].$$

The two clusters are $(V_1(q^2), W)$ and $(V_1(1), W)$. As an application, we obtain that every simple representation in $\mathcal{M}$ can be factorized into these three representations. Note that there are infinitely many simple classes in this category $\mathcal{M}$, which have the following form with $a, b \geq 0$:

$$(V_1(1))^{\otimes a} \otimes W^b \quad \text{and} \quad (V_1(q^2))^{\otimes a} \otimes W^b.$$

As another example, consider the complex Lie algebra $\mathfrak{g} = \mathfrak{sl}_3$. We define $\mathcal{M}$ as the monoidal Serre subcategory of the category $\mathcal{C}$ of finite-dimensional $\mathcal{U}_q(\widehat{\mathfrak{sl}_3})$-representations generated by the four fundamental representations

$$V_1(1), V_1(q^2), V_2(q), V_2(q^3).$$

Then we obtain a monoidal categorification

$$K(\mathcal{M}) \simeq \mathcal{A}_Q$$

with Q of type A_2 (as above) with two additional frozen variables. The total number of cluster variables is 7.

As an application, we obtain that every simple representation in $\mathcal{M}$ can be factorized into these seven representations. Note that there are infinitely many simple classes in this category $\mathcal{M}$.

4.5. Global structures

For a general simple Lie algebra $\mathfrak{g}$, a cluster algebra structure has been obtained on $K(\mathcal{C}^-)$ for $\mathcal{C}^-$ a large subcategory[2] of $\mathcal{C}$ [21]. Then Kashiwara–Kim–Oh–Park established the following conjecture in [21].

Theorem 4.5 ([25]). *The category $\mathcal{C}^-$ of finite-dimensional representations of a quantum affine algebra $\mathcal{U}_q(\widehat{\mathfrak{g}})$ is a monoidal categorification of a cluster algebra.*

As an application, we could prove for simple representations corresponding to cluster monomials, a conjecture that I formulated in 2004 [16] for non-simply laced types. Using cluster algebra methods, the result extends the results obtained by Nakajima [29] in simply laced types.

Theorem 4.6 ([14]). *The dimension (and character) of simple representations corresponding to cluster monomials can be obtained from an algorithm (à la Kazhdan–Lusztig).*

[2]This large subcategory is sufficient to describe all simple modules by using spectral parameter shifts.

Remark 4.7. The algorithm is partly based on the existence of a quantization of the Grothendieck rings, introduced in [29, 33] for simply laced types using quiver varieties (and generalized in [16] using a different method). The deformation is then identified with a natural quantization of the cluster algebra (the quantum cluster algebra). Recently, with R. Fujita [13], we proposed a construction of monoidal Jantzen filtrations for monoidal categories, with certain degenerated braidings, in order to give a purely categorical construction of such quantum Grothendieck rings. This is based on distinguished filtrations by submodules

$$M = M_0 \supset M_1 \supset \cdots \supset M_N = \{0\}$$

which lead to deformed classes

$$[M]_t = \sum_{r \geq 0} t^r [M_r / M_{r-1}]$$

and to a deformation $*$ of the product of the Grothendieck ring by considering filtrations of tensor products.

As an illustration, in the first example of Section 4.4, we have the filtrations

$$V_1(q^2) \otimes V_1(1) \supset \mathbf{1} \supset \{0\} \quad \text{and} \quad V_1(1) \otimes V_1(q^2) \supset W \supset \{0\}$$

which lead to the deformed products

$$[V_1(q^2)] * [V_1(1)] = t[W] + 1 \quad \text{and} \quad [V_1(1)] * [V_1(q^2)] = t^{-1}[W] + 1.$$

5. Shifted quantum affine algebras

Next we handle shifted quantum groups. The shifted quantum affine algebras form a new class of quantum groups closely related to (quantized K-theoretical) Coulomb branches (in the sense of Braverman–Finkelberg–Nakajima [2]). They were introduced by Finkelberg–Tsymbaliuk [6] in the study of these K-theoretical Coulomb branches.

Consider $\mathfrak{g}$ and q as above. The algebra $\mathcal{U}_q^\mu(\widehat{\mathfrak{g}})$ can be seen as a variation of the quantum affine algebra $\mathcal{U}_q(\widehat{\mathfrak{g}})$ depending on a shift parameter: a coweight μ of the Lie algebra $\mathfrak{g}$ (for example, in the $\mathfrak{sl}_2$-case, we can see μ as an integer). For $\mu = 0$, $\mathcal{U}_q^0(\widehat{\mathfrak{g}})$ is essentially $\mathcal{U}_q(\widehat{\mathfrak{g}})$. These algebras have a very interesting representation theory.

Theorem 5.1 ([17]). *The algebra $\mathcal{U}_q^\mu(\widehat{\mathfrak{g}})$ has a non-zero finite-dimensional representation if and only if μ is codominant.*

In general, $\mathcal{U}_q^{\mu}(\widehat{\mathfrak{g}})$ has an abelian category $\mathcal{O}^{\mu}$ of representations (which might be finite dimensional or infinite dimensional). Recall that n is the rank of $\mathfrak{g}$.

Theorem 5.2 ([17]). *The simple objects in $\mathcal{O}^{\mu}$ are parametrized by n-tuples of rational fractions $\mathbf{\Psi} = (\psi_i(z))_{1\leq i\leq n}$ regular at 0, with the degree condition*

$$\deg(\psi_i(z)) = \alpha_i(\mu),$$

where α_i is the simple root of $\mathfrak{g}$ attached to the vertex i of the Dynkin diagram.

Example 5.3. For $1 \leq i \leq n, a \in \mathbb{C}^*$, consider $\mathbf{\Psi}$ defined by $\psi_j(z) = 1$ if $j \neq i$ and

$$\psi_i(z) = q^{r_i}\frac{1 - zaq^{-r_i}}{1 - zaq^{r_i}},$$

with r_i the length of the simple root α_i. Then we recover the finite-dimensional fundamental representation $V_i(a)$ of the ordinary quantum affine algebra (the rational fractions $\psi_j(z)$ should not be confused with the Drinfeld polynomials).

Example 5.4. For $1 \leq i \leq n, a \in \mathbb{C}^*$, consider

$$\mathbf{\Psi} = ((1 - za)^{\delta_{i,j}})_{1\leq j\leq n}.$$

The associated simple representation of $\mathcal{U}_q^{\omega_i^\vee}(\widehat{\mathfrak{g}})$ is called the positive prefundamental representation $L_{i,a}^+$ (here, $\omega_i^\vee$ is a fundamental coweight). It is of dimension 1 !

Example 5.5. For $1 \leq i \leq n, a \in \mathbb{C}^*$, consider

$$\mathbf{\Psi} = ((1 - za)^{-\delta_{i,j}})_{1\leq j\leq n}.$$

The associated simple representation of $\mathcal{U}_q^{-\omega_i^\vee}(\widehat{\mathfrak{g}})$ is called the negative prefundamental representation $L_{i,a}^-$. It is infinite dimensional. In addition, $\mathcal{U}_q^{-\omega_i^\vee}(\widehat{\mathfrak{g}})$ contains a copy of the ordinary quantum affine Borel algebra $\mathcal{U}_q(\widehat{\mathfrak{b}})$. Then, restricted to this subalgebra, $L_{i,a}^-$ is the simple prefundamental representation constructed in [18] as a limit of simple finite-dimensional representations.

Without loss of generality, we may assume that all roots and poles of the rational fractions are integral powers of q.

One obtains an abelian category

$$\mathcal{O}^{\mathrm{sh}} = \bigoplus_{\mu \text{ coweight}} \mathcal{O}^{\mu}.$$

The (completed) sum of Grothendieck groups

$$K_0(\mathcal{O}^{\mathrm{sh}}) = \bigoplus_{\mu} K_0(\mathcal{O}^{\mu})$$

has a (topological) ring structure[3] induced by the fusion product construction (which defines a representation $V * W$ in $\mathcal{O}^{\mu+\lambda}$ from simple representations V, W in $\mathcal{O}^{\mu}$ and $\mathcal{O}^{\lambda}$, respectively). This ring can be considered as an analogue of a Grothendieck ring. In addition, it is commutative.

Consider the subcategory $\mathcal{C}^{\mathrm{sh}} \subset \mathcal{O}^{\mathrm{sh}}$ of finite-dimensional representations of shifted quantum affine algebras. Its Grothendieck ring $K(\mathcal{C}^{\mathrm{sh}})$ contains as a subring the Grothendieck ring $K(\mathcal{C})$ of finite-dimensional representations of the ordinary quantum affine algebra.

Theorem 5.6 ([17, 21, 25]). *The ring $K_0(\mathcal{C}^{\mathrm{sh}})$ is isomorphic to a cluster algebra $\mathcal{A}_{\Gamma_\infty}$, with an explicit quiver Γ_∞. The initial cluster variables are classes of positive prefundamental representations. The cluster monomials correspond to certain classes of simple modules.*

Example 5.7. For $\mathfrak{g} = \mathfrak{sl}_2$, the quiver Γ_∞ is the infinite linear quiver

$$\cdots \longrightarrow \bullet \longrightarrow \bullet \longrightarrow \cdots .$$

The initial seed is formed of classes of one-dimensional positive prefundamental representations

$$\cdots \longrightarrow L^+_{1,q^{-2}} \longrightarrow L^+_{1,1} \longrightarrow L^+_{1,q^2} \longrightarrow \cdots$$

(up to factors by invertible representations, which we omit in these notes for clarity). The first-step mutations are given by certain remarkable relations called Baxter TQ-relations

$$[L^+_{1,a}][V_1(a)] = [L^+_{1,aq^2}] + [L^+_{1,aq^{-2}}],$$

where $V_1(a)$ is the two-dimensional fundamental representation. In fact, this relation is also relevant from the point of view of quantum integrable models.

Theorem 5.8 ([15]). *The Grothendieck ring $K_0(\mathcal{O}^{\mathrm{sh}})$ is isomorphic to (a completion of) a cluster algebra $\mathcal{A}_{\Gamma'_\infty}$, with an explicit quiver Γ'_∞.*

The proof is partly based on additional symmetries of the Grothendieck ring that will be discussed in the next section.

Example 5.9. For $\mathfrak{g} = \mathfrak{sl}_2$, the new quiver Γ'_∞ is obtained from the older quiver Γ_∞ by just inverting the direction of one arrow:

$$\cdots \longrightarrow \bullet \longrightarrow \bullet \longleftarrow \bullet \longrightarrow \bullet \longrightarrow \cdots .$$

[3] It is proved in [22] that the subcategory of finite length representations is stable under the fusion product, and so it gives rise to a (non-topological) ring structure on $K_0(\mathcal{O}^{\mathrm{sh}})$.

Then the initial seed is made of positive and negative prefundamental representations

$$\cdots \longrightarrow L^+_{1,q^2} \longrightarrow L^+_{1,1} \longleftarrow L^-_{1,q^{-2}} \longrightarrow L^-_{1,q^{-4}} \longrightarrow \cdots .$$

The mutation at the vertex corresponding to $L^+_{1,1}$ yields

$$\cdots \longrightarrow L^+_{1,q^2} \longleftarrow L^-_{1,1} \longrightarrow L^-_{1,q^{-2}} \longrightarrow L^-_{1,q^{-4}} \longrightarrow \cdots$$

as we have the remarkable QQ-relation

$$[L^+_{1,1}][L^-_{1,1}] = 1 + [L^+_{1,q^2}][L^+_{1,q^{-2}}]. \tag{5.1}$$

Similarly, the mutation at the vertex corresponding to $L^-_{1,q^{-2}}$ translated the seed in the other direction:

$$\cdots \longrightarrow L^+_{1,q^2} \longrightarrow L^+_{1,1} \longrightarrow L^+_{1,q^{-2}} \longleftarrow L^-_{1,q^{-4}} \longrightarrow \cdots .$$

In general, we conjecture the following.

Conjecture 5.10 ([15]). *All cluster monomials in $\mathcal{A}_{\Gamma'_\infty}$ correspond to classes of simple objects in $\mathcal{O}$ through our isomorphism.*

This general conjecture is, for the moment, only known in the following case.

Theorem 5.11 ([15]). *The conjecture is true for $\mathfrak{g} = \mathfrak{sl}_2$.*

6. Symmetry of q-characters

We now discuss certain group symmetries of Grothendieck rings.

6.1. Classical theory

Let us briefly recall the well-known Weyl group symmetry of classical characters. Let $\mathfrak{g}$ be a complex finite-dimensional simple Lie algebra of rank n and $K_0(\mathfrak{g})$ be the Grothendieck ring of its finite-dimensional representations. The character morphism, which encodes the dimension of weight spaces of finite-dimensional representations, defines an injective ring morphism

$$\chi : K_0(\mathfrak{g}) \to \mathbb{Z}[y_i^{\pm 1}]_{1 \le i \le n}.$$

The image of χ can be characterized as an invariant subring

$$\mathrm{Im}(\chi) = (\mathbb{Z}[y_i^{\pm 1}]_{1 \le i \le n})^W,$$

where the Weyl group W is generated by the simple reflections s_i, $1 \le i \le n$, which act by the explicit formula

$$s_i(y_j) = y_j a_i^{-\delta_{ij}}, \quad \text{where } a_i = \prod_{1 \le k \le n} y_k^{C_{ji}}.$$

Here, a_i corresponds to a simple root, and the C_{ji} are coefficients of the Cartan matrix of $\mathfrak{g}$.

Example 6.1. For $\mathfrak{g} = \mathfrak{sl}_2$, the Cartan matrix $C = (2)$, and we have

$$\begin{aligned} a_1 &= y_1^2, \\ s_1(y_1) &= y_1 a_1^{-1} = y_1^{-1}, \\ s_1^2(y_1) &= y_1, \\ s_1(y_1 + y_1^{-1}) &= y_1 + y_1^{-1}, \\ \mathrm{Im}(\chi) &= (\mathbb{Z}[y_1^{\pm 1}])^W = \mathbb{Z}[y_1 + y_1^{-1}]. \end{aligned}$$

6.2. q-characters

The analogue of characters for finite-dimensional representations of quantum affine algebras $\mathcal{U}_q(\hat{\mathfrak{g}})$ are q-characters (as defined by Frenkel–Reshetikhin [12]). It defines an injective ring morphism on the Grothendieck ring $K(\mathcal{C})$ of the category $\mathcal{C}$ of finite-dimensional representations (here, we assume that the quantization parameter $q \in \mathbb{C}^*$ is not a root of unity):

$$\chi_q : K(\mathcal{C}) \to \mathcal{Y} = \mathbb{Z}[Y_{i,a}^{\pm 1}]_{1 \le i \le n,\, a \in \mathbb{C}^*}.$$

It is a refinement of characters for the action of the underlying finite-type quantum group $\mathcal{U}_q(\mathfrak{g})$ (as we can recover the ordinary by removing the spectral parameters a).

Example 6.2. For $a \in \mathbb{C}^*$, the fundamental representation $V_1(a)$ of $\mathcal{U}_q(\widehat{\mathfrak{sl}_2})$ satisfies

$$\chi_q(V_1(a)) = Y_{1,a} + Y_{1,aq^2}^{-1}.$$

If we forget the spectral characters, we recover the ordinary character $Y_1 + Y_1^{-1}$.

There are monomial analogues of simple roots (we give here the formula for simply laced types):

$$A_{i,a} = Y_{i,aq^{-1}} Y_{i,aq} \prod_{j | C_{j,i} = -1} Y_{j,a}^{-1}.$$

For example, for $\mathfrak{g} = \mathfrak{sl}_2$, we have $A_{1,a} = Y_{1,aq^{-1}} Y_{1,aq}$.

We may wonder what the analogue of the Weyl group symmetry is in this context.

6.3. Symmetry of q-characters

We introduce in [9] the following operators:

$$\Theta_i(Y_{j,a}) = Y_{j,a} A_{i,aq^{-1}}^{-\delta_{i,j}} \frac{\Sigma_{i,aq^{-3}}^{\delta_{i,j}}}{\Sigma_{i,aq^{-1}}^{\delta_{i,j}}}$$

for $1 \leq i, j \leq n$ and $a \in \mathbb{C}^*$. Here, $\Sigma_{i,a}$ is the solution of the q-difference equation

$$\Sigma_{i,a} = 1 + A_{i,a}^{-1} \Sigma_{i,aq^{-2}}$$

in a sum

$$\Pi = \bigoplus_{w \in W} \widetilde{\mathcal{Y}}^w$$

of completions $\widetilde{\mathcal{Y}}^w$ of $\mathcal{Y}$, parametrized by the Weyl group.

Example 6.3. For $\mathfrak{g} = \mathfrak{sl}_2$, we have

$$\Theta_1(Y_{1,a}) = Y_{1,aq^{-2}}^{-1} \frac{\Sigma_{1,aq^{-3}}}{\Sigma_{1,aq^{-1}}}.$$

Here, $\Sigma_{1,a}$ is the couple

$$(1 + A_{1,a}^{-1}(1 + A_{1,aq^{-2}}^{-1}(1 + \cdots)), -A_{1,aq^2}(1 + A_{1,aq^4}(1 + \cdots))).$$

It belongs to

$$\Pi = \mathcal{Y}^e \oplus \mathcal{Y}^{s_1}.$$

Remark 6.4. One component of these operators is related to operators defined in [23] on a different space.

Now $\mathcal{Y}$ embeds in Π diagonally, and we have the following analogue of the classical result.

Theorem 6.5 ([9]). *The Θ_i are involutions and define a Weyl group action (with the simple reflections acting as Θ_i). Moreover, we have the ring of invariants*

$$\mathcal{Y}^W = \mathrm{Im}(\chi_q).$$

6.4. Interpretation of the Weyl group action

Let us explain an interpretation of the Weyl group action above, which is connected to the cluster algebra symmetry.

Theorem 6.6 ([8, 17]). *The replacement of the variables*

$$Y_{i,a} = \frac{[L^+_{i,aq^{-1}}]}{[L^+_{i,aq}]}$$

in the q-character of finite-dimensional representations gives a well-defined relation in (the fraction field) of $K(\mathcal{O}^{\mathrm{sh}})$.

Note that originally, this theorem was established in [8] in the context of representations of Borel subalgebra of quantum affine algebras (as in Example 5.5, in order to prove Frenkel–Reshetikhin conjecture on spectra of quantum integrable models), but it can also be stated for representations of shifted quantum affine algebras.

Motivated by quantum integrable models and Bethe ansatz equations, we expect that the class $[L^+_{i,a}]$ can be replaced in the theorem by the class of other simple representations that we denote by $\widetilde{L}^+_{i,a}$ (see [11]). This is exactly how one can write the defining formula of the operator Θ_i, which is now interpreted as a substitution of simple classes in the Grothendieck ring.

We can now relate this picture to cluster algebra symmetry. Indeed, starting from $Q_a^{\omega_i^\vee} = [L^+_{i,a}]$, and using the Weyl group action, one can introduce for a general $w \in W$, $i \in I$,

$$Q_a^{w(\omega_i^\vee)} \in K(\mathcal{O}^{\mathrm{sh}}).$$

Theorem 6.7 ([10] Generalized QQ-systems). *The series* $Q_a^{w(\omega_i^\vee)}$ *satisfy the* QQ*-system*

$$Q_{aq}^{(ws_i)(\omega_i^\vee)} Q_{aq^{-1}}^{w(\omega_i^\vee)} - Q_{aq^{-1}}^{(ws_i)(\omega_i^\vee)} Q_{aq}^{w(\omega_i^\vee)} = \prod_{j|C_{i,j}=-1} Q_a^{w(\omega_j^\vee)}.$$

Hence the same relation is satisfied by elements in an orbit under the Weyl group action.

Example 6.8. For $\mathfrak{g} = \mathfrak{sl}_2$, this QQ-system is the quantum Wronskian relation

$$\widetilde{Q}_{aq} Q_{aq^{-1}} - \widetilde{Q}_{aq^{-1}} Q_{aq} = 1, \tag{6.1}$$

where we have denoted $Q_a = Q_a^{\omega_1^\vee}$ and $\widetilde{Q}_a = Q_a^{-\omega_1^\vee}$.

These QQ-system relations are crucial in the construction of the cluster algebra structure on $K(\mathcal{O}^{\mathrm{sh}})$ in [15], as they appear in mutations of the cluster algebra (for example, compare equations (5.1) and (6.1)). Hence, cluster symmetries are related to group symmetry of Grothendieck rings in these examples, which one could find quite intriguing.

References

[1] P. Baumann and S. Riche, Notes on the geometric Satake equivalence. In V. Heiermann and D. Prasad (eds.), *Relative aspects in representation theory, Langlands functoriality and automorphic forms. CIRM Jean–Morlet Chair, Spring 2016*, pp. 1–134, Lecture Notes in Math. 2221, Springer, Cham, 2018 Zbl 1450.22009 MR 3839695

[2] A. Braverman, M. Finkelberg, and H. Nakajima, Towards a mathematical definition of Coulomb branches of 3-dimensional $\mathcal{N} = 4$ gauge theories, II. *Adv. Theor. Math. Phys.* **22** (2018), no. 5, 1071–1147 Zbl 1479.81043 MR 3952347

[3] M. Brito and V. Chari, Tensor products and q-characters of HL-modules and monoidal categorifications. *J. Éc. polytech. Math.* **6** (2019), no. 3, 581–619 Zbl 1446.17026 MR 4014632

[4] S. Cautis and H. Williams, Cluster theory of the coherent Satake category. *J. Amer. Math. Soc.* **32** (2019), no. 3, 709–778 Zbl 1442.22022 MR 3981987

[5] N. Chriss and V. Ginzburg, *Representation theory and complex geometry*. Mod. Birkhäuser Class., Birkhäuser, Boston, MA, 1997, 495 pp. Zbl 0879.22001 MR 1433132

[6] M. Finkelberg and A. Tsymbaliuk, Multiplicative slices, relativistic Toda and shifted quantum affine algebras. In M. Gorelik, V. Hinich, and A. Melnikov (eds.), *Representations and nilpotent orbits of Lie algebraic systems*, pp. 133–304, Progr. Math. 330, Birkhäuser, Cham, 2019 Zbl 1436.17021 MR 3971731

[7] S. Fomin and A. Zelevinsky, Cluster algebras. I. Foundations. *J. Amer. Math. Soc.* **15** (2002), no. 2, 497–529 Zbl 1021.16017 MR 1887642

[8] E. Frenkel and D. Hernandez, Baxter's relations and spectra of quantum integrable models. *Duke Math. J.* **164** (2015), no. 12, 2407–2460 Zbl 1332.82022 MR 3397389

[9] E. Frenkel and D. Hernandez, Spectra of quantum KdV Hamiltonians, Langlands duality, and affine opers. *Comm. Math. Phys.* **362** (2018), no. 2, 361–414 Zbl 1397.81060 MR 3843418

[10] E. Frenkel and D. Hernandez, Extended Baxter relations and QQ-systems for quantum affine algebras. *Comm. Math. Phys.* **405** (2024), no. 8, article no. 190, 42 pp. Zbl 1556.17028 MR 4779559

[11] E. Frenkel and D. Hernandez, Weyl group symmetry of q-characters. *Selecta Math. (N.S.)* **31** (2025), no. 4, article no. 72, 36 pp. Zbl 08087587 MR 4939703

[12] E. Frenkel and N. Reshetikhin, The q-characters of representations of quantum affine algebras and deformations of $\mathcal{W}$-algebras. In N. Jing and K. C. Misra (eds.), *Recent developments in quantum affine algebras and related topics (Raleigh, NC, 1998)*, pp. 163–205, Contemp. Math. 248, American Mathematical Society, Providence, RI, 1999 Zbl 0973.17015 MR 1745260

[13] R. Fujita and D. Hernandez, Monoidal Jantzen filtrations. *Adv. Math.* **495** (2026), article no. 110963, 81 pp. MR 5061762

[14] R. Fujita, D. Hernandez, S.-j. Oh, and H. Oya, Isomorphisms among quantum Grothendieck rings and cluster algebras. 2023, arXiv:2304.02562v2

[15] C. Geiss, D. Hernandez, and B. Leclerc, Representations of shifted quantum affine algebras and cluster algebras. I. The simply laced case. *Proc. Lond. Math. Soc. (3)* **129** (2024), no. 3, article no. e12630, 75 pp. Zbl 1564.17026 MR 4793282

[16] D. Hernandez, Algebraic approach to q, t-characters. *Adv. Math.* **187** (2004), no. 1, 1–52 Zbl 1098.17009 MR 2074171

[17] D. Hernandez, Representations of shifted quantum affine algebras. *Int. Math. Res. Not. IMRN* **2023** (2023), no. 13, 11035–11126 Zbl 1532.17013 MR 4609778

[18] D. Hernandez and M. Jimbo, Asymptotic representations and Drinfeld rational fractions. *Compos. Math.* **148** (2012), no. 5, 1593–1623 Zbl 1266.17010 MR 2982441

[19] D. Hernandez and B. Leclerc, Cluster algebras and quantum affine algebras. *Duke Math. J.* **154** (2010), no. 2, 265–341 Zbl 1284.17010 MR 2682185

[20] D. Hernandez and B. Leclerc, Quantum Grothendieck rings and derived Hall algebras. *J. Reine Angew. Math.* **2015** (2015), no. 701, 77–126 Zbl 1315.17011 MR 3331727

[21] D. Hernandez and B. Leclerc, A cluster algebra approach to q-characters of Kirillov–Reshetikhin modules. *J. Eur. Math. Soc. (JEMS)* **18** (2016), no. 5, 1113–1159 Zbl 1405.17028 MR 3500832

[22] D. Hernandez and H. Zhang, Jordan–Hölder property for shifted quantum affine algebras. 2025, arXiv:2501.16859v2

[23] R. Inoue, Cluster realization of Weyl groups and q-characters of quantum affine algebras. *Lett. Math. Phys.* **111** (2021), no. 1, article no. 4, 32 pp. Zbl 1454.13033 MR 4197422

[24] S.-J. Kang, M. Kashiwara, M. Kim, and S.-j. Oh, Monoidal categorification of cluster algebras. *J. Amer. Math. Soc.* **31** (2018), no. 2, 349–426 Zbl 1460.13039 MR 3758148

[25] M. Kashiwara, M. Kim, S.-j. Oh, and E. Park, Monoidal categorification and quantum affine algebras II. *Invent. Math.* **236** (2024), no. 2, 837–924 Zbl 1551.17017 MR 4728243

[26] B. Keller, Cluster algebras and cluster monomials. In R. Latała, A. Ruciński, P. Strzelecki, J. Świątkowski, D. Wrzosek, and P. Zakrzewski (eds.), *European Congress of Mathematics*, pp. 399–413, European Mathematical Society, Zürich, 2013 Zbl 1364.13023 MR 3469134

[27] M. Khovanov and A. D. Lauda, A diagrammatic approach to categorification of quantum groups. I. *Represent. Theory* **13** (2009), 309–347 Zbl 1188.81117 MR 2525917

[28] S. Mac Lane, *Categories for the working mathematician*. 2nd edn., Grad. Texts in Math. 5, Springer, New York, 1998, 318 pp. Zbl 0906.18001 MR 1712872

[29] H. Nakajima, Quiver varieties and t-analogs of q-characters of quantum affine algebras. *Ann. of Math. (2)* **160** (2004), no. 3, 1057–1097 Zbl 1140.17015 MR 2144973

[30] H. Nakajima, Quiver varieties and cluster algebras. *Kyoto J. Math.* **51** (2011), no. 1, 71–126 Zbl 1223.13013 MR 2784748

[31] F. Qin, Triangular bases in quantum cluster algebras and monoidal categorification conjectures. *Duke Math. J.* **166** (2017), no. 12, 2337–2442 Zbl 1454.13037 MR 3694569

[32] R. Rouquier, 2-Kac–Moody algebras. 2008, arXiv:0812.5023v1

[33] M. Varagnolo and E. Vasserot, Perverse sheaves and quantum Grothendieck rings. In A. Joseph, A. Melnikov, and R. Rentschler (eds.), *Studies in memory of Issai Schur (Chevaleret/Rehovot, 2000)*, pp. 345–365, Progr. Math. 210, Birkhäuser, Boston, MA, 2003 Zbl 1162.17307 MR 1985732

David Hernandez
Université Paris Cité, Sorbonne Université, CNRS, IMJ-PRG, 75013 Paris, France;
david.hernandez@imj-prg.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/20

Forward and inverse problems in nonlinear acoustics

Barbara Kaltenbacher

Abstract. The importance of ultrasound is well established in the imaging of human tissue. In order to enhance image quality by exploiting nonlinear effects, recently techniques such as harmonic imaging and nonlinearity parameter tomography have been put forward. As soon as the pressure amplitude exceeds a certain bound, the classical linear wave equation loses its validity, and more general nonlinear versions have to be used. Another characteristic property of ultrasound propagation in human tissue is frequency power law attenuation, leading to fractional derivative damping models in the time domain. In this contribution, we will first of all dwell on modeling nonlinearity on the one hand and fractional damping on the other hand. Moreover, we will give an idea of the challenges in the analysis of the resulting PDEs and discuss some parameter asymptotics. Finally, we address a relevant inverse problem in this context, the above-mentioned task of nonlinearity parameter imaging, which leads to a coefficient identification problem for a quasilinear wave equation.

1. Introduction

The difference between linear and nonlinear acoustic wave propagation becomes apparent through the steepening of waves; see Figure 1. This is due to a pressure-dependent wave speed, which is higher in compressed areas and lower in low-density regions. This fact is visible in the wave profiles from both a space and from a time perspective and can also be read off from the PDEs; cf. (2.7) below. Another way of characterizing nonlinearity is by looking at contributions in the frequency domain in the case of a sinusoidal excitation, which in the linear case results in a response at the excitation frequency only, whereas in the nonlinear case, so-called higher harmonics are generated; see Figure 2. Also this effect is very well visible in the governing mathematical models; cf. (5.4).

Mathematics Subject Classification 2020: 35L80 (primary); 35B40, 35R30, 35R11 (secondary).
Keywords: nonlinear acoustics, parameter asymptotics, fractional attenuation, multiharmonic expansion, acoustic nonlinearity parameter tomography.

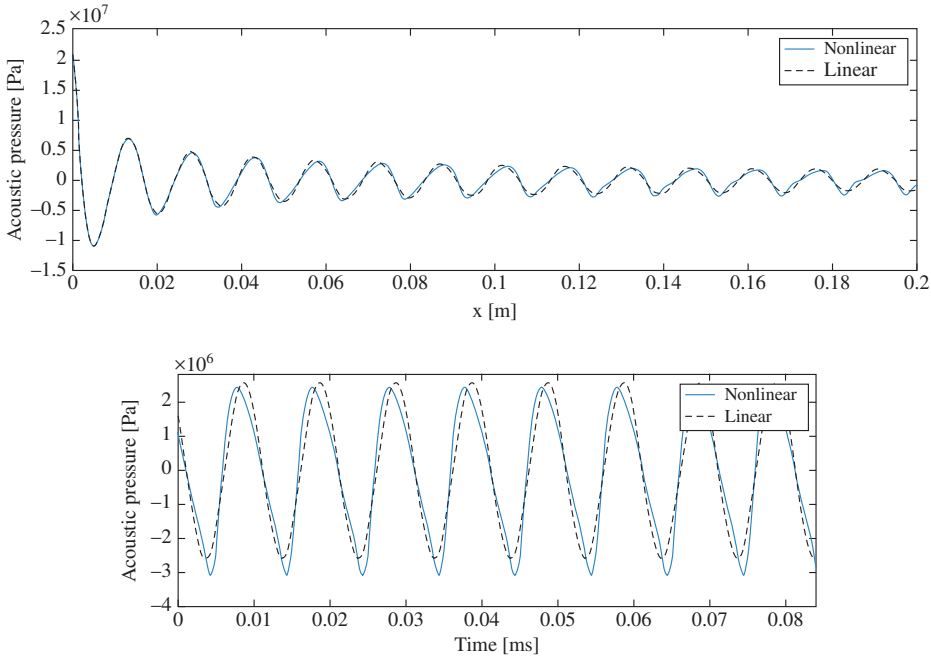


Figure 1. Linear versus nonlinear wave propagation in space (top) and time (bottom) domains.

2. Models of nonlinear acoustics

2.1. Physical principles

The main physical quantities in acoustics are

- acoustic particle velocity $\mathbf{v}$;
- acoustic pressure p;
- mass density ϱ;
- absolute temperature ϑ;
- heat flux $\boldsymbol{q}$;
- entropy η.

These are usually decomposed into mean (zero-order) and fluctuating (first-order) parts

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}_\sim = \mathbf{v}\,, \quad p = p_0 + p_\sim\,, \quad \varrho = \varrho_0 + \varrho_\sim\,, \quad \text{etc.}$$

and subject to the following physical laws:

- momentum conservation = Navier–Stokes equation (with $\nabla \times \mathbf{v} = 0$):

$$\varrho(\mathbf{v}_t + \nabla(\mathbf{v}\cdot\mathbf{v})) + \nabla p = \left(\frac{4\mu_V}{3} + \zeta_V\right)\Delta\mathbf{v};$$

- mass conservation = equation of continuity: $\varrho_t + \nabla\cdot(\varrho\mathbf{v}) = 0$;
- entropy equation: $\varrho\vartheta(\eta_t + \mathbf{v}\cdot\nabla\eta) = -\nabla\cdot\boldsymbol{q}$;

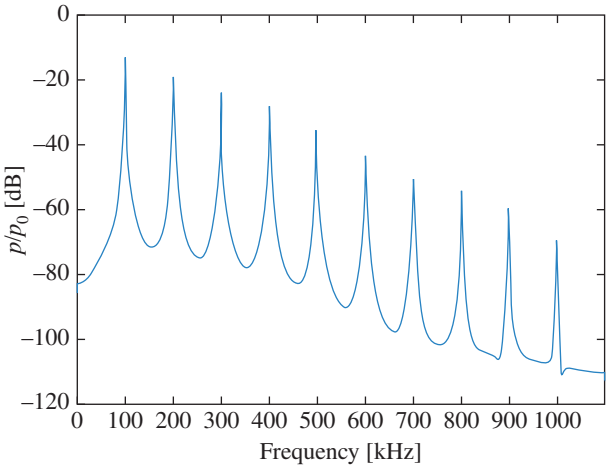


Figure 2. Higher harmonics in frequency domain.

- equation of state: $\varrho_\sim = \frac{1}{c^2} p_\sim - \frac{1}{\varrho_0 c^4} \frac{B}{2A} p_\sim{}^2 - \frac{a}{\varrho_0 c^4} \left(\frac{1}{c_V} - \frac{1}{c_p}\right) p_{\sim t}$;
- Gibbs equation: $\vartheta d\eta = c_V d\vartheta - p \frac{1}{\varrho^2} d\varrho$,

where c_p / c_V is the specific heat at constant pressure / volume and ζ_V / μ_V the bulk / shear viscosity. The missing sixth equation in this system governing the six unknown functions $\mathbf{v}$, p, ϱ, ϑ, $\boldsymbol{q}$, η is the relation between temperature and heat flux, which (like the pressure-density relation) is subject to constitutive modeling. Classically, Fourier's law

$$\boldsymbol{q} = -K \nabla \vartheta \tag{2.1}$$

(with K denoting the thermal conductivity) has been employed for this purpose. However, this is known to lead to the infinite speed of propagation paradox. For this reason, a relaxation term has been introduced that leads to the Maxwell–Cattaneo law

$$\tau \boldsymbol{q}_t + \boldsymbol{q} = -K \nabla \vartheta, \tag{2.2}$$

which contains the relaxation time τ as an additional parameter and allows for so-called "thermal waves" (also called the second sound phenomenon).

2.2. Classical models of nonlinear acoustics

In the linear case, where quadratic and higher-order terms (e.g., products of first-order terms) are skipped, it suffices to consider momentum and mass conservation along with the equation of state, and the velocity can be eliminated by taking the divergence in momentum conservation and subtracting the time derivative of mass conservation. Upon substitution of the (linearized) pressure-density relation, this

leads to the well-known second-order wave equation $p_{tt} - c^2\Delta p = 0$. In the nonlinear case, the same procedure, but retaining quadratic terms (that is, application of the so-called Blackstock's scheme in nonlinear acoustics), leads to the classical models of nonlinear acoustics.

- *Kuznetsov's equation* [46,47]:

$$p_{\sim tt} - c^2\Delta p_\sim - \delta\Delta p_{\sim t} = \left(\frac{B}{2A\varrho_0 c^2}p_\sim^2 + \varrho_0|\mathbf{v}|^2\right)_{tt},$$

  where $\varrho_0\mathbf{v}_t = -\nabla p$ for the particle velocity $\mathbf{v}$ and the pressure p, that is,

$$\psi_{tt} - c^2\Delta\psi - \delta\Delta\psi_t = \left(\frac{B}{2A\,c^2}(\psi_t)^2 + |\nabla\psi|^2\right)_t, \tag{2.3}$$

  since $\nabla\times\mathbf{v} = 0$, hence $\mathbf{v} = -\nabla\psi$ for some velocity potential ψ.

- The *Westervelt equation* [55], which is obtained from Kuznetsov's equation via the approximation $\varrho_0|\mathbf{v}|^2 \approx \frac{1}{\varrho_0 c^2}(p_\sim)^2$ that corresponds to neglecting noncumulative nonlinear effects:

$$p_{\sim tt} - c^2\Delta p_\sim - \delta\Delta p_{\sim t} = \frac{1}{\varrho_0 c^2}\left(1 + \frac{B}{2A}\right)p_{\sim tt}^2. \tag{2.4}$$

Here, $\delta = h\big(\mathrm{Pr}\big(\frac{4}{3} + \frac{\zeta_V}{\mu_V}\big) + \gamma - 1\big)$ is the diffusivity of sound (containing the thermal diffusivity h and the Prandtl number Pr) and $\frac{B}{A}$ is the nonlinearity parameter.

2.3. Advanced models of nonlinear acoustics

Taking into account further physical effects requires considering temperature, heat flux and entropy as variables as well. Modeling in nonlinear acoustics is a highly active field, and only a small portion of the resulting PDE models has been subject to rigorous mathematical investigation so far. We mention here a few examples of these.

- The *Jordan–Moore–Gibson–Thompson* (JMGT) equation [14,28,51]

$$\tau\psi_{ttt} + \psi_{tt} - c^2\Delta\psi - (\delta + \tau c^2)\Delta\psi_t = \left(\frac{B}{2Ac^2}(\psi_t)^2 + |\nabla\psi|^2\right)_t, \tag{2.5}$$

  where τ is the relaxation time, results from replacing Fourier's law (2.1) of heat conduction by (2.2); later on, we will also encounter fractional versions thereof. The mentioned second sound phenomenon can be seen by considering the auxiliary quantity $z := \psi_t + \frac{c^2}{\delta+\tau c^2}\psi$, which solves a weakly damped wave equation

$$z_{tt} - \widetilde{c}\Delta z + \gamma z_t = r(z,\psi)$$

  with $\widetilde{c} = c^2 + \frac{\delta}{\tau}$, $\gamma = \frac{1}{\tau} - \frac{c^2}{\delta+\tau c^2} > 0$ and a lower-order term $r(z,\psi)$.

- The *Blackstock–Crighton* (BCBJ) equation [5, 8, 16]

$$(\partial_t - a\Delta)(\psi_{tt} - c^2\Delta\psi - \delta\Delta\psi_t) - ra\Delta\psi_t = \left(\frac{B}{2Ac^2}(\psi_t^2) + |\nabla\psi|^2\right)_{tt}, \quad (2.6)$$

  where $a = \frac{\nu}{\mathrm{Pr}}$ is the thermal diffusivity and r another constant.

Both models can be related to the classical Kuznetsov equation (2.3) by – first of all just formally – setting the parameters τ and a to zero. We will take a look at the mathematical justification of these parameter limits in Section 3.

The amount of literature on the analysis of these and some other advanced models of nonlinear acoustics is vast, and we do not attempt to provide an overview that could claim completeness in any sense. Some selected references are [9, 17, 18, 22, 32, 33, 44, 48, 49].

Just to get an idea of the mathematical challenges arising in the analysis (referring to the above quoted literature for details), let us take a look at the question of well-posedness of a prototypical initial-boundary value problem for the Westervelt equation as the most simple of the abovementioned models:

$$\begin{aligned} u_{tt} - c^2\Delta u - b\Delta u_t &= \frac{\kappa}{2}(u^2)_{tt} && \text{in } \Omega, \\ \frac{\partial u}{\partial n} &= g && \text{on } \partial\Omega, \\ u(t=0) = u_0, \quad u_t(t=0) &= u_1 && \text{in } \Omega, \end{aligned}$$

where we use the mathematically common notation u for the PDE solution, which plays the physical role of the acoustic pressure. Differentiating out the quadratic term and a slight rearrangement leads to

$$(1 - \kappa u)u_{tt} - c^2\Delta u - b\Delta u_t = \kappa(u_t)^2$$

which reveals the fact that the coefficient of the second time derivative depends on u itself. This leads to potential degeneracy for $u \geq \frac{1}{\kappa}$ and similarly affects the other models (Kuznetsov, Jordan–Moore–Gibson–Thompson, Blackstock–Crighton). The typical approach of proving well-posedness therefore relies on smallness of the state and proceeds as follows:

- employ energy estimates to obtain a bound on u in $C(0, T; H^2(\Omega))$;
- use smallness of u in $C(0, T; H^2(\Omega))$ and the embedding $H^2(\Omega) \to L^\infty(\Omega)$ to guarantee $1 - \kappa u \geq \underline{\alpha} > 0$;
- apply a combination of these arguments in a fixed-point scheme.

This also illustrates the state dependence of the effective wave speed, since a further rearrangement allows rewriting the equation as

$$u_{tt} - \widetilde{c}^2\Delta u - \widetilde{b}(u)\Delta u_t = f(u)$$

with

$$\tilde{c}(u) = \frac{c}{\sqrt{1-\kappa u}}, \quad \tilde{b}(u) = \frac{b}{1-\kappa u}, \quad f(u) = \frac{\kappa (u_t)^2}{1-\kappa u}, \tag{2.7}$$

as long as $1 - \kappa u > 0$ (otherwise the equation loses its validity as a wave propagation model).

3. Parameter asymptotics

In this section, we point to a few analytical results on limits that interrelate the models in Section 2. We will be explicit about function space settings and mathematical tools only in part for the first example (the limit as $\tau \searrow$ in the JMGT equation; cf. Section 3.1). The other limits will be discussed briefly and collectively in Section 3.2; details can be found in the cited papers.

3.1. Vanishing relaxation time in the Jordan–Moore–Gibson–Thompson equation

Considering a family of solutions $(\psi^\tau)_{\tau\in(0,\overline{\tau}]}$ to

$$\tau\psi^\tau_{ttt} + \psi^\tau_{tt} - c^2\Delta\psi^\tau - b\Delta\psi^\tau_t = \left(\frac{B}{2Ac^2}(\psi^\tau_t)^2 + |\nabla\psi^\tau|^2\right)_t$$

with $b = \delta + \tau c^2$, the typical question to be answered is whether and in which function spaces a limit ψ^0 of ψ^τ as $\tau \searrow 0$ exists and under which conditions it can be shown to solve Kuznetsov's equation (2.3).

Some comments are in order.

- We will consider the "Westervelt-type" and the "Kuznetsov-type" equations, without and with the gradient nonlinearity $|\nabla\psi|^2_t$.
- For $\tau = 0$ (classical Westervelt and Kuznetsov equations), the reformulation of the linearization as a first-order system leads to an analytic semigroup and maximal parabolic regularity. These properties do not hold anymore with $\tau > 0$, that is, the equation loses its "parabolic nature". This is consistent with physics, corresponding to a transition from infinite to finite propagation speed. As $\tau \to 0$, the PDE changes from hyperbolic to parabolic.
- As in the classical models (2.3) and (2.4), potential degeneracy can be an issue, since we have

$$\begin{aligned} \tau\psi^\tau_{ttt} + \psi^\tau_{tt} - c^2\Delta\psi^\tau - b\Delta\psi^\tau_t &= \left(\frac{\kappa}{2}(\psi^\tau_t)^2 + |\nabla\psi^\tau|^2\right)_t \\ &= \kappa\psi^\tau_t\psi^\tau_{tt} + |\nabla\psi^\tau|^2_t \\ \iff \quad \tau\psi^\tau_{ttt} + (1-\kappa\psi^\tau_t)\psi^\tau_{tt} - c^2\Delta\psi^\tau - b\Delta\psi^\tau_t &= |\nabla\psi^\tau|^2_t. \end{aligned}$$

The plan of the analysis is to first establish well-posedness of the linearized equation along with energy estimates and use these results to prove well-posedness of the Westervelt- or Kuznetsov-type JMGT equation for $\tau > 0$ by a fixed-point argument. Relying on uniform in τ bounds obtained in the energy analysis and (weak) compactness of bounded sets in the underlying solution spaces, we then take limits as $\tau \to 0$. We refer to [6, 34, 35] for details and only sketch the above steps in a concrete function space setting for the Westervelt-type case while mentioning in passing that an extension to the Kuznetsov-type JMGT equation requires higher-order energy estimates.

We start by considering the linearized problem with variable coefficient α:

$$\begin{cases} \tau\psi_{ttt} + \alpha(x,t)\psi_{tt} - c^2\Delta\psi - b\Delta\psi_t = f & \text{in } \Omega \times (0,T), \\ \psi = 0 & \text{on } \partial\Omega \times (0,T), \\ (\psi, \psi_t, \psi_{tt}) = (\psi_0, \psi_1, \psi_2) & \text{in } \Omega \times \{0\}. \end{cases} \tag{3.1}$$

Under the assumptions

$$\alpha(x,t) \geq \underline{\alpha} > 0 \ \text{ on } \Omega \quad \text{a.e. in } \Omega \times (0,T), \tag{3.2}$$

$$\begin{aligned} &\alpha \in L^\infty(0,T; L^\infty(\Omega)) \cap L^\infty(0,T; W^{1,3}(\Omega)), \\ &f \in H^1(0,T; L^2(\Omega)), \end{aligned} \tag{3.3}$$

$$(\psi_0, \psi_1, \psi_2) \in H_0^1(\Omega) \cap H^2(\Omega) \times H_0^1(\Omega) \cap H^2(\Omega) \times H_0^1(\Omega), \tag{3.4}$$

we obtain the following result.

Theorem 3.1. *Let c^2, b, $\tau > 0$, and let $T > 0$. Let assumptions (3.2), (3.3), (3.4) hold. Then there exists a unique solution*

$$\begin{aligned} \psi \in X^W := {} & W^{1,\infty}(0,T; H_0^1(\Omega) \cap H^2(\Omega)) \\ & \cap W^{2,\infty}(0,T; H_0^1(\Omega)) \cap H^3(0,T; L^2(\Omega)). \end{aligned}$$

The solution fulfills the estimate

$$\begin{aligned} \|\psi\|_{W,\tau}^2 &:= \tau^2\|\psi_{ttt}\|_{L^2L^2}^2 + \tau\|\psi_{tt}\|_{L^\infty H^1}^2 + \|\psi_{tt}\|_{L^2H^1}^2 + \|\psi\|_{W^{1,\infty}H^2}^2 \\ &\leq C(\alpha,T,\tau)\big(|\psi_0|_{H^2}^2 + |\psi_1|_{H^2}^2 + \tau|\psi_2|_{H^1}^2 + \|f\|_{L^\infty L^2}^2 + \|f_t\|_{L^2L^2}^2\big). \end{aligned}$$

If additionally $\|\nabla\alpha\|_{L^\infty L^3} < \frac{\underline{\alpha}}{C_{H^1,L^6}^{\Omega}}$ holds, then $C(\alpha,T,\tau)$ is independent of τ.

Idea of proof. As a standard for proving well-posedness of evolutionary PDEs, we combine a Galerkin discretization with energy estimates and weak limits; however, due to the need for higher regularity, we derive energy estimates by nonstandard testing; in this case, we test by $-\Delta\psi_{tt}$ and with $\tau\psi_{ttt}$. ∎

This allows us to prove well-posedness of the Westervelt-type JMGT equation

$$\begin{cases} \tau\psi_{ttt} + (1-\kappa\psi_t)\psi_{tt} - c^2\Delta\psi - b\Delta\psi_t = 0 & \text{in } \Omega\times(0,T), \\ \psi = 0 & \text{on } \partial\Omega\times(0,T), \\ (\psi,\psi_t,\psi_{tt}) = (\psi_0,\psi_1,\psi_2) & \text{in } \Omega\times\{0\}. \end{cases}$$

Theorem 3.2. *Let c^2, $b>0$, $\kappa\in\mathbb{R}$ and $T>0$. There exist $\rho,\rho_0>0$ such that for all $(\psi_0,\psi_1,\psi_2)\in H_0^1(\Omega)\cap H^2(\Omega)\times H_0^1(\Omega)\cap H^2(\Omega)\times H_0^1(\Omega)$ satisfying*

$$\|\psi_0\|^2_{H^2(\Omega)} + \|\psi_1\|^2_{H^2(\Omega)} + \tau\|\psi_2\|^2_{H^1(\Omega)} \le \rho_0^2\,,$$

there exists a unique solution $\psi\in X^W$ and $\|\psi\|^2_{W,\tau}\le\rho^2$.

Idea of proof. We apply Banach's contraction principle to the fixed-point operator $\mathcal{T}:\phi\mapsto\psi$ solving (3.1) with $\alpha=1-\kappa\phi_t$, $f=0$. Invariance of $\mathcal{T}$ on $B_\rho^{X^W}$ can be obtained from the energy estimate in Theorem 3.1. Also contractivity $\|\mathcal{T}(\phi_1)-\mathcal{T}(\phi_2)\|_{W,\tau}\le q\|\phi_1-\phi_2\|_{W,\tau}$ can be established by means of the estimate in Theorem 3.1, since $\widehat{\psi}=\psi_1-\psi_2=\mathcal{T}(\phi_1)-\mathcal{T}(\phi_2)$ solves (3.1) with $\alpha=1-\kappa\phi_{1t}$ and $f=\kappa\widehat{\phi}_t\psi_{2tt}$, where $\widehat{\phi}=\phi_1-\phi_2$. ∎

To take limits for vanishing relaxation time, we consider the τ-independent part of the norms

$$\|\psi\|^2_{W,\tau} := \tau^2\|\psi_{ttt}\|^2_{L^2L^2} + \tau\|\psi_{tt}\|^2_{L^\infty H^1} + \|\psi_{tt}\|^2_{L^2H^1} + \|\psi\|^2_{W^{1,\infty}H^2},$$

that is,

$$\|\psi\|^2_{\bar{X}^W} := \|\psi_{tt}\|^2_{L^2H^1} + \|\psi\|^2_{W^{1,\infty}H^2}\,,$$

since these norms are uniformly bounded, independently of τ, and impose smallness of initial data in the space

$$X_0^W := H_0^1(\Omega)\cap H^2(\Omega)\times H_0^1(\Omega)\cap H^2(\Omega)\times H_0^1(\Omega).$$

This allows us to prove the following; cf [34].

Theorem 3.3. *Let c^2, b, $\kappa\in\mathbb{R}$ and $T>0$. Then there exist $\overline{\tau}$, $\rho_0>0$ such that for all $(\psi_0,\psi_1,\psi_2)\in B_{\rho_0}^{X_0^W}$, the family $(\psi^\tau)_{\tau\in(0,\overline{\tau})}$ of solutions to the Westervelt-type JMGT equation converges weakly* in $\bar{X}^W$ to a solution $\overline{\psi}\in\bar{X}^W$ of the Westervelt equation with initial conditions $\overline{\psi}(0)=\psi_0$, $\overline{\psi}_t(0)=\psi_1$ as $\tau\searrow 0$.*

Figures 3–5 illustrate these convergence results by showing some numerical simulations comparing Westervelt-type JMGT and Westervelt solutions in the following setting:

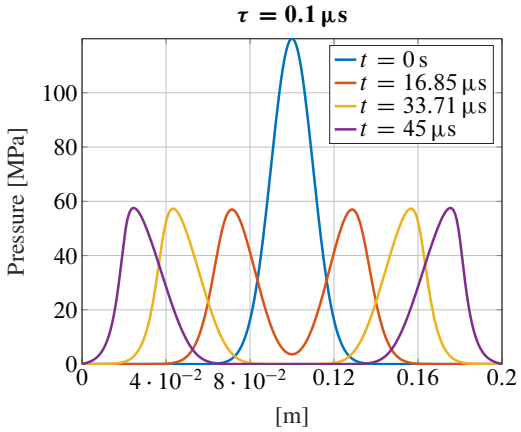


Figure 3. Snapshots of pressure $p = \varrho\psi_t$ for fixed relaxation time $\tau = 0.1\,\mu\text{s}$.

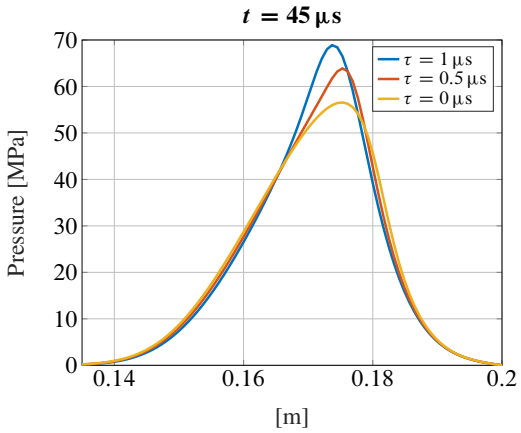
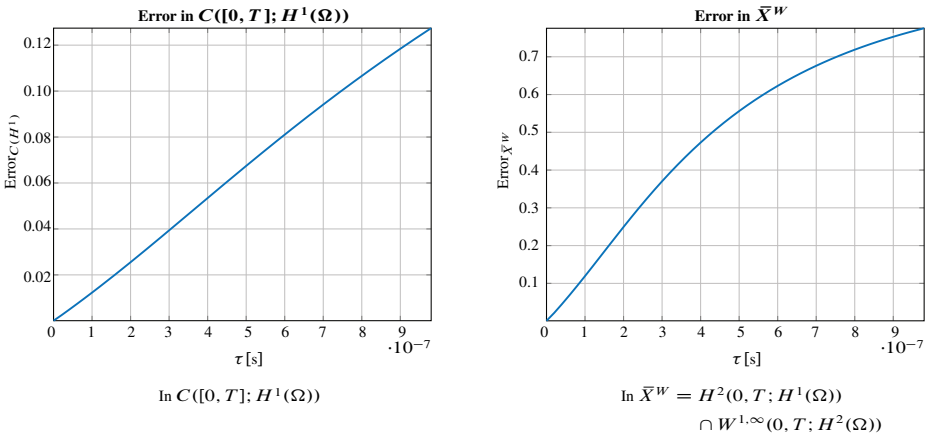


Figure 4. Pressure wave for different relaxation parameters τ at final time $t = 45\,\mu\text{s}$.

- water in a 1D channel geometry

$$c = 1500\,\text{m/s}, \quad \delta = 6 \cdot 10^{-9}\,\text{m}^2/\text{s}, \quad \varrho = 1000\,\text{kg/m}^3, \quad \frac{B}{A} = 5;$$

- space discretization with B-splines (isogeometric analysis): quadratic basis functions, globally C^2; 251 DOFs on $\Omega = [0, 0.2\,\text{m}]$;
- time discretization by Newmark scheme, adapted to third-order equation; 800 time steps on $[0, T] = [0, 45\,\mu\text{s}]$;
- initial conditions $(\psi_0, \psi_1, \psi_2) = \big(0, \mathcal{A}\exp\big(-\frac{(x-0.1)^2}{2\sigma^2}\big), 0\big)$ with amplitude $\mathcal{A} = 8 \cdot 10^4\,\text{m}^2/\text{s}^2$.



In $C([0,T];H^1(\Omega))$

In $\bar{X}^W = H^2(0,T;H^1(\Omega)) \cap W^{1,\infty}(0,T;H^2(\Omega))$

Figure 5. Relative errors as $\tau \to 0$.

3.2. Some further parameter limits

- Vanishing diffusivity of sound $\delta \searrow 0$ in Kuznetsov's equation (2.3):

$$\psi_{tt}^\delta - c^2\Delta\psi^\delta - \delta\Delta\psi_t^\delta = \left(\frac{B}{2Ac^2}(\psi_t^\delta)^2 + |\nabla\psi^\delta|^2\right)_t$$

with an undamped quasilinear wave equation

$$\psi_{tt} - c^2\Delta\psi = \left(\frac{B}{2Ac^2}(\psi_t)^2 + |\nabla\psi|^2\right)_t \tag{3.5}$$

as a limiting problem; cf. [37]. The challenge here lies in the fact that $\delta > 0$ is crucial for global-in-time well-posedness and exponential decay in $d \in \{2,3\}$ space dimensions. As a byproduct, for $\delta = 0$ we recover results (in particular on required regularity of initial data) from [19] on the undamped Westervelt equation and obtain a new result on local-in-time well-posedness of the undamped Kuznetsov equation (3.5). A similar limiting analysis for $\delta \searrow 0$ has been carried out for the JMGT equation (2.5) in [36].

- Vanishing thermal diffusivity in the Blackstock–Crighton equation (2.6):
Here, the aim is to recover solutions to Kuznetsov's equation (2.3) as a limit for $a \searrow 0$ of solutions ψ^a to

$$(\partial_t - a\Delta)(\psi_{tt}^a - c^2\Delta\psi^a - \delta\Delta\psi_t^a) - ra\Delta\psi_t^a = \left(\frac{B}{2Ac^2}(\psi_t^{a\,2}) + \left|\nabla\psi^a\right|^2\right)_{tt};$$

cf. [44]. Note that this process also requires one integration with respect to time. To obtain sufficiently regular solutions (as needed for the underlying arguments

for nondegeneracy), consistency of the initial data is thus needed:

$$\psi_2 - c^2\Delta\psi_0 - \delta\Delta\psi_1 = \frac{B}{Ac^2}\psi_1\psi_2 + 2\nabla\psi_0\cdot\nabla\psi_1.$$

- Limit $\alpha \nearrow 1$ of the differentiation order in fractional JMGT equations:
  Finally, we also consider families of time-fractional versions of the Jordan–Moore–Gibson–Thompson equation (2.5), such as

$$\tau^\alpha\partial_t^{2+\alpha}\psi^\alpha + \psi_{tt}^\alpha - c^2\Delta\psi^\alpha - (\delta + \tau^\alpha c^2)\Delta\partial_t^\alpha\psi^\alpha = \left(\frac{B}{2Ac^2}(\psi_t^\alpha)^2 + |\nabla\psi^\alpha|^2\right)_t$$

  and convergence of their solutions to one of (2.5) (with fixed positive τ); cf. [38] and see also (4.1) below. Like in the $\tau \searrow 0$ limit, the leading order time derivative changes; however, here, this happens in a continuous manner. One of the key tasks in this work was the derivation of proper models from physical balance and constitutive laws. This brings us to the topic of the next section.

4. Fractional attenuation models in ultrasonics

As well known [13], attenuation of ultrasound depends both on the tissue type (which by the way also provides a way of imaging as mentioned in Section 6) and on the frequency. In view of the semilogarithmic scaling in the graph (the attenuation is given in decibels per centimeter), the slopes of the lines correspond to exponents of the frequency in the resulting attenuation law. Since physical balance equations by their nature only contain integer derivatives, fractional (or more generally, nonlocal) modeling happens on the level of constitutive laws. In acoustics, this can be done in terms of

- pressure-density relation and/or
- temperature-heat flux relation.

Below, we will partly describe the derivation of such models in the simpler linear setting; nonlinear versions can be obtained by inserting the constitutive equations into the balance laws and applying the procedure mentioned at the beginning of Section 2.2, leading to second (or higher) order in time wave equations.

As pressure-density relations often rely on analogies to viscoelasticity (which was one of the historical origins of fractional modeling [12]), we start with a brief recap on fractional modeling there.

4.1. Fractional models of (linear) viscoelasticity

In continuum mechanics, the deformation of a body is described by

- the equation of motion (resulting from balance of forces)

$$\varrho \mathbf{u}_{tt} = \operatorname{div} \sigma + \mathbf{f};$$

- the characterization of strain as the symmetric gradient of displacements (in a geometrically linearized setting)

$$\varepsilon = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T);$$

- a constitutive model, that is, a stress-strain relation.

Here, $\mathbf{u}$ is the vector of displacements, σ and ε are the stress and strain tensors and ϱ is the mass density.

Classical examples of constitutive models are

$$\begin{aligned}
\text{Hooke's law (linear elasticity):} &\quad \sigma = b_0 \varepsilon, \\
\text{Newton model:} &\quad \sigma = b_1 \varepsilon_t, \\
\text{Kelvin–Voigt model:} &\quad \sigma = b_0 \varepsilon + b_1 \varepsilon_t, \\
\text{Maxwell model:} &\quad \sigma + a_1 \sigma_t = b_0 \varepsilon, \\
\text{Zener model:} &\quad \sigma + a_1 \sigma_t = b_0 \varepsilon + b_1 \varepsilon_t
\end{aligned}$$

with coefficients a_i, b_i. Replacing the first time derivatives in the equations by fractional ones ∂_t^β (cf. Section 4.4) with $\beta \in [0, 1)$, one arrives at (time-)fractional versions:

$$\begin{aligned}
\text{fractional Newton model:} &\quad \sigma = b_1 \partial_t^\beta \varepsilon, \\
\text{fractional Kelvin–Voigt model:} &\quad \sigma = b_0 \varepsilon + b_1 \partial_t^\beta \varepsilon, \\
\text{fractional Maxwell model:} &\quad \sigma + a_1 \partial_t^\alpha \sigma = b_0 \varepsilon, \\
\text{fractional Zener model:} &\quad \sigma + a_1 \partial_t^\alpha \sigma = b_0 \varepsilon + b_1 \partial_t^\beta \varepsilon \\
\text{which are comprised in the general model class:} &\quad \sum_{n=0}^{N} a_n \partial_t^{\alpha_n} \sigma = \sum_{m=0}^{M} b_m \partial_t^{\beta_m} \varepsilon;
\end{aligned}$$

cf. [3, 12].

4.2. Fractional models of (linear) acoustics via pressure-density relation

As we have already seen in Section 2, the fundamental laws of acoustics are first of all

- balance of momentum

$$\varrho_0 \mathbf{v}_t = -\nabla p + \mathbf{f};$$

- balance of mass

$$\varrho \nabla \cdot \mathbf{v} = -\varrho_t;$$

- combined with an equation of state, the most simple linear one is

$$\frac{\varrho_\sim}{\varrho_0} = \frac{p_\sim}{p_0}$$

and, analogously to the general viscoelastic law above, can be generalized to

$$\sum_{m=0}^{M} b_m \partial_t^{\beta_m} \frac{\varrho_\sim}{\varrho_0} = \sum_{n=0}^{N} a_n \partial_t^{\alpha_n} \frac{p_\sim}{p_0}.$$

Inserting the equation of state into a combination of the balance laws that eliminates the velocity leads to fractional acoustic wave equations. Some of the most commonly used instances are as follows [25, 52]:

- Caputo–Wismer–Kelvin wave equation (fractional Kelvin–Voigt):

$$p_{tt} - b_0 \Delta p - b_1 \partial_t^\beta \Delta p = \widetilde{f};$$

- modified Szabo wave equation (fractional Maxwell):

$$p_{tt} - a_1 \partial_t^{2+\alpha} p - b_0 \Delta p = \widetilde{f};$$

- fractional Zener wave equation:

$$p_{tt} - a_1 \partial_t^{2+\alpha} p - b_0 \Delta p + b_1 \partial_t^\beta \Delta p = \widetilde{f};$$

- general fractional model:

$$\sum_{n=0}^{N} a_n \partial_t^{2+\alpha_n} p - \sum_{m=0}^{M} b_m \partial_t^{\beta_m} \Delta p = \widetilde{f}.$$

We refer to [41, Chapter 7] for their well-posedness analysis.

4.3. Fractional models of nonlinear acoustics via temperature-heat flux relation

Recall that the classical Fourier law (2.1) leads to an infinite speed of propagation paradox, which can be amended by substituting it with the Maxwell–Cattaneo law (2.2), allowing for "thermal waves" (aka the second sound phenomenon). However, this can lead to a violation of the second law of thermodynamics and moreover is unable to reproduce the fractional power law frequency dependence of attenuation as relevant in ultrasonics. Some research in physics has therefore been devoted to

an appropriate and physically consistent modeling of the temperature-heat flux relation. This can be achieved by "interpolating" between (2.1) and (2.2), using fractional derivatives [15, 24, 50]:

$$\begin{aligned}
&\text{(GFE)} && (1+\tau^\alpha \partial_t^\alpha)\boldsymbol{q}(t) = -K\nabla\vartheta,\\
&\text{(GFE I)} && (1+\tau^\alpha \partial_t^\alpha)\boldsymbol{q}(t) = -K\tau_\vartheta^{1-\alpha}\partial_t^{1-\alpha}\nabla\vartheta,\\
&\text{(GFE II)} && (1+\tau^\alpha \partial_t^\alpha)\boldsymbol{q}(t) = -K\tau_\vartheta^{\alpha-1}\partial_t^{\alpha-1}\nabla\vartheta,\\
&\text{(GFE III)} && (1+\tau \partial_t)\boldsymbol{q}(t) = -K\tau_\vartheta^{1-\alpha}\partial_t^{1-\alpha}\nabla\vartheta.
\end{aligned}$$

Using these in place of (2.2) leads to the models

$$\begin{aligned}
\tau^\alpha\partial_t^\alpha\psi_{tt} + (1-\kappa\psi_t)\psi_{tt} - c^2\Delta\psi - \tau^\alpha c^2\partial_t^\alpha\Delta\psi - \delta\Delta\psi_t + \partial_t|\nabla\psi|^2 &= 0,\\
\tau^\alpha\partial_t^\alpha\psi_{tt} + (1-\kappa\psi_t)\psi_{tt} - c^2\Delta\psi - \tau^\alpha c^2\partial_t^\alpha\Delta\psi - \delta\partial_t^{2-\alpha}\Delta\psi + \partial_t|\nabla\psi|^2 &= 0,\\
\tau^\alpha\partial_t^\alpha\psi_{tt} + (1-\kappa\psi_t)\psi_{tt} - c^2\Delta\psi - \tau^\alpha c^2\partial_t^\alpha\Delta\psi - \delta\partial_t^{\alpha}\Delta\psi + \partial_t|\nabla\psi|^2 &= 0,\\
\tau\psi_{ttt} + (1-\kappa\psi_t)\psi_{tt} - c^2\Delta\psi - \tau c^2\Delta\psi_t - \delta\partial_t^{2-\alpha}\Delta\psi + \partial_t|\nabla\psi|^2 &= 0
\end{aligned} \tag{4.1}$$

that have been analyzed in [38].

4.4. Fractional derivatives

To clarify what we mean by ∂_t^α in the models above, we provide a brief intermission on some of the most commonly used fractional derivative concepts.

With the Abel fractional integral operator

$$I_{t_0}^\gamma f(x) = \frac{1}{\Gamma(\gamma)}\int_{t_0}^t \frac{f(s)}{(t-s)^{1-\gamma}}\,ds,$$

a fractional (time) derivative can be defined by either

$${}^{RL}_{t_0}D_t^\alpha f = \frac{d}{dt} I_{t_0}^{1-\alpha} f \quad \text{Riemann–Liouville derivative}$$

or

$${}^{DC}_{t_0}D_t^\alpha f = I_{t_0}^{1-\alpha}\frac{df}{ds} \quad \text{Djrbashian–Caputo derivative.}$$

These are nonlocal operators and have a definite starting point t_0. The R-L derivative is defined on a larger function space than the D-C one. However, the R-L derivative of a constant is nonzero; it even exhibits a singularity at initial time t_0. This is why the D-C derivative, which maps constants to zero, is often preferred in physical applications. Some recent books on fractional PDEs (also containing many classical references on fractional differentiation) are [27, 41, 45].

The nonlocal and causal character of these derivatives provides the models in which they are used with a "memory". In the context of inverse problems, it is important to note that initial values are therefore tied to later values and hence can be better reconstructed backward in time than in integer derivative models.

Some of the challenges in the analysis of fractional PDEs, in particular in deriving energy estimates, are as follows:

- The chain rule identity $u\,u_t = \frac{1}{2}\frac{d}{dt}u^2$ is lost.
- Likewise, there is no product rule of differentiation.

"Substitutes" for these devices in the fractional derivative case are

- the chain rule inequality

$$w(t)\mathrm{D}_t^\alpha w(t) \geq \frac{1}{2}(\mathrm{D}_t^\alpha w^2)(t)$$

  for $w \in W^{1,1}(0,T)$ [2];
- coercivity of the Abel integral operator

$$\int_0^T \langle \mathrm{I}^{1-\alpha} w(s), w(s)\rangle\, ds \geq \cos\left(\frac{\pi(1-\alpha)}{2}\right)\|w\|^2_{H^{-(1-\alpha)/2}(0,T)}$$

  for $w \in H^{-(1-\alpha)/2}(0,T)$ [20, 54];
- the Kato–Ponce inequality

$$\|f\,g\|_{W^{\rho,r}(0,T)} \lesssim \|f\|_{W^{\rho,p_1}(0,T)}\|g\|_{L^{q_1}(0,T)} + \|f\|_{L^{p_2}(0,T)}\|g\|_{W^{\rho,q_2}(0,T)}$$

  for f, g such that the right-hand side is finite, $0 \leq \rho \leq \overline{\rho} < 1$, $1 < r < \infty$, p_1, p_2, $q_1, q_2 \in (1,\infty]$, with $\frac{1}{r} = \frac{1}{p_i} + \frac{1}{q_i}$, $i = 1, 2$; see, e.g., [23].

5. Describing nonlinear wave propagation in the frequency domain by multiharmonic expansions

Motivated by the characteristic appearance of higher harmonics in nonlinear acoustics, cf. Figure 2, we now provide a mathematical formulation of this physical fact.

To this end, we recall the linear wave equation

$$u_{tt} - c^2\Delta u = r, \tag{5.1}$$

where using a harmonic excitation $r(x,t) = \mathrm{Re}(\hat{r}(x)e^{\imath\omega t})$ and a harmonic ansatz for u

$$u(x,t) = \mathrm{Re}(\hat{u}(x)e^{\imath\omega t})$$

leads to the Helmholtz equation

$$-\omega^2\hat{u} - c^2\Delta\hat{u} = \hat{r}.$$

Replacing (5.1) by the Westervelt equation

$$u_{tt} - c^2\Delta u - b\Delta u_t = \kappa(u^2)_{tt} + r \tag{5.2}$$

and still using a harmonic excitation $r(x,t) = \mathrm{Re}(\hat{r}(x)e^{\imath\omega t})$, we expect a response at multiples of the fundamental frequency ω. We therefore use a multiharmonic expansion of u

$$u(x,t) = \mathrm{Re}\left(\sum_{k=1}^{\infty} \hat{u}_k(x)e^{\imath k\omega t}\right), \tag{5.3}$$

which is mathematically justified by completeness of the system $(e^{\imath k\omega\cdot})_{k\in\mathbb{N}}$ in $L^2(0,T)$ as well as the fact that indeed a periodic solution to the Westervelt equation exists [29].

Theorem 5.1. *For $b, c^2, \beta, \gamma, T > 0$, $\kappa \in L^\infty(\Omega)$, there exists $\rho > 0$ such that for all $r \in L^2(0,T;L^2(\Omega))$ with $\|r\|_{L^2(0,T;L^2(\Omega))} \le \rho$, there exists a unique solution*

$$u \in X := H^2(0,T;L^2(\Omega)) \cap H^1(0;T;H^{3/2}(\Omega)) \cap L^2(0;T;H^2(\Omega))$$

of

$$\begin{cases} u_{tt} - c^2\Delta u - b\Delta u_t = \kappa(x)(u^2)_{tt} + r & \text{in } \Omega\times(0,T), \\ \beta u_t + \gamma u + \partial_\nu u = 0 & \text{on } \partial\Omega\times(0,T), \\ u(0) = u(T)\,,\ u_t(0) = u_t(T) & \text{in } \Omega, \end{cases}$$

and the solution fulfills the estimate

$$\|u\|_X \le \widetilde{C}\,\|r\|_{L^2(0,T;L^2(\Omega))}.$$

Inserting the ansatz (5.3) into (5.2) leads to the coupled system:

$$\begin{aligned} m = 1: \quad & -\omega^2\hat{u}_1 - (c^2 + \imath\omega b)\Delta\hat{u}_1 = \hat{r} - \frac{\kappa}{2}\omega^2 \sum_{k=3:2}^{\infty} \overline{\hat{u}_{\frac{k-1}{2}}}\,\hat{u}_{\frac{k+1}{2}}, \\ m = 2,3,\ldots: \quad & -\omega^2 m^2\hat{u}_m - (c^2 + \imath\omega m b)\Delta\hat{u}_m = -\frac{\kappa}{4}\omega^2 m^2 \sum_{\ell=1}^{m-1} \hat{u}_\ell\hat{u}_{m-\ell} \\ & \qquad - \frac{\kappa}{2}\omega^2 m^2 \sum_{k=m+2:2}^{\infty} \overline{\hat{u}_{\frac{k-m}{2}}}\,\hat{u}_{\frac{k+m}{2}}. \end{aligned} \tag{5.4}$$

Note that the quadratic nonlinearity corresponds to a discrete autoconvolution in time. Skipping the last terms on the right-hand side corresponds to the simplification of dropping Re in (5.3) and leads to a triangular system that can be solved by forward substitution in a sequence of Helmholtz problems.

6. Inverse problems

In this section, we mainly focus on an innovative quantitative imaging methodology, based on some of the models in the previous sections, namely acoustic nonlinearity parameter tomography ANT. The key motivation for this comes from the experimentally proven qualification of the $\frac{B}{A}$ parameter in (2.3), (2.4), (2.5) or (2.6) as a characterizing and distinguishing material property of biological tissues. Thus, viewing $\kappa = \frac{1}{\varrho c^2}\left(\frac{B}{2A} + 1\right)$ as a spatially varying coefficient in the Westervelt equation, it can be used for medical imaging. While the physical and engineering literature on ANT has already been developing over many years, with initializing work going back to the 1980s and 1990s [4,10,11,26,57], mathematical results, as they are crucial for enabling any kind of advanced medical imaging technology, have only recently been started here. In particular, we point to [1,53] for results on uniqueness from the Neumann–Dirichlet map in the Westervelt equation and to [21], which investigates and highlights the usefulness of geometric optics solutions to Westervelt equation for this imaging problem. In our own work [30, 39, 40, 42, 43], we mainly focused on uniqueness from a single boundary observation and on convergence of Newton's method in the Westervelt equation as well as in [56] on linearized uniqueness and conditional stability of the inverse problem for the BCBJ model (2.6).

With the Westervelt equation as a forward model, the inverse problem of nonlinearity parameter imaging reads as follows.

Reconstruct $\kappa(x)$ *in*

$$(u - \kappa(x)u^2)_{tt} - c_0^2\Delta u + Du = r \quad \text{in } \Omega \times (0,T),$$
$$\partial_\nu u + \gamma u = 0 \quad \text{on } \partial\Omega \times (0,T), \qquad u(0) = 0, \qquad u_t(0) = 0 \quad \text{in } \Omega$$

(with excitation r*) from observations*

$$g = u \quad \text{on } \Sigma \times (0,T)$$

on a transducer array $\Sigma \subset \overline{\Omega}$*, which we assume to be a surface or collection of discrete points.*

The attenuation term D can be defined by one of the following common fractional models:

$$\begin{aligned}
&\text{Caputo–Wismer–Kelvin:} && D = -b\Delta\partial_t^\beta && \text{with } \beta \in [0,1],\ b \geq 0,\\
&\text{fractional Zener:} && D = a\partial_t^{2+\alpha} - b\Delta\partial_t^\beta && \text{with } a > 0,\ b \geq ac^2,\\
& && && 1 \geq \beta \geq \alpha > 0,\\
&\text{space fractional Chen–Holm:} && D = b(-\Delta)^{\widetilde{\beta}}\partial_t && \text{with } \widetilde{\beta} \in [0,1],\ b \geq 0.
\end{aligned}$$

Since the speed of sound and the attenuation coefficient also depend on the underlying tissue type, they can be considered as additional unknowns in a multicoefficient inverse problem; see, e.g., [30,42].

Chances and challenges of this inverse problem result from

- the difficulties in the analysis of the forward model (as well as its numerics) that have already been pointed to in Section 2;
- the fact that the unknown coefficient $\kappa(x)$ appears in the nonlinear term;
- $\kappa(x)$ being spatially varying, whereas the data $g(t)$ is in the "orthogonal" time direction; this is well known to lead to severe ill-conditioning of the inverse problem.

On the other hand, nonlinearity helps by adding information. In particular, the generation of higher harmonics and the fact that κ appears in each of the equations in (5.4) shows that information multiplies due to nonlinearity. As a consequence, it is in fact possible to achieve enhanced uniqueness results for the inverse problem formulated above, which we here illustrate by a uniqueness result for a general spatially variable coefficient from a single observation, while in the linear setting, this type of data would at most provide identifiability of piecewise constant coefficients.

6.1. Linearized and local nonlinear uniqueness

To this end, recalling (5.4), we reformulate the inverse problem in frequency domain, using the abbreviations

$$\vec{u} = (\hat{u}_j)_{j\in\mathbb{N}} \quad B_m(\vec{u}) = \frac{1}{4}\sum_{\ell=1}^{m-1} \hat{u}_\ell \hat{u}_{m-\ell} + \frac{1}{2}\sum_{k=m+2:2}^{\infty} \overline{\hat{u}_{\frac{k-m}{2}}}\,\hat{u}_{\frac{k+m}{2}},$$

defining the forward operator by

$$F_m(\kappa,\vec{u}) := \begin{pmatrix} -(\omega^2 m^2 + (c^2 + \imath\omega m b)\Delta)\hat{u}_m + \omega^2 m^2\,\kappa\, B_m(\vec{u}) \\ \operatorname{tr}_\Sigma \hat{u}_m \end{pmatrix}$$

and the data by

$$y_m := \begin{pmatrix} \hat{r} \text{ if } m = 1 \ / \ 0 \text{ if } m \geq 2 \\ \hat{g}_m \end{pmatrix}.$$

This allows us to write the inverse problem as a nonlinear operator equation

$$\vec{F}(\kappa, \vec{u}) = y.$$

The linearized inverse problem is defined by considering the linearization of the forward operator

$$\vec{F}'(\kappa, \vec{u})(\underline{d\kappa}, \underline{d\vec{u}}) \approx F(\kappa, \vec{u}) - y,$$

where

$$F'_m(\kappa, \vec{u})(\underline{d\kappa}, \underline{d\vec{u}}) = \begin{pmatrix} -(\omega^2 m^2 + (c^2 + \imath\omega m b)\Delta)\underline{d\hat{u}}_m + \omega^2 m^2 \kappa\, B'_m(\vec{u})\underline{d\vec{u}} + \omega^2 m^2\, \underline{d\kappa}\, B_m(\vec{u}) \\ \operatorname{tr}_\Sigma \underline{d\hat{u}}_m \end{pmatrix}$$

and in particular

$$F'_m(0, \vec{u})(\underline{d\kappa}, \underline{d\vec{u}}) = \begin{pmatrix} -(\omega^2 m^2 + (c^2 + \imath\omega m b)\Delta)\underline{d\hat{u}}_m + \omega^2 m^2\, \underline{d\kappa}\, B_m(\vec{u}) \\ \operatorname{tr}_\Sigma \underline{d\hat{u}}_m \end{pmatrix}.$$

This allows us to formulate the following linearized uniqueness result that has been obtained for the Westervelt equation in [43] and for the JMGT equation in [31].

Theorem 6.1. *The homogeneous linearized (at $\kappa = 0$, $\vec{u} = \phi(x)\vec{\psi}$ with $B_m(\vec{\psi}) \neq 0$, $m \in \mathbb{N}$) inverse problem of nonlinearity coefficient imaging in frequency domain only has the trivial solution, that is, $\vec{F}'(0, \vec{u})$ is injective.*

In the JMGT model, this even enables a (local) nonlinear uniqueness result [31].

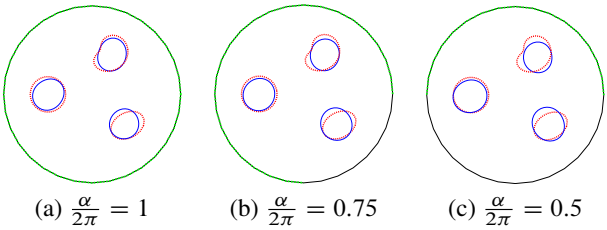
Theorem 6.2. *For $(\kappa, \vec{u})$, $(\widetilde{\kappa}, \widetilde{\vec{u}})$ in a sufficiently small $H^s(\Omega) \times h^1(L^2(\Omega))$ neighborhood of $(0, \phi(x)\vec{\psi})$ with $s \in (\frac{1}{2}, 1]$, the uniqueness result*

$$\vec{F}(\kappa, \vec{u}) = \vec{F}(\widetilde{\kappa}, \widetilde{\vec{u}}) \quad \Rightarrow \quad \kappa = \widetilde{\kappa} \text{ and } \vec{u} = \widetilde{\vec{u}}$$

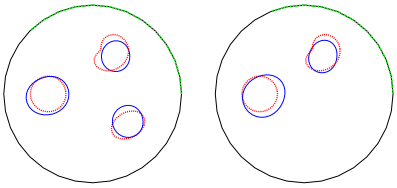
holds.

6.2. Some reconstruction results from two harmonics

Finally, we provide some images under the (practically often relevant) assumption of a piecewise constant nonlinearity coefficient, using only the information provided by the first two lines $m \in \{1, 2\}$ of the multiharmonic system (5.4). The reconstruction



(a) $\frac{\alpha}{2\pi} = 1$ (b) $\frac{\alpha}{2\pi} = 0.75$ (c) $\frac{\alpha}{2\pi} = 0.5$



(d) $\frac{\alpha}{2\pi} = 0.4$ (e) $\frac{\alpha}{2\pi} = 0.3$

Figure 6. Reconstructions from limited data: measurements are taken on the green part of the boundary; α corresponds to the observation angle.

was carried out by parametrizing the boundary of the interface curves and applying Newton's method to the nonlinear system that arises from matching measurements to simulations. Starting curves for Newton's method were found by applying a sparsity-enhancing point source insertion algorithm [7] and extending the discovered point sources to disks whose radii are determined by a mean value property of solutions to the Helmholtz equation. While the geometry in Figures 6–8 is clearly idealized, we mimic a realistic measurement scenario by adding 1% random noise to the simulated data. For details, we refer to [43], from which the pictures shown here are also taken.

In Figure 6, we consider three objects and study the impact of limited data on the reconstruction quality, which clearly decreases under loss of information. Remarkably, all three objects are still found with less than half of the full view, and only upon reduction to less than a third of the data does one of the three objects become invisible.

Figure 7 illustrates a study of the influence of the distance between two objects on their reconstructability. Obviously, imaging both of them becomes harder as they move closer to each other. Still, up to the point where the two inclusions touch, they can be identified as two objects.

Finally, in Figure 8, we visualize how attenuation compromises the reconstruction as the object moves further away from the boundary on which observations are taken.

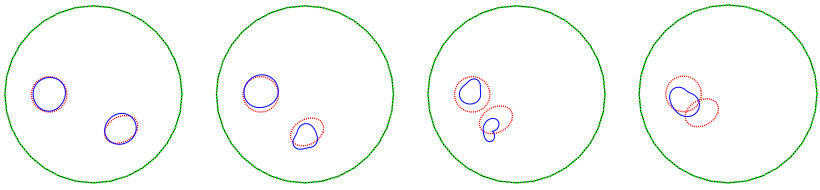


Figure 7. Reconstruction of two inclusions at different distances from each other; θ corresponds to the distance in polar coordinates.

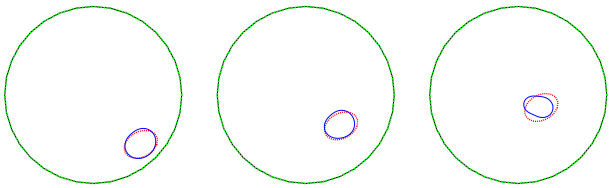


Figure 8. Reconstruction of one inclusion at different distances from the boundary.

Acknowledgments. The author thanks Benjamin Rainer, University of Klagenfurt and Austrian Institute of Technology, for providing the illustrative images in Figures 1 and 2.

Funding. This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P36318].

References

[1] S. Acosta, G. Uhlmann, and J. Zhai, Nonlinear ultrasound imaging modeled by a Westervelt equation. *SIAM J. Appl. Math.* **82** (2022), no. 2, 408–426 Zbl 1486.35458 MR 4393205

[2] A. A. Alikhanov, A priori estimates for solutions of boundary value problems for fractional-order equations (in Russian). *Differ. Uravn.* **46** (2010), no. 5, 658–664. English translation. *Differ. Equ.* **46** (2010), no. 5, 660–666 Zbl 1208.35161 MR 2797545

[3] T. M. Atanacković, S. Pilipović, B. Stanković, and D. Zorica, *Fractional calculus with applications in mechanics. Wave propagation, impact and variational principles*. Mech. Eng. Solid Mech. Ser., John Wiley & Sons, Hoboken, NJ, 2014, 406 pp. Zbl 1291.74001 MR 3242674

[4] L. Bjørnø, Characterization of biological media by means of their non-linearity. *Ultrasonics* **24** (1986), no. 5, 254–259

[5] D. T. Blackstock, *Approximate equations governing finite-amplitude sound in thermoviscous fluids*. Technical report, GD/E Report GD-1463-52, General Dynamics, Rochester, New York, 1963

[6] M. Bongarti, S. Charoenphon, and I. Lasiecka, Vanishing relaxation time dynamics of the Jordan–Moore–Gibson–Thompson equation arising in nonlinear acoustics. *J. Evol. Equ.* **21** (2021), no. 3, 3553–3584 Zbl 1480.35028 MR 4350283

[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] R. Brunnhuber and P. M. Jordan, On the reduction of Blackstock's model of thermoviscous compressible flow via Becker's assumption. *Int. J. Non-Linear Mech.* **78** (2016), 131–132

[9] F. Bucci and I. Lasiecka, Feedback control of the acoustic pressure in ultrasonic wave propagation. *Optimization* **68** (2019), no. 10, 1811–1854 Zbl 1429.49038 MR 4003786

[10] V. A. Burov, I. E. Gurinovich, O. V. Rudenko, and E. Ya. Tagunov, Reconstruction of the spatial distribution of the nonlinearity parameter and sound velocity in acoustic nonlinear tomography. *Acoustical Phys.* **40** (1994), no. 6, 816–823

[11] C. A. Cain, Ultrasonic reflection mode imaging of the nonlinear parameter B/A: I. A theoretical basis. *J. Acoustical Soc. Amer.* **80** (1986), no. 1, 28–32

[12] M. Caputo, Linear models of dissipation whose Q is almost frequency independent—II. *Geophys. J. R. Astronom. Soc.* **13** (1967), no. 5, 529–539

[13] V. Chan and A. Perlas, Basics of ultrasound imaging. In S. Narouze (ed.), *Atlas of ultrasound-guided procedures in interventional pain management*, pp. 13–19, Springer, New York, NY, 2011

[14] C. I. Christov, On frame indifferent formulation of the Maxwell–Cattaneo model of finite-speed heat conduction. *Mech. Res. Comm.* **36** (2009), no. 4, 481–486 Zbl 1258.80001 MR 2510197

[15] A. Compte and R. Metzler, The generalized Cattaneo equation for the description of anomalous transport processes. *J. Phys. A* **30** (1997), no. 21, 7277–7289 Zbl 0941.82046 MR 1603438

[16] D. G. Crighton, Model equations of nonlinear acoustics. *Annu. Rev. Fluid Mech.* **11** (1979), 11–33 Zbl 0443.76074

[17] A. Dekkers and A. Rozanova-Pierrat, Cauchy problem for the Kuznetsov equation. *Discrete Contin. Dyn. Syst.* **39** (2019), no. 1, 277–307 Zbl 1406.35194 MR 3918174

[18] F. Dell'Oro, I. Lasiecka, and V. Pata, The Moore–Gibson–Thompson equation with memory in the critical case. *J. Differential Equations* **261** (2016), no. 7, 4188–4222 Zbl 1347.35034 MR 3532069

[19] W. Dörfler, H. Gerner, and R. Schnaubelt, Local well-posedness of a quasilinear wave equation. *Appl. Anal.* **95** (2016), no. 9, 2110–2123 Zbl 1350.35128 MR 3515098

[20] P. P. B. Eggermont, On Galerkin methods for Abel-type integral equations. *SIAM J. Numer. Anal.* **25** (1988), no. 5, 1093–1117 Zbl 0665.65096 MR 0960868

[21] N. Eptaminitakis and P. Stefanov, Weakly nonlinear geometric optics for the Westervelt equation and recovery of the nonlinearity. *SIAM J. Math. Anal.* **56** (2024), no. 1, 801–819 Zbl 1532.35526 MR 4688692

[22] M. Fritz, V. Nikolić, and B. Wohlmuth, Well-posedness and numerical treatment of the Blackstock equation in nonlinear acoustics. *Math. Models Methods Appl. Sci.* **28** (2018), no. 13, 2557–2597 Zbl 1421.35217 MR 3884259

[23] L. Grafakos and S. Oh, The Kato–Ponce inequality. *Comm. Partial Differential Equations* **39** (2014), no. 6, 1128–1157 Zbl 1301.42026 MR 3200091

[24] M. E. Gurtin and A. C. Pipkin, A general theory of heat conduction with finite wave speeds. *Arch. Ration. Mech. Anal.* **31** (1968), no. 2, 113–126 Zbl 0164.12901 MR 1553521

[25] S. Holm, *Waves with power-law attenuation*. Springer, Cham, 2019, 312 pp. Zbl 1412.76001 MR 3930618

[26] N. Ichida, T. Sato, and M. Linzer, Imaging the nonlinear ultrasonic parameter of a medium. *Ultrasonic Imaging* **5** (1983), no. 4, 295–299

[27] B. Jin, *Fractional differential equations. An approach via fractional derivatives*. Appl. Math. Sci. 206, Springer, Cham, 2021, 368 pp. Zbl 1466.34002 MR 4290515

[28] P. M. Jordan, Second-sound phenomena in inviscid, thermally relaxing gases. *Discrete Contin. Dyn. Syst. Ser. B* **19** (2014), no. 7, 2189–2205 Zbl 1302.76095 MR 3253253

[29] B. Kaltenbacher, Periodic solutions and multiharmonic expansions for the Westervelt equation. *Evol. Equ. Control Theory* **10** (2021), no. 2, 229–247 Zbl 1476.35131 MR 4218100

[30] B. Kaltenbacher, Identifiability of some space dependent coefficients in a wave equation of nonlinear acoustics. *Evol. Equ. Control Theory* **13** (2024), no. 2, 421–444 Zbl 1540.35468 MR 4701200

[31] B. Kaltenbacher, Acoustic nonlinearity parameter tomography with the Jordan–Moore–Gibson–Thompson equation in frequency domain. *Inverse Probl.* **41** (2025), no. 9, article no. 095010, 27 pp. Zbl 08105152 MR 4961235

[32] B. Kaltenbacher and I. Lasiecka, An analysis of nonhomogeneous Kuznetsov's equation: local and global well-posedness; exponential decay. *Math. Nachr.* **285** (2012), no. 2–3, 295–321 Zbl 1235.35040 MR 2881283

[33] B. Kaltenbacher, I. Lasiecka, and M. K. Pospieszalska, Well-posedness and exponential decay of the energy in the nonlinear Jordan–Moore–Gibson–Thompson equation arising in high intensity ultrasound. *Math. Models Methods Appl. Sci.* **22** (2012), no. 11, article no. 1250035, 34 pp. Zbl 1257.35131 MR 2974173

[34] B. Kaltenbacher and V. Nikolić, The Jordan–Moore–Gibson–Thompson equation: well-posedness with quadratic gradient nonlinearity and singular limit for vanishing relaxation time. *Math. Models Methods Appl. Sci.* **29** (2019), no. 13, 2523–2556 Zbl 1427.35206 MR 4038340

[35] B. Kaltenbacher and V. Nikolić, Vanishing relaxation time limit of the Jordan–Moore–Gibson–Thompson wave equation with Neumann and absorbing boundary conditions. *Pure Appl. Funct. Anal.* **5** (2020), no. 1, 1–26 Zbl 1461.35156 MR 4061168

[36] B. Kaltenbacher and V. Nikolić, The inviscid limit of third-order linear and nonlinear acoustic equations. *SIAM J. Appl. Math.* **81** (2021), no. 4, 1461–1482 Zbl 1475.35282 MR 4287360

[37] B. Kaltenbacher and V. Nikolić, Parabolic approximation of quasilinear wave equations with applications in nonlinear acoustics. *SIAM J. Math. Anal.* **54** (2022), no. 2, 1593–1622 Zbl 1485.35296 MR 4392107

[38] B. Kaltenbacher and V. Nikolić, Time-fractional Moore–Gibson–Thompson equations. *Math. Models Methods Appl. Sci.* **32** (2022), no. 5, 965–1013 Zbl 1491.35433 MR 4430362

[39] B. Kaltenbacher and W. Rundell, On the identification of the nonlinearity parameter in the Westervelt equation from boundary measurements. *Inverse Probl. Imaging* **15** (2021), no. 5, 865–891 Zbl 1472.35453 MR 4301274

[40] B. Kaltenbacher and W. Rundell, On an inverse problem of nonlinear imaging with fractional damping. *Math. Comp.* **91** (2022), no. 333, 245–276 Zbl 1479.35951 MR 4350539

[41] B. Kaltenbacher and W. Rundell, *Inverse problems for fractional partial differential equations*. Grad. Stud. Math. 230, American Mathematical Society, Providence, RI, 2023, 505 pp. Zbl 1536.35002 MR 4604107

[42] B. Kaltenbacher and W. Rundell, On the simultaneous reconstruction of the nonlinearity coefficient and the sound speed in the Westervelt equation. *Inverse Probl.* **39** (2023), no. 10, article no. 105001, 18 pp. Zbl 1527.35494 MR 4646066

[43] B. Kaltenbacher and W. Rundell, Nonlinearity parameter imaging in the frequency domain. *Inverse Probl. Imaging* **18** (2024), no. 2, 388–405 Zbl 1535.35209 MR 4714558

[44] B. Kaltenbacher and M. Thalhammer, Fundamental models in nonlinear acoustics. Part I. Analytical comparison. *Math. Models Methods Appl. Sci.* **28** (2018), no. 12, 2403–2455 Zbl 1421.35226 MR 3871730

[45] A. Kubica, K. Ryszewska, and M. Yamamoto, *Time-fractional differential equations. A theoretical introduction*. SpringerBriefs Math., Springer, Singapore, 2020, 134 pp. Zbl 1485.34002 MR 4200127

[46] V. P. Kuznetsov, Equations of nonlinear acoustics. *Sov. Phys. Acoust.* **16** (1971), 467–470

[47] M. B. Lesser and R. Seebass, The structure of a weak shock wave undergoing reflexion from a wall. *J. Fluid Mech.* **31** (1968), no. 3, 501–528 Zbl 0155.54503

[48] K. Mizohata and S. Ukai, The global existence of small amplitude solutions to the nonlinear acoustic wave equation. *J. Math. Kyoto Univ.* **33** (1993), no. 2, 505–522 Zbl 0794.35108 MR 1231754

[49] V. Nikolić, Local existence results for the Westervelt equation with nonlinear damping and Neumann as well as absorbing boundary conditions. *J. Math. Anal. Appl.* **427** (2015), no. 2, 1131–1167 Zbl 1315.35059 MR 3323026

[50] Y. Z. Povstenko, Fractional Cattaneo-type equations and generalized thermoelasticity. *J. Thermal Stresses* **34** (2011), no. 2, 97–114

[51] B. Straughan, Acoustic waves in a Cattaneo–Christov gas. *Phys. Lett. A* **374** (2010), no. 26, 2667–2669 Zbl 1237.76154 MR 2645423

[52] T. L. Szabo, Time domain wave equations for lossy media obeying a frequency power law. *J. Acoustical Soc. Amer.* **96** (1994), no. 1, 491–500

[53] G. Uhlmann and Y. Zhang, An inverse boundary value problem arising in nonlinear acoustics. *SIAM J. Math. Anal.* **55** (2023), no. 2, 1364–1404 Zbl 1515.35358 MR 4580338

[54] U. Vögeli, K. Nedaiasl, and S. A. Sauter, A fully discrete Galerkin method for Abel-type integral equations. *Adv. Comput. Math.* **44** (2018), no. 5, 1601–1626 Zbl 1402.65185 MR 3874031

[55] P. J. Westervelt, Parametric acoustic array. *J. Acoustical Soc. Amer.* **35** (1963), no. 4, 535–537

[56] M. Yamamoto and B. Kaltenbacher, An inverse source problem related to acoustic nonlinearity parameter imaging. In B. Kaltenbacher, T. Schuster, and A. Wald (eds.), *Time-dependent problems in imaging and parameter identification*, pp. 413–456, Springer, Cham, 2021 Zbl 1489.35292

[57] D. Zhang, X. Gong, and S. Ye, Acoustic nonlinearity parameter tomography for biological specimens via measurements of the second harmonics. *J. Acoustical Soc. Amer.* **99** (1996), no. 4, 2397–2402

Barbara Kaltenbacher
Department of Mathematics, University of Klagenfurt, Universitätsstraße 65-67,
9020 Klagenfurt, Austria; barbara.kaltenbacher@aau.at

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/21

Introduction to inverse problems for non-linear partial differential equations

Matti Lassas

Abstract. We consider inverse problems for non-linear hyperbolic and elliptic equations and give an introduction to the method based on the multiple linearization, or on the construction of artificial sources, to solve these problems. The method is based on self-interaction of linearized waves or other solutions in the presence of non-linearities. Multiple linearization has successfully been used to solve inverse problems for non-linear equations which are still unsolved for the corresponding linear equations.

1. Introduction

In this paper, we consider inverse problems for non-linear elliptic and hyperbolic partial differential equations to demonstrate the *multiple linearization* method. In this method, the non-linearity helps one to solve inverse problems for non-linear equations which are still unsolved for the corresponding linear equations. Our aim is to give a review of recent progress on these inverse problems and discuss the main ideas of the proofs in simple cases.

In the case, where the underlying equation is linear, a standard example is the inverse problem of Calderón [15]. This problem arises in electrical impedance tomography (EIT), a method proposed for medical and industrial imaging, where the objective is to determine the electrical conductivity or impedance in a medium by making voltage-to-current measurements on its boundary. Another example of Calderón's inverse problem is the determination of an unknown potential $q(x)$ in a Schrödinger operator $\Delta + q(x)$ from boundary measurements, first solved in [63] in dimensions $n \geq 3$.

Let us next consider analogous inverse problems under non-linear settings. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^∞-smooth boundary, and consider the equation

$$\Delta u + a(x, u) = 0 \quad \text{in } \Omega.$$

Mathematics Subject Classification 2020: 35R30.
Keywords: inverse problems, non-linear PDEs, Lorentzian and Riemannian manifolds.

The Dirichlet problem for this equation is related to maintaining a temperature (or concentration or population) f on the boundary. The boundary measurements for such an equation, provided that it is well posed for some class of boundary values, may be encoded by a Dirichlet-to-Neumann map Λ_a, which maps the boundary value f to the flux (i.e., the normal derivative) $\Lambda_a(f) = \partial_\nu u|_{\partial\Omega}$ of the corresponding equilibrium state across the boundary.

Since the 1990s, inverse problems for non-linear elliptic equations have been widely studied. A standard method, introduced in [29] in the parabolic case, is to show that the first linearization of the non-linear Dirichlet-to-Neumann map is actually the Dirichlet-to-Neumann map of a linear equation and to use the theory of inverse problems for linear equations. For the semilinear Schrödinger equation $\Delta u + a(x, u) = 0$, the problem of recovering the potential $a(x, u)$ was studied in [61] in dimensions $n \geq 3$ and in [26, 30] in dimension $n = 2$. In addition, inverse problems have been studied for quasilinear elliptic equations [31, 58, 62], the degenerate elliptic p-Laplace equation [13, 59], and the fractional semilinear Schrödinger equation [45]. Certain Calderón-type inverse problems for quasilinear equations on Riemannian manifolds are considered in [51]; see also references in [1, 65].

Let us introduce the mathematical setting to study canonical examples of non-linear elliptic equations. We will denote by (M, g) a compact Riemannian manifold with C^∞ boundary ∂M, where $\dim(M) = n$, $n \geq 2$. Let $q \in C^\infty(M)$. We will consider semilinear elliptic equations of the form

$$\begin{cases} \Delta_g u + q u^m = 0 & \text{in } M, \\ u = f & \text{on } \partial M, \end{cases} \tag{1.1}$$

where $m \in \mathbb{N}$ and $m \geq 2$. Here, Δ_g is the Laplace–Beltrami operator, given in local coordinates by

$$\Delta_g u = \frac{1}{\det(g)^{1/2}} \sum_{j,k=1}^{n} \frac{\partial}{\partial x^j} \left(\det(g)^{1/2} g^{jk} \frac{\partial u}{\partial x^k} \right),$$

where $g = (g_{jk}(x))$ and $g^{-1} = (g^{jk}(x))$. As usual, we denote the Euclidean Laplace operator, corresponding to the Euclidean metric $g^{(e)} = \delta_{jk}$ in $\mathbb{R}^n$ by Δ.

The Dirichlet problem for (1.1) with a boundary condition $u|_{\partial\Omega} = f$ has a unique small solution u for sufficiently small boundary data $f \in C^s(\partial M)$, where $s > 2$ with $s \notin \mathbb{N}$. More precisely, this means that there is $\delta > 0$ such that whenever $\|f\|_{C^s(\partial M)} \leq \delta$, there is a unique solution u to (1.1) satisfying $\|u\|_{C^s(M)} \leq c\delta$. We will call u the unique small solution and denote it by u_f to indicate the boundary value. Here, C^s is the standard Hölder space for $s > 2$ with $s \notin \mathbb{N}$ (often written

as $C^{k,\alpha}$ if $s = k + \alpha$, where $k \in \mathbb{Z}$ and $0 < \alpha < 1$). Hence, the Dirichlet-to-Neumann map is defined by using the unique small solution in the following way:

$$\Lambda_{M,g,q} : \{f \in C^s(\partial M) \mid \|f\|_{C^s(\partial M)} \leq \delta\} \to C^{s-1}(\partial M),$$
$$\Lambda_{M,g,q} : f \mapsto \partial_\nu u_f|_{\partial M},$$

where ∂_ν denotes the normal derivative on the boundary ∂M. In what follows, we denote the Dirichlet-to-Neumann map by $\Lambda_{M,g}$ to when $q = 0$. When $M = \Omega \subset \mathbb{R}^n$ and g is the identity matrix, we denote the Dirichlet-to-Neumann map by Λ_q.

Next, we consider the case when $\Omega \subset \mathbb{R}^n$ for $n \geq 2$ and Δ_g is the Euclidean Laplacian.

Theorem 1.1 ([46, Reconstruction of non-linear term]). *Let $n \geq 2$, and let $\Omega \subset \mathbb{R}^n$ be a bounded domain with C^∞ boundary $\partial\Omega$. Let $q_1, q_2 \in C^\infty(\overline{\Omega})$. Assume the Dirichlet-to-Neumann maps Λ_{q_j} for equation (1.1) with potentials q_j, $j = 1, 2$, satisfy*

$$\Lambda_{q_1}(f) = \Lambda_{q_2}(f)$$

for all $f \in C^s(\partial\Omega)$ with $\|f\|_{C^s(\partial M)} < \delta$, where $\delta > 0$ is any sufficiently small number. Then $q_1 = q_2$ in Ω.

In the pioneering studies of the problem, the first- and the second-order linearizations of the non-linear Dirichlet-to-Neumann map have been used in the works [61, 62] related to non-linear equations with matrix coefficients and in [31] for a non-linear conductivity equation (see also [18]).

Next, we present the outline of the proof of Theorem 1.1. For equation (1.1) with quadratic non-linearity, that is, $m = 2$, the first linearization of the non-linear Dirichlet-to-Neumann map Λ_q, linearized at the zero boundary value, is just the Dirichlet-to-Neumann map for the standard Laplace equation

$$D|_{f=0}\Lambda_q(f) : C^s(\partial\Omega) \to C^{s-1}(\partial\Omega), \quad f \mapsto \partial_\nu v_f|_{\partial\Omega},$$

where $s > 1$ is not an integer, v_f is the unique solution of $\Delta v_f = 0$ in Ω with $v_f|_{\partial\Omega} = f$. Thus, the first linearization does not carry any information about the unknown potential q. For a quadratic non-linearity, the *second linearization* of the map $f \to \Lambda_q(f)$, denoted by $D^2|_{f=0}\Lambda_q(f)[\cdot,\cdot]$, is a symmetric bilinear form from $C^s(\partial\Omega) \times C^s(\partial\Omega)$ to $C^{s-1}(\partial\Omega)$. The second-order linearization is given by

$$D^2|_{f=0}\Lambda_q(f)[f_1, f_2] = \partial_{\varepsilon_1}\partial_{\varepsilon_2} u_{\varepsilon_1 f_1 + \varepsilon_2 f_2}|_{\varepsilon_1=\varepsilon_2=0} \quad \text{on } \partial\Omega.$$

That is, one considers boundary data $f = \varepsilon_1 f_1 + \varepsilon_2 f_2 \in C^s(\partial\Omega)$, where $\varepsilon_1, \varepsilon_2$ are sufficiently small parameters, and takes the mixed derivative

$$w(x) = \frac{\partial}{\partial\varepsilon_1}\frac{\partial}{\partial\varepsilon_2}\bigg|_{\varepsilon_1=\varepsilon_2=0} u_{\varepsilon_1 f_1 + \varepsilon_2 f_2}(x)$$

of the solutions of $u_{\varepsilon_1 f_1+\varepsilon_2 f_2}(x)$ of equation (1.1) with the coefficients q, $m = 2$ and the boundary value $f = \varepsilon_1 f_1 + \varepsilon_2 f_2$. This yields the equation

$$\Delta w = -2q v_{f_1} v_{f_2}, \tag{1.2}$$

where v_{f_k} are harmonic functions, that is, the solutions to the linearized equation $\Delta v = 0$ with the Dirichlet boundary value $v_{f_2}|_{\partial\Omega} = f_k$. Moreover,

$$D^2|_{f=0}\Lambda_q(f)[f_1, f_2] = \partial_\nu w|_{\partial\Omega}.$$

Assume next that there are two potential functions q_1 and q_2 such that

$$\Lambda_{q_1}(f) = \Lambda_{q_2}(f) \quad \text{for all } f \in C^s(\partial M),\ \|f\|_{C^s(\partial M)} \le \delta.$$

Taking the mixed second-order derivative of the Dirichlet-to-Neumann maps yields $\partial_\nu w_1 = \partial_\nu w_2$ on $\partial\Omega$, where $\Delta w_j = -2q_j v_{f_1} v_{f_2}$ and $w_j|_{\partial\Omega} = 0$. Subtracting equation (1.2) for q_1 and q_2 and integrating the resulting equation against the harmonic function v_{f_3} and using integration by parts yields the formula

$$\begin{aligned}\int_\Omega (q_1 - q_2) v_{f_1} v_{f_2} v_{f_3}\, dx &= -\frac{1}{2}\int_\Omega (\Delta w_1 - \Delta w_2) v_{f_3}\, dx \\ &= -\frac{1}{2}\int_{\partial\Omega} (\partial_\nu w_1 - \partial_\nu w_2) v_{f_3}\, dS(x) = 0. \end{aligned}\tag{1.3}$$

To apply this formula, let us denote $Q = q_1 - q_2$. When $k \in \mathbb{R}^3$, there are $\xi_1, \xi_2 \in \mathbb{C}^3$ satisfying

$$\xi_1 \cdot \xi_1 = 0, \quad \xi_2 \cdot \xi_2 = 0, \quad \xi_1 + \xi_2 = ik,$$

where $\eta \cdot \zeta = \eta_1\zeta_1 + \eta_2\zeta_2 + \eta_3\zeta_3$ for $\eta = (\eta_1, \eta_2, \eta_3), \zeta = (\zeta_1, \zeta_2, \zeta_3) \in \mathbb{C}^3$. We define

$$v_1(x) = e^{\xi_1 \cdot x}, \quad v_2(x) = e^{\xi_2 \cdot x}, \quad v_3(x) = 1.$$

These functions satisfy $\Delta v_j = 0$, and we can use formula (1.3) with $f_j = v_j|_{\partial\Omega}$, that is, $v_{f_j} = f_j$. The Fourier transform $\widehat{Q}(k)$ of $Q(x)$ (i.e., extended as a zero function in $\mathbb{R}^n \setminus \Omega$) satisfies

$$\widehat{Q}(k) = \int_\Omega Q(x) e^{ik\cdot x}\, dx = \int_\Omega Q(x) e^{\xi_1 \cdot x} \cdot e^{\xi_2 \cdot x} \cdot 1\, dx = \int_\Omega Q v_1 v_2 v_3\, dx = 0.$$

Hence, $Q = 0$, that is, $q_1 = q_2$. These arguments (with some details) prove Theorem 1.1.

Analogous inverse problems for various non-linear elliptic equations using multiple linearization has been studied in [16, 21, 35, 39–41].

We move on to describe our next result. By using higher-order linearizations, we can solve the following *simultaneous recovery* on a two-dimensional Riemannian surface and coefficients of equations.

Theorem 1.2 ([46, Simultaneous recovery of metric and potential]). *Let* (M_1, g_1) *and* (M_2, g_2) *be two compact connected manifolds with mutual* C^∞ *smooth boundaries* $\partial M_1 = \partial M_2 =: \partial M$ *and* $\dim(M_1) = \dim(M_2) = 2$. *Let* $m \geq 2$, *and let* Λ_{M_j, g_j, q_j} *be the Dirichlet-to-Neumann maps* $\Lambda_{M_j, q_j} : f \mapsto \partial_\nu u|_{\partial M}$ *of the problem*

$$\Delta_{g_j} u + q_j u^m = 0 \quad \text{in } M_j$$

for $j = 1, 2$. *Let* $s > 2$ *with* $s \notin \mathbb{N}$ *and assume that*

$$\Lambda_{M_1, g_1, q_1} f = \Lambda_{M_2, g_2, q_2} f \quad \text{on } \partial M$$

for any $f \in C^s(\partial M)$ *with* $\|f\|_{C^s(\partial M)} \leq \delta$, *where* $\delta > 0$ *is sufficiently small. Then there exist a diffeomorphism* $\Psi : M_1 \to M_2$ *and a positive smooth function* σ *on* M_1 *such that for* $x \in M_1$, *we have*

$$(\sigma \Psi^* g_2)(x) = g_1(x), \quad \sigma q_1 = q_2 \circ \Psi \quad \text{in } M_1, \tag{1.4}$$

with $\Psi|_{\partial M} = \mathrm{Id}$ *and* $\sigma|_{\partial M} = 1$.

For the linear equation $\Delta_g u + qu = 0$, an analogous result has been proved when M is a domain in $\mathbb{R}^2$ with a Riemannian metric [28], when M is a manifold and the potential q is zero [55], and when the conformal class of the manifold (M, g) is a priori known [24]. The recovery of properties of both the manifold and potential up to the gauge transformation (1.4) was a long-standing open question that was solved in 2024 in by Cârstea, Liimatainen, and Tzou (see [17]) using much more advanced techniques that the problem for the non-linear equation needs.

Next, we consider the *inverse obstacle problem* for elliptic equations with power-type non-linearities. Let $\Omega \subset \mathbb{R}^n$ be a bounded, connected open set with a C^∞-smooth boundary $\partial\Omega$. Let D be an unknown cavity in Ω. More precisely, $D \subset \Omega$ is an open set with a smooth boundary ∂D such that $\overline{D} \subset \Omega$. Moreover, let $m \in \mathbb{N}$ with $m \geq 2$. Consider the following semilinear elliptic equation, with an unknown coefficient $q = q(x)$ and an unknown relatively compact domain $D \subset \Omega$:

$$\begin{cases} \Delta u + qu^m = 0 & \text{in } \Omega \setminus \overline{D}, \\ u = 0 & \text{on } \partial D, \\ u = f & \text{on } \partial\Omega. \end{cases} \tag{1.5}$$

For $s > 1$ and $s \in \mathbb{N}$, let $f \in C^s(\partial M)$ with $\|f\|_{C^s(\partial\Omega)} < \delta$, where $\delta > 0$ is any sufficiently small number. Then by the well-posedness of (1.5) [46, Proposition 2.1] again, one can define the corresponding Dirichlet-to-Neumann map $\Lambda_{D,q}$ on $\partial\Omega$ with

$$\Lambda_{D,q} : \{f \in C^s(\partial M) \mid \|f\|_{C^s(\partial M)} \leq \delta\} \to C^{s-1}(\partial\Omega), \quad \Lambda_{D,q} : f \mapsto \partial_\nu u_f|_{\partial\Omega},$$

where $s > 1$ is not an integer and u is the unique small solution of (1.5). The inverse obstacle problem is to determine both the unknown cavity D and potential q from the Dirichlet-to-Neumann map $\Lambda_{D,q}$.

Theorem 1.3 ([47, Simultaneous recovery: Cavity and potential]). *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain for $n \geq 2$ with a C^∞-smooth boundary $\partial\Omega$. Let $D_1, D_2 \Subset \Omega$ be non-empty open cavities with C^∞-smooth boundaries and $q_j \in C^\infty(\Omega \setminus \overline{D_j})$. Let $\Lambda_{D_j,q_j} : f \mapsto \partial_\nu u|_{\partial\Omega}$, $j = 1, 2$, be the Dirichlet-to-Neumann maps for equation (1.5) with $D = D_j$ and $q = q_j$, $j = 1, 2$. Assume that $\Lambda_{D_1,q_1} f = \Lambda_{D_2,q_2} f$ for all sufficiently small $f \in C^\infty(\partial\Omega)$ on $\partial\Omega$. Then $D_1 = D_2$ and $q_1 = q_2$ in $\Omega \setminus \overline{D_1}$.*

The simultaneous recovery of an obstacle and an unknown surrounding potential in dimensions $n \geq 3$ is a long-standing open problem. This problem is closely related to the *partial data* Calderón problem studied in [27,34], where the boundary is known a priori, but the boundary measurements are done only on a subset of the boundary.

2. Inverse problems for waves

Let us consider a linear wave equation

$$\partial_t^2 u(t,x) - c(x)^2 \Delta u(t,x) = 0 \quad \text{in } \mathbb{R} \times \Omega,$$

where $\Omega \subset \mathbb{R}^n$ and $c(x)$ is the wave speed. This equation models, e.g., acoustic waves in an isotropic medium. In inverse problems, one has access to measurements of waves (the solutions $u(t,x)$) on the boundary, or in a subset of the domain Ω, and one aims to determine unknown coefficients (e.g., $c(x)$) in the interior of the domain.

We will consider *anisotropic* materials, where the wave speed depends on the direction of propagation. This means that the scalar wave speed $c(x)$, where $x = (x^1, x^2, \dots, x^n) \in \Omega$, is replaced by a positive definite symmetric matrix $(g^{jk}(x))_{j,k=1}^n$. Anisotropic materials appear frequently in applications such as seismic imaging, where one wishes to determine the interior structure of the Earth by making various measurements of waves on its surface.

To model waves in an anisotropic medium mathematically, let (N, g) be an n-dimensional Riemannian manifold with boundary and consider the wave equation

$$\begin{aligned} &\partial_t^2 u(t,x) - \Delta_g u(t,x) = 0 \quad \text{in } (0,\infty) \times N, \\ &\partial_\nu u|_{\mathbb{R}_+ \times \partial N} = f, \quad u|_{t=0} = 0, \quad \partial_t u|_{t=0} = 0, \end{aligned} \tag{2.1}$$

where Δ_g is the Laplace–Beltrami operator corresponding to a smooth time-independent Riemannian metric g on N. We denote by SN the unit sphere bundle,

$SN = \{(y, \eta) \in TN \mid \|\eta\|_g = 1\}$, and the unit speed geodesic of (N, g) emanating from a point $(y, \eta) \in SN$ by $\alpha_{y,\eta}(t) = \exp_p(t\eta)$.

The solution of (2.1), corresponding to the boundary value f (which is interpreted as a boundary source), is denoted by $u^f = u^f(t, x)$.

Let us assume that the boundary ∂N is known. The inverse problem is to reconstruct the manifold N and the metric g when we are given the boundary ∂N, as a manifold, and the *response operator*

$$\Lambda_{N,g} : f \mapsto u^f|_{\mathbb{R}_+ \times \partial N},$$

which is also called the *Neumann-to-Dirichlet map*. Physically, $\Lambda_{N,g} f$ describes the measurement of the medium response to any applied boundary source f.

It is convenient to interpret the anisotropic wave speed $(g^{jk}(x))$ also in a domain of the Euclidean space as the inverse of a Riemannian metric, thus modeling the medium as a *Riemannian manifold*. This is due to the fact that if $\Psi : N \to N$ is a diffeomorphism such that $\Psi|_{\partial N} = \mathrm{Id}$, then all boundary measurements for the metric g and the pullback metric $\Psi^* g$ coincide. Thus, to prove uniqueness results for inverse problems, one has to consider properties that are invariant in diffeomorphisms and try to reconstruct those uniquely, for example, to show that an underlying manifold structure can be uniquely determined. In practice, the inverse problem in a subset of the Euclidean space is solved in two steps. The first is to reconstruct the underlying manifold structure. The second step is to find an embedding of the constructed manifold in the Euclidean space using additional a priori information; see [32]. Below, we will concentrate on the first step.

In the 1990s, the combination of the boundary control method developed by Belishev [6] and generalized for Riemannian manifolds in [7] and Tataru's unique continuation theorem [64] gave a solution to the inverse problem of determining the isometry type of a Riemannian manifold (N, g) with given boundary ∂N and the Dirichlet-to-Neumann map $\Lambda_{N,g}$.

Theorem 2.1 ([7,64]). *Let (N_1, g_1) and (N_2, g_2) be compact Riemannian manifolds with boundary. Assume that there is a diffeomorphism $\Phi : \partial N_1 \to \partial N_2$ such that*

$$\Phi^*(\Lambda_{N_1,g_1} f) = \Lambda_{N_2,g_2}(\Phi^* f) \quad \text{for all } f \in C_0^\infty(\mathbb{R}_+ \times \partial N_1).$$

Then (N_1, g_1) and (N_2, g_2) are isometric Riemannian manifolds.

Above, $\Phi^* f(t, x) = f(t, \Phi(x))$ is the pullback of f in Φ or, more precisely, that of $\mathrm{Id}_{\mathbb{R}} \times \Phi : \mathbb{R} \times \partial N_1 \to \mathbb{R} \times \partial N_2$.

The original method to prove Theorem 2.1 was based on combining the spectral theory for the Laplace operator and control theory for the wave equation. An

alternative way to prove Theorem 2.1 is to show that the hyperbolic Dirichlet-to-Neumann map $\Lambda_{N,g}$ can be used to determine a sequence of Dirichlet boundary values f_j for which the corresponding waves $u^{f_j}(x,T)$ focus at time $T > 2\,\mathrm{diam}(N)$; see [10, 19, 36].

Theorem 2.2 ([36, Focusing of waves]). *Let $T > 2\,\mathrm{diam}(N)$, $z \in \partial N$, $s > 0$, and $y \in N$ be a point on the normal geodesic emanating from z, $y = \alpha_{z,\nu}(s)$ such that $\mathrm{dist}(y, \partial N) = s$. Using the Neumann-to-Dirichlet map $\Lambda_{N,g}$, one can construct a sequence of boundary sources $f_j \in C_0^\infty((0,T) \times \partial N)$, $j = 1, 2, \ldots$, such that for some $C_y \neq 0$ at the time T*

$$\lim_{j\to\infty} \begin{pmatrix} u^{f_j}(T,x) \\ \partial_t u^{f_j}(T,x) \end{pmatrix} = \begin{pmatrix} 0 \\ C_y \delta_y(x) \end{pmatrix} \quad \text{for } x \in N,$$

where the limit takes place in the space of generalized functions, that is, in $\mathcal{D}'(N) \times \mathcal{D}'(N)$.

Let $G(t,x;y)$ be Green's function of the wave operator, that is, the wave produced by a point source at the point y at the time zero:

$$\begin{aligned} (\partial_t^2 - \Delta_g) G(t,x;y) &= \delta_{(0,y)}(t,x) \quad \text{on } (t,x) \in \mathbb{R}_+ \times N, \\ G|_{(t,x)\in\mathbb{R}_+\times\partial N} &= 0, \quad G|_{t=0} = 0, \quad \partial_t G|_{t=0} = 0, \end{aligned}$$

where Δ_g operates in the x coordinate. Theorem 2.2 implies that the waves $u^{f_j}(t,x)$ converge in the sense of distributions to $G(t-T,x;y)$ in the domain $x \in N$, $t > T$. This means that using the Dirichlet-to-Neumann map, it is possible to construct functions $(\Lambda_{N,g} f_j)(x, t-T)$ that converge, as $j \to \infty$, to the boundary values $\partial_\nu G(\cdot,\cdot;y)|_{\mathbb{R}_+\times\partial N}$ of the waves $G(\cdot,\cdot;y)$ produced by point sources $C_y\delta_y(x)$ at the time $t = 0$. The first time $t > 0$ when the wave $G(\cdot,\cdot;y)$ is observed to be non-vanishing at the boundary point $x \in \partial N$ is defined to be

$$\begin{aligned} \mathbf{t}_y(x) = \sup\{t \in \mathbb{R} \mid &\text{the interval } \{z\} \times (0,t) \text{ has a neighborhood} \\ &U \subset \mathbb{R} \times \partial N \text{ such that } \partial_\nu G(\cdot,\cdot,y)|_U = 0\}. \end{aligned}$$

Using the finite propagation speed of waves and Tataru's unique continuation theorem [64], it turns out that $\mathbf{t}_y(x) = d_g(x,y)$; see [54]. Thus, Theorem 2.2 implies that the Dirichlet-to-Neumann map $\Lambda_{N,g}$ for the wave equation determines the functions

$$\mathbf{t}_{\alpha_{z,\nu}(s)}(x) = d_g(x, \alpha_{z,\nu}(s)) \quad \text{for } x, z \in \partial N \text{ and } s > 0 \text{ such that } \mathrm{dist}\,(x,z) = s. \tag{2.2}$$

Later, we will return to these data, but before that, we consider the consequences of Theorem 2.1.

Theorem 2.1 can also be used to prove the uniqueness of several other inverse problems. The equivalence of inverse spectral problems and several inverse boundary value problems for wave and heat equations is considered in [33]. In particular, the above results and [33] imply the following theorem.

Theorem 2.3 ([33]). *Suppose* (N, g) *is a compact smooth Riemannian manifold with boundary* ∂N. *Then the boundary* ∂N *and Neumann-to-Dirichlet map* $\Lambda : \partial_\nu u|_{\mathbb{R}_+\times\partial N} \mapsto u|_{\mathbb{R}_+\times\partial N}$, *for the heat equation* $(\partial_t - \Delta_g)u = 0$ *or for the Schrödinger equation* $(i\partial_t - \Delta_g)u = 0$, *with vanishing initial data* $u|_{t=0} = 0$ *determine the Neumann-to-Dirichlet map for the wave equation, and therefore the manifold* (N, g) *up to an isometry.*

The stability of the solutions of the above inverse problems has been analyzed in [4, 5, 12, 14, 42, 52, 60].

Without making strong assumptions about the geometry of the manifold, the existing uniqueness results for the inverse problems for the linear hyperbolic equations with vanishing initial data are limited to equations whose coefficients are time independent or real-analytic in time (see, e.g., [2, 4, 7, 20, 25, 32, 37, 38, 44, 53]). The reason for this is that these results are based on Tataru's unique continuation theorem [64]. This sharp unique continuation result does not work for general wave equations whose coefficients are not real-analytic in time, as shown by Alinhac [3]. Solving the inverse problem of finding the Riemannian metric (up to an isometry), that is, proving Theorem 2.1, can be reduced to determining the metric from passive observations from sources inside the domain, and due to this, we next consider passive observations from the point sources.

2.1. Inverse problems with passive measurements

2.1.1. Lorentzian light observation sets and waves on a Riemannian manifold. Let us consider a body in which there spontaneously appear point sources that create propagating waves. In this subsection, we assume that the velocity of the waves is independent of time and depends only on the spatial variables and model the wave speed using a Riemannian metric. We consider the inverse problem where we detect such waves either outside or at the boundary of the body and aim to determine the unknown wave speed inside the body. As an example of such a situation, one can consider the micro-earthquakes that appear very frequently near active faults. The related inverse problem is whether the surface observations of elastic waves produced by the micro-earthquakes can be used in the geophysical imaging of Earth's subsurface, that is, to determine the speed of the elastic waves in the studied volume. Below, we consider a highly idealized version of the above inverse problem. We consider the problem on an n-dimensional manifold N with a Riemannian metric $h = h_{ij}(x)$ that

corresponds to the travel time of a wave between two points (below, we will consider Lorentzian metric tensors for which we reserve the symbol g and use now the letter h for the Riemannian metric). The Riemannian distance of points $x, y \in N$ is denoted by $d_h(x, y)$. For simplicity, we assume that the manifold N is compact and has no boundary. To do this, we assume that the manifold contains a known part F and the metric is unknown outside this set, that is, in $N \setminus F$. When a spontaneous point source produces a wave at some unknown point $x \in N \setminus F$ at some unknown time $t \in \mathbb{R}$, the produced wave is observed at the point $z \in F$ at time $T_{x,t}(z) = d_h(z, x) + t$. Note that these functions are analogous to functions (2.2) (that were defined for the manifolds with boundary) with the source points $x = \alpha_{z,v}(s)$. These observation times at two points $z_1, z_2 \in F$ determine the *distance difference function*

$$D_x(z_1, z_2) = T_{x,t}(z_1) - T_{x,t}(z_2) = d_h(z_1, x) - d_h(z_2, x).$$

Physically, this function corresponds to the difference of times when the waves produced by the point source at (x, t) are observed at z_1 and z_2. Below, we assume that there are a large number of point sources and that we are given the set F and the family of functions $\{D_x \mid x \in X\} \subset C(F \times F)$, where $X \subset N$ is either the whole set $N \setminus F$ or a dense subset of X.

Next, let (N_1, h_1) and (N_2, h_2) be compact and connected Riemannian manifolds without boundary. Let $d_{h_j}(x, y)$ denote the Riemannian distance of points $x, y \in N_j$, $j = 1, 2$. For $x \in N_j$ and $z_1, z_2 \in F_j$, we denote $D_x^j(z_1, z_2) = d_{h_j}(x, z_1) - d_{h_j}(x, z_2)$.

To prove the uniqueness of the passive imaging inverse problem, we assume the following:

(i) The map $\phi : F_1 \to F_2$ is an isometry, that is, $d_{h_1}(z, w) = d_{h_2}(\phi(z), \phi(w))$ for all $z, w \in F_1$.

(ii) The collections $\mathcal{D}_j(N_j) = \{D_x^j(\cdot, \cdot) \mid x \in N_j\} \subset C(F_j \times F_j)$ with $j = 1, 2$ are equivalent, that is, $\{D_x^1(\cdot, \cdot) \mid x \in N_1\} = \{D_y^2(\phi(\cdot), \phi(\cdot)) \mid y \in N_2\}$.

Theorem 2.4 ([54, Distance difference functions determine the manifold]). *Let (N_1, h_1) and (N_2, h_2) be compact and connected Riemannian manifolds of dimension $n \geq 2$ without boundary. Let $F_j \subset N_j$, $j = 1, 2$, be closures of non-empty open sets having a smooth manifold. Suppose the above conditions* (i) *and* (ii) *are valid. Then the manifolds (N_1, h_1) and (N_2, h_2) are isometric.*

We will not consider here the proof of Theorem 2.4 which is based on Riemannian geometry. Instead, we discuss next how the distance difference functions on a Riemannian manifold (N, h) determine the earliest light observation sets on the Lorentzian manifold $M = \mathbb{R} \times N$ with the product Lorentzian metric $g = -dt^2 + h$ and consider the determination of the Lorentzian metric using the knowledge of these

sets. Let (N, h) be a Riemannian manifold. We denote the unit speed geodesic of (N, h) emanating from a point $(y, \eta) \in SN$ by $\alpha^h_{y,\eta}(t) = \exp_p(t\eta)$. Here, $SN = \{(y, \eta) \in TN \mid \|\eta\|_h = 1\}$ is the unit sphere bundle of N. Using the Riemannian manifold (N, h), we define a product manifold $M = \mathbb{R} \times N$. When $U \subset N$ is an open set having coordinates $X : U \to \mathbb{R}^n$, $X(x) = (x^1(x), x^2(x), \ldots, x^n(x))$, we define a Lorentzian metric $g = -dt^2 + h(x)$ that is given in local coordinates $(t, x^1, \ldots, x^n) \in \mathbb{R} \times U$ by

$$g = -dt^2 + \sum_{j,k=1}^{n} h_{jk}(x^1, \ldots, x^n) dx^j dx^k.$$

Below, we denote the time coordinate also by $x^0 = t$. Let $p = (t_0, y) \in \mathbb{R} \times N$ and $\xi = \frac{\partial}{\partial t} + \eta$ be a vector at the point p, where $(y, \eta) \in SN$. Observe that $g(\xi, \xi) = -1 + \|\eta\|_h^2 = 0$, that is, ξ is a future-pointing light-like vector. The curve $\gamma_{p,\xi} : \mathbb{R} \to \mathbb{R} \times N$,

$$\gamma_{p,\xi}(s) = (t_0 + s, \alpha^h_{y,\eta}(s)),$$

is a geodesic of the Lorentzian manifold (M, g). Also, let $U \subset \mathbb{R} \times F$ be an open set where we consider observations. For the point $p \in M$, we define the light observation set with the observation set $U \subset M$ by setting

$$\mathcal{P}_U(p) = \left\{ \gamma_{p,\xi}(s) \in M \mid s \geq 0,\ \xi = \frac{\partial}{\partial t} + \eta,\ \|\eta\|_h = 1 \right\} \cap U.$$

Let $\kappa : SN \to \mathbb{R} \cup \{\infty\}$ be the cut locus function on the manifold (N, h), defined by

$$\kappa(y, \eta) = \sup\{s \geq 0 \mid \operatorname{dist}_h(\alpha_{y,\eta}(s), y) = s\}.$$

Then the earliest light observation set of the point $p = (t_0, y)$ with the observation set $U \subset M$ is

$$\mathcal{E}_U(p) = \left\{ \gamma_{p,\xi}(s) \in M \mid 0 \leq s \leq \kappa(y, \eta),\ \xi = \frac{\partial}{\partial t} + \eta,\ \|\eta\|_h = 1 \right\} \cap U.$$

Next, we consider the earliest light observation sets on general Lorentzian manifolds, which may not be of the product type.

2.2. Inverse problems on Lorentzian manifolds

2.2.1. Notations in a Lorentzian spacetime. Let (M, g) be a $(1 + n)$-dimensional time-oriented Lorentzian manifold of signature $(-, +, +, \ldots, +)$, where $n \geq 2$. On M, we use local coordinates $(x^0, x^1, \ldots, x^n)$, where $x^0 = t$ is the time coordinate.

Let $L_pM = \{\xi \in T_pM \setminus \{0\} \mid g(\xi, \xi) = 0\}$ be the set of light-like vectors in the tangent space T_pM.

Let $\exp_p : T_pM \to M$ be the exponential map on (M, g). The geodesic starting at p in the direction $\xi \in T_pM \setminus \{0\}$ is the curve $\gamma_{p,\xi}(t) = \exp_p(\xi t)$, $t \geq 0$.

Let $\mu : [-1, 1] \to M$ be a smooth future-pointing time-like geodesic and $U \subset M$ be an open neighborhood of $\mu([-1, 1])$.

Let $q \in M$. The set of future-pointing *light-like* vectors at q is denoted by

$$L_q^+M = \{\theta \in T_qM \setminus 0 \mid g(\theta, \theta) = 0,\ \theta \text{ is future pointing}\}.$$

A vector $\theta \in T_qM$ is *time-like* if $g(\theta, \theta) < 0$ and *space-like* if $g(\theta, \theta) > 0$. *Causal vectors* are the collection of time-like and light-like vectors, and a curve γ is time-like (light-like, causal, future pointing) if the tangent vectors $\dot\gamma$ are time-like (light-like, causal, future pointing).

For $p, q \in M$, the notation $p \ll q$ means that p, q can be joined by a future-pointing time-like curve. The *chronological future* and *past* of $p \in M$ are

$$I^+(p) = \{q \in M \mid p \ll q\}, \quad I^-(p) = \{q \in M \mid q \ll p\}.$$

Similarly, for $p, q \in M$, the notation $p < q$ means that $p \neq q$ and p, q can be joined by a future-pointing causal (i.e., time-like or light-like) curve. We denote $p \leq q$ if $p < q$ or $p = q$. The *causal future* and *past* of $p \in M$ are

$$J^+(p) = \{q \in M \mid p \leq q\}, \quad J^-(p) = \{q \in M \mid q \leq p\}.$$

We will denote throughout the paper the causal diamond sets as

$$I(p, q) = I^+(p) \cap I^-(q), \quad J(p, q) = J^+(p) \cap J^-(q).$$

To emphasize the Lorentzian structure of (M, g), we sometimes write $I^\pm_{M,g}(p) = I^\pm(p)$ and $J^\pm_{M,g}(p) = J^\pm(p)$.

A time-oriented Lorentzian manifold (M, g) is *globally hyperbolic* if there are no closed causal paths in M, and for any $p, q \in M$, the set $J(p, q)$ is compact [9]. According to [8], a globally hyperbolic manifold is isometric to the product manifold $\mathbb{R} \times N$ with the Lorentzian metric given by

$$g = -\beta(t, y)dt^2 + \kappa(t, y), \tag{2.3}$$

where $\beta : \mathbb{R} \times N \to \mathbb{R}_+$ and κ is a Riemannian metric on N depending on t.

2.3. Passive measurements on a Lorentzian manifold

2.3.1. The set of the earliest light observations. Let $s_-, s_+ \in (-1, 1)$, $s_- < s_+$, and $p^\pm = \mu(s_\pm)$. Next, define the light observation sets and consider in particular the case when $W \subset I^-(p^+) \setminus J^-(p^-)$ is a relatively compact open set; see Figure 1 (left).

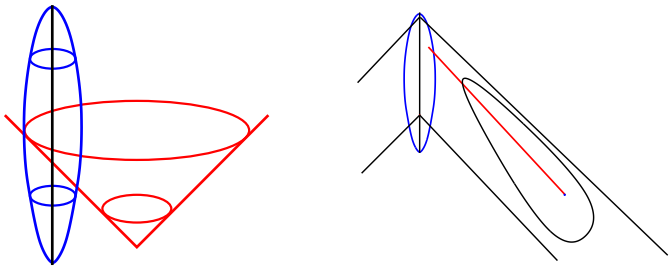


Figure 1. Left: When there are no cut points, the earliest light observation set $\mathcal{E}_U(q)$ is the intersection of the cone and the open set U. The cone is the union of future-pointing light-like geodesics from q, and the ellipsoid depicts U. Right: The unknown set $W \subset I^-(p_+) \setminus J^-(p_-)$ contains source points. Light cones emanating from these points (the red segment) are observed in the set U.

Definition 2.5. Let (M, g) be a globally hyperbolic Lorentzian manifold.

(i) For $q \in M$, let

$$\mathcal{L}^+(q) = \exp_q(L_q^+ M) \cup \{q\} = \{\gamma_{q,\xi}(s) \in M \mid \xi \in L_q^+ M,\, s \geq 0\} \subset M$$

be the future-directed light cone emanating from the point q.
The light observation set of q in the observation set U is $\mathcal{P}_U(q) = \mathcal{L}^+(q) \cap U \in 2^U$.

(ii) The earliest light observation set of $q \in M$ in U is

$$\begin{aligned}\mathcal{E}_U(q) = \{x \in \mathcal{P}_U(q) \mid &\text{there exist no point } y \in \mathcal{P}_U(q)\\ &\text{and a future-pointing time-like path } \alpha : [0,1] \to U\\ &\text{such that } \alpha(0) = y \text{ and } \alpha(1) = x\} \subset U.\end{aligned}$$

(iii) Let $W \subset M$ be open. The collection of the earliest light observation sets with source points at W is

$$\mathcal{E}_U(W) = \{\mathcal{E}_U(q) \mid q \in W\} \subset 2^U.$$

Note $\mathcal{E}_U(W)$ is defined as an unindexed set, that is, for an element $\mathcal{E}_U(q) \in \mathcal{E}_U(W)$, we do not know what the corresponding point q is.

Above, $2^U = \{U' \mid U' \subset U\}$ is the power set of U. Below, when we say that the set U is given as a differentiable manifold, we mean that we are given the set U and the local coordinate charts on it for which the corresponding transition maps are C^∞-smooth.

The set $\mathcal{P}_U(q)$ can be viewed as a model of a measurement where light emitted by a point source at q is recorded in U. As gravitational wave packets propagate at the

speed of light, $\mathcal{P}_U(q)$ could also correspond to an observation where a gravitational wave is generated at q and detected in U.

The following theorem says, roughly speaking, that observations of a large number of point sources in a region W determine the structure of the spacetime in W, up to a conformal factor.

Theorem 2.6 ([43]). *Let (M_j, g_j), where $j = 1, 2$, be two open globally hyperbolic Lorentzian manifolds of dimension $1 + n$, $n \geq 2$. Let $\mu_j : [-1, 1] \to M_j$ be a future-pointing time-like path, let $U_j \subset M_j$ be a neighborhood of $\mu_j([-1, 1])$, and let $W_j \subset M_j$ be such that $W_j \subset I^-_{M_j,g_j}(\mu_j(1)) \setminus J^-_{M_j,g_j}(\mu_j(-1))$ are open and relatively compact sets. Assume that there is a conformal diffeomorphism $\phi : U_1 \to U_2$ such that $\phi(\mu_1(s)) = \mu_2(s)$, $s \in [0, 1]$, and*

$$\{\phi(\mathcal{E}_{U_1}(q)) \mid q \in W_1\} = \{\mathcal{E}_{U_2}(q) \mid q \in W_2\}.$$

Then there is a diffeomorphism $\Psi : W_1 \to W_2$ and a strictly positive function $\alpha \in C^\infty(W_1)$ such that $\Psi^* g_2 = \alpha g_1$ and $\Psi|_{U_1 \cap W_1} = \phi$.

Next, we consider an alternative formulation of the above inverse problem where the data are given in terms of functions, not as a collection of sets.

Assume that the open set $U \subset M$, where observations are made, is of the form $U = \bigcup_{a \in A} \mu_a(-1, 1)$, where $\mu_a : (-1, 1) \to M$, $a \in \overline{A} \subset \mathbb{R}^k$ are time-like curves that depend continuously on a.

Let $p_1, p_2 \in \mu_{a_0}$, $a_0 \in A$. Also, let $W \subset J^-(p_2) \setminus J^-(p_1)$ be an open, relatively compact set. The observation time function $F_q : \overline{A} \to \mathbb{R}$ for a point $q \in W$ is

$$F_q(a) = \min\{s \in \mathbb{R} \mid \text{there exists a future-directed light-like geodesic from } q \text{ to } \mu_a(s)\}.$$

That is, $F_q(a)$ is the first time when we observe on μ_a light that comes from q.

By Definition 2.5, the set $\mathcal{E}_U(q)$ determines the function F_q in A via the formula

$$F_q(a) = s, \quad \text{where } s \in [-1, 1] \text{ is such that } \mu_a(s) \in \mathcal{E}_U(q),\ a \in A.$$

As $F_q : \overline{A} \to \mathbb{R}$ is continuous, this determines $F_q(a)$ for all $a \in \overline{A}$. On the other hand, F_q determines $\mathcal{E}_U(q)$ via the formula

$$\mathcal{E}_U(q) = \{\mu_a(F_q(a)) \mid a \in A\}.$$

Due to this, Theorem 2.6 has the following equivalent formulation.

Theorem 2.7 ([43]). *Let (M, g) be a globally hyperbolic Lorentzian manifold of dimension $1 + n \geq 3$. Assume that $\mu_a(-1, 1) \subset M$, $a \in A \subset \mathbb{R}^m$ are time-like paths,*

$U = \bigcup_{a\in A} \mu_a$ is open, and $p_1, p_2 \in \mu_{a_0}$. Let $W \subset J^-(p_2) \setminus J^-(p_1)$ be a relatively compact open set and $W' \subset W$ be dense. Then $(U, g|_U)$ and the collection of the observation time functions,

$$\{F_q : A \to \mathbb{R} \mid q \in W'\} \subset C(A),$$

determine the topological and differentiable type of the manifold W and the conformal class of the metric g in W.

We will outline the ideas of the proof of Theorem 2.7 below in Section 3.

2.4. Active measurements: Inverse problems for non-linear wave equations

Let (M, g) be a $(1 + n)$-dimensional globally hyperbolic Lorentzian manifold and assume, without loss of generality, that $M = \mathbb{R} \times N$ with a metric of form (2.3). Let $t_0 > 0$ and consider the semilinear wave equations

$$\Box_g u(x) + a(x)u(x)^m = f(x) \quad \text{for } x \in (-\infty, t_0) \times N, \tag{2.4}$$

$$u = 0 \quad \text{in } (-\infty, 0) \times N, \tag{2.5}$$

where $m \geq 2$ is an integer. Here, $a \in C^\infty(M)$ is a nowhere vanishing function and $\Box_g$ is the wave operator, or the Lorentzian Laplace operator:

$$\Box_g u = \sum_{j,k=0}^{n} |\det(g)|^{-1/2} \frac{\partial}{\partial x^j} \left(|\det(g)|^{1/2} g^{jk} \frac{\partial}{\partial x^k} u \right).$$

Let $\mu : [-1, 1] \to M$ be a time-like curve and U be its open, relatively compact neighborhood. The solution of (2.4)–(2.5) exists when the source f is supported in U and satisfies $\|f\|_{C^k(\overline{U})} < \varepsilon$, where $k \in \mathbb{Z}_+$ is sufficiently large and $\varepsilon > 0$ is sufficiently small. For such sources f, we define the measurement operator

$$L_U : f \mapsto u|_U.$$

Theorem 2.8 ([22, 43]). *Let (M, g) be a globally hyperbolic $(n + 1)$-dimensional Lorentzian manifold, $n \geq 2$. Let μ be a time-like path containing p^+ and p^-. Let $U \subset M$ be a neighborhood of μ, and let $a : M \to \mathbb{R}$ be a nowhere vanishing C^∞-smooth function. Then $(U, g|_U)$ and the measurement operator L_U determine the topology, differentiable structure, and the conformal class of the metric g in the double cone $I_{M,g}(p^-, p^+)$.*

We will outline the ideas of the proof of Theorem 2.8 below in Section 3.

The proof of Theorem 2.8 uses the results on the inverse problem for passive measurements for point sources, described below, and the non-linear interaction

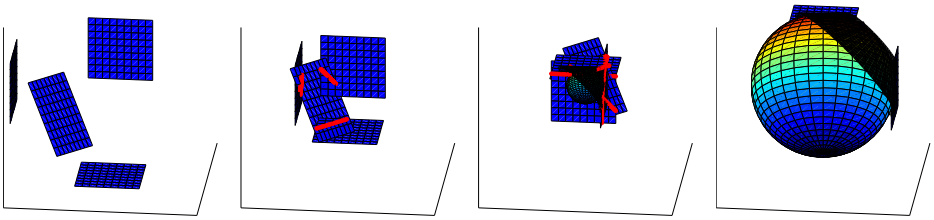


Figure 2. Four plane waves propagate in space. When the planes intersect, the non-linearity of the hyperbolic system produces new waves. Left: Plane waves before interacting. Middle left: The two-wave interactions (red line segments) appear but do not cause singularities propagating to new directions. Middle right and Right: All plane waves have intersected, and new waves have appeared. The three-wave interactions cause conic waves (the black surface). Only one such wave is shown in the figure. The interaction of four waves causes a microlocal point source that sends a spherical wave in all future light-like directions.

of waves having conormal singularities (see Figure 2). There are many results on such non-linear interaction, starting with the studies of Bony [11] and Melrose and Ritter [56].

2.4.1. Stability results and numerical reconstruction. As discussed below, any globally hyperbolic Lorentzian manifold M is isometric to a product manifold $\mathbb{R}\times N$ equipped with the product metric (2.3). Let $\Omega\subset N$ be a smooth submanifold of dimension n with smooth boundary and denote the lateral boundary of $[0,T]\times\Omega\subset M$ by $\Sigma:=[0,T]\times\partial\Omega$. We denote the Sobolev space of functions having s weak derivatives on Σ by $H^s(\Sigma)$, and the closure of $C_0^\infty(\Sigma)$ in $H^s(\Sigma)$ is denoted by $H_0^s(\Sigma)$.

We consider the non-linear wave equation

$$\begin{cases}\Box_g u(t,x)+q(t,x)u(t,x)^m=0 & \text{in } [0,T]\times\Omega,\\ u=f & \text{on } [0,T]\times\partial\Omega,\\ u(0,x)=\partial_t u(0,x)=0 & \text{on } \Omega,\end{cases}\tag{2.6}$$

where we assume that the exponent m is an integer greater than or equal to 4.

Let $\mathcal{W}$ be a compact set belonging to both the chronological future $I^+(\Sigma)$ and past of $I^-(\Sigma)$ of the lateral boundary $\Sigma=[0,T]\times\partial\Omega$:

$$\mathcal{W}\subset I^-(\Sigma)\cap I^+(\Sigma)\cap([0,T]\times\Omega).$$

Theorem 2.9 ([50, Stability estimate]). *Suppose (M,g), $M=\mathbb{R}\times N$, is a $(1+n)$-dimensional globally hyperbolic Lorentzian manifold.*

Let $T>0$, and let $\Omega\subset N$ be an open set with smooth non-empty boundary. Let $m\geq 4$ be an integer, $s\in\mathbb{N}$ with $s+1>\frac{(n+1)}{2}$, and $r\in\mathbb{R}$ with $r\leq s$. Let $j=1,2$.

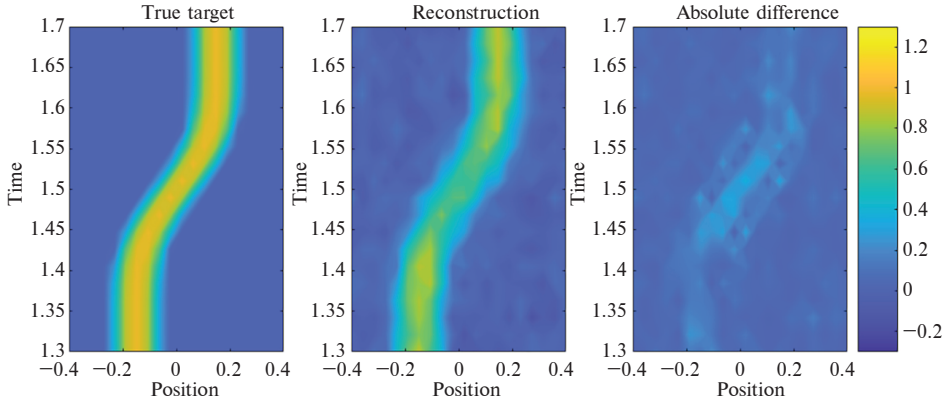


Figure 3. Reconstruction of the coefficient $q(x,t)$ in equation $\partial_t^2 u(x,t) - \partial_x^2 u(x,t) + q(x,t)u(x,t)^2 = 0$, $x \in [0,1]$, $t \in [0,T]$, from the data $f \to \langle \psi_0, \Lambda_q f \rangle_{L^2(\partial\Omega\times[0,T])}$, where Λ_q is the Dirichlet-to-Neumann map, and ψ_0 is a suitably chosen 'device function'. Above, the data have additive Gaussian noise, and the signal-to-noise ratio is 12 dB. Left: Ground truth of the potential $q(x,t)$. Center: Numerical reconstruction. Right: Error in reconstruction. For numerical results, see [49], and for mathematical analysis, see [48,50].

Assume that $q_j \in C^{s+1}(\mathbb{R}\times\Omega)$ *satisfy* $\|q_j\|_{C^{s+1}} \le c_0$, $j = 1,2$, *for some* $c_0 > 0$. *Let* $\Lambda_j : H_0^{s+1}(\Sigma) \to H^r(\Sigma)$ *be the corresponding Dirichlet-to-Neumann maps of the non-linear wave equation* (2.6).

Let $\varepsilon_0 > 0$, $L > 0$, *and* $\delta \in (0,L)$ *be such that*

$$\|\Lambda_1(f) - \Lambda_2(f)\|_{H^r(\Sigma)} \le \delta$$

for all $f \in H_0^{s+1}(\Sigma)$ *with* $\|f\|_{H^{s+1}(\Sigma)} \le \varepsilon_0$. *Then there exists a constant* $C > 0$, *independent of* q_1, q_2 *and* $\delta > 0$, *such that*

$$\|q_1 - q_2\|_{L^\infty(\mathcal{W})} \le C\delta^{\sigma(s,m)},$$

where $\sigma(s,m) = \frac{8(m-1)}{2m(m-1)(8s-n+13)+2m-1}$.

Numerical reconstructions based on Theorem 2.9 in the $(1+1)$-dimensional case are shown in Figure 3.

3. Ideas for proofs and reconstruction methods

3.1. Light observation sets and the inverse problem for passive observations

In this subsection, we outline the proof of Theorem 2.7. Below, let $\overline{W} = \mathrm{cl}(W)$ be the closure of W in M such that $\overline{W} \subset J^-(p^+) \setminus I^-(p^-)$. The set $\overline{A} \times \overline{W}$ is compact, and the map $\mathcal{F} : \overline{W} \to C(\overline{A})$, $\mathcal{F}(q) = F_q$, is continuous.

Next, we will show that the map $\mathcal{F} : \overline{W} \to \mathcal{F}(\overline{W})$, $\mathcal{F}(q) = F_q$, is injective. To prove this, we assume the opposite: Assume that there are $q_1, q_2 \in W$ that satisfy $F_{q_1} = F_{q_2}$. Then

$$\mathcal{E}_U(q_1) = \mathcal{E}_U(q_2).$$

This implies that all light-like geodesics from q_1 to U go through q_2, and hence, q_1 is in the past of q_2, or vice versa. Without loss of generality, we can assume that q_1 is in the past of q_2. Let $a_1 \in A$ and $p_1 = \mu_{a_1}(F_{q_1}^+(a_1)) \in \mu_{a_1}([-1,1])$. Let now $\gamma_{q_1,\xi_1}([0,\ell_1])$ be a light-like geodesic from q_1 to p_1, that is, $p_1 = \gamma_{q_1,\xi_1}(\ell_1) \in \mu_{a_1}([-1,1])$. Next, let $a_2 \in A$ be such that $\gamma_{q_1,\xi_1} \cap \mu_{a_2}([-1,1]) = \emptyset$. Let $p_2 = \mu_{a_2}(f_{q_1}^+(a_2)) \in \mu_{a_2}([-1,1])$ and $\gamma_{q_1,\xi_2}([0,\ell_2])$ be a light-like geodesic from q_1 to p_2, that is, $p_2 = \gamma_{q_1,\xi_2}(\ell_2) \in \mu_{a_2}([-1,1])$. Then, as $\gamma_{q_1,\xi_1} \cap \mu_{a_2}([-1,1]) = \emptyset$, ξ_2 and ξ_1 are not parallel. As all light-like geodesics from q_1 to U go through q_2, we see that there are $t_1 \in (0,\ell_1)$ and $t_2 \in (0,\ell_2)$ such that

$$q_2 = \gamma_{q_1,\xi_1}(t_1), \quad q_2 = \gamma_{q_1,\xi_2}(t_2).$$

As the vectors ξ_2 and ξ_1 are not parallel, the geodesics γ_{q_1,ξ_1} and γ_{q_2,ξ_2} are not the same paths, and we see that the path $\beta = \gamma_{q_1,\xi_1}([0,t_1]) \cup \gamma_{q_1,\xi_2}([t_2,\ell_2])$ is a union of two causal paths from the point q_1 to q_2 and then from q_2 to p_2. Since the geodesics γ_{q_1,ξ_1} and γ_{q_2,ξ_2} are not the same paths with different parametrizations, the path β is not a geodesic. As the manifold (M,g) is globally hyperbolic, this implies that there is a time-like path from q_1 to p_2, and hence, it is not possible that $p_2 = \mu_{a_2}(F_{q_1}^+(a_2)) \in \mu_{a_2}([-1,1])$, as we assumed above; see Figure 4. This contradiction implies $F_{q_1} \neq F_{q_2}$. Thus, the map $\mathcal{F} : \overline{W} \to \mathcal{F}(\overline{W})$, $\mathcal{F}(q) = F_q$, is injective.

The map $\mathcal{F} : \overline{W} \mapsto C(\overline{A})$ is continuous and injective. As $\overline{W}$ is compact, these imply that the map $\mathcal{F} : \overline{W} \mapsto C(\overline{A})$ is a homeomorphism. Thus, $\mathcal{F}(\overline{W}) \subset C(\overline{A})$ and $\overline{W}$ are homeomorphic. Furthermore, as W is an open subset of the manifold M, we can see that $\mathcal{F}(W) \subset C(\overline{A})$ and W are homeomorphic, so we can consider $\mathcal{F}(W)$

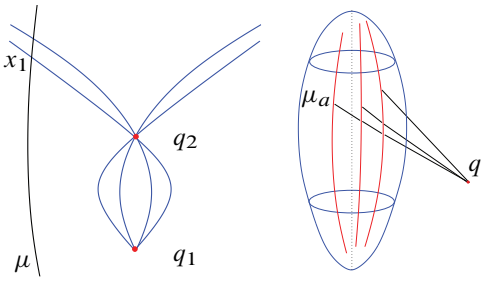


Figure 4. Left: Picture on proving that $q_1 = q_2$ by contradiction. Right: The function $X_{\vec{a}} : q \to (F_q(a_j))_{j=1}^{n+1} = (f_{a_j}^+(q))_{j=1}^{n+1} \in \mathbb{R}^{n+1}$ defines coordinates near a point q on M.

as a topological manifold which is a homeomorphic copy of the original manifold W. Thus, we have reconstructed W as a topological manifold.

Next, we consider the reconstruction of W as a differentiable manifold. Below, we will only consider how the coordinate functions on $\mathcal{F}(W)$ are constructed and omit the proof that this construction gives a differentiable structure on $\mathcal{F}(W)$ which makes it diffeomorphic to W. For details on this, we cite [43].

For $a \in A$, let us define the evaluation function $E_a[F] = F(a)$. For $q_0 \in W$, let us choose $n+1$ parameters $a_j \in A$, $j = 1, 2, \ldots, n+1$. When the parameters $(a_1, a_2, \ldots, a_{n+1})$ are chosen in a generic way, that is, this $(n+1)$-tuple is in an open and dense subset of A^{n+1}, the function $X_{\vec{a}}(q) = (F_q(a_j))_{j=1}^{n+1} = (f_{a_j}^+(q))_{j=1}^{n+1} \in \mathbb{R}^{n+1}$ defines smooth coordinates on M near q_0 (here, we omit this part of the proof; see [43] on details). This implies that for any $\mathcal{F}(q_0) \in \mathcal{F}(W)$ and generic parameters $a_j \in A$, $j = 1, 2, \ldots, n+1$, there is a neighborhood $\mathcal{V} \subset \mathcal{F}(W)$ such that the function $\mathbf{X}_{\vec{a}} : F \to (E_{a_j}[F])_{j=1}^{n+1} \in \mathbb{R}^{n+1}$ defines coordinates in $\mathcal{W}$ such that the map $\mathcal{F}$ is a local diffeomorphism near the point q_0 from the manifold W to $\mathcal{F}(W)$.

Next, we will consider the construction of the Lorentzian metric on $\mathcal{F}(W)$ which makes it conformal to $(W, g|_W)$. To this end, let us consider a point $f_0 = \mathcal{F}(q_0) \in \mathcal{F}(W)$ and its open neighborhood $\mathcal{W} \subset \mathcal{F}(W)$, where $\vec{a} = (a_j)_{j=1}^{n+1}$ defines coordinates $\mathbf{X}_{\vec{a}} : \mathcal{W} \to \mathbb{R}^{n+1}$ near f_0.

When $B \subset A$ is a sufficiently small neighborhood of the point a_1, the function $a \to f_a^+(q_0)$ is smooth for $a \in B$. Let df_a^+ denote the co-derivative of the map $f_a^+ : W \to \mathbb{R}$. We see that the co-vectors $df_a^+|_{q_0} \in T_{q_0}^* M$, $\quad a \in B$ are light-like co-vectors at q_0. Let $G = \mathcal{F}_* g|_W$ be the metric $g|_W$ pushed forward in the map $\mathcal{F}$ to the manifold $\mathcal{F}(W)$ and $E_a : \mathcal{F}(W) \to \mathbb{R}$ be the evaluation maps, $E_a : f \to f(a)$. Then we see that the co-derivatives

$$dE_a|_{\mathcal{F}(q_0)} \in T^*_{\mathcal{F}(q_0)} \mathcal{F}(W), \quad a \in B$$

are light-like co-vectors with respect to the metric G.

As $a \in B$ is above arbitrary, we see that the set $\{s\, dE_a|_{\mathcal{F}(q_0)} \in T^*_{\mathcal{F}(q_0)} \mathcal{F}(W) \mid a \in B,\ s > 0\}$ contains an open subset of the light cone in the space $(T^*_{\mathcal{F}(q_0)} \mathcal{F}(W), G)$. As the light cone is a quadratic variety, this set determines the whole light cone in $(T^*_{\mathcal{F}(q_0)} \mathcal{F}(W), G)$. These sets determine the conformal class of the metric G on $\mathcal{F}(W)$ at the point $\mathcal{F}(q_0)$.

3.2. Non-linear interaction of waves and inverse problems for active measurements

In this subsection, we outline the proof of Theorem 2.8. To do this, we consider the non-linear equation $\Box u + u^4 = f$ in $\mathbb{R}^{1+3}$, that is, we introduce the techniques by considering the Minkowski space $\mathbb{R}^{1+3}$ and the wave operator $\Box = \partial_{x^0}^2 - \sum_{j=1}^3 \partial_{x^j}^2$.

Let us consider the solutions $u_{\vec{\varepsilon}}(x)$ of the non-linear wave equation

$$\Box u_{\vec{\varepsilon}} + (u_{\vec{\varepsilon}})^4 = 0$$

that depend on parameters $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) \in \mathbb{R}^4$. Let us assume that $u_{\vec{\varepsilon}}|_{\vec{\varepsilon}=0} = 0$. Then the linearized waves

$$u_j(x) = \partial_{\varepsilon_j} u_{\vec{\varepsilon}}|_{\vec{\varepsilon}=0}, \quad j = 1, 2, 3, 4,$$

satisfy $\Box u_j = 0$. Moreover, the fourth-order derivative

$$w = \partial_{\varepsilon_1}\partial_{\varepsilon_2}\partial_{\varepsilon_3}\partial_{\varepsilon_4} u_{\vec{\varepsilon}}(x)|_{\vec{\varepsilon}=0} \tag{3.1}$$

satisfies

$$\Box w = S, \quad S = -24 u_1 u_2 u_3 u_4.$$

We say that S is an artificial source produced by the 4-wave interaction.

We use coordinates $x = (t, y_1, y_2, y_3) \in \mathbb{R}^{1+3}$ and consider, as an example, the linearized waves $u_j(t, y_1, y_2, y_3)$ given by

$$u_1(t, y) = \delta(t - y_1), \quad u_2(t, y) = \delta(t - y_2),$$
$$u_3(t, y) = \delta(t - y_3), \quad u_4(t, y) = \delta\left(t - \frac{(y_1 + y_2)}{\sqrt{2}}\right).$$

Then

$$u_1 u_2 u_3 u_4 = \delta_p(t, y), \quad \text{where } p = (0, 0, 0, 0) \in \mathbb{R}^{1+3},$$

and the 4-wave interaction, $w(x)$, is the solution of the wave equation

$$\Box w = S, \quad S = -24 u_1 u_2 u_3 u_4 = -24\delta_p. \tag{3.2}$$

In physical terms, S corresponds to an (artificial) point source in the spacetime.

We now turn our attention to a general globally hyperbolic, $(1 + 3)$-dimensional Lorentzian manifold when $m = 4$. The proof of Theorem 2.8 is based on the analysis of the interaction of four waves. Analogously to (3.1), the derivative $w = \partial_{\varepsilon_1}\partial_{\varepsilon_2}\partial_{\varepsilon_3}\partial_{\varepsilon_4} u_{\vec{\varepsilon}}|_{\vec{\varepsilon}=0}$ of the solution $u_{\vec{\varepsilon}}$ of (2.4)–(2.5) with the source

$$f_{\vec{\varepsilon}}(x) = \sum_{j=1}^{4} \varepsilon_j f_j(x), \quad \vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4),$$

is given by formula (3.2), where $u_j(x)$ are solutions to the wave equation $\Box_g u = f_j$.

When f_j is a C^∞-smooth function outside the point $p_j \in M$, the linearized wave $u_j = \Box_g^{-1} f_j$ produced by the source f_j is microlocally similar to Green's functions (i.e., the fundamental solution of the wave equation) corresponding to a point source

at the point p_j. Let us denote the light cones in M emanating from the point p_j to future by K_j, and, for simplicity, assume that the light-like geodesics of M have no cut or conjugate points. Then, in $M \setminus \{p_j\}$, the linearized solutions u_j are conormal distributions that are non-smooth on surfaces K_j (see [23, 57]). Then the two waves u_1 and u_2 are singular on hyper-surfaces K_1 and K_2, respectively, and these singularities interact on $K_1 \cap K_2$. This interaction creates only singularities that propagate along K_1 and K_2 on which the original singularities u_1 and u_2 propagate on. The interaction of the three waves u_1, u_2, and u_3 happens on the intersection $K_{123} = K_1 \cap K_2 \cap K_3$ which is a curve in the spacetime. As N^*K_{123} contains light-like directions that are not in the union $N^*K_1 \cup N^*K_2 \cup N^*K_3$, this interaction produces interesting singularities that start to propagate. These singularities correspond to the black conic wave in Figure 3 (we note that these singularities produced by 3-wave interaction can also be used to determine the light observation sets; see [22]). Finally, singularities of all four waves u_j, $j = 1, 2, 3, 4$, interact at the point $\{q\} = \bigcap_{j=1}^4 K_j$. The singularities from the point q propagate along the light cone emanating from this point, and with suitably chosen sources f_j, the wave w is singular at any given point of the light cone $\mathcal{L}(q)$. Thus, S_{1234} can be considered as a microlocal point source that sends singularities in all directions as a point source located at the point q, and these singularities are observed in the set U. In this way, the non-linear interaction of waves determines the intersection of the light cone $\mathcal{L}(q)$ and the observation domain U. Hence, we can determine the earliest light observation sets $\mathcal{E}_U(q)$ for an arbitrary point $q \in I(p^-, p^+)$. By letting q vary through the set $I(p^-, p^+)$, we can apply Theorem 2.6 to recover the topology, differentiable structure, and the conformal class of g in $I(p^-, p^+)$.

Funding. The author was partially supported by the AdG project 101097198 of the European Research Council, the Centre of Excellence of Research Council of Finland, and the FAME flagship of the Research Council of Finland (grant 359186). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the funding agencies or the EU.

References

[1] J. P. Agnelli, A. Çöl, M. Lassas, R. Murthy, M. Santacesaria, and S. Siltanen, Classification of stroke using neural networks in electrical impedance tomography. *Inverse Problems* **36** (2020), no. 11, article no. 115008, 26 pp. Zbl 1451.92178 MR 4173583

[2] S. Alexakis, A. Feizmohammadi, and L. Oksanen, Lorentzian Calderón problem under curvature bounds. *Invent. Math.* **229** (2022), no. 1, 87–138 Zbl 1491.35446 MR 4438353

[3] P. S. Alinhac, Non-unicité du problème de Cauchy. *Ann. of Math. (2)* **117** (1983), no. 1, 77–108 Zbl 0516.35018 MR 0683803

[4] M. Anderson, A. Katsuda, Y. Kurylev, M. Lassas, and M. Taylor, Boundary regularity for the Ricci equation, geometric convergence, and Gel'fand's inverse boundary problem. *Invent. Math.* **158** (2004), no. 2, 261–321 Zbl 1177.35245 MR 2096795

[5] G. Bao and H. Zhang, Sensitivity analysis of an inverse problem for the wave equation with caustics. *J. Amer. Math. Soc.* **27** (2014), no. 4, 953–981 Zbl 1325.35274 MR 3230816

[6] M. I. Belishev, An approach to multidimensional inverse problems for the wave equation (in Russian). *Dokl. Akad. Nauk SSSR* **297** (1987), no. 3, 524–527. English translation. *Dokl. Math.* **36** (1988), no. 3, 481–484 Zbl 0661.35084 MR 0924687

[7] M. I. Belishev and Y. V. Kurylev, To the reconstruction of a Riemannian manifold via its spectral data (BC-method). *Comm. Partial Differential Equations* **17** (1992), no. 5–6, 767–804 Zbl 0812.58094 MR 1177292

[8] A. N. Bernal and M. Sánchez, Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes. *Comm. Math. Phys.* **257** (2005), no. 1, 43–50 Zbl 1081.53059 MR 2163568

[9] A. N. Bernal and M. Sánchez, Globally hyperbolic spacetimes can be defined as 'causal' instead of 'strongly causal'. *Classical Quantum Gravity* **24** (2007), no. 3, 745–750 Zbl 1177.53065 MR 2294243

[10] K. Bingham, Y. Kurylev, M. Lassas, and S. Siltanen, Iterative time-reversal control for inverse problems. *Inverse Probl. Imaging* **2** (2008), no. 1, 63–81 Zbl 1141.35468 MR 2375323

[11] J.-M. Bony, Second microlocalization and propagation of singularities for semi-linear hyperbolic equations. In *Hyperbolic equations and related topics (Katata and Kyoto, 1984)*, pp. 11–49, Academic Press, Boston, MA, 1986 Zbl 0669.35073 MR 0925240

[12] R. Bosi, Y. Kurylev, and M. Lassas, Reconstruction and stability in Gelfand's inverse interior spectral problem. *Anal. PDE* **15** (2022), no. 2, 273–326 Zbl 1487.35439 MR 4409880

[13] T. Brander, B. Harrach, M. Kar, and M. Salo, Monotonicity and enclosure methods for the p-Laplace equation. *SIAM J. Appl. Math.* **78** (2018), no. 2, 742–758 Zbl 1516.35500 MR 3773773

[14] D. Burago, S. Ivanov, M. Lassas, and J. Lu, Quantitative stability of Gel'fand's inverse boundary problem. *Anal. PDE* **18** (2025), no. 4, 963–1035 Zbl 1562.35515 MR 4885877

[15] A. P. Calderón, On an inverse boundary value problem. *Comput. Appl. Math.* **25** (2006), no. 2–3, 133–138 Zbl 1182.35230 MR 2321646

[16] C. I. Cârstea, A. Feizmohammadi, Y. Kian, K. Krupchyk, and G. Uhlmann, The Calderón inverse problem for isotropic quasilinear conductivities. *Adv. Math.* **391** (2021), article no. 107956, 31 pp. Zbl 1479.35943 MR 4300916

[17] C. I. Cârstea, T. Liimatainen, and L. Tzou, The Calderón problem on Riemannian surfaces and of minimal surfaces. 2024, arXiv:2406.16944v1

[18] C. I. Cârstea, G. Nakamura, and M. Vashisth, Reconstruction for the coefficients of a quasilinear elliptic partial differential equation. *Appl. Math. Lett.* **98** (2019), 121–127 Zbl 1423.35437 MR 3964222

[19] M. F. Dahl, A. Kirpichnikova, and M. Lassas, Focusing waves in unknown media by modified time reversal iteration. *SIAM J. Control Optim.* **48** (2009), no. 2, 839–858 Zbl 1186.35238 MR 2486096

[20] G. Eskin, Inverse problems for general second order hyperbolic equations with time-dependent coefficients. *Bull. Math. Sci.* **7** (2017), no. 2, 247–307 Zbl 1381.35240 MR 3671738

[21] A. Feizmohammadi, Y. Kian, and L. Oksanen, Rigidity of inverse problems for nonlinear elliptic equations on manifolds. *Bull. Lond. Math. Soc.* **56** (2024), no. 9, 2802–2823 Zbl 1553.35236 MR 4795979

[22] A. Feizmohammadi, M. Lassas, and L. Oksanen, Inverse problems for nonlinear hyperbolic equations with disjoint sources and receivers. *Forum Math. Pi* **9** (2021), article no. e10, 52 pp. Zbl 1479.35946 MR 4336252

[23] A. Greenleaf and G. Uhlmann, Recovering singularities of a potential from singularities of scattering data. *Comm. Math. Phys.* **157** (1993), no. 3, 549–572 Zbl 0790.35112 MR 1243710

[24] C. Guillarmou and L. Tzou, Calderón inverse problem with partial data on Riemann surfaces. *Duke Math. J.* **158** (2011), no. 1, 83–120 Zbl 1222.35212 MR 2794369

[25] T. Helin, M. Lassas, L. Oksanen, and T. Saksala, Correlation based passive imaging with a white noise source. *J. Math. Pures Appl. (9)* **116** (2018), 132–160 Zbl 1401.35343 MR 3826551

[26] O. Imanuvilov and M. Yamamoto, Unique determination of potentials and semilinear terms of semilinear elliptic equations from partial Cauchy data. *J. Inverse Ill-Posed Probl.* **21** (2013), no. 1, 85–108 Zbl 1273.35318 MR 3023500

[27] O. Y. Imanuvilov, G. Uhlmann, and M. Yamamoto, The Calderón problem with partial data in two dimensions. *J. Amer. Math. Soc.* **23** (2010), no. 3, 655–691 Zbl 1201.35183 MR 2629983

[28] O. Y. Imanuvilov, G. Uhlmann, and M. Yamamoto, Partial Cauchy data for general second order elliptic operators in two dimensions. *Publ. Res. Inst. Math. Sci.* **48** (2012), no. 4, 971–1055 Zbl 1260.35253 MR 2999548

[29] V. Isakov, On uniqueness in inverse problems for semilinear parabolic equations. *Arch. Ration. Mech. Anal.* **124** (1993), no. 1, 1–12 Zbl 0804.35150 MR 1233645

[30] V. Isakov and A. I. Nachman, Global uniqueness for a two-dimensional semilinear elliptic inverse problem. *Trans. Amer. Math. Soc.* **347** (1995), no. 9, 3375–3390 Zbl 0849.35148 MR 1311909

[31] K. Kang and G. Nakamura, Identification of nonlinearity in a conductivity equation via the Dirichlet-to-Neumann map. *Inverse Problems* **18** (2002), no. 4, 1079–1088 Zbl 1043.35146 MR 1929283

[32] A. Katchalov, Y. Kurylev, and M. Lassas, *Inverse boundary spectral problems*. Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math. 123, Chapman and Hall/CRC, Boca Raton, FL, 2001, 290 pp. Zbl 1037.35098 MR 1889089

[33] A. Katchalov, Y. Kurylev, M. Lassas, and N. Mandache, Equivalence of time-domain inverse problems and boundary spectral problems. *Inverse Problems* **20** (2004), no. 2, 419–436 Zbl 1073.35209 MR 2065431

[34] C. E. Kenig, J. Sjöstrand, and G. Uhlmann, The Calderón problem with partial data. *Ann. of Math. (2)* **165** (2007), no. 2, 567–591 Zbl 1127.35079 MR 2299741

[35] Y. Kian, K. Krupchyk, and G. Uhlmann, Partial data inverse problems for quasilinear conductivity equations. *Math. Ann.* **385** (2023), no. 3–4, 1611–1638 Zbl 1512.35665 MR 4566701

[36] A. Kirpichnikova, J. Korpela, M. J. Lassas, and L. Oksanen, Construction of artificial point sources for a linear wave equation in unknown medium. *SIAM J. Control Optim.* **59** (2021), no. 5, 3737–3761 Zbl 1483.35303 MR 4325739

[37] K. Krupchyk, Y. Kurylev, and M. Lassas, Inverse spectral problems on a closed manifold. *J. Math. Pures Appl. (9)* **90** (2008), no. 1, 42–59 Zbl 1146.58020 MR 2435213

[38] K. Krupchyk and M. Lassas, An inverse problem for a hyperbolic system on a vector bundle and energy measurements. *Math. Ann.* **354** (2012), no. 4, 1431–1464 Zbl 1260.58017 MR 2993001

[39] K. Krupchyk and G. Uhlmann, A remark on partial data inverse problems for semilinear elliptic equations. *Proc. Amer. Math. Soc.* **148** (2020), no. 2, 681–685 Zbl 1431.35246 MR 4052205

[40] K. Krupchyk and G. Uhlmann, Partial data inverse problems for semilinear elliptic equations with gradient nonlinearities. *Math. Res. Lett.* **27** (2020), no. 6, 1801–1824 Zbl 1459.35399 MR 4216606

[41] K. Krupchyk and G. Uhlmann, Inverse problems for nonlinear magnetic Schrödinger equations on conformally transversally anisotropic manifolds. *Anal. PDE* **16** (2023), no. 8, 1825–1868 Zbl 1530.35364 MR 4657146

[42] Y. Kurylev, M. Lassas, J. Lu, and T. Yamaguchi, Inverse spectral problems for collapsing manifolds I: uniqueness and stability. [v1] 2012, [v3] 2024, arXiv:1209.5875v3

[43] Y. Kurylev, M. Lassas, and G. Uhlmann, Inverse problems for Lorentzian manifolds and non-linear hyperbolic equations. *Invent. Math.* **212** (2018), no. 3, 781–857 Zbl 1396.35074 MR 3802298

[44] Y. Kurylev, L. Oksanen, and G. P. Paternain, Inverse problems for the connection Laplacian. *J. Differential Geom.* **110** (2018), no. 3, 457–494 Zbl 1415.53024 MR 3880231

[45] R.-Y. Lai and Y.-H. Lin, Global uniqueness for the fractional semilinear Schrödinger equation. *Proc. Amer. Math. Soc.* **147** (2019), no. 3, 1189–1199 Zbl 1406.35468 MR 3896066

[46] M. Lassas, T. Liimatainen, Y.-H. Lin, and M. Salo, Inverse problems for elliptic equations with power type nonlinearities. *J. Math. Pures Appl. (9)* **145** (2021), 44–82 Zbl 1460.35395 MR 4188325

[47] M. Lassas, T. Liimatainen, Y.-H. Lin, and M. Salo, Partial data inverse problems and simultaneous recovery of boundary and coefficients for semilinear elliptic equations. *Rev. Mat. Iberoam.* **37** (2021), no. 4, 1553–1580 Zbl 1473.35653 MR 4269409

[48] M. Lassas, T. Liimatainen, L. Potenciano-Machado, and T. Tyni, Uniqueness, reconstruction and stability for an inverse problem of a semi-linear wave equation. *J. Differential Equations* **337** (2022), 395–435 Zbl 1497.35521 MR 4473034

[49] M. Lassas, T. Liimatainen, L. Potenciano-Machado, and T. Tyni, An inverse problem for a semi-linear wave equation: a numerical study. *Inverse Probl. Imaging* **18** (2024), no. 1, 62–85 Zbl 1529.65047 MR 4685800

[50] M. Lassas, T. Liimatainen, L. Potenciano-Machado, and T. Tyni, Stability and Lorentzian geometry for an inverse problem of a semilinear wave equation. *Anal. PDE* **18** (2025), no. 5, 1065–1118 Zbl 1565.35369 MR 4904383

[51] M. Lassas, T. Liimatainen, and M. Salo, The Poisson embedding approach to the Calderón problem. *Math. Ann.* **377** (2020), no. 1–2, 19–67 Zbl 1442.35138 MR 4099622

[52] M. Lassas, J. Lu, and T. Yamaguchi, Inverse spectral problems for collapsing manifolds II: quantitative stability of reconstruction for orbifolds. 2024, arXiv:2404.16448v1

[53] M. Lassas and L. Oksanen, Inverse problem for the Riemannian wave equation with Dirichlet data and Neumann data on disjoint sets. *Duke Math. J.* **163** (2014), no. 6, 1071–1103 Zbl 1375.35634 MR 3192525

[54] M. Lassas and T. Saksala, Determination of a Riemannian manifold from the distance difference functions. *Asian J. Math.* **23** (2019), no. 2, 173–199 Zbl 1419.53038 MR 3978249

[55] M. Lassas and G. Uhlmann, On determining a Riemannian manifold from the Dirichlet-to-Neumann map. *Ann. Sci. Éc. Norm. Super. (4)* **34** (2001), no. 5, 771–787 Zbl 0992.35120 MR 1862026

[56] R. Melrose and N. Ritter, Interaction of nonlinear progressing waves for semilinear wave equations. *Ann. of Math. (2)* **121** (1985), no. 1, 187–213 Zbl 0575.35063 MR 0782559

[57] R. B. Melrose and G. A. Uhlmann, Lagrangian intersection and the Cauchy problem. *Comm. Pure Appl. Math.* **32** (1979), no. 4, 483–519 Zbl 0396.58006 MR 0528633

[58] C. Muñoz and G. Uhlmann, The Calderón problem for quasilinear elliptic equations. *Ann. Inst. H. Poincaré C Anal. Non Linéaire* **37** (2020), no. 5, 1143–1166 Zbl 1457.35093 MR 4138229

[59] M. Salo and X. Zhong, An inverse problem for the p-Laplacian: boundary determination. *SIAM J. Math. Anal.* **44** (2012), no. 4, 2474–2495 Zbl 1251.35191 MR 3023384

[60] P. Stefanov and G. Uhlmann, Stable determination of generic simple metrics from the hyperbolic Dirichlet-to-Neumann map. *Int. Math. Res. Not. IMRN* **2005** (2005), no. 17, 1047–1061 Zbl 1088.53027 MR 2145709

[61] Z. Sun, On a quasilinear inverse boundary value problem. *Math. Z.* **221** (1996), no. 2, 293–305 Zbl 0843.35137 MR 1376299

[62] Z. Sun and G. Uhlmann, Inverse problems in quasilinear anisotropic media. *Amer. J. Math.* **119** (1997), no. 4, 771–797 Zbl 0886.35176 MR 1465069

[63] J. Sylvester and G. Uhlmann, A global uniqueness theorem for an inverse boundary value problem. *Ann. of Math. (2)* **125** (1987), no. 1, 153–169 Zbl 0625.35078 MR 0873380

[64] D. Tataru, Unique continuation for solutions to PDE's; between Hörmander's theorem and Holmgren's theorem. *Comm. Partial Differential Equations* **20** (1995), no. 5–6, 855–884 Zbl 0846.35021 MR 1326909

[65] G. Uhlmann, Electrical impedance tomography and Calderón's problem. *Inverse Problems* **25** (2009), no. 12, article no. 123011, 39 pp. Zbl 1181.35339 MR 3460047

Matti Lassas
Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), 00014 Helsinki, Finland; matti.lassas@helsinki.fi

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/22

Cheeger–Müller and Bismut–Zhang theorem on singular spaces

Ursula Ludwig

Abstract. In this article, we present the Cheeger–Müller/Bismut–Zhang theorem for singular spaces with isolated conical singularities proved recently by the author. The classical Cheeger–Müller theorem, the comparison of analytic and topological torsion of a smooth compact manifold, is one of the most important comparison theorems in global analysis. Its most general version for smooth compact manifolds has been achieved by Bismut and Zhang by combining local index techniques and the Witten deformation.

The question of a Cheeger–Müller theorem for spaces with isolated conical singularities has been raised by Dar in the late 1980s, very early in the development of the (global) analysis on these spaces. The Cheeger–Müller/Bismut–Zhang theorem for singular spaces with isolated conical singularities presented here answers the question by generalising Bismut and Zhang's approach to the singular setting. An important step in this programme is hence the generalisation of the Morse–Thom–Smale complex and the Witten deformation to the singular setting, achieved by the author, and which we also review here.

This article is a written and extended version of the author's invited talk at the 9ECM, Sevilla, 2024.

1. Introduction

An important comparison theorem in global analysis is the comparison of analytic (or Ray–Singer) and topological (or Reidemeister) torsion, aka the Cheeger–Müller theorem, for smooth compact Riemannian manifolds equipped with a unitary flat vector bundle. It has been conjectured by Ray and Singer [67] and has been independently proved by Cheeger [23] and Müller [63]. Müller [64] extended the result to unimodular bundles. At the same time, Bismut and Zhang [11, 12] achieved the most general version of the comparison theorem of torsions for smooth compact manifolds, by a very different approach, namely combining the Witten deformation and local index

Mathematics Subject Classification 2020: 58J52 (primary); 58J28, 58J35, 35S35 (secondary).
Keywords: Ray–Singer analytic torsion, Cheeger–Müller theorem, anomaly formulas, conical singularities.

techniques. Bismut and Zhang compare the analytic torsion with the Milnor torsion, a Morse theoretic version of the topological torsion defined in [62]. If the flat vector bundle is not unitary or unimodular, the two torsions are no longer equal, and the difference between them can be expressed in terms of the Mathai–Quillen current. In this note, we refer to this most general version of the comparison theorem of torsions as the Bismut–Zhang theorem. The Cheeger–Müller/Bismut–Zhang theorem has been extended to the equivariant setting [12, 46, 47] and for manifolds with boundary [18, 19, 47, 74].

In the last decade, a lot of progress has been made in the study of the Cheeger–Müller theorem for spaces with singularities of different types. Here, we only recall results in the conical world, which is the setting of this article. The study of analytic and topological torsion on *singular spaces with conical singularities* started with the work of A. Dar in [27]. She proved that on singular spaces with isolated conical singularities, the analytic torsion is well defined. She also defined the Reidemeister torsion for stratified pseudomanifolds in the context of intersection homology (with any perversity in the sense of Goresky and MacPherson [30, 31]) and proved that for an even-dimensional oriented space with isolated conical singularities, due to a duality argument, both analytic and intersection Reidemeister torsion are trivial, hence equal.

An approach to the Cheeger–Müller theorem for spaces with isolated conical singularities, suggested in [44, Problem 5.3], is to reduce the problem via gluing formulas to a comparison of torsions on the truncated cone. The gluing formula for the analytic torsion for manifolds with boundary equipped with an arbitrary flat bundle, with and without product structure assumptions near the boundary, is due to Brüning and Ma [19]. Later, Lesch [45] established gluing formulas in the context of spaces with singularities, under product assumptions, by generalising a technique of Vishik [74]. The analytic torsion of a truncated cone has been studied in [34–36, 65, 73]. A problem faced in this approach is the missing interpretation of one of the analytic torsion-like terms appearing in these explicit computations.

In [61] Mazzeo and Vertman prove the well-definedness of analytic torsion for *wedge spaces* (also called incomplete edge spaces in the literature), i.e., spaces with a singular stratum of positive dimension and a cone-like metric near the singular stratum. The authors also prove invariance of the analytic torsion under variation of the wedge metric for an odd-dimensional space with an odd-dimensional singular stratum. A Cheeger–Müller theorem on an odd-dimensional wedge space with odd-dimensional singular stratum equipped with a unimodular bundle satisfying an additional acyclicity condition has been proved in [6]. The strategy in [6] consists in the study of analytic torsion via degeneration of smooth metrics into conical metrics, a strategy already applied successfully by the authors to the study of torsion on spaces with cuspidal singularities [4, 5]. The assumptions made in [6] exclude the

case of odd-dimensional spaces with isolated conical singularities; they also exclude the case of the trivial bundle.

In [56, 58], the author proposed yet another approach to the study of torsion on singular spaces, namely by generalising the approach of Bismut and Zhang to the singular setting. This has led to an answer to the long open question of [27], by providing a Cheeger–Müller/Bismut–Zhang formula in the following singular setting (see Section 2 for the precise definitions): Let (X, g^{TX}) be a space with isolated conical singularities; the singular set is denoted by $\mathrm{Sing}(X)$. We equip the smooth locus $X \setminus \mathrm{Sing}(X)$ with a Hermitian flat vector bundle (F, ∇^F, g^F). We denote by $\theta(F, g^F)$ the closed 1-form defined in (2.2), associated with the Hermitian flat vector bundle (F, ∇^F, g^F), whose cohomology class measures the obstruction to the existence of a flat volume form on $\det F$. In particular, $\theta(F, g^F) = 0$ in case (F, ∇^F, g^F) is a unitary bundle. For a perversity $\overline{q}$ in the sense of Goresky and MacPherson, we denote by $\det IH^\bullet_{\overline{q}}(X, F)$ the determinant line of the intersection cohomology of X with perversity $\overline{q}$ and values in the flat bundle F. We denote by $\overline{m}$ (resp. $\overline{n}$) the lower (resp. upper) middle perversity. We denote by $\Psi(TX, \nabla^{TX})$ the $n-1$-dimensional Mathai–Quillen current on TX with values in the orientation line $o(TX)$, as constructed in [60, Section 7] (see also [11, Section III]).

Theorem 1.1. *Let (X, g^{TX}) be a space with isolated conical singularities, and let $\overline{q} \in \{\overline{m}, \overline{n}\}$ denote one of the middle perversities. Let (F, ∇^F, g^F) be a Hermitian flat vector bundle over $X \setminus \mathrm{Sing}(X)$ as in Section 2.2. Let Y be an anti-radial gradient-like vector field on X (see Section 3). Then*

$$\log\left(\frac{\| \ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)}}{\| \ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F)}}\right)^2 = -\int_X \theta(F, g^F) Y^* \Psi(TX, \nabla^{TX}), \tag{1.1}$$

where $\| \ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)}$ (resp. $\| \ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F)}$) denotes the Ray–Singer (resp. the Bismut–Zhang) metric (see Section 5).

If (F, ∇^F, g^F) is unitary, we have that $\theta(F, g^F) = 0$. Hence, in this case, the right-hand side of formula (1.1) vanishes, and Theorem 1.1 is a generalisation of the classical Cheeger–Müller theorem [23, 63] to spaces with isolated conical singularities. Profiting from the strength of the Bismut–Zhang approach, here we provide a comparison formula in the case of a general Hermitian flat coefficient bundle, thus also generalising the Bismut–Zhang theorem in [11] to the singular setting. Note that the integral on the right-hand side in (1.1) is well defined and corresponds to the right-hand side in the smooth Bismut–Zhang theorem.

In the case that (X, g^{TX}) is a smooth compact manifold, Theorem 1.1 recovers the smooth Cheeger–Müller/Bismut–Zhang theorem. For X an even-dimensional oriented space with isolated conical singularities equipped with a unitary vector bundle, Theorem 1.1 recovers the comparison result of torsions already known by Dar [27]. Note that for non-Witt spaces, Theorem 1.1 provides two formulas for the two middle perversities $\overline{q} \in \{\overline{m}, \overline{n}\}$, which are compatible with Poincaré duality for intersection cohomology.

The treatment of the Cheeger–Müller/Bismut–Zhang formula in [56, 58] is complete: In [58], we also describe the dependence of all three terms in formula (1.1) on the choice of metrics g^{TX}, g^F on TX, F.

Indeed, our results can be extended to a slightly more general setting: namely, to spaces with isolated asymptotically conical singularities; we can also allow metrics g^F on the flat bundle F, which are smooth up to the "boundary" but do not necessarily satisfy the product-type assumption in Section 2.2.

The main actors of the proof of the smooth Bismut–Zhang theorem in [11] are local index techniques and the Witten deformation. Hence, the first step in the programme of generalising the approach of Bismut and Zhang to the singular setting is the generalisation to singular spaces of the Morse–Thom–Smale complex in [54, 59] and of the Witten deformation in [55]. They have been achieved in a more general singular setting and will be reviewed here as well. The second main actor in the proof of Theorem 1.1 is the use of local index techniques: We combine the smooth local index techniques in [11] with a local index result on the infinite cone in [57]; we will not comment on these in this article.

This article is conceived as an overview article, and we will hence focus on conveying the main constructions and main ideas only; for the proofs, we refer the reader to the original articles. In this presentation, we restrict ourselves mostly to the case of spaces with isolated singularities. Part of the results presented here have been achieved in a much more general context, which we only mention shortly here without giving all technical details.

The article is organised as follows. In Section 2, we recall basic definitions and results on singular spaces with isolated conical singularities. In Section 3, we recall the definition of anti-radial/radial Morse functions and anti-radial/radial gradient-like vector fields as well as the generalisation of the famous Morse–Thom–Smale complex to singular spaces, anti-radial gradient-like vector fields and intersection (co-) homology. In Section 4, we review the generalisation of the Witten deformation to singular spaces with conical and iterated conical singularities. In Section 5, we recall the definition of the two torsions appearing in the Cheeger–Müller/Bismut–Zhang theorem for singular spaces with isolated conical singularities, Theorem 1.1. Moreover, we also present anomaly formulas for all three terms appearing in formula (1.1).

2. Preliminaries on singular spaces with isolated conical singularities

2.1. Singular spaces with isolated conical singularities

For a smooth manifold L and $\delta > 0$, we denote by

$$c_\delta L := ([0,\delta) \times L)/_{(0,x)\sim(0,y)}$$

the (open) truncated cone over L.

Let X be a connected topological space, $\mathrm{Sing}(X) \subset X$ a finite set of points such that $X_{\mathrm{sm}} := X \setminus \mathrm{Sing}(X)$ is a smooth manifold of dimension $n \geq 2$. We denote by TX (resp. by T^*X) the tangent bundle (resp. the cotangent bundle) of X_{sm}. Let g^{TX} be a Riemannian metric on X_{sm}. We assume that (X, g^{TX}) is a space with isolated conical singularities of dimension n, i.e.,

(1) For $p \in \mathrm{Sing}(X)$, there exist an open neighbourhood $B_\delta(p)$ of p, a smooth compact connected manifold L_p of dimension $\dim L_p = n-1$ and a diffeomorphism $\varphi_p : B_\delta(p) \setminus \{p\} \simeq c_\delta L_p \setminus \{0\}$. The diffeomorphism φ_p extends to a homeomorphism, still denoted by φ_p,

$$\varphi_p : B_\delta(p) \simeq c_\delta L_p \quad \text{and} \quad g^{TX}_{|B_\delta(p)\setminus\{p\}} = \varphi_p^*(dr^2 + r^2 g^{TL_p}), \tag{2.1}$$

where r is the radial coordinate and g^{TL_p} is a Riemannian metric on the manifold L_p (not depending on r).

(2) The set

$$X \setminus \Big(\bigcup_{p\in\mathrm{Sing}(X)} B_\delta(p) \Big)$$

is a smooth compact manifold of dimension n with boundary $\bigcup_{p\in\mathrm{Sing}(X)} L_p$.

The set $\mathrm{Sing}(X)$ is called the singular set of X. For $p \in \mathrm{Sing}(X)$, the manifold L_p is called the link of X at p. The assumption on the connectivity of the link manifolds L_p, $p \in \mathrm{Sing}(X)$, can be removed easily and is included here to simplify the presentation; it is equivalent to the topological normality of the space.

Let us emphasise that the radial coordinate r in (2.1) is fixed throughout this article.

2.2. Flat vector bundles over X

Let (F, ∇^F) be a flat vector bundle over X_{sm} with canonical flat connection ∇^F. Since the punctured neighbourhood of $p \in \mathrm{Sing}(X)$ is contractible to the link L_p, the restriction of (F, ∇^F) to the punctured neighbourhood can be identified with the pull-back of a flat vector bundle $(F_{L_p}, \nabla^{F_{L_p}})$ on the link L_p, i.e., $(F, \nabla^F)_{|B_\delta(p)\setminus\{p\}} \simeq$

$\pi^*(F_{L_p}, \nabla^{F_{L_p}})$, where $\pi : B_\delta(p) \setminus \{p\} \simeq (0, \delta) \times L_p \to L_p$ is the canonical projection.

We equip the flat vector bundle (F, ∇^F) with a (not necessarily flat) Hermitian metric g^F such that, near $p \in \mathrm{Sing}(X)$, g^F is the pull-back of a Hermitian metric $g^{F_{L_p}}$ on the flat vector bundle F_{L_p} on L_p.

We denote by F^* the flat bundle dual to F and by $F^*_{L_p}$ its restriction to L_p, $p \in \mathrm{Sing}(X)$.

Let $\omega(F, g^F)$ be the 1-form on X with values in the self-adjoint endomorphisms of F:

$$\omega(F, g^F) = (g^F)^{-1} \nabla^F g^F.$$

Note that $\omega(F, g^F) = 0$ in case g^F is flat. By taking the pointwise trace, denoted Tr, we get a closed 1-form $\theta(F, g^F)$ on X:

$$\theta(F, g^F) = \mathrm{Tr}[\omega(F, g^F)]. \tag{2.2}$$

The cohomology class $[\theta(F, g^F)]$ measures the obstruction to the existence of a flat volume form on F. By our assumption, near $p \in \mathrm{Sing}(X)$, the form $\theta(F, g^F)$ does not depend on the radial coordinate r.

2.3. Local model near $p \in \mathrm{Sing}(X)$

Let $p \in \mathrm{Sing}(X)$. We denote by $cL_p := ([0, \infty) \times L_p)/_{(0,x)\sim(0,y)}$ the infinite cone over L_p, by 0 the cone tip and by

$$Z_p := cL_p \setminus \{0\} \simeq \mathbb{R}_{>0} \times L_p$$

the punctured infinite cone. We write $x \in Z_p$ in its polar coordinates $x = (r, y)$, where r is the radial coordinate and y is the coordinate on the link. We equip Z_p with the conical metric $g^{TZ_p} = dr^2 + r^2 g^{TL_p}$. The flat bundle $(F_{L_p}, \nabla^{F_{L_p}}, g^{F_{L_p}})$ can be extended in a trivial way to a flat bundle $(F_{Z_p}, \nabla^{F_{Z_p}}, g^{F_{Z_p}})$ over Z_p. If no confusion can occur, we still write F for F_{Z_p}.

2.4. Intersection (co-)homology

An important topological invariant for singular spaces is the intersection homology and cohomology introduced by Goresky and MacPherson [30, 31]. For a perversity $\overline{q}$ in the sense of Goresky and MacPherson, we denote by $IH^{\overline{q}}_\bullet(X, F)$ (resp. by $IH^\bullet_{\overline{q}}(X, F)$) the intersection homology (resp. cohomology) with perversity $\overline{q}$ and coefficients in the local system associated with the flat bundle F.

For the most part of this article, we will focus on the two middle perversities: the lower middle perversity $\overline{m}$ and the upper middle perversity $\overline{n}$. A perversity $\overline{q}$ is a tuple of nonnegative natural numbers. For a space with isolated singularities, however, the

only relevant information is the last entry in this tuple. Hence, by slight abuse of notation, for a space with isolated singularities, we will identify a perversity $\overline{q}$ with this last entry; more concretely,

$$\overline{q} = \begin{cases} \left\lfloor \frac{n}{2} \right\rfloor - 1 & \text{for } \overline{q} = \overline{m}, \\ \left\lfloor \frac{n-1}{2} \right\rfloor & \text{for } \overline{q} = \overline{n}. \end{cases}$$

For an even-dimensional space with isolated singularities, lower and upper middle perversities coincide, and hence also the two intersection cohomologies are the same. More generally, if X is a Witt space, i.e., if $H^{\frac{n-1}{2}}(L_p, F_{L_p}) = 0$ for all $p \in \mathrm{Sing}(X)$, then by [70, Theorem 3.4],

$$IH^{\bullet}_{\overline{m}}(X, F) \simeq IH^{\bullet}_{\overline{n}}(X, F).$$

In this paper, we will not assume the Witt condition, and hence in the odd-dimensional case, we have to distinguish two cases.

We denote by $IH^{\overline{q}}_{\bullet}(cL_p, F)$ (resp. $IH^{\overline{q}}_{\bullet}(cL_p, L_p, F)$) the absolute (resp. relative) homology with perversity $\overline{q}$ of the cone cL_p; similarly, $IH^{\bullet}_{\overline{q}}(cL_p, F)$ (resp. $IH^{\bullet}_{\overline{q}}(cL_p, L_p, F)$) denotes the absolute (resp. relative) cohomology of the cone cL_p. Since it will appear frequently in this article, let us shortly recall the local calculation for the relative intersection cohomology of a cone (see [31, Section 2.4]):

$$IH^{k}_{\overline{q}}(cL_p, L_p, F) = \begin{cases} H^{k-1}(L_p, F_{L_p}) & \text{for } k \geq n - \overline{q}, \\ 0 & \text{else}, \end{cases} \tag{2.3}$$

and an analogous formula holds for the intersection homology.

2.5. L^2-cohomology

Throughout this article, we use the language of Hilbert complexes as introduced by Brüning and Lesch in [16].

We denote by $\langle\ ,\ \rangle$ the L^2-inner product on the space of sections of $\Lambda(T^*X) \otimes F$ induced from the metrics g^{TX}, g^F. We denote by $\Omega^{\bullet}_c(X, F)$ (resp. by $\Omega^{\bullet}(X, F)$) the graded vector space of smooth compactly supported sections (resp. of smooth sections) of $\Lambda(T^*X) \otimes F$. We denote by $L^2(\Lambda(T^*X) \otimes F)$ the space of L^2-sections. We denote by d_c the exterior derivative acting on $\Omega^{\bullet}_c(X, F)$.

The de Rham complex of compactly supported forms $(\Omega^{\bullet}_c(X, F), d_c, \langle\ ,\ \rangle)$ admits several closed extensions (in the Hilbert space of L^2-forms) into a Hilbert complex, a choice of which is called an ideal boundary condition (shortly *ibc*) by Cheeger. Here, we focus on the maximal (resp. minimal) extension, denoted by $(\mathcal{C}^{\bullet}_{\overline{m}}, d_{\overline{m}}, \langle\ ,\ \rangle) := (\mathcal{C}^{\bullet}_{\max}, d_{\max}, \langle\ ,\ \rangle)$ (resp. $(\mathcal{C}^{\bullet}_{\overline{n}}, d_{\overline{n}}, \langle\ ,\ \rangle) := (\mathcal{C}^{\bullet}_{\min}, d_{\min}, \langle\ ,\ \rangle)$),

where $d_{\overline{m}} := d_{\max}$ (resp. $d_{\overline{n}} := d_{\min}$) denotes the maximal (resp. the minimal) closed extension of d_c in $\mathrm{L}^2(\Lambda(T^*X) \otimes F)$. By [17, Theorems 3.7 and 3.8],

$$\mathrm{dom}(d_{\max})/\mathrm{dom}(d_{\min}) \simeq \bigoplus_{p \in \mathrm{Sing}(X)} H^{\frac{n-1}{2}}(L_p, F_{L_p}). \tag{2.4}$$

Hence, for a Witt space and in particular for an even-dimensional space, the extension of the de Rham complex $(\Omega_c^\bullet(X, F), d_c, \langle\ ,\ \rangle)$ into a Hilbert complex is unique, $d_{\min} = d_{\max}$.

The cohomology of the maximal Hilbert complex

$$(\mathcal{C}^\bullet_{\overline{m}}, d_{\overline{m}}, \langle\ ,\ \rangle) = (\mathcal{C}^\bullet_{\max}, d_{\max}, \langle\ ,\ \rangle)$$

is called the L^2-cohomology of X with values in F:

$$H^\bullet_{(2)}(X, F) := H^\bullet_{(2),\overline{m}}(X, F) := H^\bullet((\mathcal{C}^\bullet_{\max}, d_{\max}, \langle\ ,\ \rangle)). \tag{2.5}$$

We can also define the cohomology of the minimal extension $(\mathcal{C}^\bullet_{\overline{n}}, d_{\overline{n}}, \langle\ ,\ \rangle) = (\mathcal{C}^\bullet_{\min}, d_{\min}, \langle\ ,\ \rangle)$:

$$H^\bullet_{(2),\overline{n}}(X, F) := H^\bullet((\mathcal{C}^\bullet_{\min}, d_{\min}, \langle\ ,\ \rangle)). \tag{2.6}$$

The two cohomologies introduced here only depend on the quasi-isometry class of the Riemannian metric. From (2.4), we have for a Witt space $H^\bullet_{(2),\overline{m}}(X, F) = H^\bullet_{(2),\overline{n}}(X, F)$.

2.6. L^2-Hodge–de Rham theorem

By a result of Cheeger, Goresky and MacPherson [26, Section 3.4], integration of L^2-forms over intersection chains induces a de Rham isomorphism

$$H^\bullet_{(2),\overline{q}}(X, F) \simeq IH^\bullet_{\overline{q}}(X, F), \quad \overline{q} \in \{\overline{m}, \overline{n}\}. \tag{2.7}$$

Note that this shows, in particular, that the L^2-cohomologies (2.5) and (2.6) are indeed topological invariants of X.

For $\overline{q} \in \{\overline{m}, \overline{n}\}$, we denote by $\Delta^{\overline{q}}$ the Laplace operator associated with the Hilbert complex $(\mathcal{C}^\bullet_{\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle)$ in the sense of [16], and by $\Delta^{\overline{q},(i)}$ its restriction to i-forms. The Laplacian $\Delta^{\overline{q}}$ is a discrete operator. By a standard result on Hilbert complexes (see [16, Lemma 2.2 and Corollary 2.5]), the following L^2-Hodge isomorphism holds:

$$\ker(\Delta^{\overline{q}}) =: \mathcal{H}^\bullet_{(2),\overline{q}}(X, F) \simeq H^\bullet_{(2),\overline{q}}(X, F), \quad \overline{q} \in \{\overline{m}, \overline{n}\}, \tag{2.8}$$

a result first established by Cheeger [24, Section 1 and Theorem 5.1].

3. Radial/anti-radial gradient-like vector fields, radial/anti-radial Morse functions, the "singular" Morse–Thom–Smale complex computing intersection homology

The famous Morse inequalities establish a relation between the singular cohomology of a smooth compact manifold and the number of critical points of a smooth Morse function. A way of proving the Morse inequalities is to show the existence of a combinatorial complex, which is generated by the critical points of the Morse function and whose (co-)homology is isomorphic to the singular (co-)homology of the manifold. Such a complex was first established by Thom [72] and Smale [71], using the unstable cell decomposition of the manifold for the negative gradient flow of the Morse function with respect to a generic Riemannian metric. The boundary operator in this complex is given by counting trajectories between critical points of index difference 1. The Morse–Thom–Smale complex has seen a revival in the 1990s, starting with the work of Floer [29], where the idea of counting trajectories between critical points has been exploited in an infinite-dimensional setting.

In [55, Section 6] and [54], the Morse–Thom–Smale complex has been generalised for spaces with isolated singularities equipped with an anti-radial gradient-like vector field and intersection homology with arbitrary perversity $\overline{q}$. The construction also works more generally for all mezzo-perversities in the sense of [3]. This Morse theoretic description of intersection homology plays an important role in the generalisation of the Cheeger–Müller/Bismut–Zhang theorem in Section 5: In particular, the "singular" Morse–Thom–Smale complex is one of the two ingredients used in the definition of the Bismut–Zhang torsion (see Definition 5.4), providing the "topological" side in the comparison theorem of torsions in this singular setting.

In this section, we recall from [48, 54, 55] the definition of radial/anti-radial gradient-like vector fields and Morse functions, as well as the construction and main properties of the "singular" Morse–Thom–Smale complex for spaces with isolated singularities. The construction has been extended to stratified pseudomanifolds in [59], which we shortly mention in Section 3.3.

3.1. Radial/anti-radial gradient-like vector fields and radial/anti-radial Morse functions

Radial/anti-radial gradient-like vector fields and radial/anti-radial Morse functions have been introduced in [48, 54, 55]. One inspiration stems from radial vector fields introduced by Marie-Hélène Schwartz in [68, 69], in the context of Whitney stratified spaces, in her study of characteristic classes on singular spaces via obstruction theory. Radial Morse functions increase in the direction normal to the singular stratum. This is also the idea behind Goresky's π-fibre Morse functions in [33, Section 9].

Let us consider a vector field Y on X_{sm} satisfying the following conditions:

(a) Near a singular point $p \in \mathrm{Sing}(X)$, the vector field Y has the following normal form in the local coordinates (2.1):

$$Y = Y^{\pm} = \pm r \frac{\partial}{\partial r}. \tag{3.1}$$

(b) To any singular point $p \in X_{\mathrm{sm}}$ of Y (i.e., $Y(p) = 0$), there exist an open neighbourhood $U(p)$ of p and local coordinates $x_1, \ldots, x_n$, as well as $k \in \{0, \ldots, n\}$ such that

$$Y_{|U(p)} = -x_1 \frac{\partial}{\partial x_1} - \cdots - x_k \frac{\partial}{\partial x_k} + x_{k+1} \frac{\partial}{\partial x_{k+1}} + \cdots + x_n \frac{\partial}{\partial x_n}. \tag{3.2}$$

We call k the index of Y at p. We denote by $\mathrm{Crit}(Y)$ (resp. by $\mathrm{Crit}_k(Y), k = 0, \ldots, n$) the set of singular points of Y in X_{sm} (resp. the set of singular points of Y of index k).

The vector field $-Y$ induces a well-defined smooth flow on X_{sm}, which extends to a continuous flow $\Phi : X \times \mathbb{R} \to X$.

Definition 3.1. A smooth vector field Y over X_{sm} is called a standard radial (resp. anti-radial) gradient-like vector field, if it satisfies conditions (a) and (b) above as well as the following conditions:

(c) The set of non-wandering points of the flow Φ equals $\mathrm{Crit}(Y) \cup \mathrm{Sing}(X)$.

(d) The Morse–Smale transversality condition holds, i.e., all intersections of stable and unstable manifolds (w.r.t. the flow Φ) are transverse.

We call the vector field radial (resp. anti-radial) gradient-like vector field, if instead of (b), we only assume that, locally near $p \in \mathrm{Crit}(Y)$, the vector field can be written as the gradient vector field of a Morse function with respect to some Riemannian metric.

Note that the Morse–Smale transversality condition is a generic condition (see, e.g., [37, Theorem 7.2.13]) and can always be achieved by perturbing a vector field which satisfies (a)–(c) outside a neighbourhood of $\mathrm{Crit}(Y) \cup \mathrm{Sing}(X)$. Existence of radial/anti-radial gradient-like vector fields (in a more general context) has been proved in [48, Proposition 6.4].

Definition 3.2. A continuous function $f : X \to \mathbb{R}$ is called a radial (resp. an anti-radial) Morse function, if the following two conditions hold:

(a) The restriction $f_{\mathrm{sm}} := f_{|X_{\mathrm{sm}}}$ is a smooth Morse function.

(b) Near a singular point $p \in \mathrm{Sing}(X)$, the function f has the following normal form in the local coordinates (2.1):

$$f(r, y) = f^{\pm}(r, y) = f(p) \pm \frac{1}{2} r^2. \tag{3.3}$$

For a radial/anti-radial Morse function f, we denote by $\mathrm{Crit}(f_{\mathrm{sm}})$ (resp. by $\mathrm{Crit}_k(f_{\mathrm{sm}})$, $k = 0, \ldots, n$) the set of critical points of f_{sm} (resp. the set of critical points of f_{sm} of index k). Set

$$\mathrm{Crit}(f) := \mathrm{Crit}(f_{\mathrm{sm}}) \cup \mathrm{Sing}(X).$$

By a straightforward generalisation of a result of Smale [71, Theorem B], we have the following relation between the two notions just introduced: Let Y be a standard radial (resp. anti-radial) gradient-like vector field on X_{sm}. Then there exists a radial (resp. an anti-radial) Morse function $f : X \to \mathbb{R}$ and a Riemannian metric g'^{TX} on X_{sm} such that $Y = \nabla_{g'^{TX}} f$ and, in an open neighbourhood of $\mathrm{Sing}(X)$, we have $g'^{TX} = g^{TX}$. Note that, to achieve (3.2) in a neighbourhood of $\mathrm{Crit}(f_{\mathrm{sm}}) = \mathrm{Crit}(Y)$, we choose Morse coordinates and g'^{TX} to be the Euclidean metric in the Morse coordinates.

Let us again emphasise that the radial coordinate r near $p \in \mathrm{Sing}(X)$ is fixed and the same in (2.1), (3.1) and (3.3).

3.2. The "singular" Morse–Thom–Smale complex computing intersection cohomology

Let us now recall the construction and main results on the "singular" Morse–Thom–Smale complex from [54,55]: For a singular point $p \in \mathrm{Sing}(X)$, we denote by $o(TL_p)$ the orientation bundle of L_p. Moreover, we have the following notation:

$$\Omega^{\bullet}(L_p, F^*_{L_p} \otimes o(TL_p))$$

denotes the space of smooth de Rham forms on L_p with values in the flat bundle $F^*_{L_p} \otimes o(TL_p)$, and $H^{\bullet}(L_p, F^*_{L_p} \otimes o(TL_p))$ denotes the singular cohomology of L_p. Let

$$\Xi^k_p \subset \Omega^k(L_p, F^*_{L_p} \otimes o(TL_p)) \tag{3.4}$$

be a set of closed forms on L_p, whose cohomology classes form a basis of

$$H^k(L_p, F^*_{L_p} \otimes o(TL_p)), \mathrm{span}(\Xi^k_p) \simeq H^k(L_p, F^*_{L_p} \otimes o(TL_p)).$$

In Section 4.3, we will choose these representatives as follows: By the smooth Hodge theorem for the compact manifold (L_p, g^{TL_p}), we can identify $H^{\bullet}(L_p, F^*_{L_p} \otimes o(TL_p))$ with the vector space of harmonic forms $\mathcal{H}^{\bullet}(L_p, F^*_{L_p} \otimes o(TL_p))$ and

choose Ξ_p^k as an ONB of $\mathcal{H}^\bullet(L_p, F_{L_p}^* \otimes o(TL_p))$ (w.r.t. the L^2-metric induced from $g^{TL_p}, g^{F_{L_p}}$).

Let Y be a standard anti-radial gradient-like vector field. For $p \in \mathrm{Crit}(Y)$, we denote by $W^u(p)$ (resp. $W^s(p)$) the unstable (resp. stable) set of p with respect to the flow Φ; its intersections with X_{sm} are smooth manifolds. For $p, q \in \mathrm{Crit}(Y)$ with $\mathrm{ind}(p) - \mathrm{ind}(q) = 1$, by Morse–Smale transversality, the space of trajectories of the flow Φ (induced from $-Y$) starting in p and ending in q is a finite set, which we denote by $\Gamma(p, q)$. Choosing orientations on the unstable manifolds of points in $\mathrm{Crit}(Y)$ and using the flow Φ induce orientations on $\Gamma(p, q)$ (see [40, Section (c)] for more details). With other words, to each trajectory $\gamma \in \Gamma(p, q)$ we can assign $n_\gamma(p, q) \in \{\pm 1\}$. We denote by $\tau_\gamma : F_p^* \to F_q^*$ the map induced from parallel transport with respect to the flat connection along the trajectory γ.

To a standard anti-radial gradient-like vector field Y and the sets Ξ_p^{n-k}, $p \in \mathrm{Sing}(X)$, $k \geq n - \overline{q}$, one can associate a geometric complex, which computes the intersection homology of X with values in F^* and perversity $\overline{q}$, $IH_\bullet^{\overline{q}}(X, F^*)$, as follows.

Definition/Proposition 3.3. We denote by $(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet)$ the following complex:

$$C_k^{\overline{q}}(X, Y, F^*) = \begin{cases} \left(\bigoplus_{p \in \mathrm{Crit}_k(Y)} \langle [W^u(p)] \rangle \otimes F_p^* \right) \oplus \left(\bigoplus_{p \in \mathrm{Sing}(X)} \mathrm{span}(\Xi_p^{n-k}) \right) & \text{if } k \geq n - \overline{q}, \\ \bigoplus_{p \in \mathrm{Crit}_k(Y)} \langle [W^u(p)] \rangle \otimes F_p^* & \text{else.} \end{cases}$$

The boundary operator $\partial_\bullet$ is defined as follows: For $p \in \mathrm{Crit}_k(Y)$, $h \in F_p^*$:

$$\partial_k([W^u(p)] \otimes h) = \sum_{q \in \mathrm{Crit}_{k-1}(Y)} \sum_{\gamma \in \Gamma(p,q)} n_\gamma(p, q) \cdot [W^u(q)] \otimes \tau_\gamma(h) \in C_{k-1}^{\overline{q}}(X, Y, F^*).$$

For $p \in \mathrm{Sing}(X)$, $k \geq n - \overline{q}$, $\xi_p^{n-k} \in \Xi_p^{n-k}$:

$$\partial_k[\xi_p^{n-k}] = \sum_{q \in \mathrm{Crit}_{k-1}(Y)} \left(\int_{W^s(q) \cap L_p} \xi_p^{n-k} \right) \cdot [W^u(q)] \in C_{k-1}^{\overline{q}}(X, Y, F^*).$$

The complex $(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet)$ is well defined, i.e., $\partial_\bullet^2 = 0$.

The trajectories of the flow Φ induced by $-Y$ can leave the stratum only at time $t = -\infty$. This is a crucial feature used to prove the well-definedness of the singular Morse–Thom–Smale complex. The singular Morse–Thom–Smale complex does

depend on the choice of representatives $\Xi_p^\bullet$, $p \in \mathrm{Sing}(X)$, in a mild way; see [54, Section 5.2].

The main result regarding the "singular" Morse–Thom–Smale complex is the following; see [54, Theorem 6.2].

Theorem 3.4. *The complex* $(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet)$ *computes the intersection homology of* X*:*

$$H_\bullet(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet) \simeq IH_\bullet^{\overline{q}}(X, F^*). \tag{3.5}$$

Theorem 3.4 is proved in [54] by constructing a natural embedding of the complex $(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet)$ into the intersection chain complex and showing that this embedding is indeed a quasi-isomorphism. There are different approaches to the smooth Morse–Thom–Smale complex. In [54], we rely on the approach of Laudenbach [40], which gives an interpretation of the smooth Morse–Thom–Smale complex in terms of currents. All constructions in [54] assume that the anti-radial gradient-like vector field is standard; one can remove this latter assumption by a perturbation argument as in [39, Proposition 5.1].

From Theorem 3.4 (and using Poincaré duality), we can deduce Morse inequalities for the intersection homology of the singular space X with perversity $\overline{q}$ and values in F^*. For $k = 0, \ldots, n$, we denote by $Ib_k^{\overline{q}}(X, F^*) := \dim IH_k^{\overline{q}}(X, F^*)$ the intersection Betti numbers of X. We denote by $c_k(f_{\mathrm{sm}}) := \#\mathrm{Crit}_k(f_{\mathrm{sm}})$ the number of critical points of Morse index k of the Morse function f_{sm} and by

$$c_k^{\overline{q}}(f) := \begin{cases} \mathrm{rk}(F) \cdot c_k(f_{\mathrm{sm}}) + \dim IH_k^{\overline{q}}(cL_p, F^*) & \text{for } f \text{ radial,} \\ \mathrm{rk}(F) \cdot c_k(f_{\mathrm{sm}}) + \dim IH_k^{\overline{q}}(cL_p, L_p, F^*) & \text{for } f \text{ anti-radial.} \end{cases} \tag{3.6}$$

where $\mathrm{rk}(F)$ is the rank of the flat vector bundle F.

Corollary 3.5 (Morse inequalities). *For a radial (resp. anti-radial) Morse function on a singular space with isolated singularities, we have*

$$\sum_{i=0}^{k} (-1)^{k-i} c_i^{\overline{q}}(f) \geq \sum_{i=0}^{k} (-1)^{k-i} Ib_i^{\overline{q}}(X, F^*) \quad \textit{for all } 0 \leq k < n,$$

$$\sum_{i=0}^{n} (-1)^i c_i^{\overline{q}}(f) = \sum_{i=0}^{n} (-1)^i Ib_i^{\overline{q}}(X, F^*).$$

The above Morse inequalities have long been known and can also be deduced using Goresky and MacPherson's deformation lemmas in the framework of stratified Morse theory [32, Part I]. For the case of singular homology – for a topologically normal space, this corresponds to intersection homology with top perversity – they are

explicitly mentioned in [33, Section 9]. Let us emphasise again that the main interest of the "singular" Morse–Thom–Smale complex presented here does not lie in giving yet another proof of Morse inequalities but in its application to the comparison results in Sections 4.3 and 5.

Let $(C_\bullet^{\mathrm{sm}}, \partial_\bullet)$ denote the subcomplex of $(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet)$ generated by $\mathrm{Crit}(Y)$. There is an exact sequence of complexes

$$0 \to (C_\bullet^{\mathrm{sm}}, \partial_\bullet) \to (C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet) \to ((C^{\overline{q}}(X, Y, F^*)/C^{\mathrm{sm}})_\bullet, \partial_\bullet) \to 0. \quad (3.7)$$

Note that the boundary operator in the quotient complex vanishes. By the local calculation for intersection homology (2.3) and Poincaré duality on the link manifold L_p, we have

$$H_\bullet((C^{\overline{q}}(X, Y, F^*)/C^{\mathrm{sm}})_\bullet, \partial_\bullet) \simeq \bigoplus_{p \in \mathrm{Sing}(X)} IH_\bullet^{\overline{q}}(cL_p, L_p, F^*). \quad (3.8)$$

For a complex vector space V of dimension 1, we denote by V^{-1} its dual. For a vector space V, we denote by $\det V$ the maximal exterior power of V. For a graded vector space $V^\bullet = \bigoplus_{k=0}^n V^k$, we denote by $\det V^\bullet$ the complex line

$$\det V^\bullet := \bigotimes_{k=0}^n (\det V^k)^{(-1)^k}.$$

From (3.5), (3.7) and (3.8), we get a natural isomorphism

$$\begin{aligned} \det H_\bullet(C_\bullet^{\mathrm{sm}}, \partial_\bullet) \otimes \det \Big(\bigoplus_{p \in \mathrm{Sing}(X)} IH_\bullet^{\overline{q}}(cL_p, L_p, F^*) \Big) \\ \simeq \det H_\bullet(C_\bullet^{\overline{q}}(X, Y, F^*), \partial_\bullet) \simeq \det IH_\bullet^{\overline{q}}(X, F^*) \\ \simeq (\det IH_{\overline{q}}^\bullet(X, F))^{-1}. \end{aligned} \quad (3.9)$$

We denote by $(C_{\overline{q}}^\bullet(X, Y, F), \partial^\bullet)$ the dual complex of the singular Morse–Thom–Smale complex, which computes the intersection cohomology $IH_{\overline{q}}^\bullet(X, F)$.

3.3. Morse–Bott-type complex for stratified pseudomanifolds computing intersection cohomology

In [59], the construction of the Morse–Thom–Smale complex in Section 3.2 has been extended to the case of general compact stratified pseudomanifolds equipped with so-called *stratified anti-radial gradient-like vector fields*. In this more general singular context, we construct a cohomological complex, relying on the generalisation of the smooth Morse–Thom–Smale complex to smooth Morse–Bott functions. Several approaches to the smooth Morse–Bott complex can be found in the literature, most

of them having Floer theoretical applications in mind; see [9] for an overview and a comparison of the different approaches. The paper [59] makes use of the Morse–Bott cochain complex of Austin and Braam [8], defined in terms of the de Rham complexes of the critical submanifolds of the Morse–Bott function and the trajectory spaces between them.

An important property of stratified anti-radial gradient-like vector fields ξ is that they can be lifted to a smooth vector field $\widetilde{\xi}$ on the smooth unfolding $\widetilde{X}$ of X. Singularities of a stratified anti-radial gradient-like vector field ξ are isolated, but the smooth vector field $\widetilde{\xi}$ on $\widetilde{X}$ has Morse–Bott singularities. The flow associated with ξ satisfies control conditions; trajectories do not leave a stratum in finite time, and moreover, in positive infinite time, flow lines can only go from a larger into a smaller stratum. This allows us to adapt the construction in [8] and to construct, for any given perversity $\overline{q}$ in the sense of Goresky and MacPherson [30, 31], a Morse–Bott-type cochain complex $(C^\bullet_{\overline{q}}(X,\xi), \partial^\bullet)$ associated with the stratified anti-radial gradient-like vector field ξ on X, and which computes the intersection cohomology $IH^\bullet_{\overline{q}}(X)$ of X with perversity $\overline{q}$. The main idea of the construction in [59] is to replace the de Rham complex of critical submanifolds of $\widetilde{\xi}$ in the construction of the smooth Morse–Bott cochain complex of Austin and Braam [8] with the truncated complex of liftable intersection differential forms. The latter has been introduced by Brasselet, Hector and Saralegi in [14] and is a subcomplex of the more notorious complex of intersection differential forms of Brylinski, Goresky and MacPherson [21].

4. Witten deformation on singular spaces with conical singularities using radial and anti-radial Morse functions

In [75], Witten, inspired by ideas from quantum field theory, proposed a new analytic proof of the famous Morse inequalities for a smooth Morse function on a smooth manifold. His ideas met a lot of enthusiasm in the mathematical world. In [39], Helffer and Sjöstrand provided rigorous proofs for Witten's method. Generalisations to manifolds with boundary [22, 38, 41, 42], to complete manifolds [1, 76] and to the holomorphic setting [28] have been studied since.

The power of the Witten deformation goes far beyond just proving Morse inequalities: For example, the holomorphic Morse inequalities of Demailly [28] provide an effective tool in complex analysis for the study of vanishing theorems. The Witten deformation combined with local index techniques is a leitmotiv in Bismut's work on analytic torsion in its different flavours. In particular, it plays a prominent role in Bismut and Zhang's extension of the Cheeger–Müller theorem in [11, 12].

In a series of works [49–53, 55], the author has studied generalisations of the Witten deformation to singular spaces with conical and iterated conical singularities using different classes of Morse functions on these spaces.

In Sections 4.1–4.3, we recall from [55] the generalisation of the Witten deformation to spaces with isolated conical singularities equipped with radial (resp. anti-radial) Morse functions. It is one of the key ingredients in the generalisation of the Cheeger–Müller/Bismut–Zhang theorem presented in Section 5. In Sections 4.4 and 4.5, we shortly discuss further generalisations of the Witten deformation for more general singularities and using a different class of Morse functions, respectively.

4.1. The Witten deformation and the spectral gap theorem

The Witten deformation on a smooth compact Riemannian manifold (M, g^{TM}) is a deformation of the de Rham complex $(\Omega^\bullet(M, F), d)$ using a smooth Morse function $f : M \to \mathbb{R}$. We denote by $T \geq 0$ the deformation parameter. The deformation produces a deformed complex $(\Omega^\bullet(M, F), d_T)$, where $d_T := e^{-Tf} d e^{Tf}$, and a deformed Laplacian, the famous Witten Laplacian $\widetilde{\Delta}_T$. The easy part of the Witten deformation, the spectral gap theorem for the Witten Laplacian $\widetilde{\Delta}_T$, states that for sufficiently large $T \gg 0$, the spectrum of $\widetilde{\Delta}_T$ splits into two parts: exponentially small eigenvalues, which are in one-to-one correspondence with the critical points of the Morse function, and the large eigenvalues which – in case that $\nabla_g f$ is standard near critical points of f – grow at least linearly in T. An analytic proof of the smooth Morse inequalities can be easily deduced from the spectral gap theorem.

Let us now come to our singular setting. For a space with isolated conical singularities (X, g^{TX}), the main idea for the generalisation of the Witten deformation is to deform the complex of L^2-forms (instead of the de Rham complex) using an appropriate class of Morse functions. Here, we use radial (resp. anti-radial) Morse functions $f : X \to \mathbb{R}$; see Definition 3.2. As a byproduct, we get an analytic proof of the Morse inequalities for intersection (co-)homology with middle perversities $\overline{q} \in \{\overline{m}, \overline{n}\}$ in Corollary 3.5.

Let us explain this in more technical terms now: We deform the de Rham complex of compactly supported forms $(\Omega_c^\bullet(X, F), d_c, \langle\ ,\ \rangle)$ by deforming the differential d_c via

$$d_{T,c} := e^{-Tf} d_c e^{Tf}.$$

The complex $(\Omega_c^\bullet(X, F), d_{T,c}, \langle\ ,\ \rangle)$ admits a maximal and a minimal extension into a Hilbert complex $(\widetilde{\mathcal{C}}^\bullet_{T,\overline{m}/\overline{n}}, d_{T,\overline{m}/\overline{n}}, \langle\ ,\ \rangle) := (\widetilde{\mathcal{C}}^\bullet_{T,\max/\min}, d_{T,\max/\min}, \langle\ ,\ \rangle)$, where $d_{T,\overline{m}} := d_{T,\max}$ (resp. $d_{T,\overline{n}} := d_{T,\min}$) denotes the maximal (resp. minimal) extension of $d_{T,c}$ in $\mathrm{L}^2(\Lambda(T^*X) \otimes F)$. The deformed Hilbert complex $(\widetilde{\mathcal{C}}^\bullet_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$ still computes the L^2-cohomology,

$$H^\bullet((\widetilde{\mathcal{C}}^\bullet_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)) \simeq H^\bullet_{(2),\overline{q}}(X, F), \quad \overline{q} \in \{\overline{m}, \overline{n}\}. \tag{4.1}$$

We denote by $\delta_{T,c}$ the (formal) adjoint of $d_{T,c}$ with respect to the L^2-inner product $\langle\ ,\ \rangle$; it admits a minimal (resp. maximal) extension $\delta_{T,\overline{m}} := \delta_{T,\min}$ (resp. $\delta_{T,\overline{n}} := \delta_{T,\max}$).

The Witten Laplacian

$$\widetilde{\Delta}_T^{\overline{q}} := (d_{T,\overline{q}} + \delta_{T,\overline{q}})^2 \tag{4.2}$$

is the second-order operator associated with the Hilbert complex $(\widetilde{\mathcal{C}}^{\bullet}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$. It follows from the Hodge theorem for Hilbert complexes, see [16, Lemma 2.2 and Corollary 2.5], that

$$\ker(\widetilde{\Delta}_T^{\overline{q}}) \simeq H^{\bullet}(\widetilde{\mathcal{C}}^{\bullet}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle). \tag{4.3}$$

We denote by $L_{\nabla f}$ the Lie derivative in direction of ∇f and by $L^*_{\nabla f}$ its adjoint with respect to the L^2-inner product $\langle\ ,\ \rangle$. The operator $L_{\nabla f} + L^*_{\nabla f}$ acting on $L^2(\Lambda(T^*X) \otimes F)$ is a bounded operator of order 0. This follows from the smooth theory combined with the local model near $\mathrm{Sing}(X)$; see Section 4.2 for more details.

We have, see [11, Proposition 5.5], [55, Propositions 5.1],

$$\widetilde{\Delta}_T^{\overline{q}} = \Delta^{\overline{q}} + T(L_{\nabla f} + L^*_{\nabla f}) + T^2|\nabla f|^2, \quad \mathrm{dom}(\widetilde{\Delta}_T^{\overline{q}}) = \mathrm{dom}(\Delta^{\overline{q}}). \tag{4.4}$$

Let $(\widetilde{\mathcal{S}}^{\bullet}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$ denote the subcomplex of $(\widetilde{\mathcal{C}}^{\bullet}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$ generated by all eigenforms of the Witten Laplacian $\widetilde{\Delta}_T^{\overline{q}}$ to eigenvalues in the interval $[0, 1]$. The following spectral gap theorem for the Witten Laplacian $\widetilde{\Delta}_T^{\overline{q}}$ has been proved in [55, Theorem I], generalising Witten's smooth spectral gap theorem.

Theorem 4.1 (Spectral gap theorem). *Let (X, g^{TX}) be a space with isolated conical singularities equipped with a Hermitian flat vector bundle (F, ∇^F, g^F) as in Section 2.2. Let $\overline{q} \in \{\overline{m}, \overline{n}\}$ be one of the two middle perversities. Let $f : X \to \mathbb{R}$ be a radial (resp. an anti-radial) Morse function on X such that $\nabla_g f$ is standard near $\mathrm{Crit}(f)$, i.e., condition (b) in Section 3.1 holds. We assume, moreover, that the Hermitian metric g^F is flat near $\mathrm{Crit}(f_{\mathrm{sm}})$. Then:*

(a) *There exist constants $C_1, C_2, C_3 > 0$ and $T_0 > 0$ such that for $T \geq T_0$,*

$$\mathrm{Spec}(\widetilde{\Delta}_T^{\overline{q}}) \cap (C_1 e^{-C_2 T}, C_3 T) = \emptyset.$$

(b) *Moreover, for $k = 0, \ldots, n$, and $T \geq T_0$,*

$$\dim \widetilde{\mathcal{S}}^k_{T,\overline{q}} = c_k^{\overline{q}}(f),$$

where $c_k^{\overline{q}}(f)$ is as defined in (3.6).

Combining the spectral gap theorem, (4.1), the deformed Hodge theorem (4.3) and the Cheeger–Goresky–MacPherson theorem (2.7), we get an analytic proof for

the Morse inequalities for intersection (co-)homology with lower and upper middle perversities in Corollary 3.5.

There is a second, equivalent way of describing the Witten deformation: The de Rham complex $(\Omega_c^\bullet(X,F), d_c, \langle\ ,\ \rangle)$ can also be deformed by deforming the L^2-inner product $\langle\ ,\ \rangle$ via

$$\langle \alpha, \beta \rangle_T := \int_X \langle \alpha, \beta \rangle_{\Lambda(T^*X)\otimes F}(x) e^{-2Tf(x)} d\,\mathrm{vol}_X(x);$$

here, $d\,\mathrm{vol}_X$ denotes the Riemannian volume form on (X, g^{TX}).

The deformed complex $(\Omega_c^\bullet(X,F), d_c, \langle\ ,\ \rangle_T)$ also admits a maximal and minimal extension into a Hilbert complex $(\mathcal{C}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T)$, where $d_{\overline{m}} = d_{\max}$ and $d_{\overline{n}} = d_{\min}$. We denote by $\delta'_{T,c}$ the (formal) adjoint of d_c with respect to the twisted L^2-inner product $\langle\ ,\ \rangle_T$ and by $\delta'_{T,\overline{m}}$ (resp. $\delta'_{T,\overline{n}}$) its minimal (resp. its maximal) extension. We denote by $\Delta_T^{\overline{q}} := (d_{\overline{q}} + \delta'_{T,\overline{q}})^2$ the second-order self-adjoint operator associated with the Hilbert complex $(\mathcal{C}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T)$.

The two descriptions of the Witten deformation are equivalent; more precisely, we have natural isomorphisms of Hilbert complexes

$$e^{Tf} : (\widetilde{\mathcal{C}}^\bullet_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle) \longrightarrow (\mathcal{C}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T), \quad \overline{q} \in \{\overline{m}, \overline{n}\}. \tag{4.5}$$

We denote by $(\mathcal{S}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T)$ the subcomplex of $(\mathcal{C}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T)$ generated by eigenforms of $\Delta_T^{\overline{q}}$ to eigenvalues in $[0,1]$.

Note that the Hodge star operator $*^F$ associated with the metrics g^{TX}, g^F induces an isomorphism of the Hilbert complex associated with the Morse function f, the flat vector bundle F and the perversity $\overline{q}$, and the dual of the Hilbert complex associated with the Morse function $-f$, the flat vector bundle $F^* \otimes o(TX)$ and the dual perversity.

4.2. The model Witten Laplacian

For this section, we restrict our considerations to the case where $f : X \to \mathbb{R}$ is an anti-radial Morse function; the case of a radial Morse function can be treated analogously.

At the heart of the proof of the spectral gap theorem lies the phenomenon of "concentration near critical points of the Morse function f", $\mathrm{Crit}(f) = \mathrm{Crit}(f_{\mathrm{sm}}) \cup \mathrm{Sing}(X)$. This allows us to reduce the problem to a study of the local model situation and the model Witten Laplacian near $\mathrm{Crit}(f)$. In this section, we explain the local model situation near $\mathrm{Crit}(f)$: We will see, in particular, that indeed the spectral gap theorem holds in the model situation! Let us mention, moreover, that the model Witten Laplacian on the infinite cone (more precisely its analytic torsion) is a crucial ingredient in the definition of the Bismut–Zhang torsion (see Definition 5.4), the "topological" side in the Cheeger–Müller/Bismut–Zhang theorem.

For sufficiently small $0 < \varepsilon < \delta$, we identify a neighbourhood $B_\varepsilon(p)$ of $p \in \mathrm{Crit}(f_{\mathrm{sm}})$ (resp. of $p \in \mathrm{Sing}(X)$) with a neighbourhood $B_\varepsilon(0) \subset \mathbb{R}^n$ (resp. $c_\varepsilon L_p \subset cL_p$).

Using this identification, the action of the Witten Laplacian $\widetilde{\Delta}_T^{\overline{q}}$, $T > 0$, on forms with support in a neighbourhood of $p \in \mathrm{Crit}(f)$ can be identified with the action of the model Witten Laplacian, which we now recall.

Let $p \in \mathrm{Crit}_k(f_{\mathrm{sm}})$ be a smooth critical point. By Morse's lemma, we have Morse coordinates such that $f = f(p) + \frac{1}{2}(-x_1^2 - \cdots - x_k^2 + x_{k+1}^2 + \cdots + x_n^2)$. We have assumed that the Riemannian metric g^{TX} is Euclidean in the Morse coordinates near $\mathrm{Crit}(f_{\mathrm{sm}})$. Recall that, by our assumption, the Hermitian metric g^F is also flat near p. We trivialise the flat bundle F over $B_\varepsilon(p)$ by using the connection ∇^F, so that the fibres of $F_{|B_\varepsilon(p)}$ are identified with F_p. For a multi-index $I = (i_1, \ldots, i_l)$, set $dx_I := dx_{i_1} \wedge \cdots \wedge dx_{i_l}$. Denote by N^+ (resp. by N^-) the following number operators:

$$N^+ dx_I := \#\{s \mid k+1 \le i_s \le n\} \cdot dx_I, \quad N^- dx_I := (l - N^+) \cdot dx_I.$$

The model Witten Laplacian $\mathbf{\Delta}_T^p$ is an operator acting on sections of $\Lambda((\mathbb{R}^n)^*) \otimes F_p$ over $\mathbb{R}^n$, which is closely related to the quantum harmonic oscillator (see [39,75], [11, Proposition 8.2]):

$$\mathbf{\Delta}_T^p := \sum_{j=1}^n \left(-\frac{\partial^2}{\partial x_j^2} + T^2 x_j^2 \right) + T(2k - n) + 2T(N^+ - N^-). \tag{4.6}$$

Note that we get (4.6) by evaluating formula (4.4) in our model situation.

Set

$$r := \left(\sum_{i=1}^n x_i^2 \right)^{1/2} \quad \text{and} \quad \varphi_{p,T} := \left(\frac{T}{\pi} \right)^{n/4} e^{-Tr^2/2} dx_1 \wedge \cdots \wedge dx_k.$$

We have, for $T > 0$ (see [39, 75], [11, Proposition 8.3]),

$$\mathrm{Spec}(\mathbf{\Delta}_T^p) = 2T\mathbb{N}_0 \quad \text{and} \quad \ker \mathbf{\Delta}_T^p = \ker \mathbf{\Delta}_T^{p,(k)} = \langle \varphi_{p,T} \rangle \otimes F_p.$$

Let $\chi_p : \mathbb{R}^n \to \mathbb{R}$ be a cut-off function with $\mathrm{supp}\, \chi_p \subset B_\varepsilon(0)$ and $\chi_{p|B_{\varepsilon/2}(0)} \equiv 1$. Set $\omega_{p,T} := \chi_p \varphi_{p,T}$. The forms $\omega_{p,T} \otimes h$, $h \in F_p$, can be seen as elements of $\Omega^\bullet(X, F)$ supported in a neighbourhood of p. For $i = 0, \ldots, n$, define

$$E_p^i(T) := \begin{cases} \langle \omega_{p,T,l} := \omega_{p,T} \otimes h_l \mid h_l \text{ ON basis of } F_p \rangle & \text{for } i = k, \\ 0 & \text{for } i \neq k. \end{cases} \tag{4.7}$$

Let now $p \in \mathrm{Sing}(X)$ and $T \ge 0$. The model Witten Laplacian $\mathbf{\Delta}_T^{p,\overline{q}}$ is an operator acting on forms on the (punctured) infinite cone Z_p and depends on the perversity

$\overline{q} \in \{\overline{m}, \overline{n}\}$. We denote by N the number operator acting by multiplication with the form degree. We denote by $\mathbf{\Delta}_T^p$ the following operator acting on compactly supported forms on Z_p with values in the bundle F_{Z_p}:

$$\mathbf{\Delta}_T^p := \mathbf{\Delta}^p + T(n - 2N) + T^2 r^2; \tag{4.8}$$

here, $\mathbf{\Delta}^p$ denotes the Laplacian acting on compactly supported forms on Z_p; note again that we get (4.8) by using formula (4.4) for the infinite cone Z_p equipped with the model function $f^p = \frac{-r^2}{2}$. The reader is invited to compare (4.8) with (4.6) in the case of Morse index $k = n$.

The model Witten Laplacian $\mathbf{\Delta}_T^{p,\overline{q}}, \overline{q} \in \{\overline{m}, \overline{n}\}$, is the self-adjoint closed extension of (4.8), with boundary conditions at $r = 0$ inherited from the boundary conditions near $p \in \mathrm{Sing}(X)$ of the Witten Laplacian $\widetilde{\Delta}_T^{\overline{q}}$. By completeness of the infinite cone, at $r = \infty$, we do not need to specify boundary conditions. For $T > 0$, the operator $\mathbf{\Delta}_T^{p,\overline{q}}$ is a discrete operator. Using Cheeger's separation of variables technique, one can compute its spectrum and eigenforms explicitly in terms of the spectral data of the transversal Laplacian on the link. We denote by $b^i(L_p, F_{L_p}) := \dim H^i(L_p, F_{L_p})$, $i = 0, \ldots, n-1$, the Betti numbers for the cohomology of the link L_p. For $i \geq n - \overline{q}$, let

$$\{\eta_{p,l}^{i-1} \mid l = 1, \ldots, b^{i-1}(L_p, F_{L_p})\} \subset \mathcal{H}^{i-1}(L_p, F_{L_p}) \simeq H^{i-1}(L_p, F_{L_p})$$

be an orthonormal basis of harmonic forms on the link manifold, dual to the basis $\{[\xi_{p,l}^{n-i}] \mid \xi_{p,l}^{n-i} \in \Xi_p^{n-i}\}$ of $H^{n-i}(L_p, F_{L_p}^* \otimes o(TL_p)) \simeq \mathcal{H}^{n-i}(L_p, F_{L_p}^* \otimes o(TL_p))$ in (3.4). For $l = 1, \ldots, b^{i-1}(L_p, F_{L_p})$, we define

$$\varphi_{p,T,l}^i := \left(\frac{T}{\pi}\right)^{i/2-n/4} \sqrt{\mathrm{vol}(S^{2i-n-1})} \cdot e^{-Tr^2/2} r^{2i-n-1} \eta_{p,l}^{i-1} dr \in \Omega^i(Z_p, F).$$

One has, for $T > 0$ (see [55, Theorem 4.2]),

$$\ker \mathbf{\Delta}_T^{p,\overline{q},(i)} = \begin{cases} \langle \varphi_{p,T,l}^i \mid l = 1, \ldots, b^{i-1}(L_p, F_{L_p}) \rangle & \text{for } i \geq n - \overline{q}, \\ 0 & \text{else.} \end{cases} \tag{4.9}$$

Note that $\{\varphi_{p,T,l}^i \mid l = 1, \ldots, b^{i-1}(L_p, F_{L_p})\}$ is an orthonormal basis for $\ker \mathbf{\Delta}_T^{p,\overline{q},(i)}$, $i \geq n - \overline{q}$. From the local calculation for intersection cohomology [31, Section 2.4], recalled in (2.3) and (4.9), we have

$$\ker \mathbf{\Delta}_T^{p,\overline{q}} \simeq IH_{\overline{q}}^\bullet(cL_p, L_p, F). \tag{4.10}$$

Let χ_p be a cut-off function $\chi_p : cL_p \to \mathbb{R}$ with $\mathrm{supp}\, \chi_p \subset c_\varepsilon L_p$, $\chi_{p|c_{\varepsilon/2}L_p} \equiv 1$. Then

$$\omega_{p,T,l}^i := \chi_p \varphi_{p,T,l}^i \in \Omega^i(X, F).$$

For $i = 0, \dots, n$, define

$$E^i_{p,\overline{q}}(T) := \begin{cases} \langle \omega^i_{p,T,l} \mid l = 1, \dots, b^{i-1}(L_p, F_{L_p}) \rangle & \text{for } i \geq n - \overline{q}, \\ 0 & \text{else.} \end{cases} \tag{4.11}$$

4.3. Comparison between the Witten complex and the "singular" Morse–Thom–Smale complex

The second step of Witten's programme, the hard part, aims at comparing the Witten complex of eigenforms to small eigenvalues of the Witten Laplacian with the Morse–Thom–Smale complex: Witten suggested that, in case that the gradient vector field of the Morse function satisfies the Morse–Smale transversality condition, in some appropriate sense, the Witten complex should "converge" to the (dual) Morse–Thom–Smale complex as $T \to \infty$. Let us emphasise that the second part of Witten's programme is more delicate than the first part. It has also been proved by Helffer and Sjöstrand in [39] using semi-classical analysis, but here one needs more refined estimates on the decay of eigenforms of the Witten Laplacian than in the first part. A second proof has been given by Bismut and Zhang in [12] using the interpretation due to Laudenbach [40] of the Morse–Thom–Smale complex as a complex of currents.

Let us now come back to our singular setting: Here, we will compare the Witten complex with the "singular" Morse–Thom–Smale complex constructed in Section 3.2. In this section, $f : X \to \mathbb{R}$ always denotes an anti-radial Morse function such that $\nabla_g f$ is a standard anti-radial gradient-like vector field. We moreover assume that the Hermitian metric g^F is flat near $\mathrm{Crit}(f_{\mathrm{sm}})$.

For $i = 0, \dots, n$ and $T > 0$, using (4.7) and (4.11), we define the vector space

$$E^i_{\overline{q}}(T) := \Bigg(\bigoplus_{p \in \mathrm{Crit}(f_{\mathrm{sm}})} E^i_p(T) \Bigg) \oplus \Bigg(\bigoplus_{p \in \mathrm{Sing}(X)} E^i_{p,\overline{q}}(T) \Bigg) \subset \Omega^i(X, F).$$

We define a map of graded vector spaces $J_{T,\overline{q}} : C^\bullet_{\overline{q}}(X, \nabla_g f, F) \to E^\bullet_{\overline{q}}(T)$ as follows:

$$\begin{aligned} J_{T,\overline{q},i} : C^i_{\overline{q}}(X, \nabla_g f, F) &\longrightarrow E^i_{\overline{q}}(T), \\ [W^u(p)]^* \otimes h &\mapsto \omega_{p,T} \otimes h \quad \text{for } p \in \mathrm{Crit}_i(f_{\mathrm{sm}}),\ h \in F_p, \\ [\xi^{n-i}_{p,l}]^* &\mapsto \omega^i_{p,T,l} \qquad \text{for } p \in \mathrm{Sing}(X),\ i \geq n - \overline{q}, \\ & \qquad\qquad l \in \{1, \dots, b^{i-1}(L_p, F_{L_p})\}. \end{aligned}$$

We denote by $\widetilde{P}^{[0,1]}_{T,\overline{q}} : \widetilde{\mathcal{C}}^\bullet_{T,\overline{q}} \to \widetilde{\mathcal{S}}^\bullet_{T,\overline{q}}$ the orthogonal projection. The isomorphism of complexes (4.5) induces an isomorphism of complexes

$$e^{Tf} : (\widetilde{\mathcal{S}}^\bullet_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle) \longrightarrow (\mathcal{S}^\bullet_{T,\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle_T).$$

We define the map of graded vector spaces $e_{T,\overline{q}}$ as the following composition:

$$e_{T,\overline{q}} : C^{\bullet}_{\overline{q}}(X, \nabla_g f, F) \xrightarrow{J_{T,\overline{q}}} E^{\bullet}_{\overline{q}}(T) \xrightarrow{\widetilde{P}^{[0,1]}_{T,\overline{q}}} \widetilde{\mathcal{S}}^{\bullet}_{T,\overline{q}} \xrightarrow{e^{Tf}} \mathcal{S}^{\bullet}_{T,\overline{q}}.$$

Using the embedding $(C^{\overline{q}}_{\bullet}(X, \nabla_g f, F^*), \partial_{\bullet}) \to (IC^{\overline{q}}_{\bullet}(X, F^*), \partial_{\bullet})$ into the intersection chain complex, constructed in [54, Section 6], the integration of L^2-forms on intersection chains does induce a map of chain complexes $\int_{T,\overline{q}}$, $T \geq 0$ (see [55, Proposition 6.13] and [54, Section 6.3]):

$$\int_{T,\overline{q}} : (\mathcal{S}^{\bullet}_{T,\overline{q}}, d, \langle\ ,\ \rangle_T) \longrightarrow (C^{\bullet}_{\overline{q}}(X, \nabla_g f, F), \partial^{\bullet}).$$

The number operator acting on the graded vector space $C^{\bullet}_{\overline{q}}(X, \nabla_g f, F)$ by multiplication with the degree is denoted by N. We denote by $\mathcal{F}$ the operator which, for $p \in \mathrm{Crit}(f_{\mathrm{sm}})$, does act on $[W^u(p)]^* \otimes F_p$ by multiplication with $f(p)$; for $p \in \mathrm{Sing}(X)$, the operator $\mathcal{F}$ acts by multiplication with $f(p)$ on $(\mathrm{span}(\Xi^{\bullet}_p))^*$. The next theorem, proved in [55, Theorem II], is the analogue of the comparison theorem between Witten and Morse–Thom–Smale complex for a smooth compact manifold proved in [12, Theorem 6.11].

Theorem 4.2. *Let (X, g^{TX}) be a space with isolated conical singularities equipped with a Hermitian flat vector bundle (F, ∇^F, g^F) as in Section 2.2. Let $\overline{q} \in \{\overline{m}, \overline{n}\}$ be one of the two middle perversities. Let $f : X \to \mathbb{R}$ be an anti-radial Morse function on X such that $\nabla_g f$ is a standard anti-radial gradient-like vector field. We assume, moreover, that the Hermitian metric g^F is flat near* $\mathrm{Crit}(f_{\mathrm{sm}})$. *Then, as $T \to \infty$,*

$$\int_{T,\overline{q}} \circ\, e_{T,\overline{q}} = e^{T\mathcal{F}} \left(\frac{\pi}{T}\right)^{N/2-n/4} (\mathcal{A}_{\overline{q}} + \mathcal{O}(e^{-cT})),$$

where $\mathcal{A}_{\overline{q}}$ is the diagonal matrix with entries

$$\mathcal{A}_{\sigma} = \begin{cases} 1 & \text{for } \sigma = [W^u(p)]^* \otimes h,\ p \in \mathrm{Crit}(f_{\mathrm{sm}}),\ h \in F_p, \\ \dfrac{1}{\sqrt{\mathrm{vol}(S^{2i-n-1})}} & \text{for } \sigma = [\xi^{n-i}_{p,l}]^*,\ p \in \mathrm{Sing}(X),\ \xi^{n-i}_{p,l} \in \Xi^{n-i}_p. \end{cases}$$

In particular, for large T, the map $\int_{T,\overline{q}}$ is an isomorphism of complexes.

Using Theorem 4.2 and the fact that the integration map $\int_{T,\overline{q}}$ is a chain map, we get the following comparison between the boundary operators in the Witten and the Morse–Thom–Smale complex, which generalises [12, Theorem 6.12].

Theorem 4.3. *With the assumptions and notation as in Theorem 4.2, we get*

$$e^{-1}_{T,\overline{q}} \circ d \circ e_{T,\overline{q}} = \sqrt{\frac{T}{\pi}} (\mathcal{A}^{-1}_{\overline{q}} + O(e^{-cT})) e^{-T\mathcal{F}} \partial^{\bullet} e^{T\mathcal{F}} (\mathcal{A}_{\overline{q}} + O(e^{-cT})).$$

4.4. Witten deformation on Thom–Mather stratified pseudomanifolds using anti-radial and radial Morse functions

The easy part of the Witten deformation has been generalised in [55] in the more general context of spaces with "iterated" conical singularities. Let X be an n-dimensional compact stratified pseudomanifold; see, e.g., [55] for the exact definition. For $i = 0, \dots, n$, we denote by $X_{(i)}$ the union of all strata of X of dimension i. We denote by $\mathrm{Sing}(X) := \bigcup_{i=0}^{n-1} X_{(i)}$ the singular set.

We equip the top stratum $X_{\mathrm{sm}} := X_{(n)} = X \setminus \mathrm{Sing}(X)$ with a compatible Riemannian metric. The notion of a compatible Riemannian metric on a stratified pseudomanifold X was first studied in [66, Definition 4.1] and also appeared in [15, Remark 3.3]; in [2, Definition 2.3], such metrics are called rigid edge metrics. Compatible metrics are defined inductively as follows: If $\mathrm{depth}(X) = 0$, then X is a compact smooth manifold, and a compatible Riemannian metric g^{TX} is just a smooth Riemannian metric. Let now $\mathrm{depth}(X) > 0$. A compatible Riemannian metric on X is a Riemannian metric g^{TX} on the top stratum $X_{\mathrm{sm}} = X \setminus \mathrm{Sing}(X)$ such that for each point x of the stratum $X_{(i)}$ of dimension $0 \leq i < n$, there exists a neighbourhood $U(x) \subset X$ such that in an admissible chart $\varphi : U(x) \simeq \mathbb{R}^i \times cL_x$, we have

$$g^{TX}_{|U(x)\cap X_{\mathrm{sm}}} = g^{T\mathbb{R}^i}(z) + dr^2 + r^2 g^{TL_x}(z, y), \quad (z, (r, y)) \in \mathbb{R}^i \times cL_x, \quad (4.12)$$

where r is the radial coordinate, $g^{T\mathbb{R}^i}$ is a Riemannian metric on $\mathbb{R}^i$ and g^{TL_x} is a compatible Riemannian metric on the top stratum $L_{x,\mathrm{sm}}$ of the link L_x – which is a stratified pseudomanifold of lower depth than X; the metric g^{TL_x} depends smoothly on the coordinate z in $\mathbb{R}^i$.

The following definition generalises the definitions in Section 3.1 to stratified pseudomanifolds.

Definition 4.4. Let $f : X \to \mathbb{R}$ be a continuous function on X, and let g^{TX} be a compatible Riemannian metric on the compact stratified pseudomanifold X. We call the pair (f, g^{TX}) a standard *radial (resp. anti-radial) Morse pair* on the stratified space X if the following conditions are satisfied:

(1) For each i, the restriction of $f_{|X_{(i)}}$ to $X_{(i)}$ is a smooth Morse function. We denote by $\mathrm{Crit}_k(f_{|X_{(i)}})$ the set of its critical points of Morse index k.

(2) For each critical point $p \in \mathrm{Crit}_k(f_{|X_{(i)}})$, there exists an open neighbourhood $U(p)$ of p in X as well as local coordinates $z_1, \dots, z_i$ on $X_{(i)} \cap U(p)$ such that one can write

$$f_{|U(p)} = f^{\pm} = f(p) + \frac{1}{2}(-z_1^2 - \cdots - z_k^2 + z_{k+1}^2 + \cdots + z_i^2 \pm r^2),$$

and

$$g^{TX}_{|U(p)} = dz_1^2 + \cdots + dz_i^2 + dr^2 + r^2 g^{TL_p}, \quad (4.13)$$

where r is the radial coordinate, g^{TL_p} is a fixed compatible Riemannian metric on the top stratum of the link L_p. (Note that in contrast to (4.12), here g^{TL_p} does not depend on the coordinates in the stratum $z_1, \dots, z_i$).

(3) The gradient $\nabla f := \nabla_g f_{|X_{\mathrm{sm}}}$ is bounded on $X_{\mathrm{sm}} = X \setminus \mathrm{Sing}(X)$. Moreover, there exists $C > 0$ such that

$$|\nabla f(x)| > C \quad \text{for all } x \in X_{\mathrm{sm}} \setminus \Big(\bigcup_{p \in \mathrm{Crit}(f)} U(p) \Big),$$

where $\mathrm{Crit}(f) = \bigcup_{i=0}^{n} \mathrm{Crit}(f_{|X_{(i)}})$.

(4) The operator $\mathcal{L}_{\nabla f} + \mathcal{L}^*_{\nabla f}$ extends to a bounded operator on the space $\mathrm{L}^2(\Lambda(T^*X) \otimes F)$.

A compatible Riemannian metric on a stratified pseudomanifold X can always be perturbed to achieve the product form (4.13) in a small neighbourhood of finitely many points. For a space X with isolated singularities, conditions (3) and (4) of Definition 4.4 follow already from (1) and (2), and indeed, in this case, we recover the notions defined already in Section 3.1. Note that (3) does not depend on the choice of compatible Riemannian metric. Condition (4) can be achieved by appropriate bounds on the second-order derivatives of f.

For an n-dimensional compact stratified pseudomanifold X, whose top stratum is equipped with a compatible Riemannian metric g^{TX} and a Hermitian flat vector bundle (F, ∇^F, g^F) with uniformly bounded metric g^F, we can also study the Hilbert complexes of L^2-forms $(\mathcal{C}^\bullet_{\overline{m}/\overline{n}}, d_{\overline{m}/\overline{n}}, \langle\ ,\ \rangle) := (\mathcal{C}^\bullet_{\max/\min}, d_{\max/\min}, \langle\ ,\ \rangle)$, with associated Laplacian $\Delta^{\overline{m}/\overline{n}}$ and which compute the L^2-cohomology $H^\bullet_{(2),\overline{m}/\overline{n}}(X, F)$. The Cheeger–Goresky–MacPherson theorem (2.7) and (2.8) also holds in this context. (See [24, Theorem 6.1] and [66, Theorem 4.14]; see also [10, 13] for an overview on de Rham-type theorems on singular spaces.)

The complex of L^2-forms $(\mathcal{C}^\bullet_{\overline{q}}, d_{\overline{q}}, \langle\ ,\ \rangle)$, $\overline{q} \in \{\overline{m}, \overline{n}\}$, can be deformed using radial/anti-radial Morse functions into a deformed complex $(\widetilde{\mathcal{C}}^\bullet_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$. The following spectral gap theorem holds for the associated Witten Laplacian $\widetilde{\Delta}^{\overline{q}}_T$; see [55, Theorem I].

Theorem 4.5 (Spectral gap theorem). *Let X be an n-dimensional compact stratified pseudomanifold. Let (F, ∇^F, g^F) be a Hermitian flat vector bundle on the top stratum, with uniformly bounded metric g^F. Let $\overline{q} \in \{\overline{m}, \overline{n}\}$ be one of the two middle perversities. Let (f, g^{TX}) be a standard radial (resp. anti-radial) Morse pair on X. We assume that near $x \in \mathrm{Crit}(f_{|X_{\mathrm{sm}}})$, the metric g^F is flat, and that moreover, near $x \in \mathrm{Crit}(f_{|\mathrm{Sing}(X)})$, the vector bundle (F, ∇^F, g^F) is the extension of a flat vector*

bundle on the link L_x. Moreover, we assume that, for $T > 0$,

$$\mathrm{dom}(\widetilde{\Delta}_T^{\overline{q}}) = \mathrm{dom}(\Delta^{\overline{q}}). \tag{4.14}$$

(a) *Then there exist constants $C_1, C_2, C_3 > 0$ and $T_0 > 0$ such that for any $T \geq T_0$ and $i = 0, \ldots, n$,*

$$\mathrm{Spec}(\widetilde{\Delta}_T^{\overline{q}}) \cap (C_1 e^{-C_2 T}, C_3 T) = \emptyset.$$

(b) *Let us denote by $(\widetilde{\mathcal{S}}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$ the subcomplex of $(\widetilde{\mathcal{C}}_{T,\overline{q}}, d_{T,\overline{q}}, \langle\ ,\ \rangle)$ generated by all eigenforms of the Witten Laplacian $\widetilde{\Delta}_T^{\overline{q}}$ to eigenvalues in $[0, 1]$. Then, for $T \geq T_0$,*

$$\dim \widetilde{\mathcal{S}}^i_{T,\overline{q}} = \mathrm{rk}(F) c_i(f_{|X \setminus \mathrm{Sing}(X)}) + \sum_{p \in \mathrm{Crit}(f) \cap \mathrm{Sing}(X)} m^i_{p,\overline{q}} =: c_i^{\overline{q}}(f),$$

where the contribution of a critical point $p \in \mathrm{Sing}(X)$ of index k is given by

$$m^i_{p,\overline{q}} := \dim IH^{i-k}_{\overline{q}}(cL_p, F) \quad (\textit{resp. } m^i_{p,\overline{q}} := \dim IH^{i-k}_{\overline{q}}(cL_p, L_p, F))$$

if $f = f^+$ is radial (resp. $f = f^-$ is anti-radial) near p.

Remark 4.6. For a stratified pseudomanifold satisfying the Witt condition, one can – after possibly rescaling the compatible Riemannian metric – assume that the metric is *adapted*, i.e., that the associated de Rham operator is essentially self-adjoint. In this situation, the intersection (co-)homologies with upper middle and lower middle perversities coincide; one also has uniqueness of *ibcs*, $d_{\min} = d_{\max}$. It is mentioned in [55] that, in this situation, the technical condition (4.14) can be removed. Indeed, the technical condition (4.14) can also be removed for non-Witt spaces, with a different proof than in [55].

As a direct consequence of the spectral gap theorem, we get Morse inequalities for the L^2-cohomology of X, or dually – in view of (2.7), which also holds in this more general setting – for the intersection (co-)homology with lower and upper middle perversities. The easy part of the Witten deformation on conical singularities using a similar class of functions has also been studied, in a slightly different manner in [7].

4.5. Witten deformation using Goresky–MacPherson's stratified Morse theory

An important class of Morse functions for singular spaces is stratified Morse functions in the sense of the theory developed by Goresky and MacPherson in [32]. Stratified Morse functions are defined on Whitney stratified subsets of a smooth manifold. The main idea of the definition of stratified Morse functions is that critical

points in a lower stratum are not critical in normal directions. One of the main results in [32] is the stratified Morse inequalities for intersection homology for stratified Morse functions on singular complex (algebraic or analytic) varieties [32, Part II, Theorem 6.12]. In [49,51] both the easy and the hard parts of the Witten deformation for stratified Morse functions on singular complex curves have been generalised, thus providing, in particular, an analytic proof of the famous stratified Morse inequalities of Goresky and MacPherson in the case of (complex) dimension 1. One difficulty in generalising Witten's programme in the context of Goresky and MacPherson's stratified Morse theory to higher dimensions stems from the fact that complex algebraic varieties are rarely conical in case dim > 1, while the development of analytic tools for singular spaces has focused mainly on cone-like singularities. In [53], the easy part of the Witten deformation has been generalised for even-dimensional spaces with (asymptotically) cone-like singularities and so-called *admissible* Morse functions; the non-degeneracy condition for an admissible Morse function at singular points is inspired by stratified Morse theory.

5. Cheeger–Müller/Bismut–Zhang theorem for spaces with isolated conical singularities

In this section, we present the Cheeger–Müller/Bismut–Zhang theorem for spaces with isolated conical singularities, Theorem 1.1, achieved in [56,58]. For this section, the setting is the following: (X, g^{TX}) is a space with isolated conical singularities equipped with a Hermitian flat vector bundle (F, ∇^F, g^F) as in Section 2.2, and $\overline{q} \in \{\overline{m}, \overline{n}\}$ is one of the middle perversities.

Throughout, we adopt the point of view due to Quillen: For a smooth compact manifold, both the analytic (or Ray–Singer) torsion as well as the topological (Reidemeister and Milnor) torsion are seen as metrics on the determinant of the singular cohomology. As mentioned, here we will generalise the approach to torsion due to Bismut and Zhang: They compare the Ray–Singer metric with the Milnor metric defined in [62]; see also [11, Section I (d)]. For unitary, and more generally unimodular, vector bundles, the Milnor metric equals the Reidemeister metric and is therefore a combinatorial and topological invariant. For non-unimodular vector bundles, the Milnor metric depends on choices. Moreover, equality between Ray–Singer and Milnor metrics no longer holds: A third term appears in the Bismut–Zhang comparison theorem.

In the presence of singularities, the Ray–Singer metric and its – yet to be defined – "topological" counterpart are metrics on the determinant of the intersection cohomology with middle perversity $\overline{q} \in \{\overline{m}, \overline{n}\}$. The Cheeger–Müller/Bismut–Zhang

theorem for singular spaces with isolated conical singularities, Theorem 1.1, compares the Ray–Singer metric with the Bismut–Zhang metric, an invariant also introduced in [56, 58]. The Bismut–Zhang metric is a Milnor-like metric with an analytic correction term coming from the singularities of the space. One of the advantages of our approach is that this analytic correction can be identified: It is precisely the analytic torsion of the model Witten Laplacian on the infinite cone! While the analytic torsion of the model Witten Laplacian can be computed explicitly in terms of the spectral data of the transversal Laplacian on the link manifold (see [57, Theorem I]), the result in Theorem 1.1 does not rely on these explicit computations. The answer we give here to Dar's question is a conceptual one!

Moreover, the treatment of the problem of a Cheeger–Müller/Bismut–Zhang theorem in [56, 58] is complete: It incorporates the case of non-unitary vector bundles and moreover also discusses how the three terms in the comparison formula vary when the pair of metrics (g^{TX}, g^F) on TX and F varies, namely, the so-called anomaly formulas for the three terms.

This section is organised as follows: In Sections 5.1 and 5.2, we recall the definitions of the Ray–Singer and the Bismut–Zhang metrics, respectively, for a space with isolated conical singularities. In Section 5.3, we present anomaly formulas for all three terms in the Cheeger–Müller/Bismut–Zhang theorem (Theorem 1.1).

5.1. The Ray–Singer metric on $\det IH^{\bullet}_{\overline{q}}(X, F)$

The analytic (or Ray–Singer) torsion of a smooth compact Riemannian manifold is an important secondary spectral invariant, defined by Ray and Singer [67]; it is a positive real number associated with the spectrum of the Hodge Laplacian. The definition of the Ray–Singer torsion has been generalised to spaces with isolated conical singularities by Dar [27]. In this section, we recall its definition, rephrasing it however – in the spirit of Quillen – as a metric on the determinant line of the intersection cohomology.

Let us first recall that, as in the case of smooth compact manifolds, also for spaces with isolated conical singularities, the Laplacian $\Delta^{\overline{q}}$ is a discrete operator. Moreover, we have the following a priori estimate on the growth behaviour of its eigenvalues (see, e.g., [43, Proposition 1.4.7]): There exists $C > 0$ such that, for sufficiently large $k > k_0$ the k-th eigenvalue λ_k of $\Delta^{\overline{q}}$ satisfies

$$\lambda_k \geq C k^{\frac{1}{l}}, \tag{5.1}$$

where $l > \frac{n}{2}$.

In the following, we denote by Tr_s the supertrace of an operator. We denote by $\Delta^{\overline{q},\perp}$ the restriction of $\Delta^{\overline{q}}$ to $(\ker \Delta^{\overline{q}})^{\perp}$. We denote by N the number operator acting on sections of the bundle $\Lambda(T^*X) \otimes F$ by multiplication with the form degree. For

$s \in \mathbb{C}$, $\Re(s) > \frac{n}{2}$, set

$$\zeta_{\overline{q}}(s) := -\mathrm{Tr}_s[N(\Delta^{\overline{q},\perp})^{-s}],$$

which in view of (5.1) is a well-defined holomorphic function. By [27, Section 4], the function $\zeta_{\overline{q}}$ extends to a meromorphic function on the whole complex plane, which is holomorphic at $s = 0$. The result in [27, Section 4] has been proved in the case of unitary flat vector bundles, but can be adapted to the present more general situation; a more conceptual proof is given in [58].

The L^2-inner product on sections of $\Lambda(T^*X) \otimes F$ restricts to an inner product on the space of L^2 harmonic forms $\mathcal{H}^\bullet_{(2),\overline{q}}(X, F)$. Using the isomorphisms (2.7) and (2.8), we get an induced metric on the line $\det IH^\bullet_{\overline{q}}(X, F)$, which we denote by $|\ |^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)}$.

Definition 5.1. The Ray–Singer metric $\|\ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)}$ on the line $\det IH^\bullet_{\overline{q}}(X, F)$ is defined as

$$\|\ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)} := |\ |^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)} \exp\left(\frac{1}{2}\zeta'_{\overline{q}}(0)\right).$$

In contrast to the smooth situation, where in the case of a unitary flat vector bundle, the Ray–Singer metric $\|\ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F)}$ is a topological invariant – this is precisely the statement of the smooth Cheeger–Müller theorem – in the presence of singularities invariance of the Ray–Singer metric on the choice of the pair of metrics (g^{TX}, g^F) no longer holds true in general. We discuss anomaly formulas for the Ray–Singer metric, i.e., its dependence on the choice of the metrics g^{TX}, g^F on TX, F, in Section 5.3.

5.2. The Bismut–Zhang metric on $\det IH^\bullet_{\overline{q}}(X, F)$

The Bismut–Zhang metric $\|\ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F)}$, associated with the pair of metrics (g^{TX}, g^F) and the anti-radial gradient-like vector field Y, has been defined in [56,58], using the singular Morse–Thom–Smale complex for the perversity $\overline{q} \in \{\overline{m}, \overline{n}\}$ (recalled in Section 3.2) and the model Witten Laplacian $\mathbf{\Delta}_T^{p,\overline{q}}$ on the infinite cone cL_p, $p \in \mathrm{Sing}(X)$ (recalled in Section 4.2). The Bismut–Zhang metric is a metric on $\det IH^\bullet_{\overline{q}}(X, F)$ with Milnor-like contributions from $\mathrm{Crit}(Y)$ and with an analytic correction term for each $p \in \mathrm{Sing}(X)$; this analytic correction is nothing else but the analytic torsion of the model Witten Laplacian $\mathbf{\Delta}_T^{p,\overline{q}}$.

Let us now recall the definition of the Bismut–Zhang metric from [56] in more technical terms: For $T > 0$, we denote by $\mathbf{\Delta}_T^{p,\overline{q},\perp}$ the restriction of the model Witten Laplacian $\mathbf{\Delta}_T^{p,\overline{q}}$ to $(\ker \mathbf{\Delta}_T^{p,\overline{q}})^\perp$. Using the spectral properties of $\mathbf{\Delta}_T^{p,\overline{q}}$, one can show that, for $\Re(s) \gg 0$, the zeta function

$$s \mapsto \zeta_T^{p,\overline{q}}(s) := -\mathrm{Tr}_s[N(\mathbf{\Delta}_T^{p,\overline{q},\perp})^{-s}]$$

is a well-defined holomorphic function. Moreover, by [56, Proposition 4.14], the zeta function $\zeta_T^{p,\overline{q}}$, $T > 0$, extends to a meromorphic function on $\mathbb{C}$, which is holomorphic at $s = 0$.

The L^2-inner product on sections of $\Lambda(T^*Z_p) \otimes F$ restricts to an inner product on $\ker \boldsymbol{\Delta}_T^{p,\overline{q}}$, and via the isomorphism (4.10), we get an induced inner product on $IH_{\overline{q}}^{\bullet}(cL_p, L_p, F)$. We denote the induced metric on the line $\det IH_{\overline{q}}^{\bullet}(cL_p, L_p, F)$ by $|\ |^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F),T}$. We denote by

$$\|\ \|^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F),T} := |\ |^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F),T} \exp\left(\frac{1}{2}(\zeta_T^{p,\overline{q}})'(0)\right)$$

the Ray–Singer metric of the model Witten Laplacian $\boldsymbol{\Delta}_T^{p,\overline{q}}$. Set $\zeta_p^{\overline{q}} := \zeta_1^{p,\overline{q}}$.

Definition 5.2. For $p \in \mathrm{Sing}(X)$, the Ray–Singer metric on the determinant line $\det IH_{\overline{q}}^{\bullet}(cL_p, L_p, F)$ is defined as

$$\begin{aligned}\|\ \|^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F)} &:= \|\ \|^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F),1} \\ &= |\ |^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F),1} \exp\left(\frac{1}{2}(\zeta_p^{\overline{q}})'(0)\right).\end{aligned}$$

Remark 5.3. Using the scaling property of the model Witten Laplacian $\boldsymbol{\Delta}_T^{p,\overline{q}}$, one can prove that Definition 5.2 is independent of the choice of $T > 0$ (here we took $T = 1$); see [56, Proposition 4.15].

The metrics g^{F_p} on the fibre F_p, $p \in \mathrm{Crit}(f_{\mathrm{sm}})$, induce a metric on $(C_{\bullet}^{\mathrm{sm}}, \partial_{\bullet})$. We get an induced metric on $\det H^{\bullet}(\mathrm{Hom}((C_{\bullet}^{\mathrm{sm}}, \partial_{\bullet}), \mathbb{C}))$ (see [62], [11, Section I (d)]).

Definition 5.4. We denote by $\|\ \|^{\mathrm{BZ},Y}_{\det IH_{\overline{q}}^{\bullet}(X,F)}$ the metric on the determinant line $\det IH_{\overline{q}}^{\bullet}(X, F)$ induced via the natural isomorphisms (3.9) from the metric on

$$\det H^{\bullet}(\mathrm{Hom}((C_{\bullet}^{\mathrm{sm}}, \partial_{\bullet}), \mathbb{C}))$$

and the metrics $\|\ \|^{\mathrm{RS}}_{\det IH_{\overline{q}}^{\bullet}(cL_p,L_p,F)}$, $p \in \mathrm{Sing}(X)$. We call this metric $\|\ \|^{\mathrm{BZ},Y}_{\det IH_{\overline{q}}^{\bullet}(X,F)}$ the Bismut–Zhang metric of X which is constructed from the pair of metrics (g^{TX}, g^F) and the anti-radial gradient-like vector field Y.

For (X, g^{TX}) a smooth manifold, i.e., if for $p \in \mathrm{Sing}(X)$, the link manifold L_p is the standard sphere, g^{TX} is Euclidean near p, and g^F is flat near p, then the Bismut–Zhang metric is equal to the Milnor metric defined in [11, Definition 1.9]; see [56, Remark 2.11 (c)]. In case that (X, g^{TX}) is an oriented even-dimensional space with isolated conical singularities and (F, ∇^F, g^F) is unitary, the Bismut–Zhang metric equals the intersection Reidemeister metric defined by Dar [27] and is

hence a combinatorial invariant; see [56, Section 8]. However, in general, in the presence of singularities, even in the case of a unitary flat vector bundle, $\| \ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F)}$ is not a purely combinatorial invariant. Anomaly formulas for the Bismut–Zhang metric will be discussed in Section 5.3.

The definition of the Bismut–Zhang metric has been generalised to the case of stratified pseudomanifolds in [59, Section 6]. Moreover, [59, Theorem II] proves that for spaces with two strata and even-codimensional singular stratum, the Bismut–Zhang metric equals the intersection Reidemeister metric of Dar and is hence a combinatorial invariant. This result is a counterpart of the invariance of analytic torsion for a wedge space proved in [61].

5.3. Anomaly formulas

For a smooth compact manifold equipped with a unitary flat vector bundle, the Ray–Singer metric is a topological invariant. This, as already mentioned, is no longer true in the presence of singularities. Also, the Bismut–Zhang metric is not a topological or combinatorial invariant in general. In this section, we study anomaly formulas for all three terms in the Cheeger–Müller/Bismut–Zhang theorem for isolated conical singularities.

In this section, we make an additional assumption on a pair of metrics (g^{TX}, g^F) on TX, F, namely: For $p \in \mathrm{Sing}(X)$, we assume the following spectral gap condition on the co-closed spectrum $\mathrm{Spec}(\Delta_{L_p,ccl})$ of the transversal Laplacian Δ_{L_p} on the link L_p:

$$\begin{cases} \mathrm{Spec}(\Delta^{(\nu-1)}_{L_p,ccl}) \cap (0,1) = \emptyset & \text{if } n = 2\nu \text{ is even,} \\ (\mathrm{Spec}(\Delta^{(\nu-1)}_{L_p,ccl}) \cap (0,3/4)) & \text{if } n = 2\nu+1 \text{ is odd.} \\ \qquad \cup\, (\mathrm{Spec}(\Delta^{(\nu)}_{L_p,ccl}) \cap (0,3/4)) = \emptyset & \end{cases} \tag{5.2}$$

Condition (5.2) can be achieved by a rescaling of the metric g^{TL_p} into $c^2 g^{TL_p}$ with $c > 0$ sufficiently small.

In the following, let $\mathbb{R} \ni l \to (g^{TX}_l, g^F_l)$ be a family of metrics on TX, F depending smoothly on the parameter l. We assume, moreover, that the metrics g^{TX}_l are conical, i.e., for $p \in \mathrm{Sing}(X)$, there is a family of Riemannian metrics $l \in \mathbb{R} \to g^{TL_p}_l$ such that $g^{TX}_l = dr^2 + r^2 g^{TL_p}_l$ near p. Similarly, we assume that near $p \in \mathrm{Sing}(X)$, the Hermitian metric g^F_l on the flat vector bundle F is of the form explained in Section 2.2. In the following, we use a subscript or superscript l to denote the different notions and operators associated with the pair of metrics (g^{TX}_l, g^F_l).

We denote by $*_l$ (resp. by $*_{p,l}$) the Hodge star operator associated with the metric g^{TX}_l on TX (resp. $g^{TZ_p}_l = dr^2 + r^2 g^{TL_p}_l$ on TZ_p). Moreover, for $p \in \mathrm{Sing}(X)$,

$\overline{q} \in \{\overline{m}, \overline{n}\}$, $t > 0$, we denote by $Q_t^{p,l,\overline{q}}((r,y),(r',y')), (r,y),(r',y') \in Z_p$ the heat kernel of the model operator $\mathbf{\Delta}_l^{p,\overline{q}}$. We define the following Cheeger-type invariants:

$$c_{p,l}^{\overline{q}} := \frac{1}{2}\int_0^\infty \frac{du}{u}\int_{L_p} \mathrm{Tr}_s\left[(g_l^{F_{Z_p}})^{-1}\frac{\partial g_l^{F_{Z_p}}}{\partial l} Q_u^{p,l,\overline{q}}((1,y),(1,y))\right] d\,\mathrm{vol}_{L_p},$$
$$\widetilde{c}_{p,l}^{\overline{q}} := \frac{1}{2}\int_0^\infty \frac{du}{u}\int_{L_p} \mathrm{Tr}_s\left[*_{p,l}^{-1}\frac{\partial *_{p,l}}{\partial l} Q_u^{p,l,\overline{q}}((1,y),(1,y))\right] d\,\mathrm{vol}_{L_p}.$$

The well-definedness of the above invariants can be shown by arguments similar to those in the proof of the well-definedness of the classical Cheeger invariant, which is the contribution of the singularities of X to Cheeger's Chern–Gauss–Bonnet theorem [25, Theorem 5.1]. In [58], we deal with the Cheeger-type invariants by combining the local index techniques developed in [11] with the singular asymptotic lemma (SAL) of Brüning and Seeley [20].

For a Euclidean vector bundle (E, ∇^E) with metric connection, we denote by $e(E,\nabla^E)$ the Euler form in Chern–Weil theory. Let $\rho : X \times \mathbb{R} \to X$ be the canonical projection and $\rho^* TX$ the pull-back bundle. We equip the vector bundle $\rho^* TX$ with the metric $g^{TX,\mathrm{tot}}$, where $g^{TX,\mathrm{tot}}_{|X\times\{l\}} = g_l^{TX}$, and the compatible connection $\nabla^{TX,\mathrm{tot}}$ (see [11, Section IV (h)]). For a form α on $X \times \mathbb{R}$, we denote by $\iota_{\partial_l}\alpha$ the interior multiplication with the vector field $\frac{\partial}{\partial l}$ (on $\mathbb{R}$). The following theorem generalises the smooth anomaly formula for the Ray–Singer metric in [11, Section IV] to spaces with isolated conical singularities.

Theorem 5.5. *Let $\mathbb{R} \ni l \to (g_l^{TX}, g_l^F)$ be a smooth family of metrics on TX, F satisfying the spectral gap condition (5.2). Then, the variation of the Ray–Singer metric is given by*

$$\partial_l \log\big(\| \ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F),l}\big)^2 = \int_X \mathrm{Tr}\left[(g_l^F)^{-1}\frac{\partial g_l^F}{\partial l}\right] e(TX,\nabla_l^{TX}) + \int_X \iota_{\partial_l} e(\rho^* TX, \nabla^{TX,\mathrm{tot}})\theta(F, g_l^F) + \sum_{p\in \mathrm{Sing}(X)} \big(c_{p,l}^{\overline{q}} + \widetilde{c}_{p,l}^{\overline{q}}\big).$$

To prove Theorem 5.5, we follow the strategy of proof for the smooth anomaly formula for the Ray–Singer metric in [11, Section IV]: First we show that the variation $\partial_l \log(\| \ \|^{\mathrm{RS}}_{\det IH^\bullet_{\overline{q}}(X,F),l})^2$ is given by the coefficient of t^0 in the asymptotic expansion of the following supertrace, as $t \searrow 0$:

$$\mathrm{Tr}_s\left[\left(*_l^{-1}\frac{\partial *_l}{\partial l} + (g_l^F)^{-1}\frac{\partial g_l^F}{\partial l}\right)\exp(-\Delta_l^{\overline{q}})\right]. \tag{5.3}$$

For this, we adapt a trick explained by Cheeger in [23, Theorem 3.10] for manifolds with boundary (and absolute or relative boundary conditions at the boundary); it is here that the spectral gap assumption (5.2) comes into play.

To compute the supertrace (5.3), it is not enough to integrate the pointwise trace – obtained, as in [11], using local index techniques. In the presence of singularities, contributions of the singularities $\mathrm{Sing}(X)$ do appear, namely the Cheeger-type invariants $c^{\overline{q}}_{p,l}$, $\tilde{c}^{\overline{q}}_{p,l}$. The first two terms in the formula in Theorem 5.5 are the local contributions, well known from the smooth anomaly formula for the Ray–Singer metric in [11]; they do not depend on the chosen extension of the Laplacian and vanish in the case where n is odd.

Let Y be an anti-radial gradient-like vector field. For a family of metrics $\mathbb{R} \ni l \to (g^{TX}_l, g^F_l)$ on TX and F, we use the notation $\| \ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F),l}$ for the associated Bismut–Zhang metrics. For $p \in \mathrm{Sing}(X)$, we denote by $\Psi(TZ_p, \nabla^{TZ_p})$ the Mathai–Quillen current on the tangent space of the infinite cone TZ_p; it is an $(n-1)$-dimensional current on TZ_p with values in the orientation line $o(TZ_p)$. We have $TZ_p \simeq \mathbb{R} \times TL_p$, and we identify the orientation lines $o(TZ_p)$ and $o(TL_p)$ by fixing the orientation by $e_r = \frac{\partial}{\partial r}$ on the first factor. We denote by $\rho : Z_p \times \mathbb{R} \to Z_p$ the canonical projection and by $\Psi(\rho^* TZ_p, \nabla^{TZ_p,\mathrm{tot}})$ the Mathai–Quillen form of the pull-back bundle $\rho^* TZ_p$. We have the model Morse function $f^p : cL_p \to \mathbb{R}$ defined by $f^p = \frac{-r^2}{2}$ its gradient vector field $\nabla f^p = -r\frac{\partial}{\partial r}$ (resp. $\widehat{\nabla f^p}$) seen as section of TZ_p (resp. of $\rho^* TZ_p$).

Theorem 5.6. *Let Y be an anti-radial gradient-like vector field. Let $\mathbb{R} \ni l \to (g^{TX}_l, g^F_l)$ be a smooth family of metrics on TX, F, satisfying the spectral gap condition (5.2). Then*

$$\begin{aligned}
&\partial_l \log\Big(\big(\| \ \|^{\mathrm{BZ},Y}_{\det IH^\bullet_{\overline{q}}(X,F),l} \big)^2 \Big) \\
&\quad = \sum_{p \in \mathrm{Crit}(Y)} (-1)^{\mathrm{ind}(p)} \partial_l \log(\| \ \|^2_{\det F_p, l}) \\
&\qquad - \sum_{p \in \mathrm{Sing}(X)} \int_{L_p} \mathrm{Tr}\Big[(g^{F_{Z_p}}_l)^{-1} \frac{\partial g^{F_{Z_p}}_l}{\partial l} \Big] \wedge (\nabla f^p)^* \Psi(TZ_p, \nabla^{TZ_p}_l) \\
&\qquad + \sum_{p \in \mathrm{Sing}(X)} \int_{L_p} \theta(F_{Z_p}, g^{F_{Z_p}}_l) \wedge \iota_{\partial_l} (\widehat{\nabla f^p})^* \Psi(\rho^* TZ_p, \nabla^{TZ_p,\mathrm{tot}}) \\
&\qquad + \sum_{p \in \mathrm{Sing}(X)} \big(c^{\overline{q}}_{p,l} + \tilde{c}^{\overline{q}}_{p,l} \big).
\end{aligned} \tag{5.4}$$

Again the first term on the right-hand side of (5.4) is the local contribution, well known from the anomaly formula of the smooth Milnor metric; see [11, (7.3)].

The last three terms stem from the variation of the Ray–Singer metric of the model Witten Laplacian $\mathbf{\Delta}_{T=1}^{p,\overline{q}}$ on the infinite cone cL_p. The latter is computed following the same strategy as in the computation of the variation formula for the Ray–Singer metric in Theorem 5.5: The second and the third terms on the right-hand side of (5.4) are the "local" contributions; the Cheeger-type invariants appear again as the contributions of the singular point. Similarly to [11, Theorem 16.1], one can also describe the dependence of the Bismut–Zhang metric $\| \ \|^{\mathrm{BZ},Y}_{\det IH^{\bullet}_{\overline{q}}(X,F)}$ on the anti-radial gradient-like vector field.

In [58], we also compute variation formulas for the right-hand side of the Bismut–Zhang formula (1.1), generalising [11, Theorem 6.3] to the singular setting. Let g^{TX}, g'^{TX} be two conical metrics on X, and let g^F, g'^F be two Hermitian metrics on F satisfying the assumption explained in Section 2.2. The Levi-Civita connection associated with g'^{TX} is denoted by ∇'^{TX}, and we denote by $\nabla' f$ the gradient vector field of f with respect to the conical metric g'^{TX}. Let $\mathbb{R} \ni l \mapsto g_l^{TX}$ be a family of conical metrics on X connecting g^{TX}, g'^{TX}.

Theorem 5.7. *The following identity holds:*

$$\begin{aligned}
&\int_X \theta(F,g^F)(\nabla f)^*\Psi(TX,\nabla^{TX}) - \int_X \theta(F,g'^F)(\nabla' f)^*\Psi(TX,\nabla'^{TX}) \\
&\quad = \int_X \log\left(\frac{\| \ \|'^2_{\det F}}{\| \ \|^2_{\det F}}\right) e(TX,\nabla^{TX}) \\
&\qquad - \sum_{p\in \mathrm{Crit}(f_{\mathrm{sm}})} (-1)^{\mathrm{ind}(p)} \log\left(\frac{\| \ \|'^2_{\det F_p}}{\| \ \|^2_{\det F_p}}\right) - \int_X \theta(F,g'^F)\widetilde{e}(TX,\nabla_l^{TX}) \\
&\qquad + \sum_{p\in \mathrm{Sing}(X)} \int_{L_p} \log\left(\frac{\| \ \|'^2_{\det F_{Z_p}}}{\| \ \|^2_{\det F_{Z_p}}}\right) \wedge (\nabla f^p)^*\Psi(TZ_p,\nabla^{TZ_p}) \\
&\qquad - \sum_{p\in \mathrm{Sing}(X)} \int_{L_p} \theta(F_{Z_p},g'^F_{Z_p}) \wedge \widetilde{e}_\Psi(Z_p,\nabla_l^{TZ_p}),
\end{aligned} \tag{5.5}$$

where

$$\widetilde{e}(TX,\nabla_l^{TX}) := \int_0^1 e(\rho^*TX,\nabla^{TX,\mathrm{tot}}) \in \Omega^{n-1}(X,o(TX))$$

and

$$\widetilde{e}_\Psi(Z_p,\nabla_l^{TZ_p}) := \int_0^1 \widetilde{(\nabla f^p)}^*\Psi(\rho^*TZ_p,\nabla^{TZ_p,\mathrm{tot}}) \in \Omega^{n-2}(Z_p,o(TZ_p)).$$

Note that all integrals over X appearing in (5.5) are well defined, since the integrands vanish near $\mathrm{Sing}(X)$. The only difference between (5.5) and the corresponding

smooth formula in [11, Theorem 6.3] is the last two terms on the right-hand side, which are the contributions of the singularities of X. Note that they correspond to the two boundary terms appearing on the right-hand side of (5.4).

By comparing the three anomaly formulas in Theorems 5.5, 5.6 and 5.7, one sees that they are compatible with the Bismut–Zhang formula (1.1), as it should be!

Acknowledgements. First and foremost, I would like to thank Jean-Michel Bismut for many discussions and for supporting my project on the study of torsion on singular spaces via the Witten deformation from the start. I also express my gratitude to Jean-Paul Brasselet for his support and encouragement at the very start of this project.

The list of colleagues to thank for mathematical insight and encouragement is too long to be mentioned here in completeness, for which I apologise. I wish to express my gratitude to my colleagues at Laboratoire J.-A. Dieudonné for welcoming me so warmly in Nice. I thank the scientific committee of the 9ECM for the honour of this invitation and the organising committee for the perfect organisation of a wonderful meeting!

Funding. At the different stages of this project, the author has been supported by Deutsche Forschungsgemeinschaft and Agence nationale de la recherche (DFG Research Fellowship 5443725, ANR-DFG project Quantization, Singularities and Holomorphic Dynamics 490843120), the European Commission (Marie Curie Intra-European Fellowship 299724, within the 7th European Community Framework Programme) and the Max-Planck Institute for Mathematics in Bonn.

During the redaction of this overview article, the author has been partially supported by DFG under Germany's Excellence Strategy EXC 2044 – 390685587, Mathematics Münster: Dynamics-Geometry-Structure as well as under CRC 1442: Geometry: Deformations and Rigidity. This work has also been supported by the French government, through the UCA JEDI Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01.

References

[1] Z. M. Ahmed and D. W. Stroock, A Hodge theory for some non-compact manifolds. *J. Differential Geom.* **54** (2000), no. 1, 177–225 Zbl 1073.58001 MR 1815415

[2] P. Albin, É. Leichtnam, R. Mazzeo, and P. Piazza, The signature package on Witt spaces. *Ann. Sci. Éc. Norm. Supér. (4)* **45** (2012), no. 2, 241–310 Zbl 1260.58012 MR 2977620

[3] P. Albin, E. Leichtnam, R. Mazzeo, and P. Piazza, Hodge theory on Cheeger spaces. *J. Reine Angew. Math.* **2018** (2018), no. 744, 29–102 Zbl 1434.58001 MR 3871440

[4] P. Albin, F. Rochon, and D. Sher, Analytic torsion and R-torsion of Witt representations on manifolds with cusps. *Duke Math. J.* **167** (2018), no. 10, 1883–1950 Zbl 1408.58025 MR 3827813

[5] P. Albin, F. Rochon, and D. Sher, Resolvent, heat kernel, and torsion under degeneration to fibered cusps. *Mem. Amer. Math. Soc.* **269** (2021), no. 1314, 126 pp. Zbl 1476.58001 MR 4234679

[6] P. Albin, F. Rochon, and D. Sher, A Cheeger–Müller theorem for manifolds with wedge singularities. *Anal. PDE* **15** (2022), no. 3, 567–642 Zbl 1496.58005 MR 4442836

[7] J. A. Álvarez López and M. Calaza, Witten's perturbation on strata. *Asian J. Math.* **21** (2017), no. 1, 47–126 Zbl 1369.58001 MR 3632436

[8] D. M. Austin and P. J. Braam, Morse–Bott theory and equivariant cohomology. In H. Hofer, C. H. Taubes, A. Weinstein, and E. Zehnder (eds.), *The Floer memorial volume*, pp. 123–183, Progr. Math. 133, Birkhäuser, Basel, 1995 Zbl 0834.57017 MR 1362827

[9] A. Banyaga and D. E. Hurtubise, Morse–Bott homology. *Trans. Amer. Math. Soc.* **362** (2010), no. 8, 3997–4043 Zbl 1226.57038 MR 2608393

[10] F. Bei, General perversities and L^2 de Rham and Hodge theorems for stratified pseudomanifolds. *Bull. Sci. Math.* **138** (2014), no. 1, 2–40 Zbl 1283.55002 MR 3245491

[11] J.-M. Bismut and W. Zhang, *An extension of a theorem by Cheeger and Müller*. With an appendix by François Laudenbach, pp. 7–218, Astérisque 205, Société Mathématique de France, Paris, 1992 Zbl 0781.58039 MR 1185803

[12] J.-M. Bismut and W. Zhang, Milnor and Ray–Singer metrics on the equivariant determinant of a flat vector bundle. *Geom. Funct. Anal.* **4** (1994), no. 2, 136–212 Zbl 0830.58030 MR 1262703

[13] J.-P. Brasselet, De Rham theorems for singular varieties. In *Differential topology, foliations, and group actions (Rio de Janeiro, 1992)*, pp. 95–112, Contemp. Math. 161, American Mathematical Society, Providence, RI, 1994 Zbl 0838.55007 MR 1271830

[14] J.-P. Brasselet, G. Hector, and M. Saralegi, Théorème de De Rham pour les variétés stratifiées. *Ann. Global Anal. Geom.* **9** (1991), no. 3, 211–243 Zbl 0733.57010 MR 1143404

[15] J.-P. Brasselet and A. Legrand, Un complexe de formes différentielles à croissance bornée sur une variété stratifiée. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (4)* **21** (1994), no. 2, 213–234 Zbl 0839.55005 MR 1288365

[16] J. Brüning and M. Lesch, Hilbert complexes. *J. Funct. Anal.* **108** (1992), no. 1, 88–132 Zbl 0826.46065 MR 1174159

[17] J. Brüning and M. Lesch, Kähler–Hodge theory for conformal complex cones. *Geom. Funct. Anal.* **3** (1993), no. 5, 439–473 Zbl 0795.58003 MR 1233862

[18] J. Brüning and X. Ma, An anomaly formula for Ray–Singer metrics on manifolds with boundary. *Geom. Funct. Anal.* **16** (2006), no. 4, 767–837 Zbl 1111.58024 MR 2255381

[19] J. Brüning and X. Ma, On the gluing formula for the analytic torsion. *Math. Z.* **273** (2013), no. 3–4, 1085–1117 Zbl 1318.58018 MR 3030691

[20] J. Brüning and R. Seeley, Regular singular asymptotics. *Adv. Math.* **58** (1985), no. 2, 133–148 Zbl 0593.47047 MR 0814748

[21] J.-L. Brylinski, Equivariant intersection cohomology. In *Kazhdan–Lusztig theory and related topics (Chicago, IL, 1989)*, pp. 5–32, Contemp. Math. 139, American Mathematical Society, Providence, RI, 1992 Zbl 0803.55002 MR 1197827

[22] K. C. Chang and J. Liu, A cohomology complex for manifolds with boundary. *Topol. Methods Nonlinear Anal.* **5** (1995), no. 2, 325–340 Zbl 0848.58001 MR 1374068

[23] J. Cheeger, Analytic torsion and the heat equation. *Ann. of Math. (2)* **109** (1979), no. 2, 259–322 Zbl 0412.58026 MR 0528965

[24] J. Cheeger, On the Hodge theory of Riemannian pseudomanifolds. In *Geometry of the Laplace operator (Proc. Sympos. Pure Math., Univ. Hawaii, Honolulu, Hawaii, 1979)*, pp. 91–146, Proc. Sympos. Pure Math. 36, American Mathematical Society, Providence, RI, 1980 Zbl 0461.58002 MR 0573430

[25] J. Cheeger, Spectral geometry of singular Riemannian spaces. *J. Differential Geom.* **18** (1983), no. 4, 575–657 Zbl 0529.58034 MR 0730920

[26] J. Cheeger, M. Goresky, and R. MacPherson, L^2-cohomology and intersection homology of singular algebraic varieties. In S.-T. Yau (ed.), *Seminar on differential geometry*, pp. 303–340, Ann. of Math. Stud. 102, Princeton University Press, Princeton, NJ, 1982 Zbl 0503.14008 MR 0645745

[27] A. Dar, Intersection R-torsion and analytic torsion for pseudomanifolds. *Math. Z.* **194** (1987), no. 2, 193–216 Zbl 0605.57012 MR 0876230

[28] J.-P. Demailly, Champs magnétiques et inégalités de Morse pour la d''-cohomologie. *Ann. Inst. Fourier (Grenoble)* **35** (1985), no. 4, 189–229 Zbl 0565.58017 MR 0812325

[29] A. Floer, Witten's complex and infinite-dimensional Morse theory. *J. Differential Geom.* **30** (1989), no. 1, 207–221 Zbl 0678.58012 MR 1001276

[30] M. Goresky and R. MacPherson, Intersection homology theory. *Topology* **19** (1980), no. 2, 135–162 Zbl 0448.55004 MR 0572580

[31] M. Goresky and R. MacPherson, Intersection homology. II. *Invent. Math.* **72** (1983), no. 1, 77–129 Zbl 0529.55007 MR 0696691

[32] M. Goresky and R. MacPherson, *Stratified Morse theory*. Ergeb. Math. Grenzgeb. (3) 14, Springer, Berlin, 1988, 272 pp. Zbl 0639.14012 MR 0932724

[33] R. M. Goresky, Whitney stratified chains and cochains. *Trans. Amer. Math. Soc.* **267** (1981), no. 1, 175–196 Zbl 0476.57019 MR 0621981

[34] L. Hartmann and M. Spreafico, The analytic torsion of a cone over a sphere. *J. Math. Pures Appl. (9)* **93** (2010), no. 4, 408–435 Zbl 1222.58026 MR 2609036

[35] L. Hartmann and M. Spreafico, The analytic torsion of a cone over an odd dimensional manifold. *J. Geom. Phys.* **61** (2011), no. 3, 624–657 Zbl 1222.58027 MR 2763625

[36] L. Hartmann and M. Spreafico, On the Cheeger–Müller theorem for an even-dimensional cone (in Russian). *Algebra i Analiz* **27** (2015), no. 1, 194–217. English translation. *St. Petersburg Math. J.* **27** (2016), no. 1, 137–154 Zbl 1338.58022 MR 3443271

[37] B. Hasselblatt and A. Katok, *A first course in dynamics. With a panorama of recent developments*. Cambridge University Press, Cambridge, 2003, 424 pp. Zbl 1027.37001 MR 1995704

[38] B. Helffer and F. Nier, *Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach: the case with boundary*. Mém. Soc. Math. Fr. (N.S.) 105, Société Mathématique de France, Paris, 2006, 89 pp. Zbl 1108.58018 MR 2270650

[39] B. Helffer and J. Sjöstrand, Puits multiples en mécanique semi-classique. IV. Étude du complexe de Witten. *Comm. Partial Differential Equations* **10** (1985), no. 3, 245–340 Zbl 0597.35024 MR 0780068

[40] F. Laudenbach, *Appendix. On the Thom–Smale complex*. In *An extension of a theorem by Cheeger and Müller*, pp. 219–233, Astérisque 205, Société Mathématique de France, Paris, 1992 MR 0755164

[41] D. Le Peutrec, Local WKB construction for Witten Laplacians on manifolds with boundary. *Anal. PDE* **3** (2010), no. 3, 227–260 Zbl 1225.58012 MR 2672794

[42] D. Le Peutrec, Small eigenvalues of the Neumann realization of the semiclassical Witten Laplacian. *Ann. Fac. Sci. Toulouse Math. (6)* **19** (2010), no. 3–4, 735–809 Zbl 1213.58023 MR 2790817

[43] M. Lesch, *Operators of Fuchs type, conical singularities, and asymptotic methods*. Teubner-Texte Math. 136, B. G. Teubner, Stuttgart, 1997, 190 pp. Zbl 1156.58302 MR 1449639

[44] M. Lesch, Determinants of regular singular Sturm–Liouville operators. *Math. Nachr.* **194** (1998), no. 1, 139–170 Zbl 0924.58107 MR 1653090

[45] M. Lesch, A gluing formula for the analytic torsion on singular spaces. *Anal. PDE* **6** (2013), no. 1, 221–256 Zbl 1276.58010 MR 3068545

[46] J. Lott and M. Rothenberg, Analytic torsion for group actions. *J. Differential Geom.* **34** (1991), no. 2, 431–481 Zbl 0744.57021 MR 1131439

[47] W. Lück, Analytic and topological torsion for manifolds with boundary and symmetry. *J. Differential Geom.* **37** (1993), no. 2, 263–322 Zbl 0792.53025 MR 1205447

[48] U. Ludwig, Morse–Smale–Witten complex for gradient-like vector fields on stratified spaces. In *Singularity theory (Proc. 2005 Marseille, CIRM, Marseille, France)*, pp. 683–713, World Scientific, Singapore, 2007 Zbl 1131.58008 MR 2342935

[49] U. Ludwig, The geometric complex for algebraic curves with cone-like singularities and admissible Morse functions. *Ann. Inst. Fourier (Grenoble)* **60** (2010), no. 5, 1533–1560 Zbl 1207.58014 MR 2766222

[50] U. Ludwig, A proof of the stratified Morse inequalities for singular complex algebraic curves using the Witten deformation. *Ann. Inst. Fourier (Grenoble)* **61** (2011), no. 5, 1749–1777 Zbl 1242.32006 MR 2961839

[51] U. Ludwig, The Witten complex for singular spaces of dimension two with cone-like singularities. *Math. Nachr.* **284** (2011), no. 5–6, 717–738 Zbl 1215.58004 MR 2663764

[52] U. Ludwig, An analytic approach to the stratified Morse inequalities for complex cones. *Internat. J. Math.* **24** (2013), no. 12, article no. 1350100, 12 pp. Zbl 1317.32063 MR 3152207

[53] U. Ludwig, The Witten deformation for even dimensional conformally conic manifolds. *Trans. Amer. Math. Soc.* **365** (2013), no. 2, 885–909 Zbl 1479.58011 MR 2995377

[54] U. Ludwig, A complex in Morse theory computing intersection homology. *Ann. Inst. Fourier (Grenoble)* **67** (2017), no. 1, 197–236 Zbl 1387.55009 MR 3612329

[55] U. Ludwig, Comparison between two complexes on a singular space. *J. Reine Angew. Math.* **2017** (2017), no. 724, 1–52 Zbl 1367.58009 MR 3619103

[56] U. Ludwig, An extension of a theorem by Cheeger and Müller to spaces with isolated conical singularities. *Duke Math. J.* **169** (2020), no. 13, 2501–2570 Zbl 1458.58016 MR 4142751

[57] U. Ludwig, An index formula for the intersection Euler characteristic of an infinite cone. *Math. Z.* **296** (2020), no. 1–2, 99–126 Zbl 1460.58018 MR 4140733

[58] U. Ludwig, *Bismut–Zhang theorem and anomaly formula for the Ray–Singer metric for spaces with isolated conical singularities*. MPIM Preprint Ser. 41, Max-Planck-Institut für Mathematik, Bonn, 2022

[59] U. Ludwig, A Morse–Bott type complex and the Bismut–Zhang torsion for intersection cohomology. *Ann. Inst. Fourier (Grenoble)* **75** (2025), no. 2, 655–722 Zbl 1570.58021 MR 4921363

[60] V. Mathai and D. Quillen, Superconnections, Thom classes, and equivariant differential forms. *Topology* **25** (1986), no. 1, 85–110 Zbl 0592.55015 MR 0836726

[61] R. Mazzeo and B. Vertman, Analytic torsion on manifolds with edges. *Adv. Math.* **231** (2012), no. 2, 1000–1040 Zbl 1255.58012 MR 2955200

[62] J. Milnor, Whitehead torsion. *Bull. Amer. Math. Soc.* **72** (1966), 358–426 Zbl 0147.23104 MR 0196736

[63] W. Müller, Analytic torsion and R-torsion of Riemannian manifolds. *Adv. Math.* **28** (1978), no. 3, 233–305 Zbl 0395.57011 MR 0498252

[64] W. Müller, Analytic torsion and R-torsion for unimodular representations. *J. Amer. Math. Soc.* **6** (1993), no. 3, 721–753 Zbl 0789.58071 MR 1189689

[65] W. Müller and B. Vertman, The metric anomaly of analytic torsion on manifolds with conical singularities. *Comm. Partial Differential Equations* **39** (2014), no. 1, 146–191 Zbl 1292.58024 MR 3169782

[66] M. Nagase, L^2-cohomology and intersection homology of stratified spaces. *Duke Math. J.* **50** (1983), no. 1, 329–368 Zbl 0545.55004 MR 0700144

[67] D. B. Ray and I. M. Singer, R-torsion and the Laplacian on Riemannian manifolds. *Adv. Math.* **7** (1971), no. 2, 145–210 Zbl 0239.58014 MR 0295381

[68] M.-H. Schwartz, Champs radiaux et préradiaux associés à une stratification. *C. R. Acad. Sci. Paris Sér. I* **303** (1986), no. 6, 239–241 Zbl 0604.58006 MR 0860826

[69] M.-H. Schwartz, *Champs radiaux sur une stratification analytique*. Trav. Cours 39, Hermann, Paris, 1991, 183 pp. Zbl 0727.57026 MR 1096495

[70] P. H. Siegel, Witt spaces: a geometric cycle theory for KO-homology at odd primes. *Amer. J. Math.* **105** (1983), no. 5, 1067–1105 Zbl 0547.57019 MR 0714770

[71] S. Smale, On gradient dynamical systems. *Ann. of Math. (2)* **74** (1961), no. 1, 199–206 Zbl 0136.43702 MR 0133139

[72] R. Thom, Sur une partition en cellules associée à une fonction sur une variété. *C. R. Acad. Sci. Paris* **228** (1949), 973–975 Zbl 0034.20802 MR 0029160

[73] B. Vertman, Analytic torsion of a bounded generalized cone. *Comm. Math. Phys.* **290** (2009), no. 3, 813–860 Zbl 1210.58026 MR 2525641

[74] S. M. Vishik, Generalized Ray–Singer conjecture. I. A manifold with a smooth boundary. *Comm. Math. Phys.* **167** (1995), no. 1, 1–102 Zbl 0818.58049 MR 1316501

[75] E. Witten, Supersymmetry and Morse theory. *J. Differential Geom.* **17** (1982), no. 4, 661–692 Zbl 0499.53056 MR 0683171

[76] N. Yeganefar, Sur la L^2-cohomologie des variétés à courbure négative. *Duke Math. J.* **122** (2004), no. 1, 145–180 Zbl 1069.58013 MR 2046810

Ursula Ludwig
Laboratoire Jean-Alexandre Dieudonné Mathématique, Université Côte d'Azur, EUR Spectrum, Parc Valrose, 06108 Nice CEDEX 2, France; ursula.ludwig@univ-cotedazur.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/23

Strong convergence of unitary and permutation representations of discrete groups

Michael Magee

Abstract. We survey a research program on the strong convergence of unitary and permutation representations of discrete groups. We also take the opportunity to flesh out details that have not appeared elsewhere.

1. Definitions

Throughout the article, Γ denotes an infinite discrete group. For $n \in \mathbb{N}$, we write $\mathcal{U}(n)$ for the group of complex $n \times n$ unitary matrices and S_n for the group of permutations of $[n] \overset{\text{def}}{=} \{1, \dots, n\}$.

Definition 1.1 (Strong convergence, P**Mat**F). If $\{\rho_i : \Gamma \to \mathcal{U}(n_i)\}_{i=1}^{\infty}$ are a sequence of (possibly random) finite-dimensional unitary representations of Γ, we say that ρ_i strongly converges to the regular representation (a.s., or in probability, if ρ_i are random) if for any $z \in \mathbb{C}[\Gamma]$,

$$\lim_{i\to\infty} \|\rho_i(z)\| = \|\lambda_\Gamma(z)\|$$

(a.s., or in probability, respectively), where $\lambda_\Gamma : \Gamma \to U(\ell^2(\Gamma))$ is the left regular representation. The norms above are operator norms. We write $\rho_i \xrightarrow{\text{strong}} \lambda_\Gamma$ in this event. If Γ has such a sequence of unitary representations, then we say Γ is purely matricial field (P**Mat**F).

Remark 1.2. The following statements hold:

(1) *Weak convergence.* Some authors ask for weak convergence as part of the definition of strong convergence, but we do not. Weak convergence is the statement that for all $z \in \mathbb{C}[\Gamma]$,

$$\lim_{i\to\infty} \operatorname{tr}[\rho_i(z)] = \tau(z),$$

Mathematics Subject Classification 2020: 22D10 (primary); 05C80, 20B05, 20F36, 20F65, 46L54, 58J50 (secondary).
Keywords: strong convergence, unitary representations, permutations.

where $\mathrm{tr} = n_i^{-1}\mathrm{Tr}$ is the normalized trace on $n_i \times n_i$ matrices and $\tau(z)$ is the canonical tracial state on the reduced C^*-algebra $C^*_{\mathrm{red}}(\Gamma)$. If $C^*_{\mathrm{red}}(\Gamma)$ has a unique tracial state, then strong convergence implies weak convergence (e.g., [32, Lemma 6.1]).[1] We now know exactly when $C^*_{\mathrm{red}}(\Gamma)$ has a unique tracial state by Breuillard–Kalantar–Kennedy–Ozawa [11, Theorem 1.6]: this is so when Γ has no non-trivial normal amenable subgroup.

(2) *Nomenclature.* Let M_{n_i} denote the complex matrix C^*-algebra of dimension n_i, let $\ell^\infty\left(\prod_i M_{n_i}\right)$ denote the bounded sequences in the product, and let $\mathcal{I}$ be the closed two-sided ideal of sequences that converge to zero. If $\rho_i \xrightarrow{\text{strong}} \lambda_\Gamma$, then $\prod_i \rho_i$ descends (and extends) to an embedding

$$C^*_{\mathrm{red}}(\Gamma) \hookrightarrow \ell^\infty\left(\prod_i M_{n_i}\right)/\mathcal{I}.$$

If there exists any such embedding, then $C^*_{\mathrm{red}}(\Gamma)$ is called matricial field by Blackadar and Kirchberg in [9]. But conversely, such an embedding does not (a priori) have to factor through $\ell^\infty\left(\prod_i M_{n_i}\right)$ when restricted to $\mathbb{C}[\Gamma]$. Hence the adjective 'purely'. The concept of purely matricial field was introduced in Magee–de la Salle [34].

(3) More generally, if G is a locally compact topological group, we can extend our definition by replacing $\mathbb{C}[\Gamma]$ with the continuous compactly supported complex functions on G.

A first attempt to require strong convergence to factor through permutations would be that the ρ_i are the composition of some $\phi_i \in \mathrm{Hom}(\Gamma, S_n)$ with the (n-dimensional) permutation representations of S_n. However, since this gives rise to ρ_i with non-zero invariant vectors, this can never work if Γ is non-amenable, which is not satisfactory. So the definition is modified in the following way. Let std denote the $(n-1)$-dimensional irreducible subrepresentation of the defining representation of S_n.

Definition 1.3 (P**Perm**F). If there exists a sequence of homomorphisms $\{\phi_i : \Gamma \to S_{n_i}\}_{i=1}^\infty$ such that

$$\left\{\rho_i \overset{\text{def}}{=} \mathrm{std} \circ \phi_i\right\}_{i=1}^\infty$$

strongly converges to the regular representation, then we say Γ is purely permutation field (P**Perm**F).

[1] We first heard this observation from Benoît Collins in June 2022.

Strong convergence and the associated above properties P**Mat**F and P**Perm**F have recently found powerful applications in a surprising range of settings including, but not limited to, the spectral geometry of graphs [10] and hyperbolic manifolds [28,32,40], the theory of minimal surfaces (observed by Song [45,46]), and the Peterson–Thom conjecture [43] (as observed by Hayes [26]) on subalgebras of free group factors.

2. PMatF and PPermF

We begin with some basic properties of P**Mat**F and P**Perm**F. In this section, $\Lambda \leq \Gamma$ are discrete groups.

Lemma 2.1 (Restriction to subgroups). *If* Γ *is* P**Mat**F *(resp.* P**Perm**F*), then* Λ *is* P**Mat**F *(resp.* P**Perm**F*).*

Proof. Suppose $z \in \mathbb{C}[\Lambda] \leq \mathbb{C}[\Gamma]$. If $\rho_i \xrightarrow{\text{strong}} \lambda_\Gamma$, then $\|\rho_i(z)\| \to \|\lambda_\Gamma(z)\|$ as $i \to \infty$. But as a $\mathbb{C}[\Lambda]$-module,

$$\ell^2(\Gamma) \cong \bigoplus_{[\gamma]\in\Lambda\backslash\Gamma} \ell^2(\Lambda\gamma) \cong \bigoplus_{[\gamma]\in\Lambda\backslash\Gamma} \ell^2(\Gamma),$$

so $\|\lambda_\Gamma(z)\| = \|\lambda_\Lambda(z)\|$. Hence, if ρ_i' is the restriction of ρ_i to Λ, $\rho_i' \xrightarrow{\text{strong}} \lambda_\Lambda$. If the representations factor through S_{n_i}, then their restrictions still do. ∎

Lemma 2.2 (Induction to finite index overgroups). *If* Λ *is* P**Mat**F *and finite index in* Γ*, then* Γ *is also* P**Mat**F.

Lemma 2.2 is a special case of a more general phenomenon (induction from cocompact lattices) that will be covered later in the paper (see Theorem 3.1).

2.1. Amenable groups

The following argument about amenable groups was obtained in conversations with Mikael de la Salle.

Say that Γ is residually linear (RL) (resp. residually finite (RF)) if it embeds into a product of GL_{n_i} (resp. S_{n_i}). A theorem of Malcev [41] states that every finitely generated (f.g.) linear group is residually finite. Hence, if Γ is f.g. and RL, then it is RF.

If Γ is not RL, then there is $\gamma \in \Gamma$ that is killed by every homomorphism to some GL_n. Then considering

$$z = \gamma - \mathrm{id} \in \mathbb{C}[\Gamma],$$

we have

$$\|\rho(z)\| = 0$$

for every $\rho : \Gamma \to \mathrm{GL}_n$. But $\|\lambda(z)\|$ is not zero (or else $\lambda(z) = 0$, but $\mathbb{C}[\Gamma]$ always embeds into $C^*_{\mathrm{red}}(\Gamma)$). The upshot is the following observation.

Lemma 2.3. *A* P**Mat**F *group is residually linear. Hence, it is also either residually finite or not finitely generated.*

On the other hand, if Γ is countable and RF, then by projecting to large enough factors of the $\prod_i S_{n_i}$, we obtain a sequence of homomorphisms $\phi_j : \Gamma \to S_{N_j}$ such that each ϕ_j injects on a finite set S_j such that $\bigcup_j S_j = \Gamma$. For any $z \in \mathbb{C}[\Gamma]$, let Φ_j denote the unitary representation of Γ obtained by composing ϕ_j with the standard representation of S_{N_j}.

Because we eventually inject on the support of $(zz^*)^p$, we then have

$$\mathrm{tr}[\Phi_j(zz^*)^p] = \tau((zz^*)^p) - \frac{1}{N_j}\varepsilon((zz^*)^p),$$

where ε is the state associated with the trivial representation of Γ for $j \gg_p 1$. Since the trivial representation is weakly contained in the regular representation of Γ, we have

$$\varepsilon((zz^*)^p) \leq \|\lambda(z)\|^{2p}.$$

Hence, we also have

$$\begin{aligned}\|\Phi_j(z)\|^{2p} &= \|\Phi_j(zz^*)^p\| \geq \mathrm{tr}[\Phi_j(zz^*)^p] \to_{i\to\infty} \tau((zz^*)^p) \\ &= (\|\lambda(z)\| + o_{p\to\infty}(1))^{2p},\end{aligned}$$

where the last equality used that the trace τ is faithful. Taking p large and fixed and letting $j \to \infty$ for each large p, we obtain

$$\liminf_{j\to\infty} \|\Phi_j(z)\| \geq \|\lambda(z)\|.$$

On the other hand, for amenable groups, we have $\|\lambda(z)\| \geq \|\Phi_j(z)\|$ for all $z \in \mathbb{C}[\Gamma]$ because every finite-dimensional unitary representation of amenable groups is weakly contained in the regular representation. So, in fact,

$$\lim_{j\to\infty} \|\Phi_j(z)\| = \|\lambda(z)\|.$$

Hence, in summary *for amenable groups*,

$$\text{countable} + \text{RF} \implies \text{P}\mathbf{Perm}\text{F}$$

and

$$\begin{aligned} \text{PMatF} &\implies \text{RL}, \\ \text{f.g.} + \text{PMatF} &\implies \text{RF}. \end{aligned}$$

What is perhaps surprising is that *all* amenable groups have finite-dimensional (f.d.) *approximate representations* (in norm sense) that strongly converge to the regular representation [47]. This appears to leave open the question of whether non-countable discrete amenable groups are P**Mat**F or P**Perm**F.

2.2. Free groups

Let $\mathbf{F}_r$ denote a free non-abelian group of rank $r \geq 2$.

The original interest in property P**Mat**F for non-amenable groups comes from an observation of Voiculescu [48] that establishing '$\mathbf{F}_r$ is P**Mat**F' – albeit not in this language—would settle the then outstanding problem in operator algebras to prove the K-theoretic construct $\mathrm{Ext}(C^*_{\mathrm{red}}(\mathbf{F}_r))$ has non-invertible elements.

Motivated by this application to K-theory, Haagerup and Thorbjørnsen proved $\mathbf{F}_r$ is P**Mat**F in the main theorem of [22]. Importantly, this and all subsequent proofs of this fact rely on random matrix theory. As a result, there is a lacuna in the field: we do not know how to construct explicit f.d. representations of free groups that strongly converge to the regular representation.[2] Haagerup and Thorbjørnsen used GUE random matrices and an application of functional calculus to prove their result, and as a result, it left open the question of whether Haar-distributed random representations of $\mathbf{F}_r$ almost surely strongly converge to the regular representation. This was proved by Collins and Male in [19].

Later on, Bordenave and Collins [10] proved that $\mathbf{F}_r$ is P**Perm**F by use of random permutation representations. To explain their motivation and to explain the motivation for P**Perm**F in general, we detour to discuss Friedman's theorem.

Friedman's theorem [20], formerly conjecture of Alon [2], states informally that random regular graphs are almost optimal expanders. Alon did not state which random model to use, but besides, there are contiguity results that tell us it does not matter for the purposes of proving the above statement [50]. For simplicity of exposition, assume that the random regular graph has degree $2r$ and n vertices, with r fixed and $n \to \infty$.

If $\sigma_1, \ldots, \sigma_r$ are i.i.d. uniformly random elements of S_n, then the Schreier graph for these generators and the action of S_n on $\{1, \ldots, n\}$ has adjacency matrix

$$2r \cdot \mathrm{id}_{\mathbb{C}} \oplus [\mathrm{std}(\sigma_1) + \mathrm{std}(\sigma_1^{-1}) + \cdots + \mathrm{std}(\sigma_r) + \mathrm{std}(\sigma_r^{-1})].$$

[2] This problem seems known in the community; it is certainly not my own, and it was highlighted to me by Avi Wigderson.

The $2r \cdot \mathrm{id}_{\mathbb{C}}$ factor corresponds to the trivial eigenvalue $2r$, and in this setting, asymptotically optimal expansion amounts to

$$\|\mathrm{std}(\sigma_1) + \mathrm{std}(\sigma_1^{-1}) + \cdots + \mathrm{std}(\sigma_r) + \mathrm{std}(\sigma_r^{-1})\| \to 2\sqrt{2r-1}$$

(in probability) by a result of Alon–Boppana [42], the right-hand side is optimal. To put the above in a more symmetric form, and to connect to our other discussions, if

$$\phi(x_i) \stackrel{\mathrm{def}}{=} \sigma_i, \quad \rho(x_i) = \mathrm{std}(\phi(x_i)) = \mathrm{std}(\sigma_i),$$

then ρ_i is a random representation of the type appearing in the definition of P**Perm**F, and Friedman's theorem states that

$$\begin{aligned}\|\rho_i(x_1 + x_1^{-1} + \cdots + x_r + x_r^{-1})\| &\to \|\lambda_{\mathbf{F}_r}(x_1 + x_1^{-1} + \cdots + x_r + x_r^{-1})\| \\ &= 2\sqrt{2r-1}\end{aligned}$$

in probability. This is a very restricted version of P**Perm**F concerning only the linear element

$$z = x_1 + x_1^{-1} + \cdots + x_r + x_r^{-1} \in \mathbb{C}[\mathbf{F}_r],$$

whereas P**Mat**F and P**Perm**F ask for the same for all(!) elements of $\mathbb{C}[\Gamma]$. This is precisely what Bordenave–Collins prove in [10]: for all $z \in \mathbb{C}[\mathbf{F}_r]$,

$$\|\rho_i(z)\| \to \|\lambda_{\mathbf{F}_r}(z)\|$$

in probability. We cannot leave this discussion without remarking that very recently, a beautiful and short proof of Friedman's theorem has been obtained by Chen, Garza-Vargas, Tropp, and van Handel [18]. They also extend the proof without adding much length to give a new proof of Bordenave–Collins' theorem.

2.3. Non-free examples

Limit groups and surface groups. Motivated in part by applications to hyperbolic surfaces (see Section 3), in [32] Louder and I studied *limit groups:* finitely generated groups such that for every finite subset, there is a homomorphism to a free group that injects on that set.

Theorem 2.4 (Louder–Magee). *Every limit group is* P**Perm**F.

The full strength of this result uses a result of Sela [17,44] stating that every limit group embeds into an iterated sequence of (extensions of centralizers), beginning with a free group. This is combined with a quantitative version of a lemma of Baumslag [5] and Haagerup's inequality [21]. The following is one of the motivations.

Corollary 2.5. *The fundamental group of an orientable surface of genus at least two is* P**Perm**F.

While stated as a corollary, Sela's work is not needed here as a genus-two surface group

$$\langle a, b, c, d \mid [a, b][c, d] \rangle$$

explicitly embeds into the extension of centralizers

$$\langle a, b, t \mid [t, [a, b]] \rangle$$

via $a \mapsto a, b \mapsto b, c \mapsto b^t, d \mapsto a^t$; the injectivity of this map uses normal form for amalgamated products. Higher genus surface groups embed into this example via covering spaces.

Added in proof. Since the original submission of the article, the author, Puder, and van Handel [39] proved the analog of the Bordenave–Collins theorem for surface groups. That is, if Γ is the fundamental group of an orientable surface of genus at least two, then uniformly random elements of $\mathrm{Hom}(\Gamma, S_n)$, composed with the standard representation, strongly converge in probability to the regular representation as $n \to \infty$.

Right-angled Artin groups and related examples. A finitely generated right-angled Artin group (RAAG) is generated by finitely many generators with only relations that some subset of the pairs of generators are commuting pairs. Besides interpolating between free non-abelian and free abelian groups, they turn out to be fairly universal in that many natural families of groups virtually embed into RAAGs. That is, they embed after passing to a finite index subgroup. This includes the following classes of groups.

(1) Closed hyperbolic 3-manifold fundamental groups (Bergeron–Wise [8] and Agol [1]).

(2) Non-compact finite volume 3-manifold fundamental groups[3] (Wise [49, Theorem 14.29]).

(3) Arithmetic 'standard type' hyperbolic n-manifold groups with dimension $n \geq 4$ (Bergeron–Haglund–Wise [7]).

(4) Any Coxeter group (Haglund–Wise [24]).

(5) Any one-relator group with torsion (Wise [49, Corollary 18.1]).

(6) Any word-hyperbolic cubulated group (Agol [1]).

[3] I thank Jean Pierre Mutanguha for making me aware of this.

The above results also all rely on a result of Haglund and Wise [23] stating that fundamental groups of compact special cube complexes embed into RAAGs.

By Lemmas 2.1 and 2.2, P**Mat**F passes from RAAGs to those groups above. The 'master theorem' that RAAGs are indeed P**Mat**F was obtained in joint work with Thomas [40].

Theorem 2.6 (Magee–Thomas). *Finitely generated RAAGs are* P**Mat**F.

It is not true however that in general RAAGs are P**Perm**F. The following proposition has not appeared elsewhere in print.

Proposition 2.7. *The group* $\mathbf{F_2} \times \mathbf{F_2} \times \mathbf{F_2}$ *is not* P**Perm**F.

Proof. Let G_i denote the copy of $\mathbf{F}_2$ embedded at the ith factor, $i = 1, 2, 3$. Let $\{\rho_j\}_{j=1}^{\infty}$ be a putative sequence of permutation representations of $\mathbf{F_2} \times \mathbf{F_2} \times \mathbf{F_2}$ such that std $\circ\, \rho_i$ strongly converges to the regular representation. We interpret ρ_j as a linear representation by composition with the defining representation. Let $H_{i,j} \stackrel{\text{def}}{=} \rho_j(G_i) \leq S_{n_j}$. Since $\mathbf{F}_2$ is non-amenable, each $H_{i,j}$ has at most 1-dimensional space of invariant vectors in the defining representation. This means that each $H_{i,j}$ must be a transitive subgroup of S_{n_j} (acting transitively on $[n_j]$) for $j \gg 1$. A theorem of finite group theory[4] states that two commuting transitive permutation groups (say J_1 and J_2 in S_N) only arise in the following way: there is some auxiliary group U with $|U| = N$, and isomorphisms $\iota_1 : J_1 \cong \lambda(U)$ and $\iota_2 : J_2 \cong \rho(U)$, where $\lambda(U)$ (resp. $\rho(U)$) is the permutation group induced by multiplication by U on its left (resp. right).

In particular, in the situation above, $H_{1,j}$ and $H_{2,j}$ can be identified with $\lambda(U_j)$ and $\rho(U_j)$ for some j. There are now different ways to conclude; we note that $H_{3,j} \cong U_j$ by the same observation, but in the above model, it is U_j acting on itself by permutations and commuting with $\lambda(U_j)$ and $\rho(U_j)$. Only the identity in U_j can do this, so U_j is the trivial one-element group. Then std $\circ\, \rho_j$ is the 0-dimensional representation for $j \gg 1$. This is obviously a contradiction. ∎

As a byproduct of the above proof, one sees that any putative sequence of permutation representations of $\mathbf{F_2} \times \mathbf{F_2}$ such that std $\circ\, \rho_i$ strongly converges to the regular representation has to have quite a particular structure. However, there is a candidate that fits this structure.

[4] Indeed, J_1 must act freely on $[N]$, since J_2 preserves the fixed points of any $j \in J_1$. This identifies J_1 with $[N]$ via $j \mapsto j.1$ and makes the action of J_1 into the left regular action. Now $J_2 \leq \mathrm{Perms}(J_1)$ is in the centralizer of the left regular action of J_1, which is the right regular action of J_1 [25, Theorem 6.3.1]. But since J_2 is transitive, it must be isomorphic to J_1 and equal to the whole right regular action.

Question 2.8. For $n \in \mathbb{N}$, let θ_n and ϕ_n denote uniformly random permutation representations $\mathbf{F}_2 \to S_n$. Then

$$\lambda \circ \theta_n : \mathbf{F}_2 \to S_{n!} = \mathbf{Perm}(S_n), \quad \rho \circ \phi_n : \mathbf{F}_2 \to S_{n!} = \mathbf{Perm}(S_n)$$

are two commuting permutation representations of $\mathbf{F}_2$. Is it true that as $n \to \infty$,

$$\mathrm{std}_{n!} \circ [\lambda \circ \theta_n \times \rho \circ \phi_n] \xrightarrow{\text{strong}} \lambda_{\mathbf{F}_2 \times \mathbf{F}_2}$$

in probability?

Notice that if the above is true, it entails that as $n \to \infty$, $\lambda \circ \theta_n \xrightarrow{\text{strong}} \lambda_{\mathbf{F}_2}$ in probability. This is far from known. In fact, it is not known whether

$$\|\lambda \circ \theta_n[x_1 + x_2 + x_1^{-1} + x_2^{-1}]\| \to 2\sqrt{3} \tag{2.1}$$

in probability; nor is it known whether the same holds with '2' replaced by any $d \geq 0$ and $\sqrt{3}$ replaced by $\sqrt{2d-1}$. This is even stronger than the also open question:

'Are most fixed degree Cayley *graphs of S_n uniform expanders?'*

(Equation (2.1) suggests they are moreover almost optimal expanders.)

We note here an important and perhaps relevant result of Kassabov [30] giving the *existence* of fixed degree uniformly expanding Cayley graphs of S_n with $n \to \infty$. In this vein, there is a very recent work of E. Cassidy [15, 16] who proves the strong convergence in probability of

$$\rho_n : \mathbf{F}_r \to \mathcal{U}(N(n)),$$
$$\rho_n(x_i) \stackrel{\text{def}}{=} \pi_\lambda(\sigma_i),$$

where σ_i are as above and π_λ is the irreducible representation of S_n corresponding to a Young diagram λ with $1 \leq |\lambda| \leq n^c$ boxes, where $0 < c < 1$ is explicit. In fact, his result is uniform in this regime of λ. An analogous result for $U(n)$ (with slightly worse constant) was obtained recently with de la Salle [35]. Cassidy's result for $|\lambda|$ up to and including n would answer the above questions about Cayley graphs of S_n.

The fact that not all RAAGs are P**Perm**F does not have a direct implication to, e.g., closed hyperbolic 3-manifold groups in general because when one induces a permutation representation from an infinite index subgroup of a RAAG, the resulting permutation representation is of an infinite set and so cannot be used to potentially prove P**Perm**F for the RAAG.

2.4. Non-examples

We discuss here f.g. residually finite groups.[5] With M. de la Salle, we established [34] that $\mathrm{SL}_d(\mathbf{Z})$ is not P**Mat**F for $d \geq 4$. This leaves a curious gap at $d = 3$. The reason for this gap is that we rely on the following fact, established in [34]: every non-trivial finite-dimensional unitary representation of $\mathrm{SL}_4(\mathbf{Z})$ has a non-zero $\mathrm{SL}_2(\mathbf{Z})$-invariant vector. This in turn means that the action of $\mathrm{SL}_2(\mathbf{Z})$ in this representation has no spectral gap, and so when restricted to $\mathrm{SL}_2(\mathbf{Z})$, a putative sequence of representations of $\mathrm{SL}_4(\mathbf{Z})$ that strongly converge to the regular representation cannot converge to the regular representation of $\mathrm{SL}_2(\mathbf{Z})$ – which does have a spectral gap. In light of Lemma 2.1, this is a contradiction.

However, there are f.d. irreducible unitary representations of $\mathrm{SL}_3(\mathbf{Z})$ with dimension tending to infinity and without non-zero $\mathrm{SL}_2(\mathbf{Z})$-invariant vectors. See [34] for details – this example is due to Deligne.

2.5. Connection to the Fell topology on the unitary dual

Suppose G is a locally compact topological group. We assume general familiarity with the Fell topology on the unitary dual of G (the equivalence classes of continuous unitary representations of G) and also the notion of weak containment of representations.[6]

Proposition 2.9. *Suppose that G is any locally compact group. Suppose that $\pi_i : G \to \mathcal{U}(\mathcal{H}_i)$ are any sequence of representations of Γ that strongly converge to a unitary representation π_∞ of G in the sense that for all $f \in C_c(G)$,*

$$\lim_{i \to \infty} \|\pi_i(f)\| = \|\pi_\infty(f)\|.$$

Then for any compact subset $K \subset \widehat{G} \backslash \mathrm{support}(\pi_\infty)$, for $i > i_0(K)$, no element of K is weakly contained in π_i.

Remark 2.10. To attempt to promote Proposition 2.9 to an if and only if statement in the case $\pi_\infty = \lambda_G$, one needs to (at least) also discuss possible discrete series representations of G (e.g., in the case $G = \mathrm{PSL}_2(\mathbb{R})$, this issue is already present). The

[5] Non-finitely generated but residually linear groups have not really been considered in the context of P**Mat**F to the author's knowledge, and this might be an interesting thing to investigate further.

[6] In the Princeton seminar, I proved Proposition 2.9 in the case the compact set K was a point. M. de la Salle pointed out the extension to compacts.

point is that a diagonal matrix coefficient of an integrable discrete series[7] will act as a non-zero projection in $\pi_\infty = \lambda_G$ and therefore needs to act in a non-zero way in any π_i when i is sufficiently large.

Proof. One can show directly from the definition of Fell topology that for any $f \in C_c(G)$, the map

$$\pi \mapsto \|\pi(f)\| \tag{2.2}$$

is lower semicontinuous in the Fell topology. For completeness, we give this argument. Given $\pi \in \widehat{G}$, suppose that ξ is such that $\|\xi\| = 1$ and

$$\langle \pi(f^* * f)\xi, \xi\rangle = \langle \pi(f)\xi, \pi(f)\xi\rangle > \|\pi(f)\|^2 - \varepsilon.$$

Essentially, by the definition of Fell topology (see [6, Proposition F.2.4]), there is an open set around π consisting of π' such that

$$|\langle \pi'(g)\xi', \xi'\rangle - \langle \pi(g)\xi, \xi\rangle| < \frac{\varepsilon}{1 + \|f^* * f\|_{L^1}}$$

for some ξ' and for all g in the support of $f^* * f$. This gives $\|\xi'\|^2 > 1 - \varepsilon$ (taking $g = 1$), and integrating with $f^* * f$ weights yields

$$\begin{aligned}\|\pi'(f)\|^2(1-\varepsilon) > \|\pi'(f)\|^2\|\xi'\|^2 \geq \langle \pi'(f^* * f)\xi', \xi'\rangle > \langle \pi(f^* * f)\xi, \xi\rangle - \varepsilon \\ > \|\pi(f)\|^2 - 2\varepsilon\end{aligned}$$

for π' in this open set. This proves the lower semicontinuity of (2.2).

Now let K be as in the statement of the proposition. For every $\pi \in K$, as π is not in the support of π_∞, there exists $f_\pi \in C_c(G)$ and $\eta_\pi > 1$ such that $\|\pi(f_\pi)\| > \eta_\pi\|\pi_\infty(f_\pi)\|$. By the previous assertion, there is an open neighborhood W_π of π where this inequality still holds after adjusting η_π. By compactness of K, we then obtain a finite list of functions $f_1, \ldots, f_r \in C_c(G)$ and $\eta > 1$ such that for all $\pi \in K$,

$$\|\pi(f_j)\| > \eta\|\pi_\infty(f_j)\|$$

for some f_j. But for large enough $i > i_0$, by strong convergence, we have

$$\|\pi_i(f_j)\| < \eta\|\pi_\infty(f_j)\|$$

for all j. Combining the above two inequalities, we obtain that for some f_j, $\|\pi_i(f_j)\| < \|\pi(f_j)\|$, so π is not in the support of π_i. ∎

[7] Integrable meaning having matrix coefficients in L^1; not all discrete series have this property.

3. Applications

3.1. An induction principle

The following theorem is at the heart of applications of strong convergence to spectral geometry.

Theorem 3.1. *Suppose G is locally compact and Γ is a cocompact lattice in G. If $\rho_i \xrightarrow{\text{strong}} \rho_\infty$, then*

$$\mathrm{Ind}_\Gamma^G \rho_i \xrightarrow{\text{strong}} \mathrm{Ind}_\Gamma^G \rho_\infty.$$

In applications, one usually wants $\rho_\infty = \lambda_\Gamma$ so that $\mathrm{Ind}_\Gamma^G \rho_i \xrightarrow{\text{strong}} \mathrm{Ind}_\Gamma^G \lambda_\Gamma = \lambda_G$. In light of Proposition 2.9, and in the case where G is a semisimple Lie group, the identification of the unitary dual with spectral parameters yields a type of spectral convergence of $\mathrm{Ind}_\Gamma^G \rho_i$ to the Plancherel measure of G.

The downside of this general argument is that

(a) it does not give an effective rate of convergence of spectral parameters,

(b) it does not apply as-is to non-uniform lattices.

Both these issues have been dealt with in special instances (see [28, 40] regarding point (a) and [28] regarding point (b)).

We now address the proof of Theorem 3.1. It relies on the following type of 'matrix amplification' that is well known in the literature; e.g., [22, Section 9]. We take the chance to record an effective version of this lemma.[8]

Lemma 3.2 (Effective matrix amplification). *Let A be a C^*-algebra and $x \in M_n(A)$. For every integer p,*

$$\|x\|_{M_n(A)} \in \left[1, n^{\frac{1}{2p}}\right] \max_i \|((x^*x)^p)_{i,i}\|_A^{\frac{1}{2p}}.$$

As a result, we obtain the following proposition.

Proposition 3.3. *If Γ is discrete and $\{\rho_i\}_{i=1}^\infty$ are a sequence of unitary representations of Γ with $\rho_i \xrightarrow{\text{strong}} \lambda_\Gamma$, then for all $r \in \mathbb{N}$ and all $z \in \mathbf{Mat}_{r\times r}(\mathbb{C}) \otimes_{\mathbb{C}} \mathbb{C}[\Gamma]$,*

$$\|[\mathrm{id} \otimes \rho_i](z)\| \to \|\mathrm{id} \otimes \lambda_\Gamma(z)\|.$$

Proof of Theorem 3.1. Given $f \in C_c(G)$, $\pi(f)$ acts on $L^2(G, \rho)$ by

$$\pi(f)[\phi](h) = \int_G f(g)\phi(g^{-1}h)d\mu(g) = \int_G f(hg)\phi(g^{-1})d\mu(g),$$

[8] I thank Mikael de la Salle for pointing this out to me.

where μ is the left-invariant Haar measure. This is the same as

$$\sum_{\gamma\in\Gamma}\int_F f(h\gamma g)\phi(g^{-1}\gamma^{-1})d\mu(g) = \sum_{\gamma\in\Gamma}\int_F f(h\gamma g)\rho(\gamma^{-1})\phi(g^{-1})d\mu(g)$$
$$= \sum_{\gamma}\rho(\gamma^{-1})[A_f(\gamma)\phi|_F],$$

where $A_f(\gamma) : L^2(F,V) \to L^2(F,V)$ such that if $\psi \in L^2(F,V)$ and $h \in F$,

$$A_f(\gamma)[\psi](h) \stackrel{\text{def}}{=} \int_F f(h\gamma g)\psi(g^{-1})dg. \tag{3.1}$$

Using $L^2(F,V) \cong L^2(F)\otimes V$, we obtain a unitary conjugacy

$$\pi(f) \cong \sum_{\gamma\in\Gamma} a_f(\gamma)\otimes\rho(\gamma^{-1}),$$

where $a_f(\gamma)$ is defined by the same formula (3.1) as A_f but acting on $L^2(F)$.

Now we make two observations. Firstly, $a_f(\gamma) = 0$ unless there are h and g in F such that $h\gamma g \in \mathrm{supp}(f)$. Let $K \subset G$ be compact such that

$$G = \bigcup_{\gamma\in\Gamma}\gamma C$$

and $F \subset C$. The above event is majorized by $\gamma \subset C^{-1}\mathrm{supp}(f)C^{-1}$ which is contained in a compact set. This compact subset can meet only finitely many elements of γ. So

$$\gamma \mapsto a_f(\gamma)$$

has finite support.

Secondly, $a_f(\gamma)$ is an integral operator with kernel $K_\gamma(h,g) = f(h\gamma g^{-1})$. Since f is continuous, and the closure of F is compact, it is bounded on F and even Hilbert–Schmidt.

So the image of this conjugacy is contained in

$$\mathbb{C}[\Gamma]\otimes_{\mathbb{C}}\mathrm{HS}(F),$$

where $\mathrm{HS}(F)$ are the Hilbert–Schmidt operators on $L^2(F)$ and hence generate a C^*-subalgebra of $C^*_{\mathrm{red}}(\Gamma)\otimes\mathcal{K}(L^2(F))$, where $\mathcal{K}$ are the compact operators on a separable Hilbert space.

Now by Proposition 3.3 together with approximation of compact operators by finite rank ones, we have for all $f \in C_c(G)$, with notation as above, that if $\pi_i = \mathrm{Ind}_\Gamma^G\rho_i$ and $\rho_i \xrightarrow{\text{strong}} \rho_\infty$, then

$$\|\pi_i(f)\| = \left\|\sum_{\gamma\in\Gamma} a_f(\gamma)\otimes\rho_i(\gamma^{-1})\right\| \to \left\|\sum_{\gamma\in\Gamma} a_f(\gamma)\otimes\lambda_\Gamma(\gamma^{-1})\right\| \cong \|\mathrm{Ind}_\Gamma^G\rho_\infty\|.$$

This proves Theorem 3.1. ∎

3.2. Hyperbolic surfaces

In this section, Γ is the fundamental group of a closed orientable surface of genus $g \geq 2$. Moreover, we discretely embed $\Gamma \hookrightarrow \mathrm{PSL}_2(\mathbb{R})$ in some fixed but arbitrary way, fixing a hyperbolic structure on a genus g surface $X \stackrel{\text{def}}{=} \Gamma\backslash\mathbb{H}$. Importantly, this embedding could be arithmetic.

By Corollary 2.5, there exists a sequence $\{\phi_i \in \mathrm{Hom}(\Gamma, S_{n_i})\}_{i=1}^{\infty}$ such that the induced representations $\rho_i = \mathrm{std} \circ \phi_i$ satisfy $\rho_i \xrightarrow{\text{strong}} \lambda_\Gamma$. Hence,

$$\mathrm{Ind}_\Gamma^G \rho_i \xrightarrow{\text{strong}} \lambda_{\mathrm{PSL}_2(\mathbb{R})}$$

by Theorem 3.1.

Proceeding depends on knowing the unitary dual of $\mathrm{PSL}_2(\mathbb{R})$ and the Plancherel measure. Of interest here are the complementary series which are outside the support of the Plancherel measure. Now Proposition 2.9 implies that for any compact subset K of the complementary series, for $i \gg_K 1$ no member of K is weakly contained in

$$\mathrm{Ind}_\Gamma^G \rho_i.$$

A note (on fibered products). We now make one more observation. Because ρ_i is derived from ϕ_i, the space of $\mathrm{Ind}_\Gamma^G \rho_i$ is the same as L^2 sections of the fibered product

$$\Gamma\backslash_{\phi_i}(G \times \ell_0^2([n_i])).$$

In turn, such sections are the same as L^2 functions on

$$\Gamma\backslash_{\phi_i}(G \times [n_i])$$

that have mean zero in every fiber (the above is a covering space of $\Gamma\backslash G$).

By the relation between the Casimir operator of $\mathrm{PSL}_2(\mathbb{R})$ and the Laplacian on the hyperbolic surface

$$X_{\phi_i} \stackrel{\text{def}}{=} \Gamma\backslash_{\phi_i}(\mathbb{H} \times [n_i]) = \Gamma\backslash_{\phi_i}(G \times [n_i])/P\mathrm{SO}(2),$$

one obtains the following conclusion.

Theorem 3.4. *As $i \to \infty$,*

$$\mathrm{spec}(\Delta_{X_{\phi_i}}) \cap \left[0, \frac{1}{4} - o(1)\right] = \mathrm{spec}(\Delta_X) \cap \left[0, \frac{1}{4} - o(1)\right].$$

By choosing X so that Δ_X has no eigenvalues below $\frac{1}{4}$, one obtains a sequence of closed hyperbolic surfaces (covering X) with genus tending to ∞ and first non-zero eigenvalue tending to $\frac{1}{4}$.[9]

Remark 3.5. This result, which established a conjecture of Buser [12], was first obtained in joint work of the author with Hide [28] by a related method. At this time, we only had P**Perm**F for free groups, so we worked with non-compact surfaces with free fundamental groups and compactified at the end of the argument following Buser–Burger–Dodziuk [13]. The problem this introduced was that we did not have access to Theorem 3.1, so we had to make a more involved argument using the resolvent of the Laplacian and cusp-patching techniques. This technique also yields the following theorem, taking Bordenave–Collins as input.

Theorem 3.6 (Hide–Magee). *Let X be a finite-area non-compact hyperbolic surface such that $\pi_1(X) \cong \mathbf{F}$ for some free group $\mathbf{F}$. Let ϕ_n now be a uniform random element of* $\mathrm{Hom}(\mathbf{F}, S_n)$. *Then X_{ϕ_n} is a uniform random degree-n covering space of X. With probability tending to one as $n \to \infty$,*

$$\mathrm{spec}(\Delta_{X_{\phi_n}}) \cap \left[0, \frac{1}{4} - o(1)\right] = \mathrm{spec}(\Delta_X) \cap \left[0, \frac{1}{4} - o(1)\right].$$

This theorem forms a part of much recent activity on the spectral gaps of random hyperbolic surfaces [3, 4, 27, 29, 31, 36–38, 51].

It is an interesting question to what extent the 'induction principle' obtained above in Theorem 3.1 can be extended to general non-cocompact and even infinite covolume lattices in, e.g., reductive groups. In recent work [14], it has been shown that induction of strong convergence works well in the setting of conformally compact hyperbolic surfaces (of infinite area), and even gives resonance-free regions – a phenomenon that cannot be seen solely in the representation theory of $\mathrm{Ind}_{\Gamma}^{\mathrm{PSL}_2(\mathbb{R})} \rho_i$. These questions should be pursued in future work.

Acknowledgments. This paper is a survey of some of my results with collaborators together with things I thought about during a stay at I.A.S. during 2023–2024. I am extremely grateful to the I.A.S. for this opportunity. Some particular elements of this year have had a large effect on this document. The conversations I had with Mikael de la Salle, who patiently explained many things to me. The seminar at

[9] In fact, one can arrange so that the original X is arithmetic and thus obtain the above conclusion where all surfaces are arithmetic. Elaborating on this (also using strong convergence as an essential ingredient), one can prove that *every* $x \in \left[0, \frac{1}{4}\right]$ is a limit point of λ_1 of arithmetic hyperbolic surfaces [33].

Princeton University organized by Peter Sarnak on 'C^*-algebras and related topics' was extremely stimulating and sparked many conversations. Indeed, I also learned many interesting things from subsequent conversations with Ramon van Handel and his lectures in the seminar.

Funding. This material is based upon work supported by the National Science Foundation under grant no. DMS-1926686. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 949143). This work was funded in part by a Philip Leverhulme prize awarded by the Leverhulme Trust.

References

[1] I. Agol, The virtual Haken conjecture (with an appendix by I. Agol, D. Groves and J. Manning). *Doc. Math.* **18** (2013), 1045–1087 Zbl 1286.57019 MR 3104553

[2] N. Alon, Eigenvalues and expanders. *Combinatorica* **6** (1986), no. 2, 83–96 Zbl 0661.05053 MR 0875835

[3] N. Anantharaman and L. Monk, Spectral gap of random hyperbolic surfaces. 2024, arXiv: 2403.12576v1, to appear in *Contemp. Math.*

[4] N. Anantharaman and L. Monk, Friedman–Ramanujan functions in random hyperbolic geometry and application to spectral gaps. [v1] 2023, [v4] 2026, arXiv:2304.02678v4

[5] G. Baumslag, On generalised free products. *Math. Z.* **78** (1962), 423–438 Zbl 0104.24402 MR 0140562

[6] B. Bekka, P. de la Harpe, and A. Valette, *Kazhdan's property (T)*. New Math. Monogr. 11, Cambridge University Press, Cambridge, 2008, 472 pp. Zbl 1146.22009 MR 2415834

[7] N. Bergeron, F. Haglund, and D. T. Wise, Hyperplane sections in arithmetic hyperbolic manifolds. *J. Lond. Math. Soc. (2)* **83** (2011), no. 2, 431–448 Zbl 1236.57021 MR 2776645

[8] N. Bergeron and D. T. Wise, A boundary criterion for cubulation. *Amer. J. Math.* **134** (2012), no. 3, 843–859 Zbl 1279.20051 MR 2931226

[9] B. Blackadar and E. Kirchberg, Generalized inductive limits of finite-dimensional C*-algebras. *Math. Ann.* **307** (1997), no. 3, 343–380 Zbl 0874.46036 MR 1437044

[10] C. Bordenave and B. Collins, Eigenvalues of random lifts and polynomials of random permutation matrices. *Ann. of Math. (2)* **190** (2019), no. 3, 811–875 Zbl 1446.60004 MR 4024563

[11] E. Breuillard, M. Kalantar, M. Kennedy, and N. Ozawa, C*-simplicity and the unique trace property for discrete groups. *Publ. Math. Inst. Hautes Études Sci.* **126** (2017), 35–71 Zbl 1391.46071 MR 3735864

[12] P. Buser, Cubic graphs and the first eigenvalue of a Riemann surface. *Math. Z.* **162** (1978), no. 1, 87–99 Zbl 0371.53032 MR 0505920

[13] P. Buser, M. Burger, and J. Dodziuk, Riemann surfaces of large genus and large λ_1. In T. Sunada (Ed.), *Geometry and analysis on manifolds (Katata/Kyoto, 1987)*, pp. 54–63, Lecture Notes in Math. 1339, Springer, Berlin, 1988 Zbl 0646.53040 MR 0961472

[14] I. Calderón, M. Magee, and F. Naud, Spectral gap for random Schottky surfaces. 2024, arXiv:2407.21506v1, to appear in *Anal. PDE*

[15] E. Cassidy, Projection formulas and a refinement of Schur–Weyl–Jones duality for symmetric groups. [v1] 2023, [v2] 2025, arXiv:2312.01839v2

[16] E. Cassidy, Random permutations acting on k-tuples have near-optimal spectral gap for $k = \mathrm{poly}(n)$. [v1] 2024, [v3] 2025, arXiv:2412.13941v3

[17] C. Champetier and V. Guirardel, Limit groups as limits of free groups. *Israel J. Math.* **146** (2005), no. 1, 1–75 Zbl 1103.20026 MR 2151593

[18] C.-F. Chen, J. Garza-Vargas, J. A. Tropp, and R. van Handel, A new approach to strong convergence. *Ann. of Math. (2)* **203** (2026), no. 2, 555–602 MR 5037839

[19] B. Collins and C. Male, The strong asymptotic freeness of Haar and deterministic matrices. *Ann. Sci. Éc. Norm. Supér. (4)* **47** (2014), no. 1, 147–163 Zbl 1303.15043 MR 3205602

[20] J. Friedman, A proof of Alon's second eigenvalue conjecture and related problems. *Mem. Amer. Math. Soc.* **195** (2008), no. 910, 100 pp. Zbl 1177.05070 MR 2437174

[21] U. Haagerup, An example of a nonnuclear C^*-algebra, which has the metric approximation property. *Invent. Math.* **50** (1978), no. 3, 279–293 Zbl 0408.46046 MR 0520930

[22] U. Haagerup and S. Thorbjørnsen, A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. *Ann. of Math. (2)* **162** (2005), no. 2, 711–775 Zbl 1103.46032 MR 2183281

[23] F. Haglund and D. T. Wise, Special cube complexes. *Geom. Funct. Anal.* **17** (2008), no. 5, 1551–1620 Zbl 1155.53025 MR 2377497

[24] F. Haglund and D. T. Wise, Coxeter groups are virtually special. *Adv. Math.* **224** (2010), no. 5, 1890–1903 Zbl 1195.53055 MR 2646113

[25] M. Hall, Jr., *The theory of groups*. Reprint of the 1959 original published by Macmillan Company ed., Dover Books Math., Dover, Mineola, NY, 2018, 434 pp. Zbl 0084.02202

[26] B. Hayes, A random matrix approach to the Peterson-Thom conjecture. *Indiana Univ. Math. J.* **71** (2022), no. 3, 1243–1297 Zbl 1503.46049 MR 4448584

[27] W. Hide, Spectral gap for Weil–Petersson random surfaces with cusps. *Int. Math. Res. Not. IMRN* **2023** (2023), no. 20, 17411–17460 Zbl 1548.58016 MR 4659855

[28] W. Hide and M. Magee, Near optimal spectral gaps for hyperbolic surfaces. *Ann. of Math. (2)* **198** (2023), no. 2, 791–824 Zbl 1553.58014 MR 4635304

[29] W. Hide and J. Thomas, Large-n asymptotics for Weil–Petersson volumes of moduli spaces of bordered hyperbolic surfaces. *Comm. Math. Phys.* **406** (2025), no. 9, article no. 203, 41 pp. Zbl 08083383 MR 4940193

[30] M. Kassabov, Symmetric groups and expander graphs. *Invent. Math.* **170** (2007), no. 2, 327–354 Zbl 1191.20002 MR 2342639

[31] M. Lipnowski and A. Wright, Towards optimal spectral gaps in large genus. *Ann. Probab.* **52** (2024), no. 2, 545–575 Zbl 1537.58004 MR 4718401

[32] L. Louder and M. Magee, Strongly convergent unitary representations of limit groups (with an appendix by W. Hide and M. Magee). *J. Funct. Anal.* **288** (2025), no. 6, article no. 110803, 28 pp. Zbl 1560.22015 MR 4847195

[33] M. Magee, The limit points of the bass notes of arithmetic hyperbolic surfaces. 2024, arXiv:2403.00928v2, to appear in *Amer. J. Math.*

[34] M. Magee and M. de la Salle, $\mathrm{SL}_4(\mathbf{Z})$ is not purely matricial field. *C. R. Math. Acad. Sci. Paris* **362** (2024), 903–910 Zbl 1550.20010 MR 4805518

[35] M. Magee and M. de la Salle, Strong asymptotic freeness of Haar unitaries in quasi-exponential dimensional representations. *Geom. Funct. Anal.* **36** (2026), no. 1, 152–200 Zbl 08180053 MR 5033337

[36] M. Magee and F. Naud, Explicit spectral gaps for random covers of Riemann surfaces. *Publ. Math. Inst. Hautes Études Sci.* **132** (2020), no. 1, 137–179 Zbl 1508.58008 MR 4179833

[37] M. Magee, F. Naud, and D. Puder, A random cover of a compact hyperbolic surface has relative spectral gap $\frac{3}{16} - \varepsilon$. *Geom. Funct. Anal.* **32** (2022), no. 3, 595–661 Zbl 1498.58022 MR 4431124

[38] M. Magee and D. Puder, The asymptotic statistics of random covering surfaces. *Forum Math. Pi* **11** (2023), article no. e15, 51 pp. Zbl 1521.20016 MR 4591384

[39] M. Magee, D. Puder, and R. van Handel, Strong convergence of uniformly random permutation representations of surface groups. 2025, arXiv:2504.08988v2

[40] M. Magee and J. Thomas, Strongly convergent unitary representations of right-angled Artin groups. *Duke Math. J.* **175** (2026), no. 7, 1305–1343 Zbl 08207151 MR 5076493

[41] A. Malcev, On isomorphic matrix representations of infinite groups (in Russian). *Rec. Math. [Mat. Sbornik] N.S.* **8(50)** (1940), no. 3, 405–422 Zbl 0025.00804 MR 0003420

[42] A. Nilli, On the second eigenvalue of a graph. *Discrete Math.* **91** (1991), no. 2, 207–210 Zbl 0771.05064 MR 1124768

[43] J. Peterson and A. Thom, Group cocycles and the ring of affiliated operators. *Invent. Math.* **185** (2011), no. 3, 561–592 Zbl 1227.22003 MR 2827095

[44] Z. Sela, Diophantine geometry over groups I: Makanin–Razborov diagrams. *Publ. Math. Inst. Hautes Études Sci.* **93** (2001), no. 1, 31–105 Zbl 1018.20034 MR 1863735

[45] A. Song, Hyperbolic groups and spherical minimal surfaces. 2024, arXiv:2402.10869v1

[46] A. Song, Random harmonic maps into spheres. [v1] 2024, [v3] 2025, arXiv:2402.10287v3

[47] A. Tikuisis, S. White, and W. Winter, Quasidiagonality of nuclear C*-algebras. *Ann. of Math. (2)* **185** (2017), no. 1, 229–284 Zbl 1367.46044 MR 3583354

[48] D. Voiculescu, Around quasidiagonal operators. *Integral Equations Operator Theory* **17** (1993), no. 1, 137–149 Zbl 1048.47501 MR 1220578

[49] D. T. Wise, *The structure of groups with a quasiconvex hierarchy*. Ann. of Math. Stud. 209, Princeton University Press, Princeton, NJ, 2021, 357 pp. Zbl 1511.20002 MR 4298722

[50] N. C. Wormald, Models of random regular graphs. In J. D. Lamb and D. A. Preece (Eds.), *Surveys in combinatorics, 1999 (Canterbury)*, pp. 239–298, London Math. Soc. Lecture Note Ser. 267, Cambridge University Press, Cambridge, 1999 Zbl 0935.05080 MR 1725006

[51] Y. Wu and Y. Xue, Random hyperbolic surfaces of large genus have first eigenvalues greater than $\frac{3}{16} - \epsilon$. *Geom. Funct. Anal.* **32** (2022), no. 2, 340–410 Zbl 1487.32072 MR 4408435

Michael Magee
Department of Mathematical Sciences, Durham University, Lower Mountjoy, Durham DH1 3LE, UK; michael.r.magee@durham.ac.uk

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/24

The Poisson–Dirichlet problem in domains with Ahlfors regular boundary

Ariel Barton, Svitlana Mayboroda, and Alberto Pacati

Abstract. We present an announcement of some recent results concerning well-posedness of the Poisson–Dirichlet problem with boundary data in Besov spaces with fractional smoothness. This is a far-reaching generalization as previously known theorems concerning well-posedness of the Poisson problem in such intermediate smoothness classes were mostly restricted to the context of Lipschitz domains and coefficients satisfying strong regularity assumptions.

1. Introduction: The general context

The purpose of this note is to describe some recent results concerning the Poisson problem, and the Dirichlet problem with data of fractional smoothness, for second-order elliptic operators with real coefficients in very general domains.

Poisson's equation $-\Delta u = h$, for a given function h and an unknown function u, was historically one of the first differential equations to be studied, due in large part to its many applications in physics and beyond. A natural generalization is to replace the Laplace operator $-\Delta$ by an arbitrary divergence form elliptic operator $-\operatorname{div} A\nabla$ for some specified matrix A (which may be only bounded, measurable, and uniformly positive definite). Considering Poisson's equation not in all of Euclidean space, but only in a given open subset Ω, requires specifying the behavior of u on the boundary $\partial\Omega$; we will impose Dirichlet boundary conditions, that is, specify the values of u on the boundary. (It is also common in the theory to specify Neumann boundary conditions, that is, to specify $\nu \cdot A\nabla u$, where ν denotes the unit outward normal vector to Ω; we refer the interested reader to [3, 7, 26, 69] for some results on the Poisson–Neumann problem.)

We thus will discuss the Poisson–Dirichlet problem

$$-\operatorname{div} A\nabla u = h \quad \text{in } \Omega, \qquad u = f \quad \text{on } \partial\Omega \tag{1.1}$$

Mathematics Subject Classification 2020: 35J25 (primary); 35A01, 35A02 (secondary).
Keywords: Poisson–Dirichlet problem, Besov spaces, Ahlfors regular boundary.

for specified Ω, A, f, and h.

In a few situations, the range of spaces allowing for well-posedness of the Poisson–Dirichlet boundary problem above is very well understood. Such situations include VMO and small-BMO coefficients and domains (see [51, 68], among others), and Lipschitz domains with coefficients satisfying a Dini-type continuity condition [56] or independent of the direction transverse to the boundary [7].

In recent years, endpoint results, pertaining to boundary data f lying in $L^p(\partial\Omega)$, have been established considerably beyond Lipschitz domains: first, in chord-arc domains [22, 42], then in so-called 1-sided nontangentially accessible domains with a uniformly rectifiable boundary [37], and finally, on domains with a uniformly rectifiable boundary that satisfy a weak local John condition [5]. The latter is a sharp condition, necessary and sufficient for well-posedness results for the Laplacian [5]. Finally, boundary value problems in domains with lower-dimensional boundaries were treated in [24] (a later but much simpler treatment is in [28]).

Most of these results were proven in (or may be generalized to) the generality of coefficients satisfying variations of the so-called DKP condition, which controls oscillations of the coefficients near the boundary.

Note that t-independence is meaningless in domains that lack a distinguished direction (in particular, in domains where there is no constant direction transverse to the boundary) and that all of these "endpoint" estimates are in spaces with integer smoothness. Our results, by contrast with the endpoint estimates listed above, need neither t-independence nor the DKP condition.

The results of this note (described in Sections 4 and 5) treat well-posedness of boundary problems in intermediate spaces of fractional smoothness. We will show, for the first time, that well-posedness in a wide range of Besov spaces can be established by a "blind" mechanism, in the full generality of domains with Ahlfors regular boundary (possibly of lower codimension) that satisfy the weak local John condition, and general real elliptic equations, without any additional restrictions on the coefficients. Furthermore, given endpoint (that is, $L^p(\partial\Omega)$) results such as those above, the range of Besov spaces can be expanded in, again, a "blind" fashion, with no further requirements on the domains or coefficients beyond those above and those required by the endpoint results.

The methods are new, as [7, 49] relied on layer potentials, not available in such a general setting, [41] worked with the Laplacian on Lipschitz domains and the case $p > 1$ only, and [51, 56, 68] were confined to continuous, VMO, or near-VMO coefficients and Lipschitz domains. The full proofs will appear in [9] (the case $f = 0$, that is, $u = 0$ on $\partial\Omega$) and [8] (full generality), and in this note, we only announce the upcoming results and put them into historic context.

We remark that $L^p(\partial\Omega)$ represents the "zero smoothness" endpoint; an alternative endpoint is the "smoothness 1" endpoint, often called the regularity problem. The

regularity problem, like the L^p-Dirichlet problem, has been investigated extensively. However, our results will not make use of the regularity problem, but only the L^p-Dirichlet problem and a duality argument, to expand the range of Besov spaces. For this reason, we will not discuss known results for the regularity problem.

2. Definitions

In this section, we will provide all of the definitions needed to state our main results. An experienced reader can skip them and move on to Section 3. Some additional definitions are necessary to state the corollaries to our second main result and also to discuss the history of the Poisson problem and the Dirichlet problem with boundary data in fractional smoothness spaces; these definitions will be deferred to Section 6 or to the discussion.

2.1. Geometric conditions

Definition 2.1 (Ahlfors regularity). If $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ and $0 < \mathfrak{d} < \mathfrak{n}$, then we say that the boundary $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular if there exists a constant $\mathcal{A} > 0$ such that

$$\frac{1}{\mathcal{A}} r^{\mathfrak{d}} \leq \sigma(B(x,r) \cap \partial\Omega) \leq \mathcal{A} r^{\mathfrak{d}}$$

for all $x \in \partial\Omega$ and all r with $0 < r < \operatorname{diam}(\Omega)$; here, σ denotes the Hausdorff $\mathfrak{d}$-dimensional surface measure.

Remark 2.2. Following [20], we do not require the boundary dimension $\mathfrak{d}$ to be an integer. Our main results (Theorems 4.1 and 5.1) are known only in the case $\mathfrak{d} \leq \mathfrak{n} - 1$, that is, the classical codimension 1 case and the higher codimension case.

Definition 2.3 (Interior corkscrew condition). We say that Ω satisfies the interior corkscrew condition if there is a constant $\varkappa \in (0,1)$ such that if $x \in \partial\Omega$ and $0 < r < \operatorname{diam}\Omega$, then there is a point $A_r(x)$ (called a corkscrew point) such that

$$A_r(x) \in B(x,r) \quad \text{and} \quad B(A_r(x), \varkappa r) \subset \Omega.$$

Definition 2.4 (Weak local John domain). Let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ be open. Let $x \in \partial\Omega$, and let $y \in \Omega$. We say that ζ is a λ-carrot path from x to y if ζ is a connected rectifiable path, which we require to be parameterized by arc length, such that $\zeta : [0,T] \to \Omega \cup \{x\}$ is continuous, $\zeta(0) = x$, $\zeta(T) = y$, and $\operatorname{dist}(\zeta(t), \partial\Omega) \geq \lambda t$ for all $t \in [0,T]$.

Now, suppose in addition that $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular for some $0 < \mathfrak{d} < \mathfrak{n}$. Let $y \in \Omega$, let θ, $\lambda \in (0,1]$, and let $N \geq 2$. We say that y is a (θ, λ, N)-weak

local John point if there is a Borel set $F \subset B(y, N\,\mathrm{dist}(y, \partial\Omega)) \cap \partial\Omega$ such that $\sigma(F) \geq \theta\sigma(B(y, N\,\mathrm{dist}(y, \partial\Omega)) \cap \partial\Omega)$ and such that if $x \in F$, then there is a λ-carrot path from x to y.

We say that Ω is a weak local John domain if there exist constants $\theta \in (0, 1]$, $\lambda \in (0, 1]$, and $N \geq 2$ such that every $y \in \Omega$ is a (θ, λ, N)-weak local John point.

Remark 2.5. A domain with lower-dimensional Ahlfors regular boundary (that is, with $\mathfrak{d} < \mathfrak{n} - 1$) necessarily satisfies the interior corkscrew condition (see [20, Lemma 11.6]). It also necessarily satisfies the weak local John condition: by [20, Lemma 2.1], such a domain must satisfy the Harnack chain condition (defined in many places in the literature, or see Definition 6.2 below), and by connecting corkscrew points with Harnack chains, we can construct carrot paths to a point y in the domain from *any* nearby boundary point x.

2.2. Functions and function spaces

Definition 2.6. Let $\Gamma \subseteq \mathbb{R}^{\mathfrak{n}}$ be a $\mathfrak{d}$-Ahlfors regular set, let $0 < s < 1$, and let $1 \leq p \leq \infty$. The Slobodeckij space $\dot{\Lambda}^{s,p}(\Gamma)$ is defined to be the set of all equivalence classes (up to addition of constants and sets of measure zero) of locally integrable functions $f : \Gamma \to \mathbb{R}$ such that the norm

$$\|f\|_{\dot{\Lambda}^{s,p}(\Gamma)} = \left(\int_\Gamma \int_\Gamma \frac{|f(x) - f(y)|^p}{|x - y|^{\mathfrak{d}+ps}}\, d\sigma(y)\, d\sigma(x) \right)^{1/p}$$

(or the corresponding sup norm when $p = \infty$) is finite.

It is clear that the Slobodeckij spaces are Banach spaces.

Definition 2.7 ([31, Definition 2.1], [33, Definition 7]). Let $\Gamma \subset \mathbb{R}^{\mathfrak{n}}$ be a $\mathfrak{d}$-Ahlfors regular set, let $z \in \Gamma$, and let $\varrho > 0$. We let $\Delta(z, \varrho) = \Gamma \cap B(z, \varrho)$. A measurable function $a : \Gamma \to \mathbb{R}$ is said to be a *block in* $\Delta(z, \varrho)$ if

- a is supported in $\Delta(z, \varrho)$,
- $|a(x)| \leq \varrho^{-\mathfrak{d}/2}$ for all $x \in \Delta(z, \varrho)$ (and thus for all $x \in \Gamma$), and
- $|a(x) - a(y)| \leq |x - y|\varrho^{-\mathfrak{d}/2-1}$ for all $x, y \in \Gamma$.

Remark 2.8. In the setting of Hardy spaces, it is customary to work with atoms, which are blocks with the additional requirement $\int_\Gamma a\, d\sigma = 0$. However, the Besov spaces are closer in character to the Hardy–Sobolev spaces, that is, functions whose gradient lies in a Hardy space. The moment condition $\int_{\mathbb{R}^{\mathfrak{d}}} \nabla a = 0$ is trivially true for any compactly supported function a, and so moment conditions are not usually required of Hardy–Sobolev atoms; thus, we do not require them of our blocks.

Definition 2.9. Let $\Gamma \subset \mathbb{R}^{\mathfrak{n}}$ be a $\mathfrak{d}$-Ahlfors regular set, let $0 < s \leq 1$, and let $\mathfrak{d}/(\mathfrak{d}+s) < p \leq 1$ (so $s + \mathfrak{d} - \mathfrak{d}/p > 0$). We define the atomic space $\dot{A}^{s,p}(\Gamma)$ as

$$\dot{A}^{s,p}(\Gamma) = \left\{ \sum_{k=1}^{\infty} \lambda_k a_k : \sum_{k=1}^{\infty} |\lambda_k|^p \varrho_k^{\mathfrak{d}-\mathfrak{d}p/2-sp} < \infty, \right.$$
$$\left. \text{where each } a_k \text{ is a block in } \Delta(z_k, \varrho_k) \text{ for some } z_k \in \Gamma \right\},$$

and let the norm in $\dot{A}^{s,p}(\Gamma)$ be given by

$$\|f\|_{\dot{A}^{s,p}(\Gamma)} = \inf\left\{ \left(\sum_{k=1}^{\infty} |\lambda_k|^p \varrho_k^{\mathfrak{d}-\mathfrak{d}p/2-sp} \right)^{1/p} : f = \sum_{k=1}^{\infty} \lambda_k a_k \right.$$
$$\left. \text{for blocks } a_k \text{ in } \Delta(z_k, \varrho_k) \right\}.$$

It is straightforward to establish (and we will do so in the forthcoming paper [8]) that if $\sum_{k=1}^{\infty} |\lambda_k|^p \varrho_k^{\mathfrak{d}-\mathfrak{d}p/2-sp} < \infty$, then $\sum_{k=1}^{\infty} \lambda_k a_k$ converges in the Banach space $\dot{\Lambda}^{s+\mathfrak{d}-\mathfrak{d}/p,1}(\Gamma)$, and in fact $\dot{A}^{s,p}(\Gamma)$ embeds into $\dot{\Lambda}^{s+\mathfrak{d}-\mathfrak{d}/p,1}(\Gamma)$. Thus, the infinite sums above are meaningful.

Definition 2.10. Let $\Gamma \subset \mathbb{R}^{\mathfrak{n}}$ be a $\mathfrak{d}$-Ahlfors regular set for some $0 < \mathfrak{d} < \mathfrak{n}$, let $0 < s \leq 1$, and let $\mathfrak{d}/(\mathfrak{d}+s) < p \leq \infty$.

We define the Besov space $\dot{B}_s^{p,p}(\Gamma)$ by

$$\dot{B}_s^{p,p}(\Gamma) = \begin{cases} \dot{\Lambda}^{s,p}(\Gamma), & 1 \leq p \leq \infty, \\ \dot{A}^{s,p}(\Gamma), & \dfrac{\mathfrak{d}}{\mathfrak{d}+s} < p < 1, \end{cases}$$

with the norm in $\dot{B}_s^{p,p}(\Gamma)$ given by the norm in the Slobodeckij space $\dot{\Lambda}^{s,p}(\Gamma)$ or atomic space $\dot{A}^{s,p}(\Gamma)$.

In Euclidean spaces, the Besov spaces $\dot{B}_s^{p,q}(\mathbb{R}^{\mathfrak{d}})$ and the related Triebel–Lizorkin spaces $\dot{F}_s^{p,q}(\mathbb{R}^{\mathfrak{d}})$ (they are equal if $p = q$) encompass many known function spaces. See [65, Sections 2.3.5, 5.2.3, 5.2.4] and [52, Proposition 3]. For example, the Lebesgue spaces satisfy $L^p(\mathbb{R}^{\mathfrak{d}}) = \dot{F}_0^{p,2}(\mathbb{R}^{\mathfrak{d}})$ if $1 < p < \infty$; the Sobolev spaces $\dot{W}^{k,p}(\mathbb{R}^{\mathfrak{d}})$ (of functions f whose kth gradient $\nabla^k f$ lies in $L^p(\mathbb{R}^{\mathfrak{d}})$) satisfy $\dot{W}^{k,p}(\mathbb{R}^{\mathfrak{d}}) = \dot{F}_k^{p,2}(\mathbb{R}^{\mathfrak{d}})$; the Hardy spaces satisfy $H^p(\mathbb{R}^{\mathfrak{d}}) = \dot{F}_0^{p,2}(\mathbb{R}^{\mathfrak{d}})$ if $\mathfrak{d}/\mathfrak{d}+1 < p < \infty$; the spaces $\dot{C}^{\alpha}(\mathbb{R}^{\mathfrak{d}})$ of Hölder continuous functions satisfy $\dot{C}^{\alpha}(\mathbb{R}^{\mathfrak{d}}) = \dot{B}_{\alpha}^{\infty,\infty}(\mathbb{R}^{\mathfrak{d}})$ for $0 < \alpha < 1$; the Slobodeckij spaces $\dot{\Lambda}^{s,p}$ satisfy the above relation $\dot{B}_s^{p,p}(\mathbb{R}^{\mathfrak{d}}) = \dot{\Lambda}^{s,p}(\mathbb{R}^{\mathfrak{d}})$ for $0 < s < 1$ and $1 \leq p \leq \infty$ (and indeed

there exists a Slobodeckij-type characterization of $\dot{B}^{p,q}_s(\mathbb{R}^{\mathfrak{b}})$ for any $0 < s < 1$, any $1 \leq p < \infty$, and any $1 \leq q \leq \infty$, not only for $q = p$), and the Bessel potential spaces $L^p_s = (I - \Delta)^{-s/2}(L^p)$ satisfy $L^p_s = F^{p,2}_s$. Even with the restrictions $p = q$, $0 < s < 1$, and $\mathfrak{b}/\mathfrak{b} + s < p \leq \infty$, there are many equivalent characterizations of the Besov spaces $\dot{B}^{p,p}_s(\mathbb{R}^{\mathfrak{b}})$.

In [34], Besov spaces were generalized from the Euclidean setting to the setting of spaces of homogeneous type. An active area of research consists of studying these spaces; we mention in particular that by [59], our Slobodeckij characterization of the Besov spaces for $p \geq 1$ coincides with the characterization in [32,34], and by [33], the *inhomogeneous* Besov spaces have an atomic characterization; it is natural to expect the homogeneous Besov spaces to have an atomic characterization (specifically, the atomic characterization used above), but we have not been able to find this result in the literature on spaces of homogeneous type.

2.3. Boundary values

Definition 2.11 (Boundary trace). Following [43], if $\Omega \subset \mathbb{R}^{\mathfrak{n}}$, if $u : \Omega \to \mathbb{R}$ is locally integrable, and if $f : \partial\Omega \to \mathbb{R}$, then we say that $u = f$ on $\partial\Omega$ if

$$\lim_{r \to 0^+} \frac{1}{r^{\mathfrak{n}}} \int_{B(x,r) \cap \Omega} |u - f(x)| = 0 \tag{2.1}$$

for σ-almost every $x \in \partial\Omega$.

In most reasonably behaved domains, this coincides with most other notions of trace. We will use this definition in Theorems 4.1 and 5.1. We mention that some of the results described in Section 3 were proven under different notions of trace (most prominently traces in the sense of Sobolev spaces); however, the different notions of trace can generally be shown to be equivalent as needed.

We observe that if Ω satisfies the interior corkscrew condition, then the value of f is uniquely determined in that, for each $x \in \partial\Omega$, there can be at most one $f(x) \in \mathbb{R}$ such that formula (2.1) is valid.

2.4. Elliptic differential operators

Definition 2.12 (Ellipticity). Let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ and suppose that $\partial\Omega$ is $\mathfrak{b}$-Ahlfors regular. Let A be a matrix-valued function. We say that the operator $-\operatorname{div} A\nabla$ is elliptic, or, respectively, the matrix A is elliptic, if each $A(x)$ is an $\mathfrak{n} \times \mathfrak{n}$ matrix of real, not necessarily symmetric, coefficients that satisfy the following positive definiteness and boundedness conditions:

$$\xi \cdot A(x)\xi \geq \lambda \operatorname{dist}(x, \partial\Omega)^{\mathfrak{b}+1-\mathfrak{n}} \|\xi\|^2, \tag{2.2}$$

$$|\eta \cdot A(x)\xi| \leq \Lambda \operatorname{dist}(x, \partial\Omega)^{\mathfrak{b}+1-\mathfrak{n}} \|\xi\| \, \|\eta\| \tag{2.3}$$

for all $\eta, \xi \in \mathbb{R}^{\mathfrak{n}}$ and almost every $x \in \mathbb{R}^{\mathfrak{n}} \setminus \partial\Omega$.

The main results (Theorems 4.1 and 5.1) of the present paper hold in the full generality of real elliptic matrices, with no further restrictions on A.

Definition 2.13 (Weak solutions). We will consider solutions to $-\operatorname{div} A\nabla u = h$ in the case where h is the divergence of some vector field $\vec{G}$.

We will use the weak definition of $-\operatorname{div} A\nabla u = -\operatorname{div} \vec{G}$; that is, if $u : \Omega \to \mathbb{R}$ is a real-valued function, we say that $-\operatorname{div} A\nabla u = -\operatorname{div} \vec{G}$ in Ω in the weak sense if, for every $\varphi : \Omega \to \mathbb{R}$ that is smooth and supported in a compact subset of Ω, we have that

$$\int_\Omega \nabla\varphi \cdot A\nabla u = \int_\Omega \nabla\varphi \cdot \vec{G}.$$

Recall that A is elliptic and therefore invertible. For notational convenience, we will usually take our data to be $\vec{H} = A^{-1}\vec{G}$; thus, we consider solutions to $-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H}$, meaning functions u that satisfy

$$\int_\Omega \nabla\varphi \cdot A\nabla u = \int_\Omega \nabla\varphi \cdot A\vec{H} \quad \text{for all } \varphi \in C_0^\infty(\Omega). \tag{2.4}$$

This notational convention will considerably simplify the statement of Theorem 4.1 below.

We now state three known facts concerning solutions to $-\operatorname{div} A\nabla u = 0$ or $-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H}$ that we will reference in the statement of our main results.

First, for any domain Ω with $\mathfrak{d}$-Ahlfors regular boundary, $0 < \mathfrak{d} \leq \mathfrak{n} - 1$, and for any real elliptic operator $-\operatorname{div} A\nabla$ as above, for every function f defined on $\partial\Omega$ that is continuous and compactly supported, there is a unique solution u to the Dirichlet problem

$$\begin{cases} -\operatorname{div} A\nabla u = 0 & \text{in } \Omega, \\ u = f & \text{on } \partial\Omega, \\ u \text{ is continuous and bounded on } \overline{\Omega}. \end{cases} \tag{2.5}$$

That this holds for $\mathfrak{d} = \mathfrak{n} - 1$ is the fact that Ahlfors regular boundaries satisfy the classic Wiener criterion [67]; that it holds for $\mathfrak{d} < \mathfrak{n} - 1$ is [20, Lemma 9.4].

The second result we wish to mention is Meyers's reverse Hölder estimate for gradients. This estimate states that if A satisfies conditions (2.2)–(2.3), then there exist C and $\varepsilon > 0$ depending only on the dimension $\mathfrak{n}$ (and $\mathfrak{d}$) and on the ellipticity constants λ and Λ such that if $-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H}$ in $B(x,r) \subseteq \Omega$, and if $\vec{H} \in L^{2+\varepsilon}(B(x,r))$, then $\nabla u \in L^{2+\varepsilon}(B(x,r/2))$ and

$$\left(\fint_{B(x,r/2)} |\nabla u|^{2+\varepsilon}\right)^{1/(2+\varepsilon)} \leq C\left(\fint_{B(x,r)} |\nabla u|^2\right)^{1/2} + C\left(\fint_{B(x,r)} |\vec{H}|^{2+\varepsilon}\right)^{1/(2+\varepsilon)}. \tag{2.6}$$

The same of course is true for solutions to $-\operatorname{div} A^*\nabla u = -\operatorname{div} A^*\vec{H}$, possibly with a different value of ε.

Finally, we wish to state the boundary De Giorgi–Nash estimate. This estimate is well known in the case $\mathfrak{d} = \mathfrak{n} - 1$ and is still true if $0 < \mathfrak{d} < \mathfrak{n} - 1$; see [20, Lemmas 8.13 and 8.16]. It states that there exist an $\alpha > 0$ and $C > 0$ such that if $-\operatorname{div} A\nabla u = 0$ in Ω, if $u = 0$ on $\partial\Omega \cap B(\xi, r)$ for some $\xi \in \partial\Omega$ and some $r > 0$, and if $x \in B(\xi, r/2) \cap \Omega$, then

$$|u(x)| \leq \left(\frac{|x-\xi|}{r}\right)^{\alpha} \frac{C}{r^{\mathfrak{d}+1}} \int_{\Omega\cap B(\xi,r)} |u(y)| \operatorname{dist}(y, \partial\Omega)^{1+\mathfrak{d}-\mathfrak{n}}\, dy. \tag{2.7}$$

3. A brief history of the Poisson–Dirichlet problem in spaces with fractional smoothness

In this section, we will briefly review the history of the problems which we consider in this note. That is, we will review the history of the Poisson–Dirichlet problem, with one degree of smoothness for the solutions and between zero and one degree of smoothness of the boundary values, for second-order linear elliptic differential equations, with real coefficients and without lower-order terms, in domains in Euclidean space.

3.1. Higher codimension and energy solutions

Let $0 < \mathfrak{d} < \mathfrak{n}$, $\mathfrak{n} \geq 2$, let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ be such that $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular for some real number $\mathfrak{d}$ with $0 < \mathfrak{d} < \mathfrak{n}$, and let A satisfy the ellipticity conditions (2.2)–(2.3). The Lax–Milgram lemma yields unique solvability of the Poisson problem

$$-\operatorname{div} A\nabla u_0 = -\operatorname{div} A\vec{H}_0 \quad \text{in } \Omega, \qquad u_0 = 0 \quad \text{on } \partial\Omega \tag{3.1}$$

for all $\vec{H}_0 \in L^2(\Omega, w)$ and with the estimate

$$\|\nabla u_0\|_{L^2(\Omega,w)} \leq C\,\|\vec{H}_0\|_{L^2(\Omega,w)},$$

where $u_0 = 0$ on $\partial\Omega$ in the sense that u_0 lies in $\dot{W}_0^{1,2}(\Omega; w)$, the closure of $C_0^\infty(\Omega)$ in $\dot{W}^{1,2}(\Omega; w)$ (the Sobolev space of functions with gradient in $L^2(\Omega; w)$), and where w is the measure given by $dw = \operatorname{dist}(\cdot, \partial\Omega)^{\mathfrak{d}+1-\mathfrak{n}}\, dm$, m denoting Lebesgue measure. This method of solution is extremely general, relying only on the fact that $\dot{W}_0^{1,2}(\Omega; w)$ is a Hilbert space and on the upper and lower bounds

$$\int_\Omega \nabla\varphi \cdot A\nabla\psi \leq \Lambda\,\|\varphi\|_{\dot{W}_0^{1,2}(\Omega;w)}\|\psi\|_{\dot{W}_0^{1,2}(\Omega;w)}, \qquad \int_\Omega \nabla u \cdot A\nabla u \geq \lambda\|u\|^2_{\dot{W}_0^{1,2}(\Omega;w)}$$

on functions φ, ψ, $u \in \dot{W}_0^{1,2}(\Omega; w)$. Indeed, this construction is valid in many situations beyond that of this paper, such as systems and higher-order equations.

The estimate

$$\|\nabla u_0\|_{L^p(\Omega,w)} \leq C \|\vec{H}_0\|_{L^p(\Omega,w)} \tag{3.2}$$

is also true for all p sufficiently close to 2. This can be derived by using the fact that $\dot{W}_0^{1,p}(\Omega; w)$, $1 < p < \infty$, is an interpolation scale and using Šneĭberg's lemma.

Problem (1.1), with the notation of formula (2.4), becomes

$$-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H} \quad \text{in } \Omega, \qquad u = f \quad \text{on } \partial\Omega. \tag{3.3}$$

If $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular, $0 < \mathfrak{d} < \mathfrak{n}$, then we can extend an arbitrary $f \in \dot{B}^{p,p}_{1-1/p}(\partial\Omega)$ to a function F in $\dot{W}^{1,p}(\Omega; w)$ of comparable norm. (The extension result is valid even if $\mathfrak{d}$ is not an integer.) See [43] if $\mathfrak{d} = \mathfrak{n} - 1$, [20, Chapter 7] and [21, Chapter 8] if $p = 2$, and [8] in the general case. Letting F be this extension, u_0 the solution to the homogeneous Poisson problem (3.1) with data $\vec{H}_0 = \vec{H} - \nabla F$, and $u = u_0 + F$ yields a solution u to the Poisson–Dirichlet problem (3.3) with the estimate

$$\|\nabla u\|_{L^p(\Omega,w)} \leq C \|\vec{H}\|_{L^p(\Omega,w)} + C \|f\|_{\dot{B}^{p,p}_{1-1/p}(\partial\Omega)}. \tag{3.4}$$

(Trace theorems often allow us to clarify the sense in which $u = f$ on $\partial\Omega$; see the above references.) This construction is very well known in the codimension 1 case ($\mathfrak{d} = \mathfrak{n} - 1$) and is explained in detail in the general case in [20, Lemma 9.1].

We emphasize that up until the present work, in the higher codimension case, this construction of solutions to the Poisson–Dirichlet problem, restricted to boundary data in the particular Besov space $\dot{B}^{p,p}_{1-1/p}(\partial\Omega)$ for p sufficiently close to 2, was the only known result for the Dirichlet–Poisson problem with boundary data of fractional smoothness. All of the following results in this section apply only to coefficients that satisfy the ellipticity conditions (2.2)–(2.3) in the particular case $\mathfrak{d} = \mathfrak{n} - 1$.

3.2. The Poisson problem with Sobolev estimates

A natural line of study is to seek conditions on Ω and A such that estimate (3.2) holds for a wider range of p.

The estimate $\|\nabla u_0\|_{L^p(\Omega)} \leq C \|\vec{H}_0\|_{L^p(\Omega)}$ for all $1 < p < \infty$ has been established in the case where A satisfies (2.2)–(2.3) (with $\mathfrak{d} = \mathfrak{n} - 1$) and lies in the space VMO of vanishing mean oscillation and where $\partial\Omega$ is a bounded C^1 domain. See [4,25,47]. For any fixed $p \in (1, \infty)$, the result continues to hold if A and Ω are allowed to deviate slightly from these assumptions; see [11,12] for the details.

3.3. The Poisson–Dirichlet problem with weighted Sobolev estimates

Recall that in Section 3.1, we used the fact that a function $f \in \dot{B}^{p,p}_{1-1/p}(\partial\Omega)$ may be extended to a function F in the Sobolev space $\dot{W}^{1,p}(\mathbb{R}^n)$. It is also possible to extend functions in $\dot{B}^{p,p}_s(\partial\Omega)$ for a range of s, and not merely the particular value $s = 1 - 1/p$, if we take our extensions to lie in a *weighted* Sobolev space with an appropriate weight. In particular, if $0 < s < 1$, $1 \le p \le \infty$, and $\partial\Omega$ is $(n-1)$-Ahlfors regular, then any $f \in \dot{B}^{p,p}_s(\partial\Omega)$ may be extended to a function F with

$$\int_{\mathbb{R}^n} |\nabla F(x)|^p \operatorname{dist}(x, \partial\Omega)^{p-1-ps}\, dx \approx \|f\|^p_{\dot{B}^{p,p}_s(\partial\Omega)}.$$

See [8, 20, 46, 51, 56, 60, 61, 64, 66]. The case of bounded Lipschitz domains is [56, Proposition 4.1] and the $l = 1$ case of [46, Theorem 2.10], but the result was known earlier in smoother domains.

This suggests consideration of the Dirichlet–Poisson problem (3.3) with the estimate

$$\begin{aligned}\int_\Omega |\nabla u(x)|^p \operatorname{dist}(x, \partial\Omega)^{p-1-ps}\, dx \\ \le C \int_\Omega |\vec{H}(x)|^p \operatorname{dist}(x, \partial\Omega)^{p-1-ps}\, dx + C\|f\|^p_{\dot{B}^{p,p}_s(\partial\Omega)}.\end{aligned} \tag{3.5}$$

This problem was studied in [56] (which in fact considered the case where L is the Hodge Laplacian in a Lipschitz domain in a compact Riemannian manifold) and [51] (which in fact studied the problem for elliptic systems of arbitrary even order, not only for the second-order equations that are the focus of the present note).

The paper [51] established a result similar to those of the previous section. Specifically, if A satisfies (2.2)–(2.3) (with $\mathfrak{d} = n - 1$) and lies in VMO, and if Ω is a bounded Lipschitz domain whose unit outward normal also lies in VMO (this is a moderate generalization of the C^1 condition), then estimate (3.5) holds for any $1 < p < \infty$ and any $0 < s < 1$. Moreover, again as in the unweighted case of the previous section, for any fixed p and s in this range, if A is sufficiently close to VMO and Ω is a Lipschitz domain with sufficiently small constant, then estimate (3.5) holds.

The $p = 2$, $s = 1/2$ case of these results is the energy solutions of Section 3.1, while the $s = 1 - 1/p$ case is the previous section.

The paper [56] established estimate (3.5) in general bounded Lipschitz domains (without the VMO or near-VMO condition on the normals) but for a restricted range of s, p, and A. Specifically, they required the coefficients A to be symmetric and to satisfy the Dini-type condition

$$|A(x) - A(y)| \le \omega(|x - y|), \qquad \int_0^1 \frac{\sqrt{\omega(t)}}{t}\, dt < \infty$$

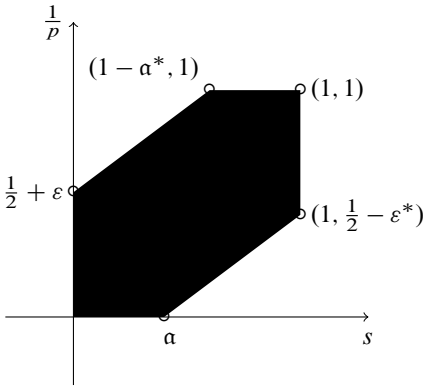


Figure 1. Values of the parameters $(s, 1/p)$ such that estimate (3.5) holds, for solutions to problem (3.3) in a Lipschitz domain, where A satisfies a Dini-type continuity condition. This region includes the edges at $p = 1$ and $p = \infty$ but does not include the edges at $s = 0$ or $s = 1$.

which is certainly stronger than lying in VMO. They were then able to establish estimate (3.5) for all s and p such that the point $(s, 1/p)$ lies in the hexagonal region illustrated in Figure 1 for some $\varepsilon, \varepsilon^*, \alpha, \alpha^* > 0$.[1]

3.4. General weights

More general weights may be considered. By [1], if $1 < p < \infty$, if ω is a Muckenhoupt A_p weight, if A is symmetric and the BMO norm of A is small enough, and if Ω is a Lipschitz domain with a sufficiently small Lipschitz constant, then solutions to the Poisson problem (3.1) satisfy the estimate

$$\|\nabla u_0\|_{L^p(\Omega;\omega)} \leq C \|\vec{H}_0\|_{L^p(\Omega;\omega)}. \tag{3.6}$$

[1]The paper [56], along with its predecessors [54,55], does not work with arbitrary second-order operators in Lipschitz domains in $\mathbb{R}^n$ but with the Hodge Laplacian Δ_g defined on a compact Riemannian manifold with metric g; more precisely, they work with the Schrödinger-type operator $L = -\Delta_g + V$ for a nonnegative potential V.

We may return to the Euclidean situation of the present paper as follows. Given a bounded Lipschitz domain Ω in $\mathbb{R}^n$, one may easily construct a compact n-dimensional manifold M with Ω contained in one coordinate chart. In local coordinates, the Hodge Laplacian is given by $\sqrt{|g|}\Delta_g = \operatorname{div}(\sqrt{|g|}g^{-1}\nabla)$, and so we can attain any real symmetric coefficient matrix A by choosing an appropriate metric on the chart containing Ω. The technical requirements on the potential V in [54–56] may be dealt with by choosing $V = 0$ in a neighborhood of Ω and $V > 0$ in appropriate subsets of $M \setminus \Omega$.

The result was generalized to quasi-convex domains in [68, Theorem 1.10]. This builds on the earlier results [2,62], among others, which proved the same result under stronger assumptions on ω.

As noted in [10, formula (2.5)] and the proof of Lemma 2.3 of [56], if $\Omega \subset \mathbb{R}^n$ is a Lipschitz domain, then $\omega(x) = \operatorname{dist}(x, \partial\Omega)^{p-1-ps}$ is a Muckenhoupt A_p weight for any $1 < p < \infty$ and any $0 < s < 1$, and so the results of the previous section fall under this umbrella.

3.5. Equivalence of norms

Inequalities (3.5) and (3.6) are in fact equivalences of norms. Observe that we may write inequality (3.5) in the form

$$\|u\|_{\dot{W}^{1,p}(\Omega;\omega_{s,p})} \leq C\,\|f\|_{\dot{B}^{p,p}_s(\partial\Omega)} + C\,\|\vec{H}\|_{L^p(\Omega;\omega_{s,p})}$$

for the weight $\omega_{s,p}(x) = \operatorname{dist}(x, \partial\Omega)^{p-1-ps}$. If u is *any* function in $\dot{W}^{1,p}(\Omega;\omega_{s,p})$, where Ω is as in the previous sections, then the boundary trace of u exists and lies in the Besov space $\dot{B}^{p,p}_s(\partial\Omega)$; indeed, we have the reverse estimate

$$\|u\big|_{\partial\Omega}\|_{\dot{B}^{p,p}_s(\partial\Omega)} \leq C\,\|u\|_{\dot{W}^{1,p}(\Omega;\omega_{s,p})}.$$

This was proven in the generality of Lipschitz domains in [56, Proposition 4.1] and as the $l = 1$ case of [46, Theorem 2.10]; similar results were already known in smoother domains (see [48,60,61,64,66]) or in the unweighted case $p - 1 - ps = 0$ (see [43]). Thus, inequality (3.5) implies that

$$\frac{1}{C}\|f\|_{\dot{B}^{p,p}_s(\partial\Omega)} \leq \|u\|_{\dot{W}^{1,p}(\Omega;\omega_{s,p})} \leq C\,\|f\|_{\dot{B}^{p,p}_s(\partial\Omega)} + C\,\|\vec{H}\|_{L^p(\Omega;\omega_{s,p})}.$$

With some care, we can also get an equivalence of norms for the Poisson data $\vec{H}$ in inequality (3.5) or $\vec{H}_0$ in inequality (3.6). If $\operatorname{div} A\vec{\Psi} = \operatorname{div} A\vec{H}_0$, then the solution u with data $\vec{\Psi}$ must equal the solution with data $\vec{H}_0$, and so estimate (3.6) may be strengthened to

$$\|\nabla u_0\|_{L^p(\Omega;\omega)} \leq C\,\inf\big\{\|\vec{\Psi}\|_{L^p(\Omega;\omega)} : \operatorname{div}(A\vec{\Psi} - A\vec{H}_0) = 0\big\}.$$

In particular, $\operatorname{div} A\vec{H}_0 = \operatorname{div} A\nabla u$, and so we may take $\vec{\Psi} = \nabla u_0$. Thus, the reverse estimate

$$\inf\big\{\|\vec{\Psi}\|_{L^p(\Omega;\omega)} : \operatorname{div}(\vec{\Psi} - \vec{H}_0) = 0\big\} \leq C\,\|\nabla u_0\|_{L^p(\Omega;\omega)}$$

is trivially satisfied. Inequality (3.6) thus implies an equivalence of norms

$$\|\nabla u_0\|_{L^p(\Omega;\omega)} \approx \inf\big\{\|\vec{\Psi}\|_{L^p(\Omega;\omega)} : \operatorname{div}(A\vec{\Psi} - A\vec{H}_0) = 0\big\}.$$

Similarly, inequality (3.5) implies an equivalence of norms

$$\|\nabla u\|_{L^p(\Omega;\omega)} \approx \|f\|_{\dot{B}^{p,p}_s(\partial\Omega)} + \inf\big\{\|\vec{\Psi}\|_{L^p(\Omega;\omega)} : \operatorname{div}(A\vec{\Psi} - A\vec{H}) = 0\big\}.$$

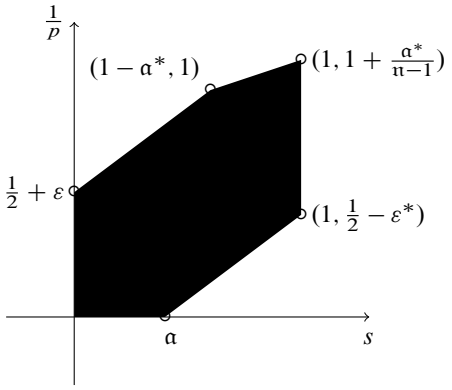


Figure 2. Values of the parameters $(s, 1/p)$ such that solutions to the Dirichlet–Poisson problem (3.3) in the domain above a Lipschitz graph, for real symmetric coefficients constant in the vertical direction, satisfy estimate (3.7).

3.6. Transversally constant coefficients and weighted averaged Sobolev estimates

Moving beyond Dini-type or VMO coefficients to another class of variable coefficient elliptic matrices, in [3, 7] the authors considered coefficients constant in the direction transverse to the boundary (but merely bounded measurable in all other directions). In [7] (written by two of the authors of the present paper), we showed that if Ω is the domain above a Lipschitz graph, and for such coefficients that in addition are real symmetric, we still have well-posedness of the Poisson–Dirichlet problem (3.3), with estimate (3.5) replaced by the estimate

$$\int_\Omega \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\nabla u|^\beta \right)^{p/\beta} \operatorname{dist}(x,\partial\Omega)^{p-1-ps}\, dx \leq C \int_\Omega \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\vec{H}|^\beta \right)^{p/\beta} \operatorname{dist}(x,\partial\Omega)^{p-1-ps}\, dx + C \|f\|^p_{\dot{B}^{p,p}_s(\partial\Omega)} \tag{3.7}$$

for β sufficiently close to 2 and for $(s, 1/p)$ in the hexagon in Figure 2. The parameters ε, ε^*, α, and α^* in that figure are all positive. (In the case where A is real, constant in the transverse direction, but not necessarily symmetric, we have well-posedness for a smaller range; namely, the range of Figure 2 for α, α^* still positive, but for ε, ε^* merely greater than $-1/2$. See the region at the bottom of Figure 4 below.)

The β-power averages were inspired by the modified nontangential maximal function of [44] and were introduced for much the same reason: if A lacks smoothness properties, then solutions u to even the homogeneous problem $\operatorname{div} A\nabla u = 0$

no longer have gradients with high degrees of integrability. The gradient ∇u may be locally in L^β only for β close to 2, and so these weighted averaged spaces are more appropriate to such differential equations. Even for p much larger than β, these spaces preserve the trace result

$$\|f\|^p_{\dot{B}^{p,p}_s(\partial\Omega)} \leq C \int_\Omega \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\nabla u|^\beta \right)^{p/\beta} \operatorname{dist}(x,\partial\Omega)^{p-1-ps}\, dx,$$

and so estimate (3.7) is an equivalence of norms. See [7, Theorem 6.3].

As mentioned above, we used layer potential techniques, which introduce harsh restrictions both on the coefficients and on the domains. Our results also crucially used the endpoint results (which may be viewed as the $s = 0$ and $s = 1$ boundaries of the hexagon in Figure 2); that is, with the techniques of [7], the results for Besov boundary values do not stand alone but require the results for boundary data in $L^p(\partial\Omega)$ (or in the Sobolev space $\dot{W}^{1,p}(\partial\Omega)$, that is, at the smoothness 1 endpoint) and the layer potential bounds.

The paper [3] applies the so-called "first-order approach" (developed in the endpoint theory of $L^p(\partial\Omega)$ boundary data) to the Dirichlet problem

$$-\operatorname{div} A\nabla u = 0 \quad \text{in } \Omega, \qquad u = f \quad \text{on } \partial\Omega,$$

with f in a fractional smoothness space, either the Besov spaces considered in this note or the related Hardy–Sobolev spaces.

Like [7], they considered coefficients constant in the direction transverse to the boundary. Again like [7], their direct well-posedness results for boundary data of fractional smoothness require endpoint results, that is, results for boundary data in $L^p(\partial\Omega)$ or $\dot{W}^{1,p}(\partial\Omega)$.

However, unlike [7], the methods of [3] encompass general complex coefficients, and even elliptic systems rather than equations. In particular, [7] is confined to equations and also to a restricted class of coefficients (specifically, those that satisfy the De Giorgi–Nash condition; this includes all real coefficients but misses many complex coefficients). Instead, [3] can build upon endpoint results for a much broader class of coefficients, including complex self-adjoint coefficients, and constant or triangular coefficients in the half-space $\mathbb{R}^{\mathfrak{n}}_+$, to produce results for Besov data.

3.7. The Poisson problem without equivalence of norms

We mention one further class of results concerning the Poisson problem with rough variable coefficients. The equivalence of norms described in Section 3.5 is a highly desirable property, in that it implies that the estimates in the preceding sections cannot be improved: the norms on the right-hand side of estimates (3.5), (3.6), and (3.7) clearly cannot be weakened without a corresponding weakening of the norms on the

left-hand side. Thus, most of the existing theory has focused on estimates with such equivalences of norms.

The homogeneous Dirichlet problem

$$-\operatorname{div} A\nabla u = 0 \quad \text{in } \Omega, \qquad u = f \quad \text{on } \partial\Omega \tag{3.8}$$

has been studied extensively. The paper [3], mentioned in the previous section, considered the Dirichlet problem with boundary data in fractional smoothness spaces; however, there is much more extensive literature on the Dirichlet problem with data f in the Lebesgue space $L^p(\partial\Omega)$ or in the boundary Sobolev space $\dot{W}^{1,p}(\partial\Omega)$ and its generalizations. The most common estimates (taking the role of estimates (3.5), (3.6), and (3.7) in the above sections) are of the forms

$$\|Nu\|_{L^p(\partial\Omega)} + \|\mathcal{A}_2(\delta\nabla u)\|_{L^p(\partial\Omega)} \le C\|f\|_{L^p(\partial\Omega)}, \tag{3.9}$$

$$\|\widetilde{N}(\nabla u)\|_{L^p(\partial\Omega)} \le C\|f\|_{\dot{W}^{1,p}(\partial\Omega)}, \tag{3.10}$$

where $\delta(y) = \operatorname{dist}(y, \partial\Omega)$, and where N, $\widetilde{N}$, and $\mathcal{A}_2$ denote the nontangential maximal function, modified nontangential maximal function, and Lusin area integral given by

$$\mathcal{A}_2(\vec{H})(x) = \left(\int_{\gamma(x)} |\vec{H}(y)|^2 \delta(y)^{-\mathfrak{n}}\, dy\right)^{1/2},$$

$$Nu(x) = \sup_{\gamma(x)} |u|, \quad \widetilde{N}(\nabla u)(x) = \sup_{y\in\gamma(x)} \left(\fint_{B(y,\delta(y)/2)} |\nabla u|^2\right)^{1/2}, \tag{3.11}$$

where $\gamma(x) = \{y \in \Omega : |x - y| < 2\delta(y)\}$. The Dirichlet problem with these estimates does generally have an equivalence of norms; that is, if $-\operatorname{div} A\nabla u = 0$ in Ω and $Nu \in L^p(\partial\Omega)$ or $\mathcal{A}_2(\delta\nabla u) \in L^p(\partial\Omega)$, then in many situations, the boundary values of u are known to exist (in the sense of nontangential limits) and lie in $L^p(\partial\Omega)$. These types of results have their origin in Fatou's celebrated result [27] that a bounded harmonic function in the unit disk has a boundary trace; for real coefficients in Lipschitz domains, these boundary results follow from [13, 17].

Major difficulties arise when we try to consider the Poisson–Dirichlet problem with Dirichlet data f in L^p and with estimates on the solution u compatible with the estimates of the homogeneous Dirichlet problem. For example, consider the function $v(x) = \eta(x)\sin(\log|\log(x_{\mathfrak{n}})|)$, where η is a smooth cutoff function supported in $B(0, 1/2)$. If we let Ω be the half-space, then it is straightforward to establish that $\mathcal{A}_2(\delta\nabla v)$ is bounded and compactly supported and thus lies in $L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)$ for all p; however, v clearly does not even possess Dirichlet boundary values in any sense. It is thus impossible to have unique solvability of the Poisson–Dirichlet problem

$$-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H} \quad \text{in } \mathbb{R}^{\mathfrak{n}}_+, \qquad u = f \quad \text{on } \partial\mathbb{R}^{\mathfrak{n}}_+,$$

with the natural estimate

$$\|\mathcal{A}_2(\delta\nabla u)\|_{L^p(\partial\mathbb{R}^n_+)} \leq C\|f\|_{L^p(\partial\mathbb{R}^n_+)} + C\|\mathcal{A}_2(\delta\vec{H})\|_{L^p(\partial\mathbb{R}^n_+)}$$

for arbitrary f and $\vec{H}$, for example, for $f = 0$, $\vec{H} = \nabla v$.

It is, however, possible to solve the Poisson–Dirichlet problem with estimates of the form

$$\|\mathcal{A}_2(\delta\nabla u)\|_{L^p(\partial\Omega)} + \|\widetilde{N}u\|_{L^p(\partial\Omega)} \leq C\|f\|_{L^p(\partial\Omega)} + C\|\vec{H}\|_{\mathfrak{X}}$$

or

$$\|\widetilde{N}(\nabla u)\|_{L^p(\partial\Omega)} \leq C\|f\|_{\dot{W}^{1,p}(\partial\Omega)} + C\|\vec{H}\|_{\mathfrak{Y}},$$

that is, with the same estimates on the Dirichlet data f and solutions u as are traditional in the theory of the homogeneous Dirichlet problem, but for function spaces $\mathfrak{X}$ and $\mathfrak{Y}$ that do not yield an equivalence of norms in the Poisson problem. Such estimates were found in the works [6, 40, 58], either implicitly as part of the method of attack on the L^p-Dirichlet problem [6, 40] or explicitly as a main result [58].

Specifically, [6, 40] established the estimates

$$\begin{aligned}
&\|\mathcal{A}_2(\delta\nabla L^{-1}(\operatorname{div} A\vec{H}))\|_{L^p(\partial\Omega)} + \|\widetilde{N}(L^{-1}(\operatorname{div} A\vec{H}))\|_{L^p(\partial\Omega)} \\
&\qquad \leq \|\widetilde{\mathfrak{C}}_1(\delta\vec{H})\|_{L^p(\partial\Omega)}, \\
&\|\widetilde{N}(\nabla L^{-1}(\operatorname{div} A\vec{H}))\|_{L^p(\partial\Omega)} \leq \|\mathcal{A}_2(\vec{H})\|_{L^p(\partial\Omega)}, \\
&\|\widetilde{N}(\nabla L^{-1}(\operatorname{div} A\vec{H}))\|_{L^p(\partial\Omega)} \leq \|\mathfrak{C}_1(\delta \operatorname{div} A\vec{H})\|_{L^p(\partial\Omega)}
\end{aligned}$$

for an appropriate range of p, for $L = -\operatorname{div} A\nabla$, and for

$$\widetilde{\mathfrak{C}}_1\vec{H}(x) = \sup_{r>0} \frac{1}{\sigma(B(x,r)\cap\partial\Omega)} \int_{B(x,r)\cap\Omega} \left(\fint_{B(y,\operatorname{dist}(y,\partial\Omega)/2)} |\vec{H}|^2 \right)^{1/2} \frac{dy}{\operatorname{dist}(y,\partial\Omega)},$$

in the case where the coefficients A are transversally independent (as in Section 3.6), where the underlying domain Ω is either the upper or lower half-space, and where $\vec{H}$ is always supported in the upper half-space. (The paper [40] considered the case of complex coefficients that satisfy the local De Giorgi–Nash–Moser condition, while [6] considered arbitrary complex coefficients and higher-order operators.)

The estimates on L^{-1} were used explicitly in [6, 40], but the above estimates can be used to establish well-posedness of the Poisson–Dirichlet problem, given well-posedness of the homogeneous Dirichlet problem. If $w = L^{-1}(\operatorname{div} A\vec{H})$ satisfies $\widetilde{N}(\nabla w) \in L^p(\partial\Omega)$, then w possesses boundary values (in the sense of nontangential limits) that necessarily lie in the boundary Sobolev space $\dot{W}^{1,p}(\partial\mathbb{R}^n_+) = \dot{W}^{1,p}(\mathbb{R}^{n-1})$; see [44]. If the homogeneous Dirichlet problem (3.8), with estimate (3.10), is well posed, then by taking $u = w + v$, where v is the solution to

the Dirichlet problem with boundary data $f - w|_{\partial\mathbb{R}^{\mathfrak{n}}_+}$, we recover a solution u to the Poisson–Dirichlet problem (3.3) that satisfies the estimate

$$\|\widetilde{N}(\nabla u)\|_{L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)} \leq C\,\|\mathcal{A}_2(\vec{H})\|_{L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)} + C\,\|f\|_{\dot{W}^{1,p}(\partial\mathbb{R}^{\mathfrak{n}}_+)}.$$

A similar but more complicated argument is valid for the Dirichlet problem with $f \in L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)$. Namely, if $\vec{H}$ is supported in $\mathbb{R}^{\mathfrak{n}}_+$, then $w = L^{-1}(\operatorname{div} A\vec{H})$ is a solution to $-\operatorname{div} A\nabla w = 0$ in the lower half-space, and so it has boundary values in $L^p(\partial\mathbb{R}^{\mathfrak{n}}_-)$ by the abovementioned Fatou-type results [13, 17]. The operator $L^{-1}\operatorname{div}$ has adequate continuity properties which implies that w takes the same values on the boundary of the upper half-space $\mathbb{R}^{\mathfrak{n}}_+$. Thus, $w\big|_{\partial\mathbb{R}^{\mathfrak{n}}_+}$ lies in $L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)$, and so by the same argument as above, well-posedness of the Dirichlet problem (3.8) with estimate (3.9) immediately yields well-posedness of the Poisson–Dirichlet problem (3.3) with the estimate

$$\|\widetilde{N}u\|_{L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)} + \|\mathcal{A}_2(\delta\nabla u)\|_{L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)} \leq C\,\|f\|_{\dot{W}^{1,p}(\partial\mathbb{R}^{\mathfrak{n}}_+)} + C\,\|\widetilde{\mathfrak{C}}_1(\delta\vec{H})\|_{L^p(\partial\mathbb{R}^{\mathfrak{n}}_+)}.$$

The paper [58] (like the present paper) works exclusively with real coefficients but in a much more general geometric setting than the half-space. Furthermore, their treatment of the Poisson–Dirichlet problem is much more explicit than that of [6, 40]. They show that if $\partial\Omega$ is $(\mathfrak{n}-1)$-Ahlfors regular, if Ω satisfies the interior corkscrew condition, and if $\mathbb{R}^{\mathfrak{n}} \setminus \Omega$ is unbounded, then solvability of the homogeneous Dirichlet problem (3.8) with estimate (3.9) is in fact equivalent to solvability of the homogeneous Poisson problem (3.1), with the estimate

$$\|\widetilde{N}u\|_{L^p(\partial\Omega)} \leq C\,\|\widetilde{\mathfrak{C}}_1(\delta\vec{H})\|_{L^p(\partial\Omega)},$$

and also to solvability of the homogeneous Poisson problem for $-\operatorname{div} A^*\nabla$, where A^* denotes the transpose of A, with the estimate

$$\|\widetilde{N}(\nabla u)\|_{L^{p'}(\partial\Omega)} \leq C\,\|\widetilde{\mathfrak{C}}_1(\delta\operatorname{div} A\vec{H})\|_{L^{p'}(\partial\Omega)}$$

for p' the Hölder conjugate of p, that is, $1/p + 1/p' = 1$. Under slightly stronger conditions on the domain and coefficients, namely uniform rectifiability of the boundary and a Dahlberg–Kenig–Pipher condition on the coefficients, they establish equivalence of the above conditions to solvability of the so-called regularity problem, that is, the homogeneous Dirichlet problem (3.8) with estimate (3.10), again with A replaced by A^* and p by p'.

As noted above, this version of the Poisson problem does not have equivalence of norms in the sense of Section 3.5, but the given estimates are the best estimates known in the context of L^p boundary data, are equivalent to quite strong conditions (indeed, the main estimate) in the theory of the homogeneous Dirichlet problem, and

thus may be reasonably conjectured to be among the best estimates possible in this setting.

3.8. The Laplace operator and other estimates

Thus far, we have focused on estimates of solutions u to the Poisson problem that satisfy local Sobolev estimates, that is, solutions where the gradient ∇u satisfies an estimate that locally reduces to membership in L^2 or L^p. We have done this because the main new results of this note (to be described in Sections 4 and 5) are phrased in these terms.

However, many important results, particularly the earliest results, instead established estimates on solutions to the Poisson problem (1.1) of the form

$$\|u\|_{\dot{B}^{p,q}_{s+1/p}(\Omega)} \leq C\|h\|_{\dot{B}^{p,q}_{s+1/p-2}(\Omega)} + C\|f\|_{\dot{B}^{p,q}_{s}(\partial\Omega)} \tag{3.12}$$

or

$$\|u\|_{\dot{F}^{p,q}_{s+1/p}(\Omega)} \leq C\|h\|_{\dot{F}^{p,q}_{s+1/p-2}(\Omega)} + C\|f\|_{\dot{B}^{p,p}_{s}(\partial\Omega)}, \tag{3.13}$$

where $\dot{B}^{p,q}_{\theta}$ and $\dot{F}^{p,q}_{\theta}$ denote the Besov and Triebel–Lizorkin spaces. (The precise statement of the estimates varies from paper to paper.)

It is often possible to pass from Besov or Triebel–Lizorkin estimates on solutions to weighted Sobolev estimates such as (3.5); see, for example, [41, Theorems 4.1 and 4.2].

Estimates of the form (3.12) or (3.13) are unsuited for solutions to $-\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H}$, or even to $-\operatorname{div} A\nabla u = 0$, for rough coefficients A. Such solutions u do not generally display any higher regularity and so cannot be expected to lie in $\dot{B}^{p,q}_{s+1/p}(\Omega)$ for $s + 1/p > 1$. See the discussion in [7, Chapter 10], or compare the two papers [55,56]; both consider the Dirichlet–Poisson problem in Lipschitz domains for coefficients A satisfying the same Dini-type condition, but [55] used the traditional estimate (3.12), while [56] used estimate (3.5), and so [55] has a somewhat artificial upper bound on $s + 1/p$ which [56] does not need to impose.

However, many important results are known in the context of estimate (3.12) or (3.13).

The paper [41], concerning only the Laplacian, is one of the foundational papers in the theory of the Poisson problem and of the Dirichlet problem with boundary data of fractional smoothness; indeed, it preceded all of the results described in Sections 3.2–3.7. That paper established estimate (3.13) in Lipschitz domains for $q = 2$ and for s and p as in Figure 1. They provided examples establishing sharpness of these results (that is, of Lipschitz domains such that estimate (3.13) fails for any $(s, 1/p)$ outside of the given region). In the case of C^1 domains, they established this

estimate for all $s \in (0,1)$ and all $p \in (1,\infty)$. (The VMO condition of [51] described in Section 3.3 is a mild generalization of the C^1 condition.)

The paper [49] removes the restriction on q in estimate (3.13) (and (3.12)) and expands the available values of p and s into the range $p < 1$, specifically to the range $1/p < 1 + s + \mathfrak{a}^* - 1/\mathfrak{n} - 1$ (in Lipschitz domains) or $0 \le 1/p < 1 + s/\mathfrak{n} - 1$ (in C^1 domains). See Figure 2. We observe that [49] precedes the papers [3, 7] described in Section 3.6 and that the paper [7] explicitly works with the same ranges of s and p as the Lipschitz case of [49].

The results of [41, 49] for C^1 domains with estimate (3.12) or (3.13) also hold in bounded convex domains; see [53].

3.9. Hölder continuous boundary data

We conclude this section with a discussion of the recent result [15]. This result considers the Dirichlet problem (2.5) with Hölder continuous boundary data; this is the $p = \infty$ (bottom) edge of the region in Figure 2 and thus falls within the ambit of boundary value problems with data of fractional smoothness.

We remark that [15] is not the only paper to treat Hölder continuous boundary values, but in the setting of real elliptic coefficients, it contains (provably) the most general possible result. In particular, they place no assumptions on the coefficients beyond the ellipticity conditions (2.2)–(2.3) (with $\mathfrak{d} = \mathfrak{n} - 1$). They showed that if $\partial\Omega$ satisfies a criterion of Wiener type, then the solution u to the continuous Dirichlet problem (2.5) satisfies the estimate

$$\|u\|_{\dot{C}^s(\Omega)} \le C \|f\|_{\dot{C}^s(\partial\Omega)} \tag{3.14}$$

for all $s > 0$ sufficiently small, where $\dot{C}^s$ denotes spaces of Hölder continuous functions. Their result is sharp; they show that their Wiener-type criterion is a necessary, as well as a sufficient, condition for estimate (3.14) for arbitrary $f \in \dot{C}^s(\partial\Omega)$.

We remark that Ahlfors regularity of $\partial\Omega$ for $\mathfrak{d} = \mathfrak{n} - 1$ (or indeed for any $\mathfrak{d} \in (\mathfrak{n} - 2, \mathfrak{n})$) implies their Wiener-type criterion and thus is a sufficient condition for Hölder well-posedness.

The $p = \infty$, $\vec{H} = 0$ analogue of estimate (3.7) is

$$\sup_{x\in\Omega} \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\nabla u|^\beta \right)^{1/\beta} \operatorname{dist}(x,\partial\Omega)^{1-s} \le C \|f\|_{\dot{C}^s(\partial\Omega)},$$

and it follows immediately (for β sufficiently close to 2) from estimate (3.14) and the Caccioppoli and Meyers estimates for solutions to $-\operatorname{div} A\nabla u = 0$.

4. The Poisson problem for very general domains and coefficients

Our first main theorem establishes well-posedness of the Poisson–Dirichlet problem for *any* second-order elliptic differential equation with real coefficients satisfying suitable upper and lower bounds and with very mild conditions on the underlying domains.

Theorem 4.1. *Let $\Omega \subseteq \mathbb{R}^{\mathfrak{n}}$, $\mathfrak{n} \geq 3$, be a connected open set whose boundary $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular in the sense of Definition 2.1 for some real number $\mathfrak{d}$ with $0 < \mathfrak{d} \leq \mathfrak{n} - 1$. If $\mathfrak{d} = \mathfrak{n} - 1$, then we additionally require that Ω satisfies the weak local John condition and the interior corkscrew condition.*[2]

We require that $\mathbb{R}^{\mathfrak{n}} \setminus \Omega$ be unbounded, that is, either Ω is bounded[3] *or both Ω and $\partial\Omega$ are unbounded.*

Let $L = -\operatorname{div} A\nabla$ be a linear differential operator associated to real, not necessarily symmetric, coefficients A that satisfy the positive definiteness and boundedness conditions (2.2)–(2.3).

There are then numbers $\mathfrak{a}, \mathfrak{a}^* \in (0, 1]$ and $\varepsilon, \delta > 0$ such that if $(s, 1/p)$ lies in the open pentagon indicated in Figure 3, and if $2 - \delta < \beta < 2 + \varepsilon$, then for all f in the Besov space $\dot{B}^{p,p}_s(\partial\Omega)$ defined in Definition 2.10 above, and all $\vec{H} \in L^{\beta}_{loc}(\Omega)$ such that the right-hand side of bound (4.2) is finite, there is a unique solution u to the Poisson–Dirichlet problem

$$\begin{cases} -\operatorname{div} A\nabla u = -\operatorname{div} A\vec{H} & \text{in } \Omega, \\ u = f & \text{on } \partial\Omega \end{cases} \tag{4.1}$$

that satisfies

$$\begin{aligned} &\int_{\Omega} \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\nabla u|^{\beta} \right)^{p/\beta} \operatorname{dist}(x,\partial\Omega)^{\mathfrak{d}-\mathfrak{n}+p-ps}\, dx \\ &\quad \leq C \|f\|^{p}_{\dot{B}^{p,p}_s(\partial\Omega)} + C \int_{\Omega} \left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\vec{H}|^{\beta} \right)^{p/\beta} \operatorname{dist}(x,\partial\Omega)^{\mathfrak{d}-\mathfrak{n}+p-ps}\, dx \end{aligned} \tag{4.2}$$

[2] If $\mathfrak{d} < \mathfrak{n} - 1$, then, as noted in Remark 2.5, the weak local John and interior corkscrew conditions are automatically satisfied.

[3] It is straightforward to establish that if $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ is bounded, then $\partial\Omega$ has dimension at least $\mathfrak{n} - 1$; in our circumstances, this means that $\mathfrak{d} = \mathfrak{n} - 1$.

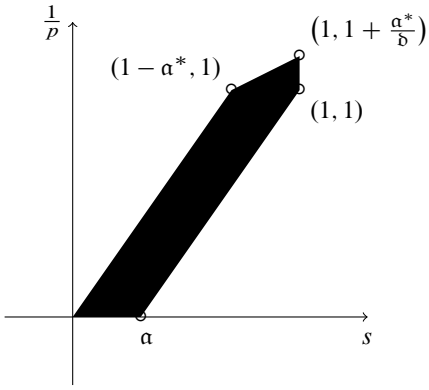


Figure 3. Values of the parameters $(s, 1/p)$ such that the $\dot{B}^{p,p}_s(\partial\Omega)$-Dirichlet problem and associated Poisson problem for L are solvable. By Remark 4.2, the illustrated region is convex.

for some constant C depending on p, s, A, and Ω but not on f or $\vec{H}$. If $p = \infty$ and $0 < s < \mathfrak{a}$, then there is a unique solution to problem (4.1) that satisfies

$$\sup_{x\in\Omega}\left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\nabla u|^\beta\right)^{1/\beta} \operatorname{dist}(x,\partial\Omega)^{1-s} \leq C\|f\|_{\dot{B}^{\infty,\infty}_s(\partial\Omega)} + C\sup_{x\in\Omega}\left(\fint_{B(x,\operatorname{dist}(x,\partial\Omega)/2)} |\vec{H}|^\beta\right)^{1/\beta} \operatorname{dist}(x,\partial\Omega)^{1-s}.$$

Furthermore, we have that $\mathfrak{a} \geq \max(\alpha, 1-\mathfrak{d})$ and $\mathfrak{a}^* \geq \max(\alpha^*, 1-\mathfrak{d})$, where α and α^* are the numbers in the boundary De Giorgi–Nash estimate (2.7).

Finally, $2+\varepsilon$ may be taken to be the number in Meyers's reverse Hölder estimate (2.6) for L, and $2-\delta$ may be taken to be the Hölder conjugate of the number in Meyers's reverse Hölder estimate for L^*.

Expressed algebraically, the pair $(s, 1/p)$ satisfies the conditions of Theorem 4.1 if either

$$\frac{\mathfrak{d}}{\mathfrak{d}+\mathfrak{a}^*} < p \leq 1 \quad \text{and} \quad \frac{\mathfrak{d}}{p} + 1 - \mathfrak{d} - \mathfrak{a}^* < s < 1$$

or

$$1 \leq p \leq \infty \quad \text{and} \quad \frac{1-\mathfrak{a}^*}{p} < s < \mathfrak{a} + \frac{1-\mathfrak{a}}{p}.$$

We observe that, with the exception of the Sobolev estimate (3.4) for p sufficiently close to 2 in Section 3.1, the Dirichlet problem with Hölder continuous boundary data of [15] (see Section 3.9), and the Poisson–Dirichlet problem without

equivalence of norms of [58] (see Section 3.7), all of the results for the Poisson and Dirichlet problems described in Section 3 were shown exclusively in fairly regular (Lipschitz or quasi-convex) domains with boundary of codimension 1, or in domains with a higher degree of smoothness, and for coefficients satisfying some regularity assumption. Our result is the first to produce well-posedness for a broad range of $(s, 1/p)$ for arbitrary real elliptic coefficients and for domains in the above generality.

Because fundamental solutions and Green's functions display different behavior in two dimensions and in higher dimensions, certain estimates in our proof of Theorem 4.1 work only if the ambient dimension n is at least three. The case $n = 2$ is currently being investigated.

Remark 4.2. Interpolation methods can generally be used to show that if the $\dot{B}_s^{p,p}$-Dirichlet problem is well posed for $(s, 1/p) = (s_0, 1/p_0)$ and $(s, 1/p) = (s_1, 1/p_1)$, then it is well posed whenever $(s, 1/p)$ lies on the line segment connecting the two points $(s_0, 1/p_0)$ and $(s_1, 1/p_1)$. Indeed, such techniques were used in [9] to prove Theorem 4.1. Thus, regions such as those in Figure 3 (or Figures 1–2) are generally expected to be convex.

The region of Figure 3 is indeed convex. To see this, observe that the given region is a polygon, and thus we need only establish convexity in a neighborhood of each vertex. The vertices at the origin, $(\alpha, 0)$, $(1, 1)$, and $(1, 1 + \alpha^*/\beta)$ clearly have convex neighborhoods. We are left with the vertex at the point $(1 - \alpha^*, 1)$.

The uppermost edge (connecting the points $(1 - \alpha^*, 1)$ and $(1, 1 + \alpha^*/\beta)$) has slope $1/\beta$. The adjacent edge connecting the point $(1 - \alpha^*, 1)$ to the origin has slope $1/(1 - \alpha^*)$. But because $\alpha^* \geq 1 - \beta$, we see that $1/(1 - \alpha^*) \geq 1/\beta$, and so there must be a convex neighborhood of the vertex $(1 - \alpha^*, 1)$ as well.

5. An extrapolation result

The result of the previous section is valid in the greatest known generality of coefficients and domains. It is essentially a consequence of the De Giorgi–Nash–Moser and Meyers inequalities which are a part of fundamental elliptic theory. As a consequence, we observe that the pentagon in Figure 3 is necessarily smaller than the hexagon in Figure 2.

The goal of our project, however, was to exploit new well-posedness results for the Dirichlet problem in L^p to obtain an enhanced range of Besov spaces. We now return to this task, and the theorem below will generalize and recover most of the results in Section 3, in particular, those in Figure 2, and will go far beyond that.

To this end, recall that the Dirichlet problem is well posed in L^p for some $1 < p < \infty$ if for any $f \in L^p(\partial\Omega)$ the solution u to the continuous Dirichlet

problem (2.5) satisfies the nontangential estimate

$$\|Nu\|_{L^q(\partial\Omega)} \leq C_q \|f\|_{L^q(\partial\Omega)} \tag{5.1}$$

and if such a solution is unique. Here, N is the nontangential maximal function of formula (3.11).

Well-posedness of the Dirichlet boundary value problem in L^p is a powerful feature. It has long been known that it is equivalent to (quantitative) absolute continuity of harmonic measure with respect to the Hausdorff measure on the boundary; see [36, Section 4] or [35] for a detailed discussion. Under suitable connectivity assumptions and assumptions on the coefficients, well-posedness of the L^p-Dirichlet problem for some p is equivalent to uniform rectifiability for $\mathfrak{d} = n - 1$; see [5,38]. We will highlight some recent work in this context in Section 7. For now, let us state our second main result.

Theorem 5.1. *Suppose that the conditions of Theorem 4.1 are satisfied.*

Suppose furthermore that there is a $q \in (1, \infty)$ such that the L^q-Dirichlet problem is well posed. That is, suppose that there is a $q \in (1, \infty)$ and a $C_q < \infty$ such that if f is continuous and compactly supported, then the solution u to problem (2.5) also satisfies estimate (5.1).

Then there is a number $\mathfrak{b}^* \in (0, 1]$ such that the conclusions of Theorem 4.1 are valid for $(s, 1/p)$ in the region indicated on the left in Figure 4.

If instead there is a $q^* \in (1, \infty)$ such that the solution u to problem (2.5), with A replaced by A^*, satisfies estimate (5.1) (with q replaced by q^*), then there is a $\mathfrak{b} \in (0, 1]$ such that the conclusions of Theorem 4.1 are valid for $(s, 1/p)$ in the region indicated on the right in Figure 4.

Furthermore, the numbers $\mathfrak{b}^*$ and $\mathfrak{b}$ satisfy the estimates

$$\mathfrak{b} \geq \alpha, \quad \mathfrak{b}^* \geq \alpha^*, \quad \mathfrak{b} \geq 1 + \frac{\mathfrak{d}}{q^*} - \mathfrak{d}, \quad \mathfrak{b}^* \geq 1 + \frac{\mathfrak{d}}{q} - \mathfrak{d},$$

where α and α^* are the boundary De Giorgi–Nash exponents for L and L^*.

(In the event that q and q^* both exist, the conclusions of Theorem 4.1 are valid for $(s, 1/p)$ in the region indicated on the bottom in Figure 4.)

Theorem 4.1 is essentially the $q = q^* = \infty$ case of Theorem 5.1; note that estimate (5.1) for $q = \infty$ reduces to the classical maximum principle. Expressed algebraically, the pair $(s, 1/p)$ satisfies the conditions of Theorem 4.1 if

$$p = \infty \quad \text{and} \quad 0 < s < \mathfrak{b},$$

$$p \leq 1 \quad \text{and} \quad \frac{\mathfrak{d}}{p} - \mathfrak{d} + 1 - \mathfrak{b}^* < s < 1,$$

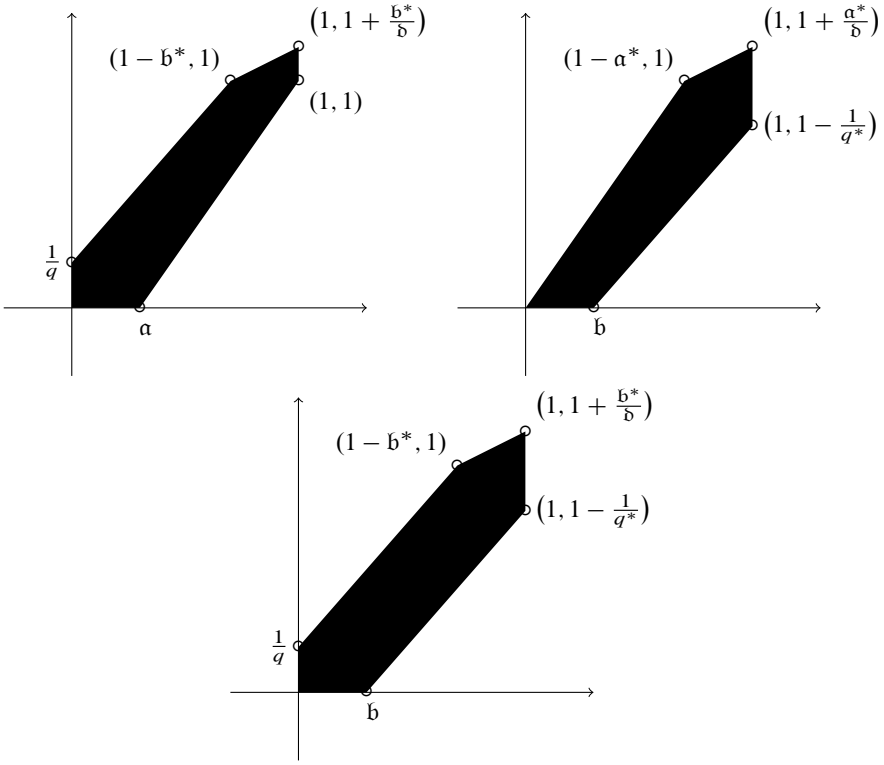


Figure 4. Values of the parameters $(s, 1/p)$ such that the $\dot{B}^{p,p}_s(\partial\Omega)$-Dirichlet problem for L is solvable, given solvability of the L^q-Dirichlet problem for $-\operatorname{div} A\nabla$ (on the left) or the L^{q^*}-Dirichlet problem for $-\operatorname{div} A^*\nabla$ (on the right) with nontangential estimates. Because $\mathfrak{b}^* \geq 1 - \mathfrak{d} + \mathfrak{d}/q$, by an argument similar to that in Remark 4.2, the illustrated region is convex. In the common case that q and q^* both exist, the $\dot{B}^{p,p}_s(\partial\Omega)$-Dirichlet problem for $-\operatorname{div} A\nabla$ is solvable for values of $(s, 1/p)$ as in the bottom picture.

or

$$1 < p < \infty, 0 < s < 1, \quad \text{and} \quad \frac{q-p}{qp-p}(1-\mathfrak{b}^*) < s < \mathfrak{b} + \frac{q^*}{pq^*-p}(1-\mathfrak{b}).$$

As in Remark 4.2, the condition $\mathfrak{b}^* \geq 1 + \mathfrak{d}/q - \mathfrak{d}$ implies that the regions in Figure 4 are convex in a neighborhood of the point $(1 - \mathfrak{b}^*, 1)$ and thus are convex.

Our methods of proof do not require that the number $\mathfrak{d}$ in Theorem 4.1 or 5.1 be an integer. However, in most of the cases in which the nontangential estimate (5.1) is known to hold for a finite q, the boundary $\partial\Omega$ is of integer dimension. (Some exceptions do exist; see in particular [18].) That is, in most of the known applications of Theorem 5.1, $\mathfrak{d}$ is an integer.

6. Further definitions

In Section 7, we will discuss some known cases in which the L^q-Dirichlet problem is well posed and thus some cases in which we may apply Theorem 5.1. It is known that the L^q-Dirichlet problem requires additional conditions on the coefficients and domains, beyond those listed in Theorem 4.1; see [14, 57]. That is, such results necessarily require more stringent conditions on the coefficients and domains. Thus, in this section, we will define some of those more stringent conditions.

6.1. Geometric conditions

Definition 6.1 (Uniformly rectifiable). Let $\mathfrak{d}$, $\mathfrak{n}$ be integers with $1 \leq \mathfrak{d} \leq \mathfrak{n} - 1$, and let F be a closed subset of $\mathbb{R}^{\mathfrak{n}}$. We say that F is uniformly rectifiable (or uniformly $\mathfrak{d}$-rectifiable) if F is $\mathfrak{d}$-Ahlfors regular and has big pieces of Lipschitz images; that is, if there exist positive constants θ and M such that if $x \in F$ and $0 < r < \operatorname{diam} F$, then there is a Lipschitz function $\varphi : B_{\mathfrak{d}}(0, r) \to \mathbb{R}^{\mathfrak{n}}$, where $B_{\mathfrak{d}}(0, r) \subset \mathbb{R}^{\mathfrak{d}}$ is the $\mathfrak{d}$-dimensional ball of radius r such that the Lipschitz constant of φ is at most M and such that

$$\sigma(F \cap B_{\mathfrak{n}}(x, r) \cap \varphi(B_{\mathfrak{d}}(0, r))) \geq \theta r^{\mathfrak{d}}.$$

Definition 6.2 (Harnack chains). Let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$. We say that Ω satisfies the *Harnack chain condition* if there are constants $M \geq 1$ and $m \in (0, 1)$ such that if $x, y \in \Omega$, then there is an integer N, depending only on the ratio

$$\frac{|x - y|}{\min(\operatorname{dist}(x, \partial\Omega), \operatorname{dist}(y, \partial\Omega))}$$

such that there is a chain of open balls $\{B_k\}_{k=1}^{N}$ with $x \in B_1$, $y \in B_N$, $B_k \cap B_{k+1} \neq \emptyset$ for all $1 \leq k < N$, and $m \operatorname{diam}(B_k) \leq \operatorname{dist}(B_k, \partial\Omega) \leq M \operatorname{diam}(B_k)$.

By [20, Lemma 2.1], and as noted in Remark 2.5, if $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular for some $0 < \mathfrak{d} < \mathfrak{n} - 1$, then Ω automatically satisfies the Harnack chain condition. However, the higher codimension condition does not guarantee uniform rectifiability.

Definition 6.3 (Chord-arc domain). Let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$. We call Ω a chord-arc domain if Ω satisfies the interior corkscrew condition (Definition 2.3) and the Harnack chain condition, if $\partial\Omega$ is $(\mathfrak{n} - 1)$-Ahlfors regular, and if Ω, in addition, satisfies the exterior corkscrew condition, that is, if the exterior domain $\mathbb{R}^{\mathfrak{n}} \setminus \overline{\Omega}$ also satisfies the interior corkscrew condition.

Remark 6.4. By connecting corkscrew points with Harnack chains, we can construct a λ-carrot path from any point $x \in \partial\Omega$ to any point $y \in \Omega$ with $|x - y| <$

$N\,\mathrm{dist}(y,\partial\Omega)$, where λ depends only on N, the dimension n, and the corkscrew and Harnack chain constants. Thus, a chord-arc domain is a weak local John domain and so satisfies the conditions of Theorem 4.1.

Remark 6.5. A chord-arc domain necessarily has uniformly rectifiable boundary; this follows from [22] or [63] combined with [39].

Definition 6.6 (Chord-arc surfaces with small constant). Let $\Gamma \subset \mathbb{R}^n$ be closed and unbounded. We say that Γ is a chord-arc surface with constant ε if, for every $x \in \Gamma$ and $r > 0$, there is a Lipschitz function $\varphi_x : \mathbb{R}^{n-1} \to \mathbb{R}$ with Lipschitz constant at most ε and a rigid motion T_x such that if

$$G = \{T_x((y,\varphi(y))) : y \in \mathbb{R}^{n-1}\},$$

then $G \cap B(x, r/2) \neq \emptyset$ and

$$\sigma(\Gamma \cap B(x,r) \setminus G) + \sigma(G \cap B(x,r) \setminus \Gamma) < \varepsilon r^{n-1}.$$

Remark 6.7. By [23, Lemma 5.4] and the following remarks, if $\Omega \subset \mathbb{R}^n$ and $\partial\Omega$ is a chord-arc surface with sufficiently small constant ε, then Ω and $\mathbb{R}^n \setminus \overline{\Omega}$ are both chord-arc domains.

6.2. Conditions on coefficients

In this section, we will state one form of the Dahlberg–Kenig–Pipher (DKP) condition on coefficients A. Early forms of this condition were proposed by Dahlberg and investigated by Kenig and Pipher in [45], and many weaker forms have been studied.

In particular, for simplicity of exposition, we will state the DKP condition only in the classical codimension 1 case (that is, $\mathfrak{d} = n - 1$). However, generalizations of these notions to the higher codimensional case have been investigated (see [19, 30]) and are also a present active area of research.

Definition 6.8 (The weak DKP condition). Let $\Omega \subset \mathbb{R}^n$ be open and such that $\partial\Omega$ is $(n-1)$-Ahlfors regular. Let A be a matrix-valued function defined on Ω that satisfies the ellipticity conditions (2.2)–(2.3) for some $\Lambda > \lambda > 0$ with $\mathfrak{d} = n - 1$.

We say that A satisfies the weak DKP condition with constant M in Ω if $A = B + C$, where

$$\sup_{x\in\partial\Omega}\sup_{r>0}\frac{1}{r^{n-1}}\int_{B(x,r)\cap\Omega}|\nabla B(y)|^2\,\mathrm{dist}(y,\partial\Omega) + |C(y)|^2\frac{1}{\mathrm{dist}(y,\partial\Omega)}\,dy = M < \infty.$$

7. Corollaries of Theorem 5.1

The L^q-Dirichlet problem has been a subject of intense study over the past few decades, and so bound (5.1) is known to be valid in many cases beyond Lipschitz domains. A complete survey of such results is beyond the scope of this note; in this section, we will describe a few such results.

7.1. Higher codimensional results

The most elegant (albeit not the most general) of the known results for the Dirichlet problem in the higher codimensional case was established in [24] (and later by another method in [28]) and is as follows.

Let $\mathfrak{d}$ be an integer with $1 \leq \mathfrak{d} \leq \mathfrak{n} - 2$. Let $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ and suppose that $\partial\Omega$ is $\mathfrak{d}$-Ahlfors regular and uniformly rectifiable. (By Remark 2.5, because $\mathfrak{d} < \mathfrak{n} - 1$, we have that Ω satisfies the conditions of Theorem 4.1.)

If $\beta > 0$, define the function $D_\beta : \Omega \to \mathbb{R}$ by

$$D_\beta(x) = \left(\int_{\partial\Omega} |x - y|^{-\mathfrak{d}-\beta} \, d\sigma(y) \right)^{-1/\beta}.$$

By [19, Lemma 5.1], $D_\beta(x) \approx \operatorname{dist}(x, \partial\Omega)$.

Let $A_\beta(x) = D_\beta(x)^{\mathfrak{d}+1-\mathfrak{n}} I$, where I is the identity matrix, and let $L_\beta = \operatorname{div} A_\beta \nabla$ (as defined by Definition 2.13). Observe that A satisfies the ellipticity conditions (2.2)–(2.3) and is also a symmetric matrix, and hence $L_\beta = L_\beta^*$.

By [28, Theorem 1.14] or [24, Theorem 1.1], the harmonic measure associated to L is in $A_\infty(\sigma)$; by [50, Theorem 4.1], this implies that the L^q-Dirichlet problem is well posed for some $q \in (1, \infty)$.

Thus, for a uniformly rectifiable boundary, and for the operator L_β, $\beta > 0$, we have that the conditions of Theorem 5.1 are satisfied, and so the solution to the Poisson–Dirichlet problem (4.1) satisfies estimate (4.2) for all $(s, 1/p)$ in the hexagon at the bottom of Figure 4.

Especially in the higher codimensional setting, the question of coefficients A for which the L^q-Dirichlet problem is well posed is an active area of research, and we expect further results of this type to be proven soon. Some (somewhat complicated) results of this type are known; we refer the interested reader to [19,30] for the details.

7.2. The Laplace operator and the DKP condition

We now turn to known results for the codimension 1 case, that is, the case $\mathfrak{d} = \mathfrak{n} - 1$.

As in the previous section, the most elegant known results are those for a particular operator; as is usual in the codimension 1 case, the favored particular operator is the Laplacian. By [5], if $\Omega \subset \mathbb{R}^{\mathfrak{n}}$ is a weak local John domain and satisfies the

interior corkscrew condition, and if $\partial\Omega$ is $(n-1)$-Ahlfors regular and uniformly rectifiable, then there is a $q \in (1,\infty)$ such that the L^q-Dirichlet problem is well posed for $L = -\Delta$, that is, for coefficients $A = I$, where I is the identity matrix.

There has been considerable investigation into the weakest possible sufficient conditions (and the strongest possible necessary conditions!) for well-posedness of the L^q-Dirichlet problem. Many forms of the DKP condition have been investigated under many conditions on domains. We will now mention one known well-posedness result for operators satisfying a quite weak version of the DKP condition (defined above); this is not the only such known result, and in fact, some results are known in more general domains.

Specifically, if $\Omega \subset \mathbb{R}^n$ is a chord-arc domain, and if A satisfies the weak DKP condition of Definition 6.8 in Ω for some finite M, then there is a $q \in (1,\infty)$ such that the L^q-Dirichlet problem for $-\operatorname{div} A\nabla$ is well posed.

This is based on known results as follows. By [29, Theorem 1.21], if $\Omega \subset \mathbb{R}^n$ is open, if $\partial\Omega$ is $(n-1)$-Ahlfors regular and uniformly rectifiable, if A satisfies the weak DKP condition, and if $-\operatorname{div} A\nabla u = 0$ and u is bounded in Ω, then

$$\int_{B(x,r)\cap\Omega} |\nabla u(y)|^2 \operatorname{dist}(y,\partial\Omega)\, dy \le C\|u\|_{L^\infty}^2 \sigma(B(x,r)\cap\partial\Omega) \tag{7.1}$$

for all $x \in \partial\Omega$ and all $r > 0$.

By [16, Theorem 1.1], if Ω is a 1-sided chord-arc domain, then the Carleson condition (7.1) implies that the harmonic measure for $-\operatorname{div} A\nabla$ lies in $A_\infty(\partial\Omega)$.

Let ω^y be the harmonic measure for $-\operatorname{div} A\nabla$ with pole at y, and let k^y be the Radon–Nikodym derivative of ω^y with respect to the surface measure. As observed just following [16, Definition 2.13], their A_∞ condition implies that there exist a $C < \infty$ and a $p \in (1,\infty)$ such that if $x \in \partial\Omega$, $0 < r < \operatorname{diam}\Omega$, and $y = A_r(x)$ is the corkscrew point to x at scale r, then

$$\left(\fint_{\Delta(x,r)} (k^y)^p\, d\sigma\right)^{1/p} \le C \fint_{\Delta(x,r)} k^y\, d\sigma.$$

It is well known (see, for example, [36, Proposition 4.5]) that this reverse Hölder estimate on the harmonic measure yields well-posedness of the L^q-Dirichlet problem for $q = p/(p-1)$, that is, for q the Hölder conjugate of p.

7.3. The small DKP condition

We conclude this discussion by mentioning a "small constant" result. That is, notice that the results of the previous two sections all imply the existence of a $q \in (1,\infty)$ such that the L^q-Dirichlet problem is well posed, without ever specifying precise values of q. By requiring more of our coefficients and domains, it is possible to choose a value of q and find situations in which the L^q-Dirichlet problem is well posed.

Specifically, let $q \in (1, \infty)$, let $\mathfrak{n} \geq 3$ be an integer, and let $\Lambda > \lambda > 0$. Then there is a $\delta > 0$ (depending on q, $\mathfrak{d}$, Λ, and λ) such that if

(1) $\Omega \subset \mathbb{R}^{\mathfrak{n}}$,

(2) $\partial\Omega$ is a chord-arc surface with constant at most δ (as defined in Definition 6.6),

(3) A satisfies the ellipticity conditions (2.2)–(2.3) (with $\mathfrak{d} = \mathfrak{n} - 1$), and

(4) A satisfies the weak Dahlberg–Kenig–Pipher condition in Ω with constant at most δ (as defined in Definition 6.8),

then the L^q-Dirichlet problem is well posed.

This result follows from known results in the literature by a line of argument similar to the previous section. First, by the recent results of [23], the given conditions imply a quantitative local A_∞-type condition on the harmonic measure.[4] Second, one can show that this A_∞-type condition implies a local reverse Hölder condition on the density (that is, the Radon–Nikodym derivative with respect to the surface measure) of the harmonic measure. Finally, as before, by [36, Proposition 4.5], this reverse Hölder estimate yields well-posedness of the L^q-Dirichlet problem.

The precise form of the A_∞-type condition varies from paper to paper, and some work must be done to relate particular forms to the required reverse Hölder condition. Moreover, in the small constant case, we must at every step ensure that our results work for the chosen value of q and not, as in the previous section, merely for sufficiently large finite q. This means that the argument for this result is somewhat more involved than in the previous section; we will give the details in the appendix of [9].

We emphasize that all of the results mentioned in Section 7 (and many other recent advancements in this spirit) extrapolate automatically to well-posedness within the subregions in Figure 4 by virtue of Theorem 5.1. In particular, in the small constant case, the region includes arbitrarily large compact subsets of the square $(0, 1) \times (1, \infty)$. We emphasize that this is not a result of the more traditional interpolation between the Dirichlet problem and the regularity problem, as in most of these cases, the regularity problem has not been treated.

Funding. Svitlana Mayboroda is partially supported by the NSF RAISE DMS 1839077 and the Simons Initiative of Geometry of Flows. Svitlana Mayboroda

[4] The weak DKP condition required by [23] is not quite that required by [29]. To lighten the exposition, we have chosen to state only the stronger of these two weak DKP conditions. It is straightforward to establish that the weak DKP condition of [23] implies the weak DKP condition of [29] in the setting of [23], that is, in domains Ω such that the Whitney boxes of [23, formula (1.2)] have volume comparable to $r^{\mathfrak{n}}$.

and Alberto Pacati are both partially supported by the Simons Collaboration on Localisation of Waves, 563916 SM.

References

[1] K. Adimurthi, T. Mengesha, and N. C. Phuc, Gradient weighted norm inequalities for linear elliptic equations with discontinuous coefficients. *Appl. Math. Optim.* **83** (2021), no. 1, 327–371 Zbl 1465.35160 MR 4221600

[2] K. Adimurthi and N. C. Phuc, An end-point global gradient weighted estimate for quasilinear equations in non-smooth domains. *Manuscripta Math.* **150** (2016), no. 1–2, 111–135 Zbl 1342.35053 MR 3483172

[3] A. Amenta and P. Auscher, *Elliptic boundary value problems with fractional regularity data. The first order approach*. CRM Monogr. Ser. 37, American Mathematical Society, Providence, RI, 2018, 152 pp. Zbl 1398.35001 MR 3753666

[4] P. Auscher and M. Qafsaoui, Observations on $W^{1,p}$ estimates for divergence elliptic equations with VMO coefficients. *Boll. Unione Mat. Ital. Ser. (8)* **5-B** (2002), no. 2, 487–509 Zbl 1173.35419 MR 1911202

[5] J. Azzam, S. Hofmann, J. M. Martell, M. Mourgoglou, and X. Tolsa, Harmonic measure and quantitative connectivity: geometric characterization of the L^p-solvability of the Dirichlet problem. *Invent. Math.* **222** (2020), no. 3, 881–993 Zbl 1453.31009 MR 4169053

[6] A. Barton, The L^p Neumann problem for higher order elliptic equations. *Mem. Differ. Equ. Math. Phys.* **96** (2025), 29–80 Zbl 08152298 MR 5012550

[7] A. Barton and S. Mayboroda, Layer potentials and boundary-value problems for second order elliptic operators with data in Besov spaces. *Mem. Amer. Math. Soc.* **243** (2016), no. 1149, 110 pp. Zbl 1378.35100 MR 3517153

[8] A. Barton, S. Mayboroda, and A. Pacati, The Dirichlet problem with boundary data of fractional smoothness in domains with Ahlfors regular boundary. 2026, in preparation

[9] A. Barton, S. Mayboroda, and A. Pacati, The Poisson problem in domains with Ahlfors regular boundary. 2026, arXiv:2603.22067v1

[10] K. Brewster and M. Mitrea, Boundary value problems in weighted Sobolev spaces on Lipschitz manifolds. *Mem. Differ. Equ. Math. Phys.* **60** (2013), 15–55 Zbl 1310.35089 MR 3288169

[11] S.-S. Byun, Elliptic equations with BMO coefficients in Lipschitz domains. *Trans. Amer. Math. Soc.* **357** (2005), no. 3, 1025–1046 Zbl 1087.35027 MR 2110431

[12] S.-S. Byun and L. Wang, Elliptic equations with BMO coefficients in Reifenberg domains. *Comm. Pure Appl. Math.* **57** (2004), no. 10, 1283–1310 Zbl 1112.35053 MR 2069724

[13] L. Caffarelli, E. Fabes, S. Mortola, and S. Salsa, Boundary behavior of nonnegative solutions of elliptic operators in divergence form. *Indiana Univ. Math. J.* **30** (1981), no. 4, 621–640 Zbl 0512.35038 MR 0620271

[14] L. A. Caffarelli, E. B. Fabes, and C. E. Kenig, Completely singular elliptic-harmonic measures. *Indiana Univ. Math. J.* **30** (1981), no. 6, 917–924 Zbl 0482.35020 MR 0632860

[15] M. Cao, P. Hidalgo-Palencia, J. M. Martell, C. Prisuelos-Arribas, and Z. Zhao, Elliptic operators in rough sets and the Dirichlet problem with boundary data in Hölder spaces. *J. Funct. Anal.* **288** (2025), no. 5, article no. 110801, 68 pp. Zbl 1567.35122 MR 4842691

[16] J. Cavero, S. Hofmann, J. M. Martell, and T. Toro, Perturbations of elliptic operators in 1-sided chord-arc domains. Part II. Non-symmetric operators and Carleson measure estimates. *Trans. Amer. Math. Soc.* **373** (2020), no. 11, 7901–7935 Zbl 1453.31011 MR 4169677

[17] B. E. J. Dahlberg, D. S. Jerison, and C. E. Kenig, Area integral estimates for elliptic differential operators with non-smooth coefficients. *Ark. Mat.* **22** (1984), no. 1–2, 97–108 Zbl 0537.35025 MR 0735881

[18] G. David, M. Engelstein, and S. Mayboroda, Square functions, nontangential limits, and harmonic measure in codimension larger than 1. *Duke Math. J.* **170** (2021), no. 3, 455–501 Zbl 1471.31001 MR 4255042

[19] G. David, J. Feneuil, and S. Mayboroda, Dahlberg's theorem in higher co-dimension. *J. Funct. Anal.* **276** (2019), no. 9, 2731–2820 Zbl 1414.42032 MR 3926132

[20] G. David, J. Feneuil, and S. Mayboroda, Elliptic theory for sets with higher co-dimensional boundaries. *Mem. Amer. Math. Soc.* **274** (2021), no. 1346, 123 pp. Zbl 1483.35099 MR 4341338

[21] G. David, J. Feneuil, and S. Mayboroda, *Elliptic theory in domains with boundaries of mixed dimension*. Astérisque 442, Société Mathématique de France, Paris, 2023, 139 pp. Zbl 1567.35123 MR 4619112

[22] G. David and D. Jerison, Lipschitz approximation to hypersurfaces, harmonic measure, and singular integrals. *Indiana Univ. Math. J.* **39** (1990), no. 3, 831–845 Zbl 0758.42008 MR 1078740

[23] G. David, L. Li, and S. Mayboroda, Small A_∞ results for Dahlberg–Kenig–Pipher operators in sets with uniformly rectifiable boundaries. *Trans. Amer. Math. Soc.* **376** (2023), no. 11, 7857–7909 Zbl 1528.35038 MR 4657223

[24] G. David and S. Mayboroda, Harmonic measure is absolutely continuous with respect to the Hausdorff measure on all low-dimensional uniformly rectifiable sets. *Int. Math. Res. Not. IMRN* **2023** (2023), no. 11, 9319–9426 Zbl 1527.35149 MR 4597210

[25] G. Di Fazio, L^p estimates for divergence form elliptic equations with discontinuous coefficients. *Boll. Unione Mat. Ital. A (7)* **10** (1996), no. 2, 409–420 Zbl 0865.35048 MR 1405255

[26] E. Fabes, O. Mendez, and M. Mitrea, Boundary layers on Sobolev–Besov spaces and Poisson's equation for the Laplacian in Lipschitz domains. *J. Funct. Anal.* **159** (1998), no. 2, 323–368 Zbl 0930.35045 MR 1658089

[27] P. Fatou, Séries trigonométriques et séries de Taylor. *Acta Math.* **30** (1906), no. 1, 335–400 JFM 37.0283.01 MR 1555035

[28] J. Feneuil, Absolute continuity of the harmonic measure on low dimensional rectifiable sets. *J. Geom. Anal.* **32** (2022), no. 10, article no. 247, 36 pp. Zbl 1514.42029 MR 4456214

[29] J. Feneuil, L. Li, and S. Mayboroda, Green functions and smooth distances. *Math. Ann.* **389** (2024), no. 3, 2637–2727 Zbl 1545.35045 MR 4753073

[30] J. Feneuil, S. Mayboroda, and Z. Zhao, The Dirichlet problem in domains with lower dimensional boundaries. *Rev. Mat. Iberoam.* **37** (2021), no. 3, 821–910 Zbl 1466.35112 MR 4236798

[31] Y. Han, S. Lu, and D. Yang, Inhomogeneous Besov and Triebel–Lizorkin spaces on spaces of homogeneous type. *Approx. Theory Appl.* **15** (1999), no. 3, 37–65 Zbl 0957.46025 MR 1746473

[32] Y. Han, D. Müller, and D. Yang, A theory of Besov and Triebel–Lizorkin spaces on metric measure spaces modeled on Carnot–Carathéodory spaces. *Abstr. Appl. Anal.* **2008** (2008), no. 1, article no. 893409, 250 pp. Zbl 1193.46018 MR 2485404

[33] Y. Han and D. Yang, Some new spaces of Besov and Triebel–Lizorkin type on homogeneous spaces. *Studia Math.* **156** (2003), no. 1, 67–97 Zbl 1032.42025 MR 1961062

[34] Y. S. Han and E. T. Sawyer, Littlewood–Paley theory on spaces of homogeneous type and the classical function spaces. *Mem. Amer. Math. Soc.* **110** (1994), no. 530, 126 pp. Zbl 0806.42013 MR 1214968

[35] S. Hofmann, Quantitative absolute continuity of harmonic measure and the Dirichlet problem: a survey of recent progress. *Acta Math. Sin. (Engl. Ser.)* **35** (2019), no. 6, 1011–1026 Zbl 1421.31010 MR 3952700

[36] S. Hofmann and P. Le, BMO solvability and absolute continuity of harmonic measure. *J. Geom. Anal.* **28** (2018), no. 4, 3278–3299 Zbl 1405.42032 MR 3881972

[37] S. Hofmann and J. M. Martell, Uniform rectifiability and harmonic measure. I. Uniform rectifiability implies Poisson kernels in L^p. *Ann. Sci. Éc. Norm. Supér. (4)* **47** (2014), no. 3, 577–654 Zbl 1302.31007 MR 3239100

[38] S. Hofmann, J. M. Martell, S. Mayboroda, T. Toro, and Z. Zhao, Uniform rectifiability and elliptic operators satisfying a Carleson measure condition. *Geom. Funct. Anal.* **31** (2021), no. 2, 325–401 Zbl 1473.35165 MR 4268304

[39] S. Hofmann, J. M. Martell, and I. Uriarte-Tuero, Uniform rectifiability and harmonic measure. II. Poisson kernels in L^p imply uniform rectifiability. *Duke Math. J.* **163** (2014), no. 8, 1601–1654 Zbl 1323.31008 MR 3210969

[40] S. Hofmann, S. Mayboroda, and M. Mourgoglou, Layer potentials and boundary value problems for elliptic equations with complex L^∞ coefficients satisfying the small Carleson measure norm condition. *Adv. Math.* **270** (2015), 480–564 Zbl 1311.35074 MR 3286542

[41] D. Jerison and C. E. Kenig, The inhomogeneous Dirichlet problem in Lipschitz domains. *J. Funct. Anal.* **130** (1995), no. 1, 161–219 Zbl 0832.35034 MR 1331981

[42] D. S. Jerison and C. E. Kenig, Boundary behavior of harmonic functions in non-tangentially accessible domains. *Adv. Math.* **46** (1982), no. 1, 80–147 Zbl 0514.31003 MR 0676988

[43] A. Jonsson and H. Wallin, Function spaces on subsets of $\mathbf{R}^n$. *Math. Rep.* **2** (1984), no. 1, 221 pp. Zbl 0875.46003 MR 0820626

[44] C. E. Kenig and J. Pipher, The Neumann problem for elliptic equations with non-smooth coefficients. *Invent. Math.* **113** (1993), no. 1, 447–509 Zbl 0807.35030 MR 1231834

[45] C. E. Kenig and J. Pipher, The Dirichlet problem for elliptic equations with drift terms. *Publ. Mat.* **45** (2001), no. 1, 199–217 Zbl 1113.35314 MR 1829584

[46] D. Kim, Trace theorems for Sobolev–Slobodeckij spaces with or without weights. *J. Funct. Spaces Appl.* **5** (2007), no. 3, 243–268 Zbl 1136.46030 MR 2352844

[47] J. Kinnunen and S. Zhou, A boundary estimate for nonlinear equations with discontinuous coefficients. *Differential Integral Equations* **14** (2001), no. 4, 475–492 Zbl 1161.35394 MR 1799417

[48] P. I. Lizorkin, Boundary properties of functions from "weight" classes (in Russian). *Dokl. Akad. Nauk SSSR* **132** (1960), no. 3, 514–517. English translation. *Soviet Math. Dokl.* **1** (1960), 589–593 Zbl 0106.30802 MR 0123103

[49] S. Mayboroda and M. Mitrea, Sharp estimates for Green potentials on non-smooth domains. *Math. Res. Lett.* **11** (2004), no. 4, 481–492 Zbl 1073.35075 MR 2092902

[50] S. Mayboroda and Z. Zhao, Square function estimates, the BMO Dirichlet problem, and absolute continuity of harmonic measure on lower-dimensional sets. *Anal. PDE* **12** (2019), no. 7, 1843–1890 Zbl 1432.35074 MR 3986543

[51] V. Maz'ya, M. Mitrea, and T. Shaposhnikova, The Dirichlet problem in Lipschitz domains for higher order elliptic systems with rough coefficients. *J. Anal. Math.* **110** (2010), no. 1, 167–239 Zbl 1199.35080 MR 2753293

[52] O. Mendez and M. Mitrea, The Banach envelopes of Besov and Triebel–Lizorkin spaces and applications to partial differential equations. *J. Fourier Anal. Appl.* **6** (2000), no. 5, 503–531 Zbl 0972.46017 MR 1781091

[53] D. Mitrea, M. Mitrea, and L. Yan, Boundary value problems for the Laplacian in convex and semiconvex domains. *J. Funct. Anal.* **258** (2010), no. 8, 2507–2585 Zbl 1201.35095 MR 2593333

[54] M. Mitrea and M. Taylor, Potential theory on Lipschitz domains in Riemannian manifolds: Sobolev–Besov space results and the Poisson problem. *J. Funct. Anal.* **176** (2000), no. 1, 1–79 Zbl 0968.58023 MR 1781631

[55] M. Mitrea and M. Taylor, Sobolev and Besov space estimates for solutions to second order PDE on Lipschitz domains in manifolds with Dini or Hölder continuous metric tensors. *Comm. Partial Differential Equations* **30** (2005), no. 1–2, 1–37 Zbl 1068.58013 MR 2130803

[56] M. Mitrea and M. Taylor, The Poisson problem in weighted Sobolev spaces on Lipschitz domains. *Indiana Univ. Math. J.* **55** (2006), no. 3, 1063–1090 Zbl 1387.35166 MR 2244598

[57] L. Modica and S. Mortola, Construction of a singular elliptic-harmonic measure. *Manuscripta Math.* **33** (1980), no. 1, 81–98 Zbl 0505.35039 MR 0596380

[58] M. Mourgoglou, B. Poggi, and X. Tolsa, Solvability of the Poisson–Dirichlet problem with interior data in $L^{p'}$-Carleson spaces and its applications to the L^p-regularity problem. *J. Eur. Math. Soc. (JEMS)* (2025), DOI 10.4171/JEMS/1660

[59] D. Müller and D. Yang, A difference characterization of Besov and Triebel–Lizorkin spaces on RD-spaces. *Forum Math.* **21** (2009), no. 2, 259–298 Zbl 1171.42013 MR 2503306

[60] S. M. Nikol'skiĭ, *Approximation of functions of several variables and imbedding theorems* (in Russian). Nauka, Moscow, 1969, 420 pp. Zbl 0185.37901 MR 0310616

[61] S. M. Nikol'skiĭ, P. I. Lizorkin, and N. V. Miroshin, Weighted function spaces and their applications to the investigation of boundary value problems for degenerate elliptic equations (in Russian). *Izv. Vyssh. Uchebn. Zaved. Mat.* (1988), no. 8, 4–30. English translation. *Soviet Math. (Iz. VUZ)* **32** (1988), no. 8, 1–40 Zbl 0679.35030 MR 0971868

[62] N. C. Phuc, Weighted estimates for nonhomogeneous quasilinear equations with discontinuous coefficients. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **10** (2011), no. 1, 1–17 Zbl 1228.35260 MR 2829320

[63] S. Semmes, Analysis vs. geometry on a class of rectifiable hypersurfaces in $\mathbf{R}^n$. *Indiana Univ. Math. J.* **39** (1990), no. 4, 1005–1035 Zbl 0796.42014 MR 1087183

[64] V. V. Shan'kov, The averaging operator with variable radius, and the inverse trace theorem (in Russian). *Sibirsk. Mat. Zh.* **26** (1985), no. 6, 141–152. English translation. *Sib. Math. J.* **26** (1985), no. 6, 891–901 Zbl 0597.46032 MR 0816512

[65] H. Triebel, *Theory of function spaces*. Monogr. Math. 78, Birkhäuser, Basel, 1983, 281 pp. Zbl 0546.46027 MR 0781540

[66] S. V. Uspenskii, Imbedding theorems for classes with weights (in Russian). *Tr. Mat. Inst. Steklova* **60** (1961), 282–303. English translation. *Amer. Math. Soc. Transl. Ser. II* **87** (1970), 121–145 Zbl 0198.46106 MR 0136980

[67] N. Wiener, The Dirichlet problem. *J. Math. Phys.* **3** (1924), no. 3, 127–146 JFM 51.0361.01

[68] S. Yang, D. Yang, and W. Yuan, Global gradient estimates for Dirichlet problems of elliptic operators with a BMO antisymmetric part. *Adv. Nonlinear Anal.* **11** (2022), no. 1, 1496–1530 Zbl 1498.35217 MR 4422226

[69] D. Z. Zanger, The inhomogeneous Neumann problem in Lipschitz domains. *Comm. Partial Differential Equations* **25** (2000), no. 9–10, 1771–1808 Zbl 0959.35041 MR 1778780

Ariel Barton
Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA; aeb019@uark.edu

Svitlana Mayboroda
Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland;
svitlana.mayboroda@math.ethz.ch

Alberto Pacati
Department of Mathematics, ETH Zurich, 8092 Zurich, Switzerland;
alberto.pacati@math.ethz.ch

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/25

Cohomology rings of character varieties

Anton Mellit

Abstract. In this talk, I give an introduction and present some recent progress towards understanding the cohomology rings of character varieties of Riemann surfaces, such as a proof of the $P = W$ conjecture and the computation of the zero-dimensional cohomological Hall algebra (COHA). In the case of punctured sphere, I present an explicit description relating the cohomology rings to the Hilbert scheme of $\mathbb{C}^2$, refining conjectures of Hausel–Letellier–Rodriguez-Villegas and Chuang–Diaconescu–Donagi–Pantev. I explain how the general case should be related to the symplectic geometry of the Hilbert scheme.

1. Classical motivation

One way that character varieties naturally arise is through the study of linear ODEs. The monodromy theorem associates to a linear ODE a representation of the fundamental group. The famous twenty-first Hilbert problem asked if any representation arises in such a way. A historical review of this subject can be found in [2].

Referring the reader for details and references to [2, Section 3], we briefly summarize the ideas here. Suppose we have a linear ODE

$$\frac{\partial}{\partial t}\vec{\varphi}(t) = A(t)\vec{\varphi}(t) \quad (t \in \mathbb{C}). \tag{1.1}$$

Here, $A(t)$ is an $n \times n$ matrix-valued meromorphic function of one complex variable t, and we are looking for vector-valued solutions $\vec{\varphi}(t)$. Let $t_1, \dots, t_k$ be the poles of $A(t)$ (including ∞ if necessary), which we call *punctures*. Choose a basepoint $*$, away from the poles of $A(t)$. In a neighborhood of $*$, we have an n-dimensional vector space of local solutions of the ODE (1.1). For any loop γ based at $*$ avoiding the punctures, solving the ODE along γ produces a linear transformation of the space

Mathematics Subject Classification 2020: 14M35 (primary); 14H60, 14D20 (secondary).
Keywords: character variety, cohomology, moduli spaces, Higgs bundles, Coulomb branch, Hilbert scheme.

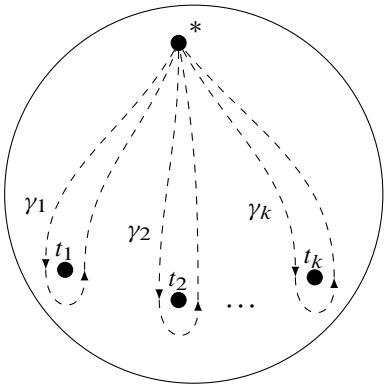


Figure 1. Fundamental group of a k-punctured sphere.

of solutions. Choosing a basis of the space of solutions, we obtain the *monodromy representation* of the fundamental group

$$\rho : \pi_1(*, \mathbb{P}^1(\mathbb{C}) \setminus \{t_1, \dots, t_k\}) \to \mathrm{GL}_n(\mathbb{C}).$$

A different choice of basis produces an isomorphic representation.

Another way to think about this is that local solutions to the ODE form a *locally constant sheaf* on $\mathbb{P}^1(\mathbb{C}) \setminus \{t_1, \dots, t_k\}$. Locally constant sheaves are also called *local systems*, and they are in a natural bijection with equivalence classes of monodromy representations; see [43, Corollary 2.6.2].

More explicitly, the fundamental group of $\mathbb{P}^1(\mathbb{C}) \setminus \{t_1, \dots, t_k\}$ is the free group on $k-1$ generators. Drawing loops around the punctures, as shown in Figure 1, one obtains the following presentation of the fundamental group:

$$\pi_1(*, \mathbb{P}^1(\mathbb{C}) \setminus \{t_1, \dots, t_k\}) = \langle \gamma_1, \dots, \gamma_k \mid \gamma_1 \cdots \gamma_k = 1 \rangle. \tag{1.2}$$

This presentation of π_1 by k generators and one relation is better than the free group presentation for the following reason. Let us assume that $A(t)$ has simple poles. In this case, analyzing the ODE locally around each puncture and finding a basis of solutions using *the Frobenius method* (see [44, Section 4.4]), one can compute the *local monodromy* matrices $\rho(\gamma_i)$, up to conjugation. For instance, the characteristic polynomial of $\rho(\gamma_i)$ coincides with the characteristic polynomial of the matrix $\exp\left(2\pi\sqrt{-1}A_i\right)$, where A_i is the residue of $A(t)$ at $t = t_i$.

Using the presentation (1.2), any monodromy representation corresponds to a k-tuple of $n \times n$ matrices $M_1, \dots, M_k$ satisfying

$$M_1 \cdots M_k = I_n, \tag{1.3}$$

where the local monodromy computation informs us about the conjugacy class of each M_i.

To summarize, the local monodromy, which is the data of the conjugacy classes of the matrices $M_1, \dots, M_k$, can be determined by an algebraic computation. The global monodromy, i.e., the representation ρ up to isomorphism, is the data of the tuple of matrices $M_1, \dots, M_k$ satisfying (1.3), up to a *simultaneous* conjugation. Unless in special cases, we know only how to compute the global monodromy by transcendental techniques. So it is probably hopeless to look for an algebraic way to compute the global monodromy exactly. On the other hand, it still makes sense to formulate the following problem.

Problem 1.1. Given n, k, and a collection of local monodromies (i.e., conjugacy classes), determine all algebraic relations among the matrix entries of the corresponding monodromy representations that are satisfied by every representation with these local monodromies.

This is the basic idea about the character variety. Before we give a precise definition, we consider an example which can be worked out explicitly.

2. The Fricke–Klein example

2.1. Equations

This example was essentially worked out in [13]. Let $k = 4$, $n = 2$. Let $a_1, \dots, a_4 \in \mathbb{C} \setminus \{-2, 2\}$. We exclude the cases -2 and 2 because in these cases the matrices M_i do not have distinct eigenvalues, and so the Jordan forms are not completely determined by the characteristic polynomials. Let us try to parameterize all 4-tuples of 2×2 matrices $M_1, M_2, M_3, M_4 \in \mathrm{SL}_2(\mathbb{C})$, up to conjugation, satisfying

$$\prod_{i=1}^{4} M_i = I_2, \quad \operatorname{tr} M_i = a_i.$$

Let $x = \operatorname{tr}(M_1 M_2)$, $y = \operatorname{tr}(M_1 M_3)$, $z = \operatorname{tr}(M_2 M_3)$. The following SAGE code computes the ideal of relations between x, y, z:

```
R.<m111,m112,m121,a1,m211,m212,m221,a2,m311,m312,m321,a3,x,y,z,a4>=QQ[]
M1=matrix(2,2,[m111,m112,m121,a1-m111])
M2=matrix(2,2,[m211,m212,m221,a2-m211])
M3=matrix(2,2,[m311,m312,m321,a3-m311])
eqns=[M1.det()-1,M2.det()-1,M3.det()-1,(M1*M2*M3).trace()-a4,(M1*M2).trace()-x, \
(M1*M3).trace()-y,(M2*M3).trace()-z]
J=R.ideal(eqns)
J.elimination_ideal([m111,m112,m121,m211,m212,m221,m311,m312,m321])
```

The computation yields the elimination ideal generated by the single cubic relation

```
Ideal (a1*a2*a3*a4 - a1*a2*x - a1*a3*y - a2*a3*z + x*y*z - a3*x*a4 - a2*y*a4 - a1*z*a4
+ a1^2 + a2^2 + a3^2 + x^2 + y^2 + z^2 + a4^2 - 4) of Multivariate Polynomial Ring in
m111, m112, m121, a1, m211, m212, m221, a2, m311, m312, m321, a3, x, y, z, a4 over
Rational Field
```

So for each 4-tuple of matrices satisfying the conditions, the corresponding $x, y, z \in \mathbb{C}$ as above produce a solution to the so-called *Markov equation* with parameters a_1, a_2, a_3, a_4:

$$xyz + x^2 + y^2 + z^2 - (a_1a_2 + a_3a_4)x - (a_1a_3 + a_2a_4)y - (a_1a_4 + a_2a_3)z + a_1a_2a_3a_4 + a_1^2 + a_2^2 + a_3^2 + a_4^2 - 4 = 0. \tag{2.1}$$

Conversely, suppose $x, y, z \in \mathbb{C}$ is a solution to the Markov equation. Let us see that for generically chosen values a_1, a_2, a_3, a_4, we can always uniquely recover the tuple of matrices, up to conjugation.

First, we observe that knowing the traces $x, y, z, a_1, a_2, a_3, a_4$, we can compute the trace of an arbitrary product of matrices $M_1, \dots, M_4$ and their inverses. Indeed, given such a product, we first replace any occurrence of M_4 by $M_3^{-1}M_2^{-1}M_1^{-1}$. Then using

$$A^{-1} = \operatorname{tr}(A)I_2 - A \quad (A \in \mathrm{SL}_2(\mathbb{C})),$$

we can reduce to the situation when no inverses appear. So we have an arbitrary product of M_1, M_2, M_3. Using

$$BA = \operatorname{tr}(BA)I_2 - A^{-1}B^{-1} \quad (A, B \in \mathrm{SL}_2(\mathbb{C}))$$

and getting rid of the inverses, we rearrange terms modulo products of shorter length until we obtain a linear combination of I_2, M_1, M_2, M_3, M_1M_2, M_1M_3, M_2M_3, $M_1M_2M_3 = M_4^{-1}$, all of whose traces are explicitly computed in terms of x, y, z and the parameters.

Knowing the traces of arbitrary products, let us find out if any 4 products form a basis of the space of 2×2 matrices. Indeed, because the trace pairing is non-degenerate, matrices $Q_1, \dots, Q_{n^2}$ form a basis of the space of $n \times n$ matrices if and only if the $n^2 \times n^2$ matrix of traces $\operatorname{tr}(Q_iQ_j)$ is non-degenerate.

Suppose products of $M_1, \dots, M_4$ do span the space of matrices, i.e., the monodromy representation is irreducible, and we choose a basis $Q_1, \dots, Q_4$. Knowing the traces allows us to express any other product as a linear combination of $Q_1, \dots, Q_4$. In particular, we recover the ring structure on the 4-dimensional space of linear combinations of $Q_1, \dots, Q_4$, and any monodromy representation provides a $\mathbb{C}$-linear isomorphism of this ring with the matrix ring. Any other monodromy representation with the same traces would produce a $\mathbb{C}$-linear automorphism of the matrix ring. All $\mathbb{C}$-linear automorphisms of the matrix ring are inner, i.e., related by a change-of-basis

transformation, by the Skolem–Noether theorem [28, p. 173], and so we conclude that the monodromy representations with the same traces are equivalent.

This was under the assumption that the monodromy representation is irreducible. If the representation is reducible, then in a suitable basis the matrices $M_1, \dots, M_4$ can be written in upper triangular form. In this situation, we cannot determine the matrices only from the trace information. Denoting the top-left diagonal entry of the upper triangular matrix M_i by λ_i, we must have $\prod_{i=1}^4 \lambda_i = 1$. On the other hand, λ_i is a root of the polynomial $t^2 - a_i t + 1$ which has 2 roots. So this kind of situation is possible only for special values of the parameters a_1, a_2, a_3, a_4.

More precisely, we call a_1, a_2, a_3, a_4 *generic* if none of the 16 products of the roots of $t^2 - a_i t + 1$ (we pick one root from each quadratic equation) equals 1. We have seen that if the parameters are generic, then all monodromy representations are irreducible and in one-to-one correspondence with the solutions of the Markov equation (2.1).

2.2. Geometry of the Markov cubic

In algebraic geometry, given an algebraic variety X, i.e., a set of solutions to a system of polynomial equations, we are usually interested in its cohomology ring $H^*(X, \mathbb{C})$. Hodge theory extended by Deligne to non-compact and singular spaces [10, 11] endows the cohomology with *mixed Hodge structure*. Structure constants of the ring typically encode interesting enumerative information. Sometimes the cohomology ring as a ring itself can be an interesting object of study, perhaps making a connection with a ring constructed in some other way in a completely different field of mathematics.

Although the ring structure in the Fricke–Klein situation turns out to be trivial, the cohomology already has a non-trivial mixed Hodge structure. We explain how to compute it.

Assuming the parameters $a_1, a_2, a_3, a_4 \in \mathbb{C} \setminus \{-2, 2\}$ are generic, the cubic X given by the Markov equation (2.1) is non-singular. Let $\overline{X} \subset \mathbb{CP}^3$ be the solution set of the homogenization of (2.1). Its affine part is $X \subset \overline{X}$, and the complement $\overline{X} \setminus X$ is given in $\mathbb{CP}^2$ as the solution set of the homogeneous equation $xyz = 0$; thus, it is a union of three lines forming a triangle. Denote this triangle by Δ. We refer to Figure 2. It is straightforward to check that homogenization of any cubic of the form

$$xyz + x^2 + y^2 + z^2 + \text{lower degree terms}$$

does not have any singularities at infinity. So $\overline{X}$ is a smooth cubic surface and therefore can be obtained as a blowup of $\mathbb{CP}^2$ at 6 points (for background, see [21]). The cubic $\overline{X}$, as any other smooth cubic surface, contains 27 projective lines. Three of them form Δ. The intersection graph of the 27 lines is well known, and one can see

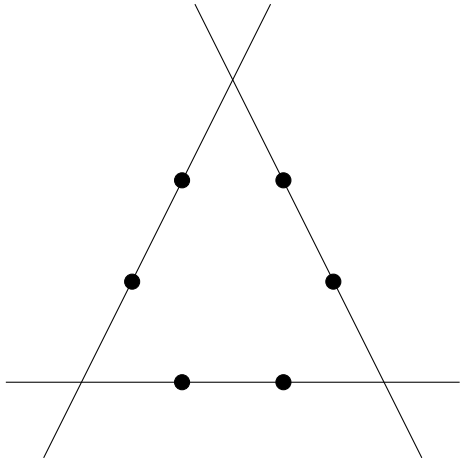


Figure 2. Markov's cubic as the complement of a triangle in the projective plane blown up at 6 points.

that for any triangle, one can choose 6 lines which do not intersect pairwise and do not contain any of the lines of the triangle. Each side of the triangle intersects 2 of the 6 lines. Blowing down these 6 lines, we obtain a map $f : \overline{X} \to \mathbb{CP}^2$. The 3 lines of Δ are isomorphically mapped to 3 lines on $\mathbb{CP}^2$. Choosing coordinates appropriately, one can assume that $f(\Delta)$ is the complement of $\mathbb{C}^* \times \mathbb{C}^* \subset \mathbb{CP}^2$. So we have

$$U = f^{-1}(\mathbb{C}^* \times \mathbb{C}^*) = \mathbb{C}^* \times \mathbb{C}^* \subset X,$$

an open subset,[1] and the complement $X \setminus U$ is a disjoint union of 6 copies of $\mathbb{C}$:

$$X = \mathbb{C}^* \times \mathbb{C}^* \sqcup \underbrace{\mathbb{C} \sqcup \cdots \sqcup \mathbb{C}}_{6 \text{ copies}}.$$

The 6 copies of $\mathbb{C}$ are divided into 3 pairs according to which of the three lines of Δ they intersect.

2.3. Cohomology

With this picture in mind, we have several ways to compute the cohomology. All cohomology groups we consider have complex coefficients. From the relative cohomology sequence for the pair $\mathbb{C}^* \times \mathbb{C}^* = U \subset X$, we obtain an exact sequence

$$\cdots \to H^1(X) \to H^1(U) \to H^2(X, U) \to H^2(X) \to H^2(U) \to \cdots . \qquad (2.2)$$

[1] It is possible to explicitly write down the map f and correspondingly the bijection between an open subset of X and $\mathbb{C}^* \times \mathbb{C}^*$, but the resulting formulas are cumbersome and do not seem to provide any useful information for our discussion, so we postpone this to another publication.

Here, since U is homotopy equivalent to the 2-torus, we have $H^0(U) = \mathbb{C}$, $H^1(U) = \mathbb{C}^2$, $H^2(U) = \mathbb{C}$. Consider the pair U, X. The complement $X \setminus U$ is the disjoint union of 6 copies of $\mathbb{C}$. Let us denote them by $Z_1, \dots, Z_6$. Since both $\mathbb{C}$ and X are non-singular, by the tubular neighborhood theorem we can find disjoint open neighborhoods $V_i \supset Z_i$, $V_i \subset X$, such that each V_i looks like a vector bundle of real rank 2 over $Z_i \cong \mathbb{C}$. By the excision theorem, we obtain

$$H^*(X, U) = \bigoplus_{i=1}^{6} H^*(V_i, V_i \setminus Z_i).$$

Using the Thom isomorphism

$$H^*(V_i, V_i \setminus Z_i) \cong H^{*-2}(Z_i)$$

and the contractibility of $Z_i \cong \mathbb{C}$, we conclude that $H^2(V_i, V_i \setminus Z_i) \cong \mathbb{C}$ and all other cohomologies vanish. So we obtain $H^2(X, U) \cong \mathbb{C}^6$ and all other cohomologies vanish. So the remaining part of the long exact sequence (2.2) vanishes, except for the isomorphism $H^0(X) \cong H^0(U) \cong \mathbb{C}$. The map $\mathbb{C}^2 = H^1(U) \to H^2(X, U) = \mathbb{C}^6$ is dual to the map $\mathbb{C}^6 \cong H_2(X, U) \to H_1(U) \cong \mathbb{C}^2$. Here, a point class on $Z_i \cong \mathbb{C}$ corresponds to a disc class in $H_2(X, U)$ whose boundary is a 1-cycle in U which gets contractible in $U \cup Z_i$. From Figure 2, we see that all 1-cycles in U become contractible in X, so the map $H_2(X, U) \to H_1(U)$ is surjective, the dual map is injective, and we have $H^1(X) = 0$, while $H^2(X)$ is an extension of $H^2(U) = \mathbb{C}$ by $H^2(X, U)/H^1(U) \cong \mathbb{C}^4$.

This is reflected in the weight filtration on $H^2(X)$. As a mixed Hodge structure, the subspace $H^2(X, U)/H^1(U) \subset H^2(X)$ is pure of weight 2, while $H^2(U)$ has weight 4. Thus, we have $W_2 = H^2(X, U)/H^1(U) \subset H^2(X)$ with $\dim W_2 = 4$, $W_3 = W_2$, and $W_4 = H^2(X)$ with $\dim W_4/W_2 = 1$.

Another way to compute the cohomology is to use the compactification. Let us denote $X_0 = \overline{X}$, $X_1 = X_0 \setminus \mathbb{CP}^1$, $X_2 = X_0 \setminus (\mathbb{CP}^1 \cup \mathbb{CP}^1)$, $X_3 = X_0 \setminus \Delta$, so that we pass from X_i to X_{i+1} by removing a line. This leads to exact sequences for each i:

$$0 \to H^1(X_{i+1}) \to H^2(X_i, X_{i+1}) \to H^2(X_i) \to H^2(X_{i+1}) \to H^3(X_i, X_{i+1}) \to 0$$
$$0 \to H^3(X_{i+1}) \to H^4(X_i, X_{i+1}) \to H^4(X_i) \to H^4(X_{i+1}) \to 0,$$

as we will see that the odd cohomologies of X_i vanish at each step. At each step, we will use that $H^k(X_i, X_{i+1}) = H^{k-2}(X_i \setminus X_{i+1})$ by the combination of the tubular neighborhood theorem, excision, and the Thom isomorphism, as detailed in the previous argument. Since $X_i \setminus X_{i+1}$ is $\mathbb{CP}^1$, $\mathbb{C}$ or $\mathbb{C}^*$, the latter cohomology is non-zero only for $k = 2, 3, 4$. Initially, we have $H^0(X_0) = \mathbb{C}$, $H^2(X_0) = \mathbb{C}^7$ (represented by 7 algebraic cycles: the line in $\mathbb{CP}^2$ and the 6 exceptional curves), $H^4(X_0) = \mathbb{C}$, all other cohomologies vanish.

For $i = 0$, we have $X_0 \setminus X_1 = \mathbb{CP}^1$, which has $H^0 = \mathbb{C}$ and $H^2 = \mathbb{C}$. These coincide with $H^2(X_0, X_1)$ and $H^4(X_0, X_1)$, respectively, and we see that in $H^2(X_1)$ (resp. $H^4(X_1)$), we have one less dimension than in $H^2(X_0)$, respectively, $H^4(X_0)$. So $H^4(X_1) = 0$, $H^2(X_1) = \mathbb{C}^6$.

For $i = 1$, we have $X_1 \setminus X_2 = \mathbb{C}$, which has only $H^0 = \mathbb{C}$. So X_2 has one dimension less in $H^2(X_2)$, while nothing else changes: $H^2(X_1) = \mathbb{C}^5$.

Finally, for $i = 2$, we have $X_2 \setminus X_3 = \mathbb{C}^*$, which has $H^0 = \mathbb{C}$ and $H^1 = \mathbb{C}$, which coincide with $H^2(X_2, X_3)$ and $H^3(X_2, X_3)$ correspondingly. Thus, two things happen. First, only 4 dimensions from $H^2(X_2)$ out of 5 contribute to $H^2(X_3)$. On the other hand, one extra dimension comes from $H^3(X_3, X_2)$. The net total is $H^2(X_3) = \mathbb{C}^5$, as in X_2, but the nature of the 5 dimensions is different, and correspondingly the weights are different. The 4 dimensions coming from $H^2(X_2)$ have weight 2, and this is reflected by the fact that they are represented by algebraic cycles. The remaining class has weight 4 as it comes from $H^3(X_3, X_2)$. The weight of $H^1(\mathbb{C}^*)$ is 2, but it gets increased by 2 when we pass to $H^3(X_3, X_2)$.

Yet another approach to describe the cohomology is by using differential forms and Chern classes. Notice that the moduli space X carries four natural line bundles $L_1, \dots, L_4$. Each point $p \in X$ corresponds to a 4-tuple of 2×2 matrices $M_1, \dots, M_4$. For each i, the matrix M_i has two one-dimensional eigenspaces, denoted E_i, E_i'. Letting p vary, one can show that the family of one-dimensional vector spaces $\mathrm{Hom}_{\mathbb{C}}(E_i, E_i')$ is locally trivial and therefore produces a line bundle, denoted L_i. Its Chern class $c_1(L_i)$ is an element of $H^2(X)$. These Chern classes $c_1(L_i)$ span $W_2H^2(X)$. To obtain one more class, we use the differential form

$$\omega = \frac{dx}{x} \wedge \frac{dy}{y}$$

on $U \cong \mathbb{C}^* \times \mathbb{C}^*$, written in the natural coordinates x, y on $\mathbb{C}^* \times \mathbb{C}^*$. Let us verify that ω extends to the whole of X. It is sufficient to consider the neighborhood of one of the exceptional lines. The blowup is locally described by the equation

$$x - 1 = zy,$$

and performing the substitution, we obtain

$$\omega = \frac{dz \wedge dy}{yz + 1},$$

which extends to $y = 0$ in a neighborhood of the line $\{(y, z) \in \mathbb{C}^2 \mid y = 0\}$.

These findings can be summarized in Table 1.

H^*	$W_i H^i$	$W_{i+2}H^i / W_i H^i$
H^2	$\mathbb{C}^4 = \langle c_1(L_i) \mid i = 1,2,3,4\rangle$	$\mathbb{C} = \langle \omega = \text{symplectic form}\rangle$
H^0	$\mathbb{C} = \langle 1\rangle$	

Table 1. Cohomology of the Markov cubic.

2.4. Mixed Hodge polynomial

The numerical information about cohomology is packaged in the *Poincaré polynomial*

$$\sum_i q^i \dim H^i(X).$$

In the case of an algebraic variety, the information about the weight filtration is encoded in the *mixed Hodge polynomial*[2]

$$P(X;q,t) = \sum_{i,j} q^i t^j \dim(W_j H^i(X)/W_{j-1}H^i(X)),$$

so that the Poincaré polynomial is its $t = 1$ specialization. From Figure 1, we obtain that the Poincaré polynomial of X is $1 + 5q^2$ and the mixed Hodge polynomial is $P(q,t) = 1 + 4q^2t^2 + q^2t^4$. We prefer to consider a slightly different version of the mixed Hodge polynomial, which encodes the same information. Let

$$\mathrm{MH}(X;q,t) := q^{\frac{1}{2}\dim_{\mathbb{C}} X} P\left(X; q^{-\frac{1}{2}}t^{-\frac{1}{2}}, t^{\frac{1}{2}}\right).$$

For instance, in our running example, we obtain

$$\mathrm{MH}(X;q,t) = q + t + 4.$$

We should also compare our calculations with the number of points on our variety over the finite field $\mathbb{F}_q$. Assuming all the lines are defined over the base field, we see that the first computation produces

$$(q-1)^2 + 6q = q^2 + 4q + 1.$$

Here, $\mathbb{C}^* \times \mathbb{C}^*$ gets replaced by $\mathbb{F}_q^* \times \mathbb{F}_q^*$, which has size $(q-1)^2$, and each of the six lines $\mathbb{C}$ becomes $\mathbb{F}_q$, which has size q. The second computation produces

$$q^2 + 7q + 1 - (q+1) - q - (q-1) = q^2 + 4q + 1.$$

[2]Normally, the mixed Hodge polynomial is defined as a polynomial in three variables. The third variable additionally records the Hodge decomposition of the pure Hodge structure $W_j H^i(X)/W_{j-1}H^i(X)$. In our situation, however, this decomposition is typically trivial, i.e., the pure Hodge structure is Tate.

Here, the number of points on $\overline{X}$, a smooth cubic surface, is q^2+7q+1. We remove subsequently $\mathbb{P}^1\mathbb{F}_q$ of size $q+1$, $\mathbb{F}_q$ of size q, and $\mathbb{F}_q^*$ of size $q-1$. In each case, one can track how the different powers of q contribute to the cohomology so that q^2 corresponds to H^0 and $4q+1$ corresponds to H^2.

In any case, for a general algebraic variety X, if the number of points is a polynomial of q, then the counting polynomial coincides with the E-polynomial, as shown by N. Katz in the appendix of [24]. Then we have

$$|X(\mathbb{F}_q)| = q^{\dim_{\mathbb{C}} X} P\left(X; -1, q^{-\frac{1}{2}}\right) = q^{\frac{1}{2}\dim_{\mathbb{C}} X} \mathrm{MH}(X; q, q^{-1}), \tag{2.3}$$

assuming MH is indeed a polynomial in q, t, which means that all the weights are even, and the cohomology is concentrated in even degrees.

3. Setup

After reviewing in detail the basic example, we turn to the general setup and precise definitions. There is no reason to consider only genus 0 complex curves, and the classical motivation directly extends to the more general case; for instance, ODEs can be studied on complex curves of arbitrary genus. So we consider a complex curve of genus g with k punctures.

Definition 3.1. Fix $g \geq 0$, $n, k \geq 1$ and Jordan forms $J_1, J_2, \ldots, J_k \in \mathrm{GL}_n$. The associated genus g *character variety* $X = \mathrm{Char}_{g;J_1,\ldots,J_k}$ is the quotient space

$$\{(A_1,\ldots,A_g,B_1,\ldots,B_g,M_1,\ldots,M_k) : M_i \sim J_i, \\ A_1B_1A_1^{-1}B_1^{-1}\cdots A_gB_gA_g^{-1}B_g^{-1}M_1\cdots \\ M_k = I\}/\mathrm{PGL}_n(\mathbb{C}),$$

where $\mathrm{PGL}_n(\mathbb{C})$ acts by simultaneously conjugating all the matrices.

In general, the quotient space is an Artin stack, but we will only consider generic conjugacy classes in which case the quotient is a smooth affine algebraic variety.

Definition 3.2 ([22]). The data $J_1,\ldots,J_k$ are *generic* if

(1) $\prod_{i=1}^k \det J_i = 1$

(2) for every $0 < n' < n$ and every block decomposition $J_i \sim \begin{pmatrix} J_i' & J_i'' \\ 0 & J_i''' \end{pmatrix}$ with $J_i' \in \mathrm{GL}_{n'}$, we have $\prod_{i=1}^k \det J_i' \neq 1$.

Theorem 3.3 ([22]). *For generic data $n, k, J_1, \dots, J_k$, the quotient space*

$$X = \mathrm{Char}_{g;J_1,\dots,J_k}$$

has the structure of a smooth affine algebraic variety.

In fact, considering functions on X of the form

$$\mathrm{tr}(M_{i_1} \text{ or } A_{i_1} \text{ or } B_{i_1} \cdots M_{i_m} \text{ or } A_{i_m} \text{ or } B_{i_m}),$$

we can realize X as an algebraic subset of $\mathbb{C}^N$ for sufficiently large N.

Basically, the same reasoning as in the Fricke–Klein situation can be applied in general. However, the number of traces one needs to consider is uncontrollably large, and it seems hopeless to be able to produce all equations by the elimination computation. Even if we do obtain a complete list of equations, it is not clear which interesting information we should be able to extract from such a list. For instance, equations in the case of genus 0, rank 2, five punctures are given in [42, Proposition 2.1]. We do hope, however, that computing the cohomology ring is an achievable goal.

Problem 3.4. Compute the cohomology ring $H^*(X, \mathbb{C})$.

4. What is known?

Having defined our basic object of study, we explain some important structural results.

4.1. Non-abelian Hodge correspondence

The non-abelian Hodge correspondence, pioneered by Hitchin [26] and developed by Simpson [39–41], is a construction of a remarkable bijection between local systems and stable Higgs bundles, which induces a diffeomorphism between the character variety and the moduli space that parameterizes stable Higgs bundles.

The construction of the character variety only depends on the topological type of the Riemann surface. Now we need to view the Riemann surface with punctures as a compact algebraic curve C over $\mathbb{C}$ with marked points. Then a *Higgs bundle* is a pair $(\mathcal{E}, \theta)$, where $\mathcal{E}$ is an algebraic vector bundle on C and θ a *Higgs field*, which is a twisted matrix-valued function on C of a certain kind. More precisely, $\theta : \mathcal{E} \to \mathcal{E} \otimes \omega_C(D)$, where $D = \sum_{i=1}^{k} [t_i]$ is the divisor on C corresponding to the marked points $t_1, \dots, t_k$. Explicitly, this means that on an open subset of C where $\mathcal{E}$ is trivialized, θ is a matrix whose entries are algebraic differential forms on C with simple residues at the marked points.

Furthermore, these Higgs bundles are required to be endowed with *parabolic structures* at the marked points. This is simply a choice of a partial flag at the fiber

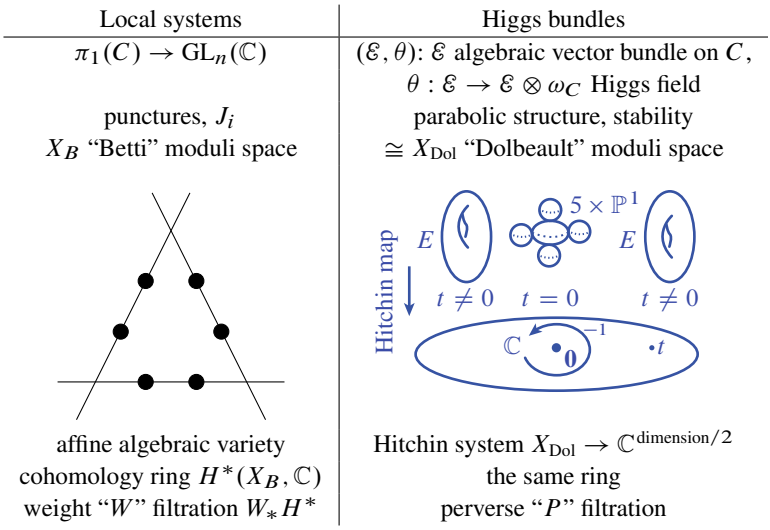


Figure 3. Non-abelian Hodge correspondence.

of $\mathcal{E}$ at each marked point, preserved by the residue $\mathrm{res}_{t_i}\,\theta$, which acts on that fiber by a linear transformation. The steps of the partial flag precisely match the dimensions of the generalized eigenspaces of the Jordan forms J_i. The data of a Higgs bundle with a parabolic structure have to satisfy a *stability condition*. The data of the eigenvalues and the Jordan blocks of J_i translate into *parabolic weights* defining the stability condition, and into the Jordan form of $\mathrm{res}_{t_i}\,\theta$ acting on the subquotients of the flag. Details are given in [39].

The moduli space of Higgs bundles in this context is referred to as the *Dolbeault moduli space* X_{Dol}, while the character variety is called the *Betti moduli space* X_B. The non-abelian Hodge correspondence provides a homeomorphism between X_B and X_{Dol}. In particular, their cohomology rings are isomorphic. We will use X when it is not important whether we take X_B or X_{Dol}. We summarize features appearing on both sides of this correspondence in the table in Figure 3.

4.2. Generators of $H^*(X)$

Generators of the cohomology ring are known. There is a general argument in [29] concerning moduli spaces of sheaves on complex surfaces. With a little modification, it can be applied to the situation of parabolic Higgs bundles. Basically speaking, certain Chern class of the complex of vector bundles on $X_{\mathrm{Dol}} \times X_{\mathrm{Dol}}$ computing the derived Hom between Higgs bundles is a multiple of the diagonal class. Moreover, there exists a compactification $\overline{X_{\mathrm{Dol}}} \supset X_{\mathrm{Dol}}$ such that the complex extends to $X_{\mathrm{Dol}} \times \overline{X_{\mathrm{Dol}}}$, and the Chern class remains a multiple of the diagonal class. Thus,

the identity map $H^*(X_{\mathrm{Dol}}) \to H^*(X_{\mathrm{Dol}})$ can be written in terms of the tautological classes.

Consider the tautological bundle[3] $\mathcal{E} \to X_{\mathrm{Dol}} \times C$. For each $i \in \{0, 1, 2\}$, $\xi \in H^i(C)$, and $m \in \mathbb{Z}_{\geq 0}$, define

$$e_m(\xi) := (\pi_X)_*(c_m(\mathcal{E}) \cup \xi) \in H^{2m+i-2}(X_{\mathrm{Dol}}). \tag{4.1}$$

Here, $c_m(\mathcal{E}) \in H^{2m}(X_{\mathrm{Dol}} \times C)$ is the Chern class and π_X is the projection $X_{\mathrm{Dol}} \times C \to X_{\mathrm{Dol}}$. If we have a marked point t_j, then each point of X_{Dol} gives a flag $0 = F_{j0} \subset F_{j1} \subset \cdots \subset F_{jr_j} = \mathcal{E}(t_j)$. Each subquotient of the flag produces a vector bundle on X_{Dol}, and we take its Chern classes $c_{m,j,\ell} := c_m(F_{j\ell}/F_{j\ell-1})$.

Theorem 4.1. *The cohomology ring $H^*(X_{\mathrm{Dol}})$ is generated by the classes $e_m(\xi)$ and $c_{m,j,\ell}$, where ξ ranges over a basis of the cohomology ring $H^*(C)$.*

The cohomology ring of the Riemann surface C has a basis $1 \in H^0(C)$, $\mathrm{pt} \in H^2(C)$, and $\sigma_1, \ldots, \sigma_{2g} \in H^1(C)$ with the product structure given by $\sigma_i\sigma_{i+g} = -\sigma_{i+g}\sigma_i = \mathrm{pt}$, and all other $\sigma_i\sigma_j$ being zero.

4.3. Explicit relations of $H^*(X_{\mathrm{Dol}})$

Explicit relations of $H^*(X_{\mathrm{Dol}})$ are known for $n = 2$, $k = 1$, $J_1 = -I$, and arbitrary g [25]. The cohomology is generated by certain elements $\varepsilon_1, \ldots, \varepsilon_{2g} \in H^1(X_{\mathrm{Dol}})$, $\alpha \in H^2(X_{\mathrm{Dol}})$, $\psi_1, \ldots, \psi_{2g} \in H^3(X_{\mathrm{Dol}})$, and $\beta \in H^4(X_{\mathrm{Dol}})$. These are certain multiples of $e_1(\sigma_i)$, $e_2(1)$, $e_2(\sigma_i)$, and $e_2(\mathrm{pt})$. The elements ε_i generate a copy of the cohomology ring of the Jacobian of C, and we have

$$H^*(X_{\mathrm{Dol}}) = H^*(\mathrm{Jac}(C)) \otimes \bigoplus_{i=0}^{g} \wedge_0^i(\psi) \otimes \mathbb{C}[\alpha, \beta, \gamma]/J_i^{g-i},$$

where $J_i^{g-i} \subset \mathbb{C}[\alpha, \beta, \gamma]$ is the ideal generated by an explicit list of polynomials in α, β, γ. Here, $\gamma = -2\sum_{i=1}^{g} \psi_i\psi_{i+g}$ and $\wedge_0^i(\psi)$ denotes the primitive part in the space of i-fold products of ψ_j-classes.

[3] In fact, the tautological bundle is often not well defined as a vector bundle. It is well defined as a projective bundle, and the "determinant" $\det \mathcal{E}$ is also well defined, but the determinant might have no n-th root and correspondingly $c_1(\det \mathcal{E}) \in H^2(X_{\mathrm{Dol}}, \mathbb{Z})$ may not be divisible by n. Since we are dealing with the cohomology with complex coefficients, this does not create a problem, and all the Chern classes are still well defined.

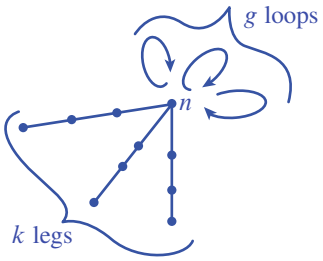


Figure 4. Comet-shaped quiver with k legs and g loops.

4.4. Weight filtration

Since X_B is an algebraic variety, its cohomology is endowed with the weight filtration constructed by Deligne in [10, 11]. The weight filtration on $H^*(X_B)$ is very easily described in terms of the tautological generators. The following was shown in [38] for the case without parabolic points and in [32] in a more general case.

Theorem 4.2. *Define the weight of the tautological generators $e_m(\xi)$ and $c_{m,j,\ell}$ to be $2m$. Then $W_m H^*(X_B)$ is spanned by products of tautological generators of total weight $\leq m$.*

4.5. The pure part

The *pure part* $H^*_{\text{pure}}(X_B) = \bigoplus_i W_i H^i(X_B)$ is naturally a subring of $H^*(X_B)$. In [22], it was conjectured that the Poincaré polynomial of $H^*_{\text{pure}}(X_B)$ matches that of the quiver variety associated to the corresponding comet-shaped quiver with g loops and k legs whose lengths depend on the conjugacy classes J_i; see Figure 4.

4.6. $P = W$

Computing the characteristic polynomial of $\theta : \mathcal{E} \to \mathcal{E} \otimes \omega_C(D)$ at all points of C produces a well-defined polynomial map $X_{\text{Dol}} \to \mathbb{C}^{\dim_{\mathbb{C}} X/2}$ called the *Hitchin map*, which is known to be proper. This induces a filtration $P_* H^*(X_{\text{Dol}})$ on $H^*(X_{\text{Dol}})$ called the *perverse Leray filtration* [9]. According to the recently proved $P = W$ conjecture [8], we have the following theorem.

Theorem 4.3 ([23, 30]). *The following equality holds:*

$$P_i = W_{2i}.$$

4.7. $q - t$ symmetry

Recall the mixed Hodge polynomial $q + t + 4$ from Section 2.4. Notice that it is symmetric in the variables q and t. The mixed Hodge polynomials were conjectured

to be symmetric in q, t in general in [22]. This was called the *curious hard Lefschetz* property. Assuming $P = W$, the curious hard Lefschetz property is a consequence of the *relative hard Lefschetz theorem* on the X_{Dol}-side. We have the following theorem.

Theorem 4.4 ([32]). *The cohomology $H^*(X_B)$ satisfies curious hard Lefschetz:*

$$\omega^i : \mathrm{Gr}^W_{\dim_{\mathbb{C}} X-2i} H^j(X_B) \to \mathrm{Gr}^W_{\dim_{\mathbb{C}} X+2i} H^{j+2i}(X_B)$$

is an isomorphism, where ω is the class of a certain explicit holomorphic 2-form.

In fact, the proof of $P = W$ from [30] relies on this result. On the other hand, the proof of $P = W$ in [23] is independent of it and therefore provides another proof of the curious hard Lefschetz.

Similar curious hard Lefschetz property holds for a certain version of the Khovanov homology of knots (conjectured in [12], proved in [16]).

4.8. Partition function

Perhaps the biggest problem in the field that still remains open is the conjectural formula for the generating function of the mixed Hodge polynomials of all character varieties given in [22] in terms of a partition function. Here, k (the number of marked points) and g are fixed. For each $i = 1, \dots, k$, let $\mathbf{x}_i := (x_{i1}, x_{i2}, \dots)$ be an infinite sequence of variables. Suppose we have a sequence $\boldsymbol{\mu}_i := (\mu_{i1}, \mu_{i2}, \dots)$ of non-negative integers with $\sum_j \mu_{ij} = n$. To this we associate a conjugacy class J_{μ_i} of the diagonal matrix with eigenvalues having multiplicities $\mu_{i1}, \mu_{i2}, \dots$. For a tuple of such sequences $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$, we choose the eigenvalues arbitrarily so that the conjugacy classes satisfy the genericity condition (Definition 3.2) and consider the corresponding character variety $X_{g;J_{\mu_1},\dots,J_{\mu_k}}$. We take its mixed Hodge polynomial (Section 2.4) with weight

$$\mathbf{x}^{\boldsymbol{\mu}} := \prod_{i=1}^{k} \prod_j x_{ij}^{\mu_{ij}}.$$

The result is a symmetric function in each sequence of variables $\mathbf{x}_i$ whose coefficients are polynomials in q and t:

$$\mathrm{MH}_{g,k}(\mathbf{x}_1, \dots, \mathbf{x}_k; q, t) := \sum_{n=1}^{\infty} \sum_{\mu_1,\dots,\mu_k \vdash n} \mathrm{MH}(X_{g;J_{\mu_1},\dots,J_{\mu_k}}; q, t)\, \mathbf{x}^{\boldsymbol{\mu}}.$$

On the other hand, we define the *partition function*

$$\Omega_{g,k}(\mathbf{x}_1,\dots,\mathbf{x}_k;q,t) := \sum_{n=0}^{\infty}\sum_{\lambda\vdash n}\widetilde{H}_\lambda(\mathbf{x}_1;q,t)\cdots\widetilde{H}_\lambda(\mathbf{x}_k;q,t) \cdot\prod_{\square\in\lambda}\frac{\left(q^{a(\square)+1/2}-t^{\ell(\square)+1/2}\right)^{2g}}{(q^{a(\square)}-t^{\ell(\square)+1})(q^{a(\square)+1}-t^{\ell(\square)})}, \tag{4.2}$$

where λ goes over all partitions (Young diagrams), $\widetilde{H}_\lambda(\mathbf{x}_k;q,t)$ denotes the *modified Macdonald polynomial*, the product goes over the cells of the Young diagram, and a,ℓ denote the usual arms and legs statistics (see [15]).

Conjecture 4.5 ([22]). *The partition function is related to the mixed Hodge polynomials as follows:*

$$\Omega_{g,k} = \mathrm{Exp}\left(-\frac{\mathrm{MH}_{g,k}}{(q-1)(t-1)}\right). \tag{4.3}$$

Here, Exp denotes the *plethystic exponential*. It can be defined for any expression involving some variables $a,b,c,\dots$ as follows:

$$\mathrm{Exp}(f(a,b,c,\dots)) := \exp\left(\sum_{n=1}^{\infty}\frac{f(a^n,b^n,c^n,\dots)}{n}\right).$$

If $a,b,c,\dots$ are monomials in some variables, then we have

$$\mathrm{Exp}(a+b+c+\cdots) = \frac{1}{1-a}\,\frac{1}{1-b}\,\frac{1}{1-c}\cdots,$$

so plethystic exponential can be thought of as the conversion between infinite sums and infinite products. There is an inverse operation Log called the *plethystic logarithm*. So if we focus on computing the mixed Hodge polynomials from (4.3), we can write

$$\mathrm{MH}_{g,k} = -(1-q)(1-t)\,\mathrm{Log}\,\Omega_{g,k}. \tag{4.4}$$

4.9. What is known?

Conjecture 4.5 was motivated by a computation performed in [22]. It was shown that its implication for the number of $\mathbb{F}_q$ points is correct (see (2.3)), by counting representations $\pi_1(C)\to \mathrm{GL}_n(\mathbb{F}_q)$ using the character theory of the finite group $\mathrm{GL}_n(\mathbb{F}_q)$.

Theorem 4.6 ([22]). *The number of points over $\mathbb{F}_q$ ($t=q^{-1}$ specialization) deduced from the conjectural equality (4.3) is correct.*

Initially, in identity (4.4), it was not at all clear that the right-hand side has coefficients which are polynomials in q, t, as they should be because the mixed Hodge polynomials on the left-hand side are polynomials by definition. Using a new identity in the combinatorics of Macdonald polynomials established in [14], the author first proved the following theorem.

Theorem 4.7 ([31]). *The coefficients of* $(q-1)(t-1)\operatorname{Log}\Omega$ *are polynomials in* q *and* t.

Then, together with an approach to count Higgs bundles over $\mathbb{F}_q$ by Schiffmann [36, 37] and a new relationship between the Macdonald polynomials and the affine Springer fibers, the author proved the following theorem.

Theorem 4.8 ([33, 34]). *The Poincaré polynomials* ($t = 1$ *specialization) deduced from the conjectural equality* (4.3) *are correct.*

With the help of this theorem, there is a following line of attack on Conjecture 4.5. Suppose we managed to construct a "candidate model" M for the cohomology ring of the character variety so that it comes with a surjective map $M \to H^*(X)$. Showing surjectivity of the map is plausible because the generators of $H^*(X)$ are known. If we furthermore manage to show that the dimension of M is exactly as predicted by Conjecture 4.5, then by Theorem 4.8 the map $M \to H^*(X)$ must be an isomorphism. Partial progress in this direction is discussed in the next sections.

5. COHAs/Coulomb branches

In order to explain our approach, we need to introduce some ideas.

5.1. Cohomological Hall algebra (COHA)

Denote by $\mathcal{H}_2$ the Lie algebra of Hamiltonian vector fields on $\mathbb{C}^2$. Here, $\mathbb{C}^2$ is viewed as a symplectic variety with symplectic form $dx \wedge dy$, where x and y are the natural coordinates. Symplectic geometry (for a quick introduction, see, e.g., [1, Chapter 8]) associates to any function $f(x, y)$ the *Hamiltonian vector field*

$$H_f := \frac{\partial f}{\partial y}\frac{\partial}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial}{\partial y}.$$

The Lie bracket of vector fields is compatible with the Poisson bracket of functions:

$$[H_f, H_g] = H_{\{f,g\}}, \quad \{f, g\} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial g}{\partial y}.$$

The vector space of functions becomes a Lie algebra with respect to the Poisson bracket, and it is this Lie algebra we consider.

Explicitly, it is the Lie algebra with basis elements $T_{r,s}$ for each pair $r, s \geq 0$, corresponding to the functions $x^r y^s$. The Lie bracket is given as follows:

$$[T_{r,s}, T_{r',s'}] = (r's - rs')T_{r+r'-1,s+s'-1} \quad (r, s \geq 0).$$

The tensor product $\mathcal{H}_2 \otimes H^*(C)$ is naturally a super-Lie algebra. Let $H^*(X)[x, y]$ be the cohomology ring with two formal variables x, y adjoined. Let n and d be the rank and the degree of the Higgs bundles parameterized by the moduli space X, and let $\chi := d + (1 - g)n$ be the Euler characteristic.

Theorem 5.1 ([23,35]). *The Lie algebra $\mathcal{H}_2 \otimes H^*(C)$ acts on $H^*(X)[x, y]$ in such a way that $T_{0,m} \otimes \xi$ satisfies*

$$\begin{aligned}((T_{0,m} \otimes \xi) f)(0, 0) &= \psi_m(\xi)(f) \\ &:= (p_m(\xi) + m(1 - g)p_{m-1}(\xi \cup \mathrm{pt})) \cup f \quad (f \in H^*(X)),\end{aligned}$$

and $T_{1,0} \otimes \mathrm{pt}$, $T_{1,0} \otimes 1$, $T_{0,1} \otimes \mathrm{pt}$, $T_{0,1} \otimes 1$ act as x, $-n\frac{\partial}{\partial y}$, y, $n\frac{\partial}{\partial x} + \chi$, respectively.

Here, $p_m(\xi)$ is defined similarly to $e_m(\xi)$ in (4.1), except that we use the power sums of the Chern roots, i.e., $m!$ times the m-th component of the Chern character. Thus, for instance, the central operator $T_{0,0} \otimes \mathrm{pt}$ acts via the multiplication by n, the rank of the Higgs bundles. The operator $T_{0,1} \otimes 1$ commutes with y and acts on elements of $H^*(X)[y]$ via the multiplication by $d + (1 - g)n = \chi$.

Note that quadratic Hamiltonians correspond to the linear action of $\mathrm{SL}(2, \mathbb{C})$ on $\mathbb{C}^2$, and so the elements $T_{2,0}$, $T_{1,1}$, and $T_{0,2}$ span a copy of $\mathfrak{sl}_2 \subset \mathcal{H}_2$. In fact, it is this $\mathfrak{sl}_2$ that was matched with the relative Lefschetz $\mathfrak{sl}_2$ in [23] in order to pin down the perverse filtration algebraically.

5.2. Coulomb branch

The construction described in [3] associates to a pair of a reductive group G and its representation N an affine Poisson algebraic variety called the Coulomb branch. We are interested in the situation where $G = \mathrm{GL}_n \times \mathbb{C}^*$ acts on the space of matrices $N = \mathbb{C}^{n \times n}$: GL_n acts by conjugation and $\mathbb{C}^*$ by the scalar multiplication.

Then Higgs bundles can be seen as maps from the curve C to the stack N/G. This is because, more generally, more or less by definition, maps from some space X to some quotient stack Y/G are pairs of a principal G-bundle X' on X and a G-equivariant map $X' \to Y$. A principal G-bundle on X is a pair of a line bundle L and a rank n bundle E, and an equivariant map $X' \to N$ corresponds to a linear map

$E \otimes L \to E$. So in the case when L is the tangent bundle of C, we can view Higgs bundles as maps $C \to N/G$.

Our hope is that the Coulomb branch has something to do with the cohomology of our moduli spaces. This is the situation of the *adjoint representation with flavor*, and the Coulomb branch is related to the Hilbert scheme [4] of $\mathbb{C}^* \times \mathbb{C}$, which is a symplectic resolution of singularities of $(\mathbb{C}^* \times \mathbb{C})^n/S_n$. It is not entirely clear why one should pass from the Hilbert scheme of $\mathbb{C}^* \times \mathbb{C}$ to that of $\mathbb{C}^2$, but our construction in [23] in fact first produces an algebra of Hamiltonian fields on $\mathbb{C}^* \times \mathbb{C}$, and only then it degenerates to the corresponding algebra for $\mathbb{C}^2$. Hopefully, this passage will be clarified by experts in Coulomb branches.

Now suppose we are on $\mathrm{Hilb}^n(\mathbb{C}^2)$. Then $\mathcal{H}_2$ acts on many natural objects such as functions, differential forms, etc. So even if we do not completely understand the Coulomb branch picture, it is a natural guess to look for our model of the cohomology ring among natural objects living on the Hilbert scheme.

5.3. Geometric engineering

An indication similar to the one above is given in [7], where a physical explanation of Conjecture 4.5 is provided via interpreting the left-hand side (the partition function) as

$$\Omega = \chi(\mathrm{Hilb}_n(\mathbb{C}^2), P^{\otimes k} \otimes (\Omega^*)^{\otimes g} \otimes O(g-1)). \tag{5.1}$$

Here, P is the Procesi bundle (S_n-equivariant bundle coming from Haiman's identification $\mathrm{Hilb}_n(\mathbb{C}^2) = S_n - \mathrm{Hilb}(\mathbb{C}^{2n})$), Ω^* is the bundle of differential forms, and χ denotes the equivariant Euler characteristic.

5.4. Proposal

One problem with (5.1) is that the right-hand side does not always have positive coefficients. So the vector bundle on the right-hand side of (5.1) has higher cohomologies. Thus, there seems to be no way to upgrade identity (4.3) to an isomorphism between vector spaces using (5.1). One can trace the problem to the fact that the factor $-\frac{1}{(1-q)(1-t)}$ is manifestly not positive, when expanded in q and t. To resolve this issue, we propose to expand in q^{-1} and t instead. The geometric meaning of this expansion is different from taking the global sections. Instead, we should take the local cohomology with respect to a Lagrangian subvariety.

Theorem 5.2 (A. Mellit, M. Romero). *Let $L \subset \mathrm{Hilb}_n(\mathbb{C}^2)$ be the Lagrangian cut out by the equations $\sum_{j=1}^n x_j^r = 0$ $(r > 0)$. Let $\mathcal{E}$ be any vector bundle on $\mathrm{Hilb}_n(\mathbb{C}^2)$. Then for any $i \neq n$, we have*

$$H^i_L(\mathrm{Hilb}_n(\mathbb{C}^2), \mathcal{E}) = 0.$$

Here, H^i_L is the functor of local cohomology (see [20]), the derived functor of taking sections supported at L. Then we expect that (4.3) has the following interpretation.

Conjecture 5.3. *The local cohomology is related to the Borel–Moore homology of a certain Higgs moduli stack as follows:*

$$H^n_L(P^{\otimes k} \otimes (\Omega^*)^{\otimes g} \otimes O(g-1))^\vee \cong H^{\mathrm{BM}}_*(\textit{Higgs moduli stack}).$$

Here, it is not clear which precise Higgs moduli stack should be taken on the right-hand side. The whole Higgs moduli stack has infinitely many irreducible components, and the expectation is that we need to take a nice open substack containing only finitely many components in each degree, possibly imposing a bound on the slope. For a bigraded vector space M, we denote by $M^\vee$ the bigraded dual vector space.

6. Main example

A special case can be worked out in full detail. We take $g = 0$, i.e., $C = \mathbb{P}^1$, and consider the $k+1$-twisted Higgs bundles, so that the Higgs field is a morphism $\mathcal{E} \to \mathcal{E}(k-1)$. Equivalently, we may consider $k+1$ marked points without any parabolic structures and conditions on the residue of the Higgs field. The corresponding specialization of the partition function is given by

$$\Omega_k(z;q,t) := \sum_{n=0}^{\infty} \sum_{\lambda \vdash n} \frac{(-1)^{(k+1)|\lambda|}\left(q^{n(\lambda')}t^{n(\lambda)}\right)^{k+1}}{\prod_{\square\in\lambda}(q^{a(\square)}-t^{\ell(\square)+1})(q^{a(\square)+1}-t^{\ell(\square)})} z^n. \tag{6.1}$$

We use the variable z to keep track of the rank. In (4.2), we did not need it because the rank of each term can be extracted as the degree of the symmetric function in each set of variables. The first terms of Ω look as follows:

$$1 + \frac{(-1)^k}{(1-t)(1-q)} z + \frac{q^{k+1}(1-t^2) - t^{k+1}(1-q^2)}{(q-t)(1-q)(1-q^2)(1-t)(1-t^2)} z^2 + \cdots,$$

and the first terms of its plethystic logarithm read

$$\operatorname{Log}\Omega_k(z;q,t) = -\frac{1}{(1-q)(1-t)}\left((-1)^{k+1}z + z^2 \sum_{\substack{i\geq 0,\, j\geq 0,\, i+j\leq k-2,\\ i+j\equiv k \pmod 2}} q^i t^j + \cdots\right). \tag{6.2}$$

In (6.1), one can recognize the localization formula for the bundle $O(k)$ on the Hilbert scheme (see [18]). We have

$$\Omega_k(z;q,t) = \sum_{n=0}^{\infty} (-1)^{kn} z^n \chi(\mathrm{Hilb}_n(\mathbb{C}^2), O(k)).$$

Let $R_x = \mathbb{C}[x_1, \dots, x_n]$, $R_{x,y} = \mathbb{C}[x_1, \dots, x_n, y_1, \dots, y_n]$. The symmetric group S_n acts on these, permuting x_i and y_i simultaneously. For an S_n-module M, we denote by M^+ and M^- the subspace of invariant and anti-invariant elements, respectively. Notice that $R^-_{x,y} \subset R_{x,y}$ is a module over $R^+_{x,y}$. By $(R^-_{x,y})^k$ we denote the linear span of k-fold products of elements of $R^-_{x,y}$ inside $R_{x,y}$. This is also a module over $R^+_{x,y}$. Haiman obtained the following results about the bundles $O(k)$.

Theorem 6.1 ([18]). *For any $k \geq 0$, we have an isomorphism of $R^+_{x,y}$-modules*

$$\Gamma(\mathrm{Hilb}_n(\mathbb{C}^2), O(k)) \cong (R^-_{x,y})^k \subset R_{x,y},$$

and the i-th cohomology of $O(k)$ vanishes for $i > 0$. The R^+_x-module $(R^-_{x,y})^k$ is free.

For instance, for $k = 0$ and $k = 1$, we have

$$\Gamma(\mathrm{Hilb}_n(\mathbb{C}^2), O) = R^+_{x,y}, \quad \Gamma(\mathrm{Hilb}_n(\mathbb{C}^2), O(1)) = R^-_{x,y}.$$

From this we can obtain the following presentation for the dual of the local cohomology.

Proposition 6.2. *Let $M_{n,k} := H^n_L(O(k))^\vee$. Let $\Delta_x := \prod_{i<j\leq n}(x_i - x_j)$. For $k \geq 0$, we have*

$$M_{n,k} = \left\{ f \in \frac{1}{\Delta_x^k} R^-_{x,y} \mid D(x_1, \dots, x_n, \partial_{y_1}, \dots, \partial_{y_n}) f \in R_{x,y} \ (\forall D \in (R^-_{x,y})^k) \right\}.$$

For $k \geq 1$, we have

$$M_{n,k} = \left\{ f \in \frac{1}{\Delta_x^{k-1}} R^+_{x,y} \mid D(x_1, \dots, x_n, \partial_{y_1}, \dots, \partial_{y_n}) f \in R_{x,y} \ (\forall D \in (R^-_{x,y})^{k-1}) \right\}.$$

For instance, we have $M_{n,0} = R^-_{x,y}$ and $M_{n,1} = R^+_{x,y}$. Equivalence of the two presentations follows from the fact that if $f \in \frac{1}{\Delta_x} R^-_{x,y}$ satisfies $Df \in R_{x,y}$ for all $D \in R^-_{x,y}$, then we have $f \in R^+_{x,y}$.

In the case $n = 2$, the module $R^-_{x,y}$ is generated over $R^+_{x,y}$ by $x_1 - x_2$ and $y_1 - y_2$. So the condition on f is equivalent to

$$(x_1 - x_2)^i (\partial_{y_1} - \partial_{y_2})^j f \in R_{x,y} \quad (i + j = k - 1).$$

Equivalently, in the expansion

$$f = \sum_{i,j} (x_1 - x_2)^i (y_1 - y_2)^j f_{ij}(x_1 + x_2, y_1 + y_2),$$

we have $f_{ij} = 0$ unless $i \geq 0$ or $i \geq -k + 1 + j$. Note that also $f_{ij} = 0$ unless $i + j \equiv k - 1 \pmod 2$. So $M_{2,k}$ has a direct sum decomposition $M_{2,k} = R_{x,y}^{(-1)^{k-1}} \oplus M'_{2,k}$, where $M_{2,k'}$ has the Hilbert series

$$\frac{1}{(1-q)(1-t)} \sum_{\substack{i \geq 1,\, j \geq 0,\, i+j \leq k-1 \\ i+j \equiv k-1 \pmod 2}} q^{-i} t^j,$$

which exactly matches the z^2-term of (6.2) after the substitution $q \to q^{-1}$.

More generally, we notice that there is a *shuffle product* $M_{n',k} \otimes M_{n'',k} \to M_{n'+n'',k}$ given by

$$f * g = \sum_{\sigma \in S_{n'+n''}/(S'_n \times S_{n''})} (-1)^{(k-1)|\sigma|} \sigma (f(x_1, \ldots, x_{n'}, y_1, \ldots, y_{n'}) \cdot g(x_{n'+1}, \ldots, x_{n'+n''}, y_{n'+1}, \ldots, y_{n'+n''})).$$

Supported by computer experiments, we formulate the following conjecture.

Conjecture 6.3. *Denote by $M_{n,k}^{\text{shuffles}} \subset M_{n,k}$ the span of all shuffle products of the form $f * g$ for $f \in M_{n',k}$ and $g \in M_{n-n',k}$, $0 < n' < n$. Let the space of primitive elements be the quotient $M_{n,k}^{\text{prim}} := M_{n,k} / M_{n,k}^{\text{shuffles}}$. We have an isomorphism of bigraded vector spaces*

$$H^*(X)\left[\sum_{i=1}^n x_i, \sum_{i=1}^n y_i\right] \cong M_{n,k}^{\text{prim}}.$$

The ring structure of the cohomology is not immediately evident from this description, but our COHA tells us where to look for it.

Proposition 6.4. *The Lie algebra $\mathcal{H}_2 \otimes H^*(\mathbb{P}^1)$ acts on $M_{n,k}$ by the following operations:*

$$(T_{r,s} \otimes \text{pt}) f = \sum_{i=1}^n x_i^s \partial_{y_i}^r f,$$

$$(T_{r,s} \otimes 1) f = \sum_{i=1}^n x_i^{s-1} \partial_{y_i}^{r-1} (-s \partial_{y_i} y_i - r x_i \partial_{x_i}) f.$$

This action is compatible with the shuffle product and therefore descends to $M_{n,k}^{\text{prim}}$.

Now we describe the ring structure. First we expect that the following is true.

Conjecture 6.5. *The space* $M_{n,k}^{\mathrm{prim}}$ *is spanned by products of* $\frac{1}{\Delta_x^{k-1}}$ *and polynomials of the form*

$$\psi_s(\mathrm{pt}) := \sum_{i=1}^{n} x_i^s, \quad \psi_s(1) := s \sum_{i=1}^{n} x_i^{s-1} y_i.$$

This conjecture implies that $M_{n,k}^{\mathrm{prim}}$ is isomorphic to the quotient of the ring of polynomials in the generators $\psi_s(\mathrm{pt})$, $\psi_s(1)$ by a certain ideal. So it endows $M_{n,k}^{\mathrm{prim}}$ with a ring structure, and it is natural to formulate the following conjecture.

Conjecture 6.6. *The isomorphism of Conjecture 6.3 is an isomorphism of rings.*

6.1. Some remarks

The Hilbert scheme has long been a source of interesting bigraded modules. For instance, there was $(n+1)^{n-1}$ conjecture and $n!$ conjecture proved by Haiman [18]. The $(n+1)^{n-1}$ conjecture is about the module of diagonal coinvariants.

Theorem 6.7 ([18]). *The following equality holds:*

$$\dim R_{x,y}/(R_{x,y}^+)_{>0} = (n+1)^{n-1}.$$

An explicit combinatorial presentation of the bigraded Frobenius character of $R_{x,y}/(R_{x,y}^+)_{>0}$ was known as a shuffle conjecture [17] and was proved in [5]. Also in this context Haiman observed that the vector space is in a certain sense generated by the differential operators $\sum_{i=1}^n x_i^{s-1} \frac{\partial}{\partial y_i}$, which led to the operator conjectures in [19]. The double coinvariants have been interpreted in terms of Borel–Moore cohomology of affine Springer fibers in [6].

Our story suggests that there are many more interesting bigraded modules to be constructed using the Hilbert scheme.

7. Poisson geometry of Coulomb branches

We would like to conclude by explaining a curious formalism which hints at how Conjectures 6.3–6.6 can be extended to surfaces of higher genus. First, observe that, in view of (5.1), we should consider the bundle of multiforms, i.e., the g-fold tensor product of the bundle of differential forms on the Hilbert scheme. How does the Lie algebra $\mathcal{H}_2 \otimes H^*(C)$ act? Quite naturally, in fact. Take a function $f \in R_{x,y}^+$,

for instance $f = \sum_{i=1}^{n} x_i^s y_i^r$. Symplectic geometry associates to f the Hamiltonian vector field H_f. Let

$$\psi_f(\mathrm{pt}) := \text{multiplication by } f,$$
$$\psi_f(1) := \text{Lie derivative along } H_f.$$

This is what we had in genus 0. But with multiforms, additional operations arise. For $i = 1, \dots, g$, set

$$\psi_f(\sigma_i) = (df)^{(i)}.$$

This is the operation of multiplication by df acting on the i-th component in the g-fold tensor product. Set

$$\psi_f(\sigma_{i+g}) = \iota_{H_f}^{(i)},$$

the operation of insertion of H_f into the i-th component. From

$$\psi_{fg}(1) = \mathrm{Lie}_{fH_g + gH_f}, \quad \mathrm{Lie}_{fH_g} = f\,\mathrm{Lie}_{H_g} + \sum_{i=1}^{g} (df)^{(i)} \iota_{H_g}^{(i)},$$

we obtain the following identity:

$$\psi_{fg}(1) = \psi_f(1)\psi_g(\mathrm{pt}) + \psi_f(\mathrm{pt})\psi_g(1) + \sum_{i=1}^{2g} \psi_f(\sigma_i)\psi_g(\sigma_i^*).$$

This formula precisely matches the Künneth decomposition of the diagonal class in $H^2(C \times C)$, which confirms the consistency of the formalism. Okounkov has observed that this formalism essentially coincides with the Clifford algebras considered in [27].

Funding. This research is supported by the Consolidator Grant No. 101001159, "Refined invariants in combinatorics, low-dimensional topology and geometry of moduli spaces" of the European Research Council.

References

[1] V. I. Arnold, *Mathematical methods of classical mechanics*. Grad. Texts in Math. 60, Springer, New York, 1978, 462 pp. Zbl 0386.70001 MR 0690288

[2] T. Bothner, On the origins of Riemann–Hilbert problems in mathematics. *Nonlinearity* **34** (2021), no. 4, R1–R73 Zbl 1462.30070 MR 4246443

[3] A. Braverman, M. Finkelberg, and H. Nakajima, Towards a mathematical definition of Coulomb branches of 3-dimensional $\mathcal{N} = 4$ gauge theories, II. *Adv. Theor. Math. Phys.* **22** (2018), no. 5, 1071–1147 Zbl 1479.81043 MR 3952347

[4] A. Braverman, M. Finkelberg, and H. Nakajima, Line bundles on Coulomb branches. *Adv. Theor. Math. Phys.* **25** (2021), no. 4, 957–993 Zbl 1495.81069 MR 4407969

[5] E. Carlsson and A. Mellit, A proof of the shuffle conjecture. *J. Amer. Math. Soc.* **31** (2018), no. 3, 661–697 Zbl 1387.05265 MR 3787405

[6] E. Carlsson and A. Oblomkov, Affine Schubert calculus and double coinvariants. [v1] 2018, [v6] 2026, arXiv:1801.09033v6

[7] W.-y. Chuang, D.-E. Diaconescu, R. Donagi, and T. Pantev, Parabolic refined invariants and Macdonald polynomials. *Comm. Math. Phys.* **335** (2015), no. 3, 1323–1379 Zbl 1367.14019 MR 3320315

[8] M. A. A. de Cataldo, T. Hausel, and L. Migliorini, Topology of Hitchin systems and Hodge theory of character varieties: the case A_1. *Ann. of Math. (2)* **175** (2012), no. 3, 1329–1407 Zbl 1375.14047 MR 2912707

[9] M. A. A. de Cataldo and L. Migliorini, The decomposition theorem, perverse sheaves and the topology of algebraic maps. *Bull. Amer. Math. Soc. (N.S.)* **46** (2009), no. 4, 535–633 Zbl 1181.14001 MR 2525735

[10] P. Deligne, Théorie de Hodge, II. *Publ. Math. Inst. Hautes Études Sci.* **40** (1971), no. 1, 5–57 Zbl 0219.14007 MR 0498551

[11] P. Deligne, Théorie de Hodge, III. *Publ. Math. Inst. Hautes Études Sci.* **44** (1974), no. 1, 5–77 Zbl 0237.14003 MR 0498552

[12] N. M. Dunfield, S. Gukov, and J. Rasmussen, The superpolynomial for knot homologies. *Exp. Math.* **15** (2006), no. 2, 129–159 Zbl 1118.57012 MR 2253002

[13] R. Fricke and F. Klein, *Vorlesungen über die Theorie der automorphen Funktionen. Band 1: Die gruppentheoretischen Grundlagen. Band II: Die funktionentheoretischen Ausführungen und die Andwendungen*. Bibliotheca Mathematica Teubneriana, Bande 3, 4, Johnson Reprint Corp., New York; B. G. Teubner Verlagsgesellschaft, Stuttgart, 1965, Band I: 634 pp.; Band II: 668 pp. MR 0183872

[14] A. Garsia and A. Mellit, Five-term relation and Macdonald polynomials. *J. Combin. Theory Ser. A* **163** (2019), 182–194 Zbl 1403.05158 MR 3885219

[15] A. M. Garsia and M. Haiman, A remarkable q, t-Catalan sequence and q-Lagrange inversion. *J. Algebra. Comb.* **5** (1996), no. 3, 191–244 Zbl 0853.05008 MR 1394305

[16] E. Gorsky, M. Hogancamp, and A. Mellit, Tautological classes and symmetry in Khovanov–Rozansky homology. *Duke Math. J.* **173** (2024), no. 13, 2481–2561 Zbl 1550.18014 MR 4800617

[17] J. Haglund, M. Haiman, N. Loehr, J. B. Remmel, and A. Ulyanov, A combinatorial formula for the character of the diagonal coinvariants. *Duke Math. J.* **126** (2005), no. 2, 195–232 Zbl 1069.05077 MR 2115257

[18] M. Haiman, Vanishing theorems and character formulas for the Hilbert scheme of points in the plane. *Invent. Math.* **149** (2002), no. 2, 371–407 Zbl 1053.14005 MR 1918676

[19] M. D. Haiman, Conjectures on the quotient ring by diagonal invariants. *J. Algebra. Comb.* **3** (1994), no. 1, 17–76 Zbl 0803.13010 MR 1256101

[20] R. Hartshorne, *Local cohomology. A seminar given by A. Groethendieck, Harvard University. Fall, 1961*. Lecture Notes in Math. 41, Springer, Berlin, 1967, 106 pp. Zbl 0185.49202 MR 0224620

[21] R. Hartshorne, *Algebraic geometry*. Grad. Texts in Math. 52, Springer, New York, 1977, 496 pp. Zbl 0367.14001 MR 0463157

[22] T. Hausel, E. Letellier, and F. Rodriguez-Villegas, Arithmetic harmonic analysis on character and quiver varieties. *Duke Math. J.* **160** (2011), no. 2, 323–400 Zbl 1246.14063 MR 2852119

[23] T. Hausel, A. Mellit, A. Minets, and O. Schiffmann, $P = W$ via $\mathcal{H}_2$. [v1] 2022, [v2] 2025, arXiv:2209.05429v2

[24] T. Hausel and F. Rodriguez-Villegas, Mixed Hodge polynomials of character varieties (with an appendix by N. M. Katz). *Invent. Math.* **174** (2008), no. 3, 555–624 Zbl 1213.14020 MR 2453601

[25] T. Hausel and M. Thaddeus, Relations in the cohomology ring of the moduli space of rank 2 Higgs bundles. *J. Amer. Math. Soc.* **16** (2003), no. 2, 303–329 Zbl 1015.14018 MR 1949162

[26] N. J. Hitchin, The self-duality equations on a Riemann surface. *Proc. Lond. Math. Soc. (3)* **55** (1987), no. 1, 59–126 Zbl 0634.53045 MR 0887284

[27] D. Kazhdan and A. Okounkov, L-function genera and applications. *Pure Appl. Math. Q.* **20** (2024), no. 5, 2231–2339 Zbl 07946805 MR 4829851

[28] F. Lorenz, *Algebra. Volume II: fields with structure, algebras and advanced topics*. Universitext, Springer, New York, 2008, 336 pp. Zbl 1130.12001 MR 2371763

[29] E. Markman, Generators of the cohomology ring of moduli spaces of sheaves on symplectic surfaces. *J. Reine Angew. Math.* **2002** (2002), no. 544, 61–82 Zbl 0988.14019 MR 1887889

[30] D. Maulik and J. Shen, The $P = W$ conjecture for GL_n. *Ann. of Math. (2)* **200** (2024), no. 2, 529–556 Zbl 07995676 MR 4792069

[31] A. Mellit, Integrality of Hausel–Letellier–Villegas kernels. *Duke Math. J.* **167** (2018), no. 17, 3171–3205 Zbl 1411.14040 MR 3874651

[32] A. Mellit, Toric stratifications of character varieties. *Publ. Math. Inst. Hautes Études Sci.* **142** (2025) 153–240 Zbl 08129093 MR 4992851

[33] A. Mellit, Poincaré polynomials of character varieties, Macdonald polynomials and affine Springer fibers. *Ann. of Math. (2)* **192** (2020), no. 1, 165–228 Zbl 1467.14086 MR 4125451

[34] A. Mellit, Poincaré polynomials of moduli spaces of Higgs bundles and character varieties (no punctures). *Invent. Math.* **221** (2020), no. 1, 301–327 Zbl 1455.14022 MR 4105090

[35] A. Mellit, A. Minets, O. Schiffmann, and E. Vasserot, Coherent sheaves on surfaces, COHAs and deformed $W_{1+\infty}$-algebras. [v1] 2023, [v2] 2026, arXiv:2311.13415v2

[36] S. Mozgovoy and O. Schiffmann, Counting Higgs bundles and type A quiver bundles. *Compos. Math.* **156** (2020), no. 4, 744–769 Zbl 1514.14011 MR 4071236

[37] O. Schiffmann, Indecomposable vector bundles and stable Higgs bundles over smooth projective curves. *Ann. of Math. (2)* **183** (2016), no. 1, 297–362 Zbl 1342.14076 MR 3432585

[38] V. Shende, The weights of the tautological classes of character varieties. *Int. Math. Res. Not. IMRN* **2017** (2017), no. 22, 6832–6840 Zbl 1405.14037 MR 3737322

[39] C. T. Simpson, Harmonic bundles on noncompact curves. *J. Amer. Math. Soc.* **3** (1990), no. 3, 713–770 Zbl 0713.58012 MR 1040197

[40] C. T. Simpson, Nonabelian Hodge theory. In *Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990)*, pp. 747–756, Math. Soc. Japan, Tokyo, 1991 Zbl 0765.14005 MR 1159261

[41] C. T. Simpson, Higgs bundles and local systems. *Publ. Math. Inst. Hautes Études Sci.* **75** (1992), no. 1, 5–95 Zbl 0814.32003 MR 1179076

[42] C. T. Simpson, An explicit view of the Hitchin fibration on the Betti side for $\mathbb{P}^1$ minus five points. In J. E. Andersen, A. Dancer, and O. García-Prada (eds.), *Geometry and physics. Volume II*, pp. 705–724, Oxford University Press, Oxford, 2018 Zbl 1418.14003 MR 3931793

[43] T. Szamuely, *Galois groups and fundamental groups*. Cambridge Stud. Adv. Math. 117, Cambridge University Press, Cambridge, 2009, 270 pp. Zbl 1189.14002 MR 2548205

[44] G. Teschl, *Ordinary differential equations and dynamical systems*. Grad. Stud. Math. 140, American Mathematical Society, Providence, RI, 2012, 356 pp. Zbl 1263.34002 MR 2961944

Anton Mellit
Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; anton.mellit@univie.ac.at

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/26

Moduli of Galois representations and generalized class groups

James Newton

Abstract. This lecture describes some recent progress on establishing the conjectural relationships between Galois representations and automorphic forms, together with the number-theoretic motivation for these results. This includes equidistribution theorems like the Sato–Tate conjecture concerning the number of points on reductions of elliptic curves modulo primes. The connection to these equidistribution statements goes back to Dirichlet's work on primes in arithmetic progressions, in which the class number formula for real quadratic fields plays an important role (at least from one point of view). We discuss how results on 'generalized class groups' defined using Galois cohomology have been crucial to more recent developments.

1. Dirichlet's theorem

1.1. Primes in arithmetic progressions

A seminal result in number theory is Dirichlet's theorem on primes in arithmetic progressions.

Theorem 1.1 ([9]). *There are infinitely many primes of the form $a + kN$, with a and N fixed coprime integers and k varying over $\mathbf{Z}$.*

This theorem (and its refinement, the Chebotarev density theorem) tells us about the distribution of the images of the primes in the multiplicative group $(\mathbf{Z}/N\mathbf{Z})^\times$. Dirichlet's proof relied on using the character theory of the group $(\mathbf{Z}/N\mathbf{Z})^\times$.

Indeed, suppose we have a homomorphism $\chi : (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times$ (as is standard, we extend by zero to view χ as a function on $\mathbf{Z}/N\mathbf{Z}$ when convenient). Orthogonality of characters tells us that we can express the indicator function of the subset $\{a\} \subset (\mathbf{Z}/N\mathbf{Z})^\times$ as a linear combination of characters:

$$\mathbf{1}_{\{a\}} = \frac{1}{\phi(N)} \sum_{\chi} \overline{\chi}(a)\chi.$$

Mathematics Subject Classification 2020: 11F55.
Keywords: Galois representations, automorphic forms, Langlands program.

Informally, to show that the primes distribute evenly over $(\mathbf{Z}/N\mathbf{Z})^\times$, it suffices to show that the expectation $\mathbf{E}(\chi(p)) = 0$ for every non-trivial character χ, which gives $\mathbf{P}(p \equiv a \bmod N) = \frac{1}{\phi(N)}$.

More precisely, if one shows that

$$\frac{\sum_{p \leq x} \chi(p)}{\#\{p \leq x\}} \to 0 \quad \text{as } x \to \infty$$

for every non-trivial χ, then one deduces the equidistribution statement

$$\frac{\#\{p \leq x : p \equiv a \bmod N\}}{\#\{p \leq x\}} \to \frac{1}{\phi(N)} \quad \text{as } x \to \infty. \tag{1.1}$$

In fact, Dirichlet proved something a little weaker than this, showing that the infinite sum over primes

$$\sum_p \frac{\chi(p)}{p^s}$$

(which converges absolutely for $\mathrm{Re}(s) > 1$) remains bounded as s approaches 1 from above. In contrast, the sum

$$\sum_p \frac{1}{p^s}$$

diverges as s approaches 1.

The behaviour of these sums is closely related to, respectively, the behaviour of the Dirichlet L-function $L(\chi, s)$ and the Riemann zeta function $\zeta(s)$ at $s = 1$. Recall that

$$L(\chi, s) = \sum_{n \geq 1} \frac{\chi(n)}{n^s} = \prod_p (1 - \chi(p)p^{-s})^{-1}$$

(again, these expressions converge absolutely for $\mathrm{Re}(s) > 1$).

A formal manipulation shows that $\sum_p \frac{\chi(p)}{p^s}$ contributes to the principal branch of the logarithm of $L(\chi, s)$:

$$\log L(\chi, s) = \sum_p \sum_{m \geq 1} \frac{\chi(p)^m}{m p^{ms}},$$

and the key input to Dirichlet's theorem is to show that $\log L(\chi, s)$ remains bounded as s approaches 1. This comes down to showing that $L(\chi, 1) \neq 0$ for our non-trivial character χ. Next, we will relate the non-vanishing of this L-value to one of the fundamental objects in algebraic number theory, the ideal class group of a number field.

1.2. Class groups and Dirichlet L-functions

We let $\zeta_N = e^{2i\pi/N}$. We may identify the multiplicative group $(\mathbf{Z}/N\mathbf{Z})^\times$ with the Galois group $\mathrm{Gal}(\mathbf{Q}(\zeta_N)/\mathbf{Q})$ using the cyclotomic character

$$\begin{aligned}\chi_N : \mathrm{Gal}(\mathbf{Q}(\zeta_N)/\mathbf{Q}) &\to (\mathbf{Z}/N\mathbf{Z})^\times, \\ \sigma &\mapsto i, \quad \text{where } \sigma(\zeta_N) = \zeta_N^i.\end{aligned}$$

From this point of view, the characters $\chi : (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times$ are *Artin representations*, i.e., complex representations of a Galois group. A special case of the Artin formalism for L-functions of Artin representations gives the product formula[1]

$$\zeta_{\mathbf{Q}(\zeta_N)}(s) = \prod_\chi L(\chi, s),$$

where the product is over all characters χ of $(\mathbf{Z}/N\mathbf{Z})^\times$ and

$$\zeta_{\mathbf{Q}(\zeta_N)}(s) = \sum_{\mathfrak{n} \lhd \mathbf{Z}[\zeta_N]} \frac{1}{\mathrm{Nm}(\mathfrak{n})^s}$$

is the Dedekind zeta function of the number field $\mathbf{Q}(\zeta_N)$.

The analytic class number formula of Dedekind gives a precise formula for the residue $\lim_{s\to 1+}(s-1)\zeta_{\mathbf{Q}(\zeta_N)}(s)$ in terms of the size of the class group of $\mathbf{Z}[\zeta_N]$ (and other arithmetic data). In particular, this residue is non-zero, so $\zeta_{\mathbf{Q}(\zeta_N)}(s)$ has a simple pole at $s = 1$.[2] We can deduce immediately from this that the $L(\chi, s)$ are non-vanishing at $s = 1$ for non-trivial χ, since the trivial character already contributes a simple pole.

2. Distribution of Frobenius elements

We can re-interpret Dirichlet's theorem, or more precisely, the equidistribution result (1.1), as a statement about the distribution of certain elements in the Galois group $\mathrm{Gal}(\mathbf{Q}(\zeta_N)/\mathbf{Q})$.

[1] Strictly speaking, the product formula is only correct up to some simple factors indexed by primes dividing N. This can be fixed by replacing each χ with its associated *primitive* character: a character of $(\mathbf{Z}/M\mathbf{Z})^\times$ for the smallest divisor M of N such that χ factors through the quotient $(\mathbf{Z}/N\mathbf{Z})^\times \to (\mathbf{Z}/M\mathbf{Z})^\times$.

[2] To make sense of this, we can rely on the fact that our L- and ζ-functions extend to meromorphic functions for $\mathrm{Re}(s) > 0$, although the argument can be rephrased to avoid this.

2.1. Frobenius elements

Let L/F be a Galois extension of number fields and suppose $\mathfrak{p}$ is a prime ideal of $\mathcal{O}_F$ which is unramified in $\mathcal{O}_L$ (i.e., $\mathfrak{p}\mathcal{O}_L$ factors as a product of distinct prime ideals in $\mathcal{O}_L$). Then for each prime $\mathfrak{q} \mid \mathfrak{p}\mathcal{O}_L$, there is a unique element $\sigma_{\mathfrak{q}|\mathfrak{p}} \in \mathrm{Gal}(L/F)$ satisfying

$$\sigma_{\mathfrak{q}|\mathfrak{p}}(x) \equiv x^{\mathrm{Nm}(\mathfrak{p})} \bmod \mathfrak{q} \quad \text{for all } x \in \mathcal{O}_L.$$

The conjugacy class of $\sigma_{\mathfrak{q}|\mathfrak{p}}$ depends only on $\mathfrak{p}$, so we often denote one of these elements by $\sigma_{\mathfrak{p}}$.

In our cyclotomic example $\mathrm{Gal}(\mathbf{Q}(\zeta_N)/\mathbf{Q})$, the unramified primes p are those coprime to N, and the Frobenius element σ_p corresponds under the cyclotomic character to the class of p in $(\mathbf{Z}/N\mathbf{Z})^\times$. So the equidistribution statement (1.1) says that the σ_p distribute uniformly over the $\phi(N)$ elements of $\mathrm{Gal}(\mathbf{Q}(\zeta_N)/\mathbf{Q})$ as p varies.

2.2. Sato–Tate conjecture

We now move on to a more complicated equidistribution result, which is similar in spirit. Suppose we have an elliptic curve E defined over $\mathbf{Q}$, given by the projective planar curve with (affine) equation

$$y^2 = x^3 + Ax + B \tag{2.1}$$

for some coefficients $A, B \in \mathbf{Z}$ satisfying $4A^3 + 27B^2 \neq 0$ (so the curve is nonsingular). Consider odd primes p with $p \nmid 4A^3 + 27B^2$ so that the reduction of equation (2.1) defines an elliptic curve over $\mathbf{F}_p$.[3]

The Sato–Tate conjecture describes the distribution of the integers

$$a_p(E) = p - \#\{(x, y) \in \mathbf{F}_p^2 : y^2 = x^3 + Ax + B\} = p + 1 - \#E(\mathbf{F}_p)$$

or, more precisely, the real numbers

$$\frac{a_p(E)}{2\sqrt{p}}$$

which lie in the interval $[-1, 1]$ by a theorem of Hasse.

See [22] for examples of histograms generated from the values $\frac{a_p(E)}{2\sqrt{p}}$ for many primes p.

We state the following conjecture for the generic case of elliptic curves E with no complex multiplication.

[3] From this point on, we will ignore the finitely many primes p dividing the discriminant $4A^3 + 27B^2$. Omitting these primes does not affect the subsequent analysis.

Conjecture 2.1. *For any subinterval* $[a, b] \subset [-1, 1]$*, the limit*

$$\lim_{X\to\infty} \frac{\#\{p \le X : \frac{a_p(E)}{2\sqrt{p}} \in [a, b]\}}{\#\{p \le X\}}$$

exists and is given by the integral

$$\int_a^b \frac{2}{\pi}\sqrt{1-t^2}dt.$$

This says that a histogram plotting the values of $\frac{a_p(E)}{2\sqrt{p}}$ for primes p up to X will approach a (rescaled) semicircle as X gets large.

This conjecture was proved in a series of works published between 2008 and 2011, using *automorphic forms* [3, 7, 11, 23]. The strategy is modelled on Dirichlet's, as we will now describe.

2.3. Sato–Tate and Frobenius elements

As in the case of Dirichlet's theorem, we can rephrase the Sato–Tate conjecture as a question about the distribution of Frobenius elements.

For each p, the integer $a_p(E)$ can be interpreted as the trace of a 2×2 matrix with determinant p. One way to construct this matrix is by l-adic approximation for a prime $l \neq p$. Firstly, we consider the reduction mod l. The l-torsion points $E[l]$ of E form a two-dimensional $\mathbf{F}_l$-vector space with an action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. One can show that the trace of the Frobenius element σ_p under this representation is equal to $a_p(E) \bmod l$ [21, Thm. V.2.3.1]. Extending this to the l^n-torsion points $E[l^n]$ for each $n \ge 1$ and taking the limit identifies $a_p(E)$ with the trace of σ_p under a Galois representation

$$\rho_{E,l} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathbf{Z}_l).$$

The Hasse bound implies that

$$x^2 - \frac{a_p(E)}{\sqrt{p}}x + 1$$

is the characteristic polynomial of a matrix $u_p(E) \in \mathrm{SU}(2)$, which is well defined up to conjugation.

A reformulation of Conjecture 2.1 is that the conjugacy classes of the $u_p(E)$ are equidistributed with respect to the Haar (i.e., translation-invariant) measure on the Lie group SU(2).

3. Symmetric power L-functions and functoriality

The proof of Dirichlet's theorem on primes relied on a non-vanishing result for certain (Dirichlet) L-functions. The same is true for the proof of the Sato–Tate conjecture. To prove the equidistribution of the conjugacy classes of the $u_p(E)$ in SU(2), we must show that

$$\lim_{X\to\infty} \frac{\sum_{p\leq X} f(u_p(E))}{\#\{p \leq X\}} = \int_{\mathrm{SU}(2)} f(u)du,$$

with du the Haar probability measure on SU(2), for any continuous class function f on SU(2).

The Peter–Weyl theorem implies that it suffices to check this equality for f running over the characters of irreducible representations of SU(2). These irreducible representations are given by $V_n = \mathrm{Sym}^n \mathbf{C}^2$ for $n \geq 0$, where the action of SU(2) on $\mathbf{C}^2$ is the standard action by matrix multiplication. Orthogonality of characters tells us that integrating the character χ_n of V_n over SU(2) gives 0 if $n \geq 1$ and 1 if $n = 0$.

We are left needing to prove that

$$\lim_{X\to\infty} \frac{\sum_{p\leq X} \chi_n(u_p(E))}{\#\{p \leq X\}} = 0 \tag{3.1}$$

for $n \geq 1$.

As in Section 1.1, the sum $\sum_p \frac{\chi_n(u_p(E))}{p^s}$ contributes the main term to the log of an L-function, the nth symmetric power L-function, defined by

$$L(E, \chi_n, s) := \prod_p \det(1 - u_p(E)p^{-s} \mid V_n).$$

To show (3.1), it suffices to prove that, for each $n \geq 1$, $L(E, \chi_n, s)$ extends to a meromorphic function on an open neighbourhood of the region $\mathrm{Re}(s) \geq 1$ and is non-zero on this half-plane. See, for example, [20, Appendix to Ch. I] for a proof using the Wiener–Ikehara theorem.

Example 3.1. The first symmetric power L-function we need to understand is for $n = 1$. In this case, $L(E, \chi_1, s)$ is essentially the Hasse–Weil L-function of the elliptic curve E. Its non-vanishing for $\mathrm{Re}(s) \geq 1$ implies that (informally) the expected value $\mathbf{E}\left(\frac{a_p(E)}{2\sqrt{p}}\right)$ is zero.

The proof of this non-vanishing combines the modularity of elliptic curves with a non-vanishing result for Hecke L-functions of modular forms.

Theorem 3.2 ([4, 24, 28]). *There is a modular form f (a cuspidal Hecke eigenform of weight 2) with*

$$L(E, \chi_1, s) = L\left(f, s + \frac{1}{2}\right).$$

The Hecke L-function $L(f,s)$ can be identified (up to some explicit factors) with the Mellin transform $\int_0^\infty f(iy)y^{s-1}dy$.

Theorem 3.3 ([18,19]). *Suppose f is a cuspidal Hecke eigenform of weight 2. Then $L(f,s) \neq 0$ for $\mathrm{Re}(s) \geq 3/2$.*

3.1. Symmetric power functoriality

In order to prove the Sato–Tate conjecture, the natural way to proceed is to generalize Example 3.1 to handle the higher ($n > 1$) symmetric power L-functions. The generalization of the modularity theorem follows from special cases of Langlands's functoriality conjectures for automorphic forms. They are formulated in terms of the modular form f and predict the existence of a related automorphic form $\mathrm{Sym}^n f$ (a special function on $\mathrm{GL}_{n+1}(\mathbf{R})$). We were able to establish this existence in joint work with Thorne.

Theorem 3.4 ([14,15,17]). *For each $n \geq 1$, there exists an automorphic form $\mathrm{Sym}^n f$ with L-function $L(E,\chi_n,s)$.*

Importantly, whilst the L-function $L(E,\chi_n,s)$ is very difficult to analyse directly, the automorphic L-function $L(\mathrm{Sym}^n f,s)$ has an integral representation, and the appropriate generalization of Theorem 3.3 was proved by Jacquet and Shalika [12].

With these ingredients, we can deduce the Sato–Tate conjecture. The original proof of the Sato–Tate conjecture instead proved something weaker about the L-functions $L(E,\chi_n,s)$: they can be represented as an alternating product of automorphic L-functions. The more precise statement of Theorem 3.4 can be used to prove an effective enhancement of the Sato–Tate conjecture [26].

4. Hints at proofs of symmetric power functoriality

We have given two different proofs of symmetric power functoriality for modular forms f. The second [17] applies more generally to Hilbert modular forms. The two proofs follow different strategies, but both work with the Galois representations

$$\mathrm{Sym}^n \rho_{E,l} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_{n+1}(\mathbf{Z}_l)$$

given by composing the nth symmetric power of the standard representation of GL_2 with $\rho_{E,l}$. From this point of view, we want to show that these Galois representations are *automorphic*, which amounts to constructing the associated automorphic form $\mathrm{Sym}^n f$.

Both strategies also build on the *modularity lifting theorems* introduced by Wiles in the proof of Fermat's last theorem, which propagate automorphy via congruences between Galois representations. However, the basic structure of the two strategies differs:

- In our first proof [14, 15], a key ingredient is a result due to Buzzard and Kilford on 2-adic modular forms [5]. It is closely related to the observation of Tate that if a Galois representation

$$\overline{\rho} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{SL}_2(\overline{\mathbf{F}}_2)$$

  is unramified at every odd prime, then its semisimplification is trivial. The proof of our main theorem is an induction on the level of the modular form (equivalently, on the conductor of the elliptic curve or associated Galois representation).
- Our later work [17] follows a strategy due to Clozel and Thorne [8] which relies on reducibility of the representation $\mathrm{Sym}^n(\overline{\mathbf{F}}_l^2)$ of GL_2 in characteristic $l \leq n$. That enables an induction on n, since the irreducible constituents of $\mathrm{Sym}^n(\overline{\mathbf{F}}_l^2)$ involve smaller symmetric powers.

On the other hand, a key ingredient in both proofs is to show that we do not have too many Galois representations.

Example 4.1. Let $F \subset \mathbf{C}$ be a number field. There are finitely many Galois extensions L/F (in $\mathbf{C}$) such that

- $\mathrm{Gal}(L/F)$ is Abelian,
- all primes of $\mathcal{O}_F$ are unramified in L.

In fact, the maximal such L is the Hilbert class field, whose Galois group over F is isomorphic to the ideal class group of $\mathcal{O}_F$.

Conjecture 4.2 ([10]). *There are finitely many isomorphism classes of representations* $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_n(\mathbf{Q}_l)$ *with bounded ramification.*[4]

We prove a weak form of this finiteness conjecture in [16], which is a crucial input to the methods of [14, 17]. This weaker statement is best interpreted in terms of a moduli space of Galois representations, so we will end the lecture by giving this interpretation.

[4] Describing the ramification condition at l involves p-adic Hodge theory.

5. Geometry of group representations

To give an idea of how to think about these moduli spaces of Galois representations, we introduce the representation and character variety of a finitely presented group

$$G = \langle g_1, \dots, g_d \rangle / \langle r_1, \dots, r_k \rangle.$$

To specify a homomorphism $G \to \mathrm{GL}_n(\mathbf{C})$, we need to give a d-tuple of matrices $A_1, \dots, A_d \in \mathrm{GL}_n(\mathbf{C})$ satisfying the relations $r_1, \dots, r_k$. These can be identified with the $\mathbf{C}$-points of an algebraic variety $R_{G,n}$, a closed subset of $(\mathrm{GL}_n)^d$. This is called the *representation variety*, whilst the quotient $A_{G,n} = R_{G,n}/\mathrm{PGL}_n(\mathbf{C})$ by the adjoint action is the *character variety*, whose points parameterize isomorphism classes of semisimple representations of G.

If we fix an irreducible representation $\rho : G \to \mathrm{GL}_n(\mathbf{C})$, the tangent space $T_\rho A_{G,n}$ can be identified with a cohomology group $H^1(G, M_n(\mathbf{C}))$, where G acts on $M_n(\mathbf{C})$ by conjugation via ρ.

Constructing an analogue $A^l_{\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}),n}$ of $A_{G,n}$ parameterizing l-adic Galois representations with bounded ramification is more involved, but this has been done by Wake and Wang-Erickson [27] building on the work of Mazur [13] and Chenevier [6].

The tangent space at an irreducible representation $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_n(\mathbf{Q}_l)$ is a subspace of $H^1(\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}), M_n(\mathbf{Q}_l))$, where the group action on the coefficient space is again conjugation via ρ. The subspace corresponds to the ramification conditions. If Conjecture 4.2 holds, the Galois moduli space is just a finite set of points, and so the tangent spaces would vanish.[5] Our weak finiteness result is the vanishing of these tangent spaces.

Theorem 5.1 ([1, 2, 16, 25]). *Under some technical assumptions, the tangent spaces to $A^l_{\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}),n}$ at points corresponding to automorphic Galois representations vanish.*

We view this vanishing result as finiteness of (the l-primary part of) a generalized class group. In fact, for $n = 1$, it follows from finiteness of the usual class group.

Acknowledgements. Many thanks to the organizers and participants at the 2024 ECM in Sevilla for a very stimulating conference.

Funding. This work was supported by a UKRI Future Leaders Fellowship, grant MR/V021931/1.

[5] More precisely, finiteness implies that the underlying reduced subspace of the Galois moduli space is a finite set of points, so we are implicitly enhancing the finiteness conjecture to include reducedness of the moduli space.

References

[1] L. A'Campo, Rigidity of automorphic Galois representations over CM fields. *Int. Math. Res. Not. IMRN* **2024** (2024), no. 6, 4541–4623 Zbl 1562.11085 MR 4721650

[2] P. B. Allen, Deformations of polarized automorphic Galois representations and adjoint Selmer groups. *Duke Math. J.* **165** (2016), no. 13, 2407–2460 Zbl 1417.11093 MR 3546966

[3] T. Barnet-Lamb, D. Geraghty, M. Harris, and R. Taylor, A family of Calabi–Yau varieties and potential automorphy. II. *Publ. Res. Inst. Math. Sci.* **47** (2011), no. 1, 29–98 Zbl 1264.11044 MR 2827723

[4] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, On the modularity of elliptic curves over $\mathbf{Q}$: wild 3-adic exercises. *J. Amer. Math. Soc.* **14** (2001), no. 4, 843–939 Zbl 0982.11033 MR 1839918

[5] K. Buzzard and L. J. P. Kilford, The 2-adic eigencurve at the boundary of weight space. *Compos. Math.* **141** (2005), no. 3, 605–619 Zbl 1187.11020 MR 2135280

[6] G. Chenevier, The p-adic analytic space of pseudocharacters of a profinite group and pseudorepresentations over arbitrary rings. In F. Diamond, P. L. Kassaei, and M. Kim (eds.), *Automorphic forms and Galois representations, Volume 1*, pp. 221–285, London Math. Soc. Lecture Note Ser. 414, Cambridge University Press, Cambridge, 2014 Zbl 1350.11063 MR 3444227

[7] L. Clozel, M. Harris, and R. Taylor, Automorphy for some l-adic lifts of automorphic mod l Galois representations. *Publ. Math. Inst. Hautes Études Sci.* **108** (2008), 1–181 Zbl 1169.11020 MR 2470687

[8] L. Clozel and J. A. Thorne, Level-raising and symmetric power functoriality. I. *Compos. Math.* **150** (2014), no. 5, 729–748 Zbl 1304.11040 MR 3209793

[9] J. P. G. L. Dirichlet, Beweis des Satzes, dass jede unbegrenzte arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthält. In *Abhandlungen der Königlich Preussischen Akademie der Wissenschaften von 1837*, pp. 45–81, Königlich Akademie der Wissenschaften, Berlin, 1839

[10] J.-M. Fontaine and B. Mazur, Geometric Galois representations. In J. Coates and S.-T. Yau (eds.), *Elliptic curves, modular forms & Fermat's last theorem (Hong Kong, 1993)*, pp. 41–78, Ser. Number Theory I, International Press, Cambridge, MA, 1995 Zbl 0839.14011 MR 1363495

[11] M. Harris, N. Shepherd-Barron, and R. Taylor, A family of Calabi–Yau varieties and potential automorphy. *Ann. of Math. (2)* **171** (2010), no. 2, 779–813 Zbl 1263.11061 MR 2630056

[12] H. Jacquet and J. A. Shalika, A non-vanishing theorem for zeta functions of GL_n. *Invent. Math.* **38** (1976), no. 1, 1–16 Zbl 0349.12006 MR 0432596

[13] B. Mazur, Deforming Galois representations. In Y. Ihara, K. Ribet, and J.-P. Serre (eds.), *Galois groups over $\mathbb{Q}$ (Berkeley, CA, 1987)*, pp. 385–437, Math. Sci. Res. Inst. Publ. 16, Springer, New York, 1989 Zbl 0714.11076 MR 1012172

[14] J. Newton and J. A. Thorne, Symmetric power functoriality for holomorphic modular forms. *Publ. Math. Inst. Hautes Études Sci.* **134** (2021), 1–116 Zbl 1503.11085 MR 4349240

[15] J. Newton and J. A. Thorne, Symmetric power functoriality for holomorphic modular forms. II. *Publ. Math. Inst. Hautes Études Sci.* **134** (2021), 117–152 Zbl 1503.11086 MR 4349241

[16] J. Newton and J. A. Thorne, Adjoint Selmer groups of automorphic Galois representations of unitary type. *J. Eur. Math. Soc. (JEMS)* **25** (2023), no. 5, 1919–1967 Zbl 1536.11082 MR 4592862

[17] J. Newton and J. A. Thorne, Symmetric power functoriality for Hilbert modular forms. *Ann. of Math. (2)* **203** (2026), no 1, 283–347 MR 5008554

[18] A. P. Ogg, On a convolution of L-series. *Invent. Math.* **7** (1969), no. 4, 297–312 Zbl 0205.50902 MR 0246819

[19] R. A. Rankin, Contributions to the theory of Ramanujan's function $\tau(n)$ and similar arithmetical functions. I. The zeros of the function $\sum_{n=1}^{\infty} \frac{\tau(n)}{n^s}$ on the line $\Re s = \frac{13}{2}$. *Math. Proc. Cambridge Philos. Soc.* **35** (1939), no. 3, 351–356 Zbl 0021.39201 MR 0000411

[20] J.-P. Serre, *Abelian l-adic representations and elliptic curves*. Res. Notes Math. 7, A K Peters, Wellesley, MA, 1998, 180 pp. Zbl 0902.14016 MR 1484415

[21] J. H. Silverman, *The arithmetic of elliptic curves*. 2nd edn., Grad. Texts in Math. 106, Springer, New York, NY, 2009, 513 pp. Zbl 1194.11005 MR 2514094

[22] A. V. Sutherland, Sato–Tate distributions in genus 1. https://math.mit.edu/~drew/g1SatoTateDistributions.html visited on 17 March 2026

[23] R. Taylor, Automorphy for some l-adic lifts of automorphic mod l Galois representations. II. *Publ. Math. Inst. Hautes Études Sci.* **108** (2008), 183–239 Zbl 1169.11021 MR 2470688

[24] 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

[25] J. A. Thorne, On the vanishing of adjoint Bloch–Kato Selmer groups of irreducible automorphic Galois representations. *Pure Appl. Math. Q.* **18** (2022), no. 5, 2159–2202 Zbl 1511.11052 MR 4538044

[26] J. Thorner, Effective forms of the Sato–Tate conjecture. *Res. Math. Sci.* **8** (2021), no. 1, article no. 4, 21 pp. Zbl 1465.11140 MR 4198758

[27] P. Wake and C. Wang-Erickson, Deformation conditions for pseudorepresentations. *Forum Math. Sigma* **7** (2019), article no. e20, 44 pp. Zbl 1470.11143 MR 3987305

[28] 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

James Newton
Mathematical Institute, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK;
james.newton@maths.ox.ac.uk

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/27

Differential calculus and optimization in persistence module categories

Steve Oudot

Abstract. Persistence modules are representations of products of totally ordered sets in the category of vector spaces. They appear naturally in the representation theory of algebras, but in recent years, they have also found applications in other areas of mathematics, including symplectic topology, complex analysis, and topological data analysis, where they arise from filtrations of topological spaces by the sublevel sets of real-valued functions. Two fundamental properties of persistence modules make them useful in such contexts: (1) the fact that they are stable under perturbations of the originating functions and (2) the fact that they can be approximated, in the sense of relative homological algebra, by classes of indecomposable modules with an elementary structure. In this text, we give an introduction to the theory of persistence modules, then we explain how the above properties can be leveraged to build a framework for differential calculus and optimization with convergence guarantees in persistence module categories.

1. Introduction

Broadly speaking, persistence modules are representations of posets in the category of vector spaces over a fixed field. While these objects have been studied for decades, particularly in representation theory, the name *persistence modules* was coined by the topological data analysis (TDA) community at the turn of the 21st century [55] to emphasize the role played by these modules in the study of the topology of sublevel sets of functions defined on data. Meanwhile, TDA has shed a different light on these objects, raising new questions for mathematicians: a whole new theory of persistence modules has emerged, called *persistence theory* [48], centered around the definition of invariants for persistence modules that can be effectively computed from data, that are provably stable under perturbations of the data, and whose constructions can be

Mathematics Subject Classification 2020: 55N31 (primary); 16G20, 49J52, 90C15 (secondary).
Keywords: topological data analysis, poset representations, nonsmooth analysis, o-minimal geometry, stochastic gradient descent.

differentiated so that their parameters can be optimized in machine or deep learning contexts.

This text is a walk-through of persistence theory, including its basic results and some of its latest developments. By no means is it an exhaustive account of the theory or of its connections to other areas of mathematics; for such accounts, the reader can refer to recent textbooks on the subject [24, 49, 50]. The story arc here is the study of a particular class of invariants for persistence modules called *barcodes*, which play a central role in the 1-parameter instance of the theory – focused on the sublevel sets of $\mathbb{R}$-valued functions – and which have recently been adapted as *signed barcodes* to the multi-parameter instance of the theory – focused on $\mathbb{R}^n$-valued functions. Our goal is to unveil the meaning of these invariants in terms of the structure of persistence modules and to show how, ultimately, their construction can be differentiated and optimized in data analysis contexts.

The text is organized as follows: Section 2 introduces persistence modules, and Section 3 connects them to the topology of sublevel sets of functions; Sections 4 and 5 study their structure and stability properties, which are leveraged in Sections 6 and 7 to build a framework for differential calculus and optimization with persistence modules; the text concludes with Section 8, giving some directions for possible future developments on the subject.

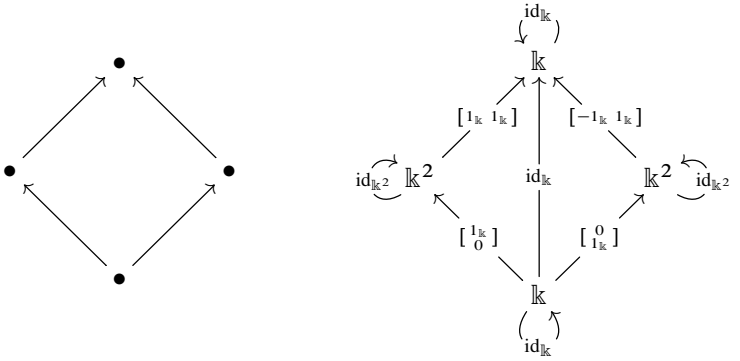
2. Persistence modules

Informally, a persistence module is a diagram of vector spaces and linear maps indexed by a poset. Formally, we fix a poset $(P, \leq)$ and a field $\mathbb{k}$. The order relation $\leq$ is omitted in the notations whenever it is obvious from the context. We make no assumption on $\mathbb{k}$; in particular it does not have to be algebraically closed, and its characteristic can be arbitrary. For now, we make no assumption on P either.

Definition 2.1. A *persistence module* over P is a functor $M : P \to \mathsf{vect}_{\mathbb{k}}$, where $\mathsf{vect}_{\mathbb{k}}$ denotes the category of finite-dimensional vector spaces over $\mathbb{k}$, and where P is identified with the category having one object for each element $p \in P$ and a unique morphism for each pair of comparable elements $p \leq q \in P$. A *morphism* of persistence modules $M \to N$ is a natural transformation between functors $M, N : P \to \mathsf{vect}_{\mathbb{k}}$.

Below is an example of a poset P, represented by its Hasse diagram (left), and of a persistence module M over P (right). The diagram on the right-hand side commutes

by functoriality of M.



From now on, unless otherwise specified, we assume our poset P to be a finite product of totally ordered sets, as is the case for instance in the above example:

$$P = \prod_{i=1}^{n} T_i, \quad \text{where each } T_i \text{ is totally ordered.}$$

Here, $n \in \mathbb{N}^*$ is fixed and considered as the number of parameters of the persistence modules over P; thus, we speak of *n-parameter persistence modules*. Note that the totally ordered sets T_i can be arbitrary; in particular, they are not required to be finite.

We write $\mathsf{vect}_{\Bbbk}^{P}$ for the category of persistence modules over P. As a category of functors from a small category to an abelian category, $\mathsf{vect}_{\Bbbk}^{P}$ is itself abelian. This means in particular that it has a zero object, called the *zero* (or *trivial*) persistence module and denoted by 0. The category also has pointwise finite direct sums, that is, direct sums $\bigoplus_{j\in J} M_j$ such that the set $\{j \mid M_j(p) \neq 0\}$ is finite for every $p \in P$. Moreover, Hom-sets are $\Bbbk$-vector spaces, with composition being bilinear, and every morphism has a kernel and a cokernel. The other properties of an abelian category will be used implicitly in this text.

Persistence modules as modules. As their name suggests, persistence modules can be viewed as modules over a certain algebra $\Bbbk P$ defined from the indexing poset P. This algebra is freely generated by the pairs of comparable elements in P:

$$\Bbbk P := \Bbbk^{(\{(p,q)\mid p\leq q\in P\})},$$

with multiplication given by

$$(p,q)\cdot(p',q') := \begin{cases} (p,q') & \text{if } q = p', \\ 0 & \text{otherwise,} \end{cases}$$

and extended by linearity.

There is an embedding of $\mathsf{vect}_{\Bbbk}^{P}$ into the category $\mathrm{Mod}_{\Bbbk P}$ of right modules over $\Bbbk P$, given by the fully faithful additive functor $\mathsf{vect}_{\Bbbk}^{P} \to \mathrm{Mod}_{\Bbbk P}$ defined on objects by $M \mapsto \bigoplus_{p \in P} M_p$ with right multiplication:

$$\left(\sum_{p \in P} m_p \right) \cdot (q, r) = M(q \leq r)(m_q).$$

Thus, persistence modules can be viewed as right modules over the algebra $\Bbbk P$. When P is finite, the embedding $\mathsf{vect}_{\Bbbk}^{P} \hookrightarrow \mathrm{Mod}_{\Bbbk P}$ induces an equivalence of abelian categories $\mathsf{vect}_{\Bbbk P} \simeq \mathrm{mod}_{\Bbbk P}$, where $\mathrm{mod}_{\Bbbk P}$ is the category of finite-dimensional right $\Bbbk P$-modules.

3. Sublevel sets topology

The typical use case of persistence modules is in the study of the topology of sublevel sets. Given a topological space X and a function $f = (f_1, \ldots, f_n) \colon X \to \mathbb{R}^n$, both of which can be arbitrary, one considers the sublevel sets of f, written $F(t)$ for every $t = (t_1, \ldots, t_n) \in \mathbb{R}^n$ and defined as follows – see Figure 1 for an illustration:

$$F(t) := \bigcap_{i=1}^{n} f_i^{-1}((-\infty, t_i]).$$

Letting $P = \mathbb{R}^n$, viewed as the product of n copies of the totally ordered real line and equipped with the induced product order $\leq$, one gets

$$F(s) \subseteq F(t) \quad \forall s \leq t \in \mathbb{R}^n.$$

This yields a functor $F \colon P \to \mathsf{Top}$, where Top denotes the category of topological spaces, defined as above on objects and sending each morphism $s \leq t \in \mathbb{R}^n$ to the

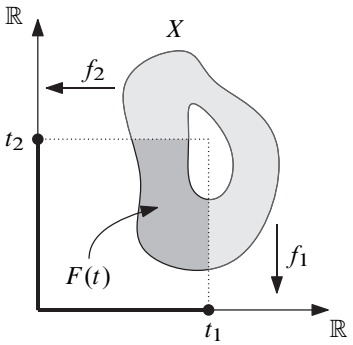


Figure 1. The $t = (t_1, t_2)$-sublevel set of the function $f = (f_1, f_2) \colon X \to \mathbb{R}^2$ defined by the orthogonal projections onto the coordinate axes.



Figure 2. Filtration of the distance function (excerpt with 10 different levels t, starting at $t = 0$).

inclusion map $F(s) \hookrightarrow F(t)$. This functor is called the *filtration* of the sublevel sets of f.

Assuming (as we do from now on) that the sublevel sets of f each have finite-dimensional singular homology in every degree, one post-composes F by the singular homology functor H_* in some fixed degree $*$ with coefficients in $\Bbbk$ to get a persistence module $H_*(f)\colon \mathbb{R}^n \to \mathsf{vect}_\Bbbk$ called the *persistent homology of f in degree $*$*.

Let us give two classic examples in the context of TDA.

Example 3.1. Let $A \subset \mathbb{R}^d$ be a finite set of data points. Take $X = \mathbb{R}^d$ and define $f\colon \mathbb{R}^d \to \mathbb{R}$ as the distance to the data

$$f(x) := \min_{a \in A} \|x - a\|_2.$$

Its sublevel sets are unions of Euclidean balls of the same radius around the data points, as illustrated in Figure 2. The corresponding 1-parameter persistence modules in singular homology of degrees 0 and 1 over $\Bbbk$, denoted by $H_0(f)$ and $H_1(f)$, respectively, are shown below (restricted to the same 10 indices t as in Figure 2).

$$H_0(f)\colon \Bbbk^{84} \twoheadrightarrow \Bbbk^9 \twoheadrightarrow \Bbbk^3 \twoheadrightarrow \Bbbk^2 \twoheadrightarrow \Bbbk \xrightarrow{\simeq} \Bbbk \xrightarrow{\simeq} \Bbbk \xrightarrow{\simeq} \Bbbk \xrightarrow{\simeq} \Bbbk \xrightarrow{\simeq} \Bbbk\,,$$

$$H_1(f)\colon 0 \longrightarrow \Bbbk^3 \xrightarrow{0} \Bbbk \xrightarrow{0} \Bbbk \xrightarrow{0} \Bbbk^2 \xrightarrow{[1\ 0]} \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0\,.$$

As singular homology in degree 0 captures path-connected components, the pointwise dimension of $H_0(f)$ gives the evolution of the number of connected components in the union of balls: initially 84 (since there are 84 data points in A), it decreases monotonically as the radius increases and the balls grow, until it stabilizes at 1, at which point the union of balls forms a single connected component. The arrows in $H_0(f)$ encode how the connected components merge as the radius increases.

Similarly, singular homology in degree 1 captures holes in the union of balls: initially 0, it becomes non-trivial when holes appear in the filtration. Of notable interest is the fifth arrow in the sequence, which is nonzero and encodes the fact that the main hole present at the fifth step of the filtration of Figure 2 is still present at the sixth step, while the other holes do not persist across steps.

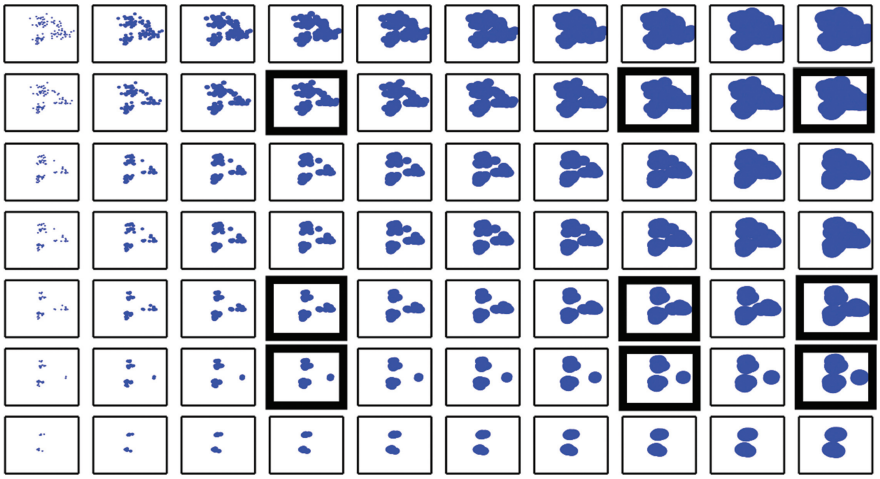


Figure 3. Bi-filtration of the distance function and co-density estimator (excerpt with 10 different levels for the distance and 7 for the co-density). Distance levels are represented horizontally and co-density levels vertically. The indices highlighted by thick rectangles arranged in a 3×3 grid exhibit three connected components that merge pairwise in three different ways, then all together.

Example 3.2. Let again $A \subset \mathbb{R}^d$ be a finite set of data points. Take $X = \mathbb{R}^d$ and define $f = (f_1, f_2)\colon \mathbb{R}^d \to \mathbb{R}^2$ as the combination of the distance to the data with some co-density estimator:

$$f_1(x) := \min_{a \in A} \|x - a\|_2,$$
$$f_2(x) := \#\{a \in A \mid \|x - a\|_2 > \varepsilon\} \quad \text{for some fixed parameter } \varepsilon \geq 0.$$

Its sublevel sets are unions of Euclidean balls of the same radius around data points with high enough density, as illustrated in Figure 3. The corresponding 2-parameter persistence module in singular homology of degree 0 over $\Bbbk$ restricts to the following persistence module over the 3×3 grid highlighted in Figure 3.

$$\begin{array}{ccccc}
\Bbbk^2 & \xrightarrow{\begin{bmatrix}1 & 1\end{bmatrix}} & \Bbbk & \xrightarrow{\mathrm{id}} & \Bbbk \\
{\scriptstyle \begin{bmatrix}1&0&0\\0&1&1\end{bmatrix}}\uparrow & & \uparrow{\scriptstyle \begin{bmatrix}1 & 1\end{bmatrix}} & & \uparrow{\scriptstyle \mathrm{id}} \\
\Bbbk^3 & \xrightarrow{\begin{bmatrix}1&0&1\\0&1&0\end{bmatrix}} & \Bbbk^2 & \xrightarrow{\begin{bmatrix}1 & 1\end{bmatrix}} & \Bbbk \\
{\scriptstyle \mathrm{id}}\uparrow & & \uparrow{\scriptstyle \begin{bmatrix}1&0&1\\0&1&0\end{bmatrix}} & & \uparrow{\scriptstyle \begin{bmatrix}1 & 1\end{bmatrix}} \\
\Bbbk^3 & \xrightarrow[\mathrm{id}]{} & \Bbbk^3 & \xrightarrow[\begin{bmatrix}1&1&0\\0&0&1\end{bmatrix}]{} & \Bbbk^2
\end{array} \tag{3.1}$$

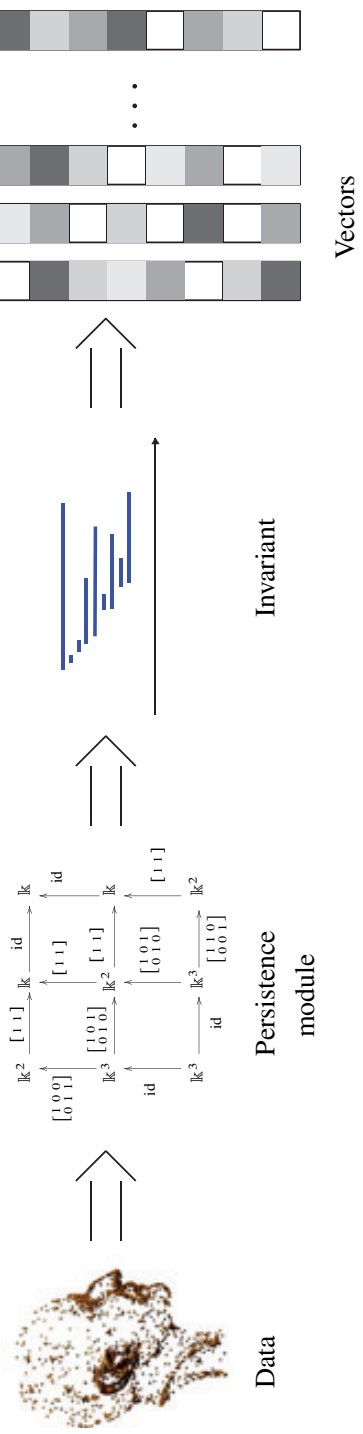


Figure 4. The topological data analysis pipeline.

More broadly, in the context of TDA, one builds persistence modules from data, similarly to Examples 3.1 and 3.2. Then one computes invariants on these modules: a variety of invariants have been proposed, and in this text, they all take the form of *barcodes*, that is, multisets of intervals in the indexing poset. Then one turns these barcodes into vectors in order to feed them to some machine or deep learning architecture. The corresponding pipeline is illustrated in Figure 4. Two fundamental properties of persistence modules make this pipeline relevant for data analysis:

(Structure) Persistence modules can be decomposed, or approximated in the sense of homological algebra, by certain classes of modules with an elementary structure called *interval modules*; this is where their barcodes are defined from.

(Stability) Persistence modules and their barcodes are stable with respect to perturbations of their originating functions; this enables the use of barcodes as stable features for data.

The next two sections are devoted to fleshing out these properties: Section 4 for the structure, Section 5 for the stability.

4. Structure of persistence modules

As before, let $(P, \leq)$ be a poset and $\mathbb{k}$ a field. Unless otherwise stated, we make no assumption on P. There are a variety of ways to encode the structure of persistence modules over P, and here, we report on two of them: via direct sum decompositions (Section 4.1) and via homological approximations (Section 4.2). In both cases, interval modules play a central part in the story, so we begin by giving their definition.

Definition 4.1. An *interval of* P is a subset $I \subseteq P$ that is both convex and connected with respect to the order $\leq$ on P, that is,

$$\text{(Convex)} \quad s,t \in I \Longrightarrow u \in I \ \forall\, s \leq u \leq t,$$
$$\text{(Connected)} \quad s,t \in I \Longrightarrow \exists r \in \mathbb{N}^*, \exists \{u_i\}_{i=0}^{r} \subseteq I : s = u_0 \leq u_1 \geq \cdots \geq u_r = t.$$

Definition 4.2. Given an interval $I \subseteq P$, the corresponding *interval module* $\mathbb{k}_I$ is defined by

$$\mathbb{k}_I(t) = \begin{cases} \mathbb{k} & \text{if } t \in I, \\ 0 & \text{otherwise,} \end{cases} \qquad \mathbb{k}_I(s \leq t) = \begin{cases} \mathrm{id}_{\mathbb{k}} & \text{if } s,t \in I, \\ 0 & \text{otherwise.} \end{cases}$$

Interval modules are useful for applications because they are entirely characterized by their support, which is an interval of P and therefore a geometric object. This makes it possible for practitioners to think purely in terms of the geometry of these intervals, thus forgetting about the algebra. Examples will be given in the following.

4.1. Direct sum decompositions

A persistence module M is said to be *decomposable* if it decomposes as $M \simeq N \oplus N'$ for some modules $N, N' \neq 0$. Otherwise, if no such decomposition exists, M is called *indecomposable*. Note that interval modules are indecomposable: this is because their endomorphism ring is isomorphic to $\Bbbk$, hence local. Therefore, interval modules can serve as building blocks for decomposing arbitrary persistence modules. The hope is that no other types of indecomposables appear in the decomposition, which is what happens for instance when the poset P is totally ordered – an important special case that encompasses the study of the topology of the sublevel sets of real-valued functions as in Example 3.1.

Theorem 4.3 ([11, 22]). *Suppose $(P, \leq)$ is totally ordered. Then any $M \in \mathsf{vect}_{\Bbbk}^{P}$ decomposes as follows:*

$$M \simeq \bigoplus_{j \in J} \Bbbk_{I_j},$$

where J is a potentially infinite or even uncountable indexing set, and where each I_j is an interval of P. This decomposition is unique up to isomorphism and reordering of the terms.

When P is finite, this is just Gabriel's theorem [27] specialized to the case of A_n-type quivers. The version of the result stated here, due to Botnan and Crawley-Boevey, makes no assumption on P (besides being totally ordered) nor on M.

Example 4.4. The persistence modules from Example 3.1 decompose as follows.

$$\begin{aligned}
H_0(f) \simeq\ & \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \\
& \oplus\ \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \oplus\ \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \oplus \Big(\Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \Big)^6 \\
& \oplus\ (\Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0)^{75}, \\
H_1(f) \simeq\ & 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \Bbbk \xrightarrow{\mathrm{id}_\Bbbk} \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \oplus\ 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \oplus\ 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \oplus\ 0 \longrightarrow 0 \longrightarrow \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \\
& \cdot (0 \longrightarrow \Bbbk \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0 \longrightarrow 0)^3.
\end{aligned}$$

Interval decompositions, when they exist, provide a direct interpretation of the structure of the persistence modules: in Example 4.4, for instance, the supports of the interval summands of $H_0(f)$ encode the lifespan of each path-connected component in the union of balls as the radius grows; meanwhile, all but one of the interval summands of $H_1(f)$ are supported on singletons, indicating that the corresponding holes appearing in the growing union of balls have an ephemeral existence, while the non-singleton interval has a corresponding hole (the main one in Figure 2) that lasts over two levels in the excerpt. This motivates the use of interval decompositions – when they exist – in applications. Or, rather, it motivates the use of their associated *barcodes*, which encode the (unique) geometric information contained in each decomposition, that is, the supports of the interval summands.

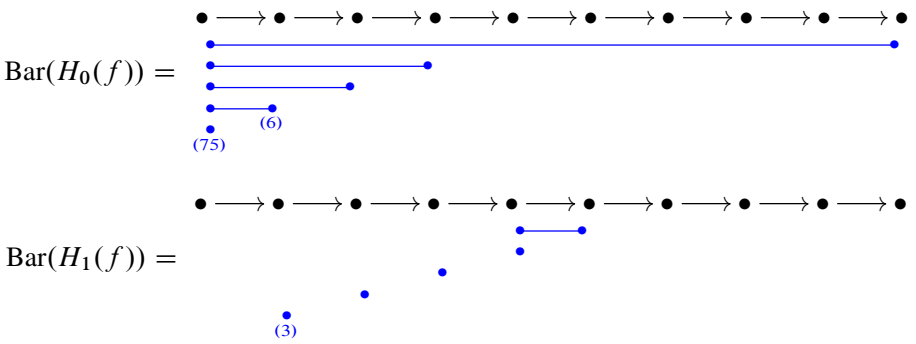
Definition 4.5. Suppose a persistence module M over an arbitrary poset P decomposes as a direct sum of interval modules:

$$M \simeq \bigoplus_{j \in J} \Bbbk_{I_j}, \quad \text{where each } I_j \text{ is an interval of } P.$$

Then M is called *interval decomposable*, and by uniqueness of the decomposition up to isomorphism and reordering of the terms, the multiset of intervals involved in the decomposition is unique. This multiset is called the *barcode* of M, denoted by $\mathrm{Bar}(M)$:

$$\mathrm{Bar}(M) := \{I_j \mid j \in J\}.$$

Example 4.6. The interval decompositions of Example 4.4 lead to the following barcodes (in blue).



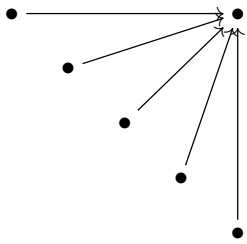
Unfortunately, when P is not totally ordered, for instance, when P is a finite product of $n > 1$ totally ordered sets, interval decomposability is no longer guaranteed, even though every persistence module M still decomposes into indecomposable summands.

Theorem 4.7 ([11]). *For any poset $(P, \leq)$, any $M \in \mathsf{vect}_{\Bbbk}^{P}$ decomposes as follows:*

$$M \simeq \bigoplus_{j \in J} M_j,$$

where each summand M_j is indecomposable with local endomorphism ring. This decomposition is unique up to isomorphism and reordering of the terms.

The summands M_j appearing in the decomposition may not be interval modules: for instance, the persistence module (3.1) is indecomposable yet not an interval module, being not of pointwise dimension ≤ 1. And such summands do appear in applications, since (3.1) is coming from Example 3.2. From the algebraic standpoint, the situation is even worse: not only may persistence modules have non-interval summands, but in fact the algebra $\Bbbk P$ is generally of *wild representation type*, since for instance one can embed the representation category of the following wild-type quiver into the representation category of $\Bbbk P$ when P contains a large enough n-dimensional ($n > 1$) grid.



For us, this means that, unless P is very small (not more than a handful of elements), when $n > 1$, there is no canonical classification of all the indecomposable persistence modules over the poset P; therefore, parametrizing the indecomposables to build up unique barcodes from direct sum decompositions is an impossible task.

4.2. Homological approximations

Given the above shortcoming, and the importance of working with invariants that can be interpreted in application contexts, the TDA community has made significant efforts in the past decade to propose alternative invariants from which meaningful barcodes can be built. These invariants are necessarily incomplete, considering the wild representation type of $\Bbbk P$. Here, we describe one class of such invariants, based on homological approximations of the modules.

From now on, we assume that the poset P is finite, so its algebra $\Bbbk P$ is finite dimensional. We can then rely on the well-established theory of representations of Artin algebras. The indecomposable projectives in this setting are the interval modules $\Bbbk_{[p,\infty)}$, where $[p, \infty) := \{q \in P \mid q \geq p\}$ is the principal upset at p in P. Thus, by

taking successive projective approximations, we can build left resolutions of any persistence module M by interval modules of this type. Moreover, assuming M admits finite projective presentations, the minimal such resolution has finite length, bounded by the cardinality of P, or, more tightly, by the number n of its parameters when P is a product of n finite totally ordered sets, by Hilbert's syzygy theorem.

$$0 \longrightarrow \bigoplus_{p\in J_n} \Bbbk_{[p,\infty)} \longrightarrow \bigoplus_{p\in J_{n-1}} \Bbbk_{[p,\infty)} \longrightarrow \cdots \longrightarrow \bigoplus_{p\in J_0} \Bbbk_{[p,\infty)} \twoheadrightarrow M.$$

As the minimal resolution is unique up to isomorphism, the supports of the summands in the various terms of the resolution are uniquely defined – so are the multisets J_i for $0 \leq i \leq n$. We can then gather the supports to form a barcode: more precisely, for reasons that we will explain afterward, we build a composite object called the *signed barcode* of M, denoted by $\mathrm{Bar}_{\llcorner}(M)$, composed of two regular barcodes; the first one (called the *positive part* and denoted by $\mathrm{Bar}^{+}_{\llcorner}(M)$) gathers the supports coming from even degrees in the resolution, and the second one (called the *negative part* and denoted by $\mathrm{Bar}^{-}_{\llcorner}(M)$) gathers the supports coming from odd degrees:

$$\mathrm{Bar}_{\llcorner}(M) := (\mathrm{Bar}^{+}_{\llcorner}(M),\, \mathrm{Bar}^{-}_{\llcorner}(M)),$$
$$\text{where } \begin{cases} \mathrm{Bar}^{+}_{\llcorner}(M) := \bigsqcup_{i\in 2\mathbb{N}} \{[p,\infty) \mid p \in J_i\}, \\ \mathrm{Bar}^{-}_{\llcorner}(M) := \bigsqcup_{i\in 2\mathbb{N}+1} \{[p,\infty) \mid p \in J_i\}. \end{cases} \tag{4.1}$$

Compared to regular barcodes, signed barcodes put signs on their intervals based on the parity of the degree in the resolution in which they appear. This twist is motivated by what happens in the Grothendieck group K_0, where the resolution yields the following signed decomposition of the equivalence class $[M]$ of M:

$$[M] = \sum_{i\in\mathbb{N}} (-1)^i \left(\sum_{p\in J_i} [\Bbbk_{[p,\infty)}] \right) = \sum_{p\in \bigsqcup_{i \text{ even}} J_i} [\Bbbk_{[p,\infty)}] - \sum_{p\in \bigsqcup_{i \text{ odd}} J_i} [\Bbbk_{[p,\infty)}].$$

Note that some terms may cancel out in this alternating sum, in which case they disappear from the decomposition of $[M]$ in K_0. By contrast, we keep them in the positive and negative parts of the signed barcode (4.1), thus preserving all the information contained in the terms of the minimal projective resolution of M. In practice, this is more costly to compute, since we have to go through computing the resolution as opposed to directly decomposing M in K_0. But it yields a stronger invariant, with provable stability properties as we shall see in the next section.

The biggest limitation of this approach is that it does not make use of the full set of intervals in P to approximate the modules, restricting itself to using the subset of principal upsets. In order to enable the use of larger collections of intervals to build our resolutions, we rely on the theory of relative homological algebra for Artin

algebras developed in [3]. Let thus $\mathcal{I}$ be a set of intervals in P. In view of the above discussion, we assume that $\mathcal{I}$ contains all the principal upsets $[p, \infty)$ for $p \in P$. The bottom line is to make the interval modules $\Bbbk_I$ for $I \in \mathcal{I}$ projective. For this, we consider only those short exact sequences on which all the Hom functors $\mathrm{Hom}(\Bbbk_I, -)$ for $I \in \mathcal{I}$ are exact. Let $\mathcal{E}_{\mathcal{I}}$ denote the collection of such sequences:

$$\mathcal{E}_{\mathcal{I}} := \{0 \to L \to M \to N \to 0 \text{ exact} \mid 0 \to \mathrm{Hom}(\Bbbk_I, L) \to \mathrm{Hom}(\Bbbk_I, M) \to \mathrm{Hom}(\Bbbk_I, N) \to 0 \text{ exact for all } I \in \mathcal{I}\}.$$

Proposition 4.8 ([3]). *Suppose P is finite and $\mathcal{I}$ contains the principal upsets in P. Then:*

(i) $\mathsf{vect}_{\Bbbk}^P$ *equipped with $\mathcal{E}_{\mathcal{I}}$ has the structure of an exact category,*

(ii) $\mathsf{vect}_{\Bbbk}^P$ *has enough projectives relative to $\mathcal{E}_{\mathcal{I}}$,*

(iii) *the projectives relative to $\mathcal{E}_{\mathcal{I}}$ are the finite direct sums of interval modules of type $\Bbbk_I$ for $I \in \mathcal{I}$.*

Item (i) implies that we may consider projective resolutions of persistence modules relative to $\mathcal{E}_{\mathcal{I}}$, when they exist. Item (ii) comes from our assumption that $\mathcal{I}$ contains all the principal upsets in P, so the projectives relative to $\mathcal{E}_{\mathcal{I}}$ include in particular the usual projectives. This item ensures that every persistence module does admit projective resolutions relative to $\mathcal{E}_{\mathcal{I}}$. Finally, item (iii) guarantees that the terms in such resolutions are finite direct sums of interval modules supported on elements of $\mathcal{I}$ as desired. We can then take projective resolutions of any persistence module M relative to $\mathcal{E}_{\mathcal{I}}$. More precisely, we take the minimal such resolution, as shown below.

$$0 \longrightarrow \bigoplus_{I \in J_r} \Bbbk_I \longrightarrow \bigoplus_{I \in J_{r-1}} \Bbbk_I \longrightarrow \cdots \longrightarrow \bigoplus_{I \in J_1} \Bbbk_I \longrightarrow \bigoplus_{I \in J_0} \Bbbk_I \twoheadrightarrow M.$$

Provided this resolution has finite length, we can form a signed barcode as before, but this time relative to our choice of family $\mathcal{I}$ of intervals:

$$\mathrm{Bar}_{\mathcal{I}}(M) := (\mathrm{Bar}_{\mathcal{I}}^+(M), \mathrm{Bar}_{\mathcal{I}}^-(M)), \quad \text{where } \begin{cases} \mathrm{Bar}_{\mathcal{I}}^+(M) := \bigsqcup_{i \in 2\mathbb{N}} J_i, \\ \mathrm{Bar}_{\mathcal{I}}^-(M) := \bigsqcup_{i \in 2\mathbb{N}+1} J_i. \end{cases}$$

Again, the signs in the barcode match with the ones in the decomposition of the equivalence class $[M]$ of M in the Grothendieck group $K_0^{\mathcal{E}_{\mathcal{I}}}$ relative to $\mathcal{E}_I$, and the difference is that the signed barcode keeps the terms that cancel out in $K_0^{\mathcal{E}_{\mathcal{I}}}$:

$$[M] = \sum_{i \in \mathbb{N}} (-1)^i \left(\sum_{I \in J_i} [\Bbbk_I] \right) = \sum_{I \in \bigsqcup_{i \text{ even}} J_i} [\Bbbk_I] - \sum_{I \in \bigsqcup_{i \text{ odd}} J_i} [\Bbbk_I]. \tag{4.2}$$

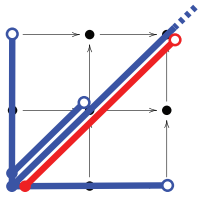
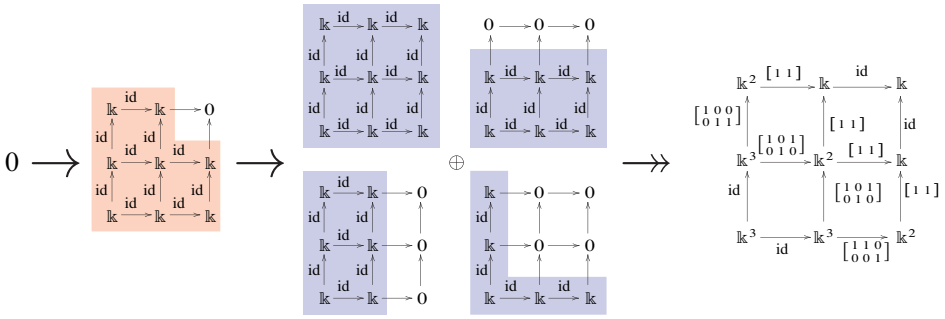


Figure 5. The signed barcode $\mathrm{Bar}_{\mathcal{I}}(M)$ of the indecomposable module M of Example 4.9, represented as an actual barcode in the Euclidean plane, where, by convention, each interval $[p,\infty) \setminus [q,\infty)$ gives rise to a copy of the line segment $[p,q)$, while each interval $[p,\infty)$ gives rise to a copy of the diagonal ray starting at p. The bars in the barcode are colored according to their sign: blue for positive, red for negative.

Example 4.9. Let us consider the indecomposable module M from (3.1). Let $\mathcal{I} = \{[p,\infty) \setminus [q,\infty) \mid p < q \in P \cup \{\infty\}\}$. The minimal projective resolution of M relative to $\mathcal{E}_{\mathcal{I}}$, which has length 1, is shown below without specifying the differentials (which are irrelevant to our invariant) and with the supports of the summands in the resolution colored blue or red depending on the parity of their degree in the resolution.



The blue supports form the positive part of the signed barcode $\mathrm{Bar}_{\mathcal{I}}(M)$, while the red supports form the negative part; see Figure 5 for an alternative representation of $\mathrm{Bar}_{\mathcal{I}}(M)$ as an actual barcode in the Euclidean plane. Note that if we replaced M by $M \oplus \Bbbk_{I_1}$, where $\Bbbk_{I_1}$ is the interval module appearing in degree 1 in the above resolution, then the minimal projective resolution would gain one copy of $\Bbbk_{I_1}$ in degree 0, which would be added to the positive part of the signed barcode but would cancel out with the copy of $\Bbbk_{I_1}$ in degree 1 in the relative Grothendieck group $K_0^{\mathcal{E}_{\mathcal{I}}}$.

The intervals considered in the above example are called *hooks*, because when P has two parameters, they have the shapes of (possibly degenerate) hooks, as we saw in the example. We therefore write $\mathcal{I} = \{\text{hooks}\}$ for their family. These intervals are of particular interest to us because their corresponding interval modules are quotients of two usual indecomposable projectives, so in this sense, {hooks} is the simplest family

of intervals to consider beyond the principal upsets. Another related consideration is that the intervals that appear in the direct sum decomposition of finitely presented modules in the 1-parameter setting (Theorem 4.3) are right open, and therefore, they are precisely the 1-dimensional hooks.

For the previous construction to make sense, it is important that the minimal projective resolution of M relative to $\mathcal{E}_{\mathcal{I}}$ has finite length, for otherwise the alternating sum in (4.2) is not defined. Another motivation is that, in practice, one can only store finite barcodes on a computer. This is why bounding the global dimension $\mathrm{gldim}^{\mathcal{E}_{\mathcal{I}}}(\mathsf{vect}_{\Bbbk}^{P})$ of the persistence module category relative to the chosen set $\mathcal{I}$ of intervals has become a prominent question for the TDA community in recent years. Here are some of the results obtained.

Theorem 4.10 ([2, 9, 12, 13, 16, 44]). *For any finite poset $(P, \leq)$, one has that $\mathrm{gldim}^{\mathcal{E}_{\mathcal{I}}}(\mathsf{vect}_{\Bbbk}^{P}) < \infty$ for the following sets $\mathcal{I}$ of intervals:*

- $\mathcal{I} = \{\textit{hooks}\}$,
- $\mathcal{I} = \{\textit{segments}\} := \{[p, q] \mid p \leq q \in P\}$,
- $\mathcal{I} = \{\textit{upsets}\} := \{[Q, \infty) \mid Q \subseteq P\}$, *where* $[Q, \infty) := \{s \in P \mid \exists q \in Q \text{ s.t. } s \geq q\}$,
- $\mathcal{I} = \{\textit{single-source intervals}\} := \{[p, \infty) \cap (-\infty, Q] \mid p \in P,\ Q \subseteq P\}$, *where* $(-\infty, Q] := \{s \in P \mid \exists q \in Q \text{ s.t. } s \leq q\}$,
- $\mathcal{I} = \{\textit{intervals}\}$.

Almost all the upper bounds on the global dimension obtained so far depend on the size of the poset P; therefore, they do not yield any interesting upper bound when P is infinite, for instance, when $P = \mathbb{R}^n$. A notable exception is the bound obtained for hooks, which depends only on the number n of parameters of P. More precisely, it was proven in [13] that $\mathrm{gldim}^{\mathcal{E}_{\{\text{hooks}\}}}(\mathsf{vect}_{\Bbbk}^{P})$ is exactly $2n - 2$ when $P = \prod_{i=1}^{n} T_i$ with each T_i finite and totally ordered. This linear dependence in n of the global dimension is remarkable in that it is of the same order as its counterpart in the usual projective setting, where Hilbert's syzygy theorem implies that $\mathrm{gldim}(\mathsf{vect}_{\Bbbk}^{P}) = n$. The result was then naturally extended to the case $P = \mathbb{R}^n$, assuming the modules under consideration are finitely presented and denoting by $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ the corresponding full subcategory.

Theorem 4.11 ([13]). $\mathrm{gldim}^{\mathcal{E}_{\{\textit{hooks}\}}}(\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}) = 2n - 2$.

Note that this result recovers the interval decomposition theorem (Theorem 4.3) for finitely presented 1-parameter persistence modules as a special case: indeed, when $n = 1$, the global dimension relative to hooks is $2 - 2 = 0$, which means that every finitely presented persistence module over $\mathbb{R}$ is projective relative to $\mathcal{E}_{\{\text{hooks}\}}$

and hence interval decomposable with right-open (i.e., hook) intervals. A more general and systematic treatment of the connection between the finite and infinite poset settings can be found in [1, 10].

5. Stability of persistence modules

Throughout this section, we assume the ground poset P to be $\mathbb{R}^n$ equipped with the product order, denoted by $\leq$.

5.1. Interleaving distance

Expressing the stability of persistence modules requires first equipping their category $\mathsf{vect}_{\Bbbk}^{\mathbb{R}^n}$ with a metric. Among the many choices of metrics available, one turns out to be universal in the context of sublevel sets topology: the *interleaving distance*. In some sense, this distance measures the defect of isomorphism between persistence modules. In order to define it, we first introduce the *ε-shift functor* $-[\varepsilon]$, parametrized by $\varepsilon \geq 0$ and defined on objects by

$$\begin{aligned} M[\varepsilon](t) &:= M(t + \varepsilon\mathbb{1}), \\ M[\varepsilon](s \leq t) &:= M(s + \varepsilon\mathbb{1} \leq t + \varepsilon\mathbb{1}), \\ \text{where } \mathbb{1} &:= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n, \end{aligned}$$

and likewise on morphisms.

For any object M, there is a natural morphism $\eta_M^\varepsilon \colon M \to M[\varepsilon]$ given by $\eta_M^\varepsilon(t) := M(t \leq t + \varepsilon\mathbb{1})$. Furthermore, we have the identities $M[\varepsilon][\varepsilon'] = M[\varepsilon + \varepsilon']$ and $\eta^\varepsilon_{M[\varepsilon']} = \eta_M^\varepsilon[\varepsilon']$ for any $\varepsilon, \varepsilon' \geq 0$.

Definition 5.1. Two persistence modules $M, N \colon \mathbb{R}^n \to \mathsf{vect}_{\Bbbk}$ are *ε-interleaved* if there exist morphisms $f \colon M \to N[\varepsilon]$ and $g \colon N \to M[\varepsilon]$ such that the following diagram commutes.

$$\begin{array}{ccccc} M & \xrightarrow{\eta_M^\varepsilon} & M[\varepsilon] & \xrightarrow{\eta_M^\varepsilon[\varepsilon]} & M[2\varepsilon] \\ & \searrow^{f} \quad \nearrow_{g} & & \searrow^{f[\varepsilon]} \quad \nearrow_{g[\varepsilon]} & \\ N & \xrightarrow[\eta_N^\varepsilon]{} & N[\varepsilon] & \xrightarrow[\eta_N^\varepsilon[\varepsilon]]{} & N[2\varepsilon] \end{array} \tag{5.1}$$

Such a pair of morphisms (f, g) is called an *ε-interleaving* of the modules M, N. The *interleaving distance* $\mathrm{d_i}(M, N)$ is defined by

$$\mathrm{d_i}(M, N) := \inf\{\varepsilon \geq 0 \mid M, N \text{ are } \varepsilon\text{-interleaved}\}.$$

When $\varepsilon = 0$, diagram (5.1) collapses to the standard commutative diagram for isomorphism $M \simeq N$. In this sense, the interleaving distance measures the defect of isomorphism between M and N. It satisfies the triangle inequality, because interleavings compose to interleavings. It may take infinite values when no interleaving between the modules exists. Finally, while being isomorphic is equivalent to being 0-interleaved and therefore implies being at interleaving distance 0, the converse is not true: for instance, some modules (e.g., $\Bbbk_{\{t\}}$ for any singleton $\{t\} \subset \mathbb{R}^n$) are called *ephemeral modules* because they are at interleaving distance 0 from the zero module without being isomorphic to it. Thus, the interleaving distance is an extended pseudo-distance on isomorphism classes of persistence modules. It has been shown to be an extended distance on the quotient of the persistence module category by the full subcategory of ephemeral modules [7, 17].

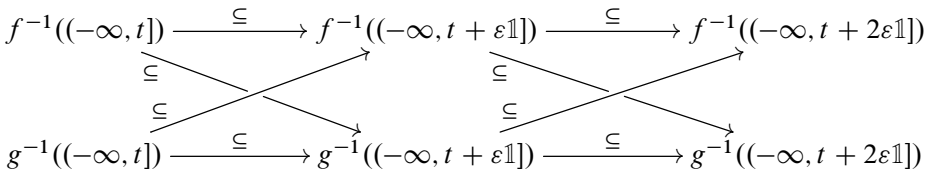
The next result justifies the use of the interleaving distance in the context of sublevel sets topology. It says indeed that $\mathrm{d_i}$ is a stable metric, in fact the most discriminative stable metric, in that context.

Proposition 5.2. *For any topological space X, functions $f, g\colon X \to \mathbb{R}^n$, and homology degree $*$,*

$$\mathrm{d_i}(H_*(f),\ H_*(g)) \leq \|f - g\|_\infty := \sup_{x \in X} \|f(x) - g(x)\|_\infty.$$

Moreover, assuming either that $n = 1$ or that $\Bbbk$ is a prime field, $\mathrm{d_i}$ is the maximum extended pseudo-distance on $\mathsf{vect}_\Bbbk^{\mathbb{R}^n}$ that satisfies the above inequality for any space X, functions f, g, and homology degree $*$.

The second part of the proposition is due to Lesnick [37]. The first part follows immediately from the commutativity of the diagram below for any $\varepsilon \geq \|f - g\|_\infty$ and any $t \in \mathbb{R}^n$, which, by functoriality of H_*, induces a commutative diagram like (5.1) at the singular homology level, hence an ε-interleaving between $H_*(f)$ and $H_*(g)$.



5.2. Bottleneck distance

Definition 5.1 does not provide any clear avenue for computing the interleaving distance. This is due to the size of the search space, which, up to shifts, is the product of the Hom-spaces between the two modules under consideration. The idea behind the bottleneck distance is to turn the computation into a combinatorial assignment problem, whose search space is much smaller. For this, it leverages the direct sum decompositions of the modules and restricts its focus to those interleavings that factor through these decompositions, considering partial matchings between the direct summands of the two modules under consideration.

In the following, M, N are two persistence modules over $\mathbb{R}^n$, and $M \simeq \bigoplus_{j\in J} M_J$ and $N \simeq \bigoplus_{l\in L} N_l$ are their direct sum decompositions as per Theorem 4.7.

Definition 5.3. A *partial matching* between the summands of M and N is a bijection $J' \to L'$ between some subsets $J' \subseteq J$ and $L' \subseteq L$. The modules M, N are *bottleneck ε-interleaved* if there is a partial matching $J \supseteq J' \xrightarrow{\gamma} L' \subseteq L$ such that

$$\begin{cases} \mathrm{d_i}(M_j,\, N_{\gamma(j)}) \leq \varepsilon & \forall\, j \in J', \\ \mathrm{d_i}(M_j,\, 0) \leq \varepsilon & \forall\, j \in J \setminus J', \\ \mathrm{d_i}(N_l,\, 0) \leq \varepsilon & \forall l \in L \setminus L'. \end{cases}$$

We call such a partial matching a *bottleneck ε-interleaving* of the modules M, N. The *bottleneck distance* $\mathrm{d_b}(M,\, N)$ is defined by

$$\mathrm{d_b}(M,\, N) := \inf\{\varepsilon \leq 0 \mid M, N \text{ are bottleneck } \varepsilon\text{-interleaved}\}.$$

From a bottleneck ε-interleaving of M, N, it is straightforward to build an ε-interleaving of M, N, by taking the direct sum of all the interleaving morphisms between summands of M and of N. By construction, this ε-interleaving factors through the direct sum decompositions of M and N in that it restricts to ε-interleavings between their matched summands on the one hand, between their unmatched summands and the zero module on the other hand. Hence the name *bottleneck ε-interleaving*, and the following immediate inequality:

$$\forall M, N : \mathbb{R}^n \to \mathsf{vect}_{\Bbbk}, \quad \mathrm{d_i}(M,\, N) \leq \mathrm{d_b}(M,\, N). \tag{5.2}$$

The converse inequality does not hold in general. Indeed, not every ε-interleaving of M, N factors through their direct sum decompositions, nor can it be turned into one that does without increasing ε. See, e.g., [4]. However, the converse inequality does hold in the 1-parameter setting, which yields the so-called *isometry theorem*.

Theorem 5.4 (Isometry [6,18,21,37]). *For all $M, N \in \mathsf{vect}_{\Bbbk}^{\mathbb{R}}$, $\mathrm{d_i}(M,\, N) = \mathrm{d_b}(M,\, N)$.*

From a practical standpoint, this result, combined with Theorem 4.3, implies that, for persistence modules over $\mathbb{R}$, computing the bottleneck distance is a way of computing the interleaving distance $\mathrm{d_i}$ exactly and efficiently, since it boils down to a mere assignment problem between interval modules, whose pairwise interleaving distances are easy to compute. For persistence modules over $\mathbb{R}^n$ with $n > 1$, by (5.2) the bottleneck distance still provides an upper bound on $\mathrm{d_i}$, and in the special case of interval decomposable modules, it is still computable in polynomial time [25]. There also exist non-trivial lower bounds on $\mathrm{d_i}$ that can be computed in polynomial time, for instance, the *matching distance* [33, 36]. However, computing the interleaving distance itself, whether exactly or within any factor less than 3, is known to be NP-hard over $\mathbb{R}^n$ when $n > 1$, even for interval decomposable persistence modules [8].

5.3. Signed bottleneck distance

Since the bottleneck distance $\mathrm{d_b}$ is computable and relevant on interval decomposable modules, and since signed barcodes are made of (pairs of multisets of) intervals, it is a natural idea to adapt $\mathrm{d_b}$ to them. Let $\mathcal{I}$ be a fixed set of intervals of $\mathbb{R}^n$. Given any persistence modules M, N over $\mathbb{R}^n$ that have well-defined signed barcodes $\mathrm{Bar}_{\mathcal{I}}(M)$ and $\mathrm{Bar}_{\mathcal{I}}(N)$ relative to $\mathcal{I}$, we define their *signed bottleneck distance* as follows:

$$\widehat{\mathrm{d}}_{\mathrm{b}}^{\mathcal{I}}(M,\, N) := \mathrm{d_b}\big(\Bbbk_{\mathrm{Bar}_{\mathcal{I}}^{+}(M)} \oplus \Bbbk_{\mathrm{Bar}_{\mathcal{I}}^{-}(N)},\ \Bbbk_{\mathrm{Bar}_{\mathcal{I}}^{+}(N)} \oplus \Bbbk_{\mathrm{Bar}_{\mathcal{I}}^{-}(M)}\big), \tag{5.3}$$

where, for any multiset Bar of intervals of $\mathbb{R}^n$, $\Bbbk_{\mathrm{Bar}}$ denotes the interval decomposable module $\bigoplus_{I \in \mathrm{Bar}} \Bbbk_I$. The intuition behind this definition comes from geometric measure theory: if we view multisets of intervals as point measures on the space of intervals of $\mathbb{R}^n$, then $\mathrm{d_b}$ is the Wasserstein-∞ distance for a specific choice of cost function on transport plans, and (5.3) is its signed version for signed measures. Note that the signed version does not satisfy the triangle inequality; however, it satisfies a certain universality property that makes it canonical among the dissimilarities on signed barcodes [13, 47].

Assume now that $\mathcal{I}$ contains all the principal upsets of R^n, and let $\mathcal{E}_{\mathcal{I}}$ be the class of short exact sequences on which the Hom functors $\mathrm{Hom}(\Bbbk_I, -)$ are exact for all $I \in \mathcal{I}$.

Theorem 5.5 ([47, Theorem 8.4]). *Assume that:*

(i) $\{t + \varepsilon\mathbb{1} \mid t \in I\} \in \mathcal{I}$ *for all* $I \in \mathcal{I}$ *and* $\varepsilon \geq 0$,

(ii) $m := \mathrm{gldim}^{\mathcal{E}_{\mathcal{I}}}\big(\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}\big) < \infty$,

(iii) *there is a constant* $c \geq 0$ *such that* $\mathrm{d_b}(\mathrm{Bar}(M), \mathrm{Bar}(N)) \leq c\, \mathrm{d_i}(M,\, N)$ *for all* $M, N \in \mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ *that are* $\mathcal{E}_{\mathcal{I}}$*-projective.*

Then, for all $M, N \in \mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$,

$$\widehat{\mathrm{d}}_{\mathrm{b}}^{\mathcal{I}}(M,\, N) \leq c(m+1)\, \mathrm{d}_{\mathrm{i}}(M, N).$$

Note that assumption (i) here differs from assumption (a) in [47, Theorem 8.4]: it is in fact a sufficient condition for (a) to hold in our context, and one that is more readily verifiable. For instance, it clearly holds when $\mathcal{I} = \{\text{principal upsets}\}$ or $\mathcal{I} = \{\text{hooks}\}$.

For $\mathcal{I} = \{\text{principal upsets}\}$, Hilbert's syzygy theorem provides (ii) with $m = n$, while Bjerkevik's stability theorem for free modules [4] provides (iii) with $c = n - 1$ (assuming $n > 1$), hence we have $\widehat{\mathrm{d}}_{\mathrm{b}}^{\mathcal{I}} \leq (n^2 - 1)\, \mathrm{d}_{\mathrm{i}}$ in this setting. For $\mathcal{I} = \{\text{hooks}\}$, Theorem 4.11 provides (ii) with $m = 2n - 2$, while an adaptation of Bjerkevik's stability proof to the case of hook modules (see [13]) provides (iii) with $c = 2n - 1$, hence we have $\widehat{\mathrm{d}}_{\mathrm{b}}^{\mathcal{I}} \leq (2n-1)^2\, \mathrm{d}_{\mathrm{i}}$ in this setting.

Example 5.6. Let $M, N : : \mathbb{R}^2 \to \mathsf{vect}_{\Bbbk}$ be the two interval modules whose supports and minimal projective resolutions relative to hooks are depicted in Figures 6 and 7, respectively. Their signed bottleneck distance, realized by the partial matching between their signed barcodes shown in Figure 8, is equal to their interleaving

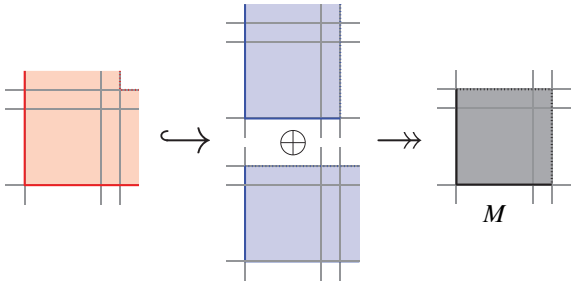


Figure 6. Minimal $\mathcal{E}_{\{\text{hooks}\}}$-projective resolution of the interval module M of Example 5.6.

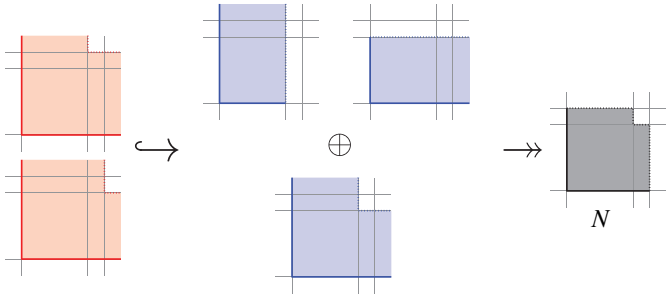


Figure 7. Minimal $\mathcal{E}_{\{\text{hooks}\}}$-projective resolution of the interval module N of Example 5.6.

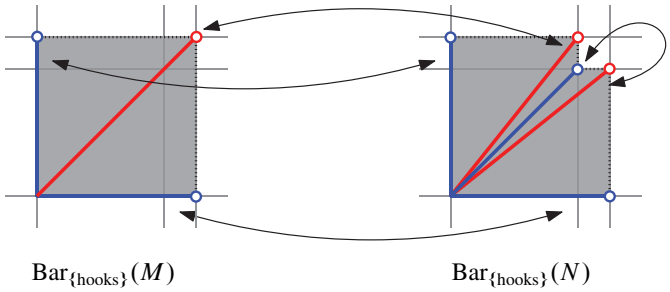


Figure 8. Partial matching realizing $\hat{\mathrm{d}}_{\mathrm{b}}^{\{\mathrm{hooks}\}}(M,\ N)$ in Example 5.6.

distance and so bounded from above by $(2 \times 2 - 1)^2 = 9$ times their interleaving distance.

6. Differential calculus

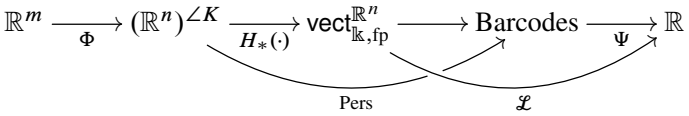
Let $\mathbb{R}^n$ be our poset, equipped with the product order, and let $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ be our persistence module category of interest. The finite presentedness assumption is motivated by the fact that, in TDA applications, persistence modules come from finite data. Note that $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ is an essentially small category, so it makes sense to talk about set maps from or to it, provided we consider its objects up to isomorphism – which we do from now on. Our goal is to characterize what it means for a map from or to $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ to be differentiable and to define derivatives for such maps so that the chain rule applies. The main obstruction is that the category does not come equipped with a differential structure on its object set, so we must define one.

6.1. Setup

With TDA applications in mind, we place ourselves in the context of sublevel sets topology, described in Section 3. Let X be a fixed topological space, which we assume to be the underlying space of some finite simplicial complex K. We consider functions $X \to \mathbb{R}^n$ that are defined simplexwise, meaning that they are the geometric realizations in X of functions $f = (f_1, \ldots, f_n)\colon K \to \mathbb{R}^n$. We further assume that every f considered is *monotonic* in the sense that each f_i is monotonic with respect to the face inclusion order on K, that is, $f_i(\sigma) \leq f_i(\tau)$ for all simplices $\sigma \subseteq \tau \in K$. These monotonic functions $K \to \mathbb{R}^n$ form a closed convex affine cone in $(\mathbb{R}^n)^{\#K}$, denoted by $(\mathbb{R}^n)^{\angle K}$, which is the intersection of the closed half-spaces $\{f = (f_1, \ldots, f_n)\colon K \to \mathbb{R}^n \mid f_i(\sigma) \leq f_i(\tau)\}$ for all pairs of simplices $\sigma \subseteq \tau \in K$ and all coordinates $i \in [\![1, n]\!]$. In the following, for the sake of simplicity of the exposition, we identify the functions $f \in (\mathbb{R}^n)^{\angle K}$ with their geometric realization in X.

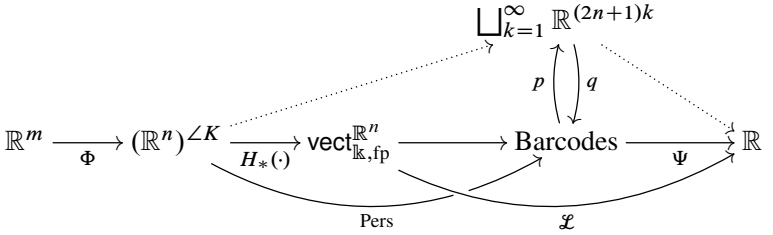
Every function $f \in (\mathbb{R}^n)^{\angle K}$ has a well-defined persistent (singular) homology $H_*(f)$ as a function $X \to \mathbb{R}^n$. Meanwhile, thanks to its monotonicity as a function $K \to \mathbb{R}^n$, f also has a well-defined persistent (simplicial) homology, which is isomorphic to $H_*(f)$. Since K is finite, $H_*(f)$ is finitely presented, and there is an upper bound, depending only on n and $\#K$, on the size of its minimal presentation.

In practice, the monotonic functions $K \to \mathbb{R}^n$ are built from data, so there is a map $\Phi\colon \mathbb{R}^m \to (\mathbb{R}^n)^{\angle K}$ from some parameter space that we assume to be the Euclidean space $\mathbb{R}^m$ (or some subanalytic subset): for instance, $\mathbb{R}^m$ can be the space of labeled point clouds with r points in $\mathbb{R}^d$ (in which case $m = dr$) or the parameter space of some neural network architecture. This setup is summarized in the first part of the following pipeline.



The second part of the pipeline is for maps from the persistence module category. Generally speaking, one may consider maps to any Euclidean space (or any subanalytic subset), but anticipating the next section on optimization, we take $\mathbb{R}$ as the codomain. In order to leverage the theory developed in the previous sections, we assume that the map $\mathscr{L}\colon \mathsf{vect}^{\mathbb{R}^n}_{\Bbbk,\mathrm{fp}} \to \mathbb{R}$ under consideration factors through some metric space of invariants. In our setting, it is either the space of finite barcodes equipped with the bottleneck distance (when $n = 1$) or the space of finite signed barcodes equipped with the signed bottleneck distance (when $n > 1$); in either case, we write Barcodes for the space of invariants. To complete the picture, we call *persistent homology map* and write Pers for the composite map $(\mathbb{R}^n)^{\angle K} \to \mathsf{vect}^{\mathbb{R}^n}_{\Bbbk,\mathrm{fp}} \to \mathrm{Barcodes}$, and we write Ψ for the map Barcodes $\to \mathbb{R}$ through which $\mathscr{L}$ factors.

As the space of barcodes does not come equipped with a differential structure, we lift it to a disjoint union of Euclidean spaces as illustrated below.



(6.1)

Specifically, we map each proper hook $[s,t)$ with $s < t \in \mathbb{R}^n$ to the vector $(s,t,1) \in \mathbb{R}^{2n+1}$ and each principal upset $[t,\infty)$ to the vector $(t,t,1) \in \mathbb{R}^{2n+1}$. Given a fixed total order on the bars (proper hooks and principal upsets) in $\mathbb{R}^n$, we now map each

barcode with k bars to the vector in $\mathbb{R}^{(2n+1)k}$ whose coordinates are given by the list of the coordinates of the images of its bars in $\mathbb{R}^{2n+1}$, sorted according to the total order on bars. In the following, we take the total order on bars induced by the lexicographic order on the coordinates of their images in $\mathbb{R}^{2n+1}$. For example, in the case $n = 1$, the barcode

$$\{[0, \infty),\ [-1, 3.2),\ [0, 2.5)\}$$

is sent to the vector

$$(-1, 3.2, 1, 0, 0, 1, 0, 2.5, 1) \in \mathbb{R}^9;$$

in the case $n = 2$, the barcode

$$\{[(0, 0), (3, 2)),\ [(-1, 1), (2, 2)),\ [(2, 3), \infty)\}$$

is sent to the vector

$$(-1, 1, 2, 2, 1, 0, 0, 3, 2, 1, 2, 3, 2, 3, 1) \in \mathbb{R}^{15}.$$

For signed barcodes $\mathrm{Bar} = (\mathrm{Bar}^+, \mathrm{Bar}^-)$, we map Bar^+ to $\mathbb{R}^{(2n+1)\#\mathrm{Bar}^+}$ as above, and we map Bar^- to $\mathbb{R}^{(2n+1)\#\mathrm{Bar}^-}$ by sending each of its proper hooks $[s, t)$ with $s < t \in \mathbb{R}^n$ to $(s, t, -1)$ and each of its principal upsets $[t, \infty)$ to $(t, t, -1)$, keeping the same total order on bars as before. In other words, the last coordinate of the image of each bar encodes its sign in the signed barcode. Then we concatenate the two vectors in $\mathbb{R}^{(2n+1)(\#\mathrm{Bar}^+ + \#\mathrm{Bar}^-)}$. For instance, in the case $n = 2$, the signed barcode

$$(\{[(0, 0)(0, 1)),\ [(0, 0), (1, 1))\},\ \{[(0, 0), (1, 0))\})$$

is sent to the vector

$$(0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, -1) \in \mathbb{R}^{15}.$$

Let us write $p\colon \mathrm{Barcodes} \to \bigsqcup_{k=1}^{\infty} \mathbb{R}^{(2n+1)k}$ for the resulting lifting map on barcodes and signed barcodes. This map clearly admits a left inverse q, since every (signed or regular) barcode can be reconstructed from its vectorization. Note that p is not onto, given the special form of the vectors obtained in $\mathbb{R}^{2n+1}$. Note also that p is not continuous either, because barcodes with different numbers of bars are sent to different terms of the disjoint union $\bigsqcup_{k=1}^{\infty} \mathbb{R}^{(2n+1)k}$. However, p is good enough for our purposes, as we will see in the next subsection.

As mentioned previously, there is a global upper bound, depending only on n and $\#K$, on the size of the minimal presentations of the persistent homology $H_*(f)$ of the monotonic maps $f\colon K \to \mathbb{R}^n$. As a consequence, there is also a global upper bound on the size of their (signed or regular) barcode. Thus, in effect, Barcodes is lifted to a finite disjoint union of Euclidean spaces, which, in the following, we view as a subanalytic subset of some real analytic manifold.

6.2. Results

Proposition 5.2, combined with Theorem 5.4 (case $n = 1$) or 5.5 (case $n > 1$), implies that the persistent homology map Pers is Lipschitz continuous. In practice, the maps Φ and Ψ are chosen by the user, and the most common choices are known to be locally Lipschitz continuous. Under this assumption, the composite map $\Psi \circ \text{Pers} \circ \Phi \colon \mathbb{R}^m \to \mathbb{R}$ is also locally Lipschitz continuous, and therefore differentiable almost everywhere in $\mathbb{R}^m$, by Rademacher's theorem [20, Theorem 3.4.1 and Remark 3.4.2]. In the following, we aim at getting some control over the singular set of the map in $\mathbb{R}^m$ and at deriving a chain rule to compute the map's derivatives at its regular points. For this, we assume that the maps Φ and $\Psi \circ q$ are subanalytic,[1] an assumption that is verified by most common choices of maps Φ and Ψ in practice. The key point is that $p \circ \text{Pers}$ has an elementary structure, described next.

Proposition 6.1 ([15, 40, 51]). *There is an affine Whitney stratification of* $(\mathbb{R}^n)^{\angle K}$ *such that the map* $p \circ \text{Pers}$ *is affine on each stratum.*

In effect, the stratification is given by the restriction, to the cone $(\mathbb{R}^n)^{\angle K}$, of the arrangement of hyperplanes $\{f = (f_1, \ldots, f_n) \colon K \to \mathbb{R}^n \mid f_i(\sigma) = f_i(\tau)\}$ for all pairs of simplices $\sigma \neq \tau \in K$ and all coordinates $i \in [\![1, n]\!]$. The intuition behind the result is that, for any monotonic functions $f, g \colon K \to \mathbb{R}^n$ belonging to the same stratum, the coordinate functions f_i and g_i induce the same pre-order on the simplices of K, so the persistent homologies $H_*(f)$ and $H_*(g)$ are isomorphic persistence modules up to a monotonic piecewise affine reparametrization of each coordinate axis in $\mathbb{R}^n$; as a consequence, the (signed or regular) barcodes of $H_*(f)$ and $H_*(g)$ are the same up to the reparametrization of the coordinates of their bars' endpoints.

Corollary 6.2. *Assume* Φ *and* $\Psi \circ q$ *are subanalytic. Then* $p \circ \text{Pers} \circ \Phi$ *and* $\Psi \circ \text{Pers} \circ \Phi$ *are subanalytic as well, and there is a finite* C^1 *Whitney stratification of* $\mathbb{R}^m$ *such that* $p \circ \text{Pers} \circ \Phi$ *and* $\Psi \circ \text{Pers} \circ \Phi$ *are differentiable on top-dimensional strata,* $\Psi \circ q$ *is differentiable on the images of these strata via* $p \circ \text{Pers} \circ \Phi$*, and the following chain rule applies:*

$$\nabla_x(\Psi \circ \text{Pers} \circ \Phi) = \nabla_{(p \circ \text{Pers} \circ \Phi)(x)}(\Psi \circ q)\, \mathbf{J}_x(p \circ \text{Pers} \circ \Phi).$$

That $p \circ \text{Pers} \circ \Phi$ is subanalytic follows from Proposition 6.1 and the assumption that Φ itself is subanalytic. Then the composition $\Psi \circ q \circ p \circ \text{Pers} \circ \Phi = \Psi \circ \text{Pers} \circ \Phi$

[1] Strictly speaking, we see $\Psi \circ q$ as a map from the finite disjoint union of Euclidean spaces in which barcodes coming from monotonic maps $K \to \mathbb{R}^n$ are lifted, and as mentioned previously, we see this finite disjoint union as a subanalytic subset of some real analytic manifold.

is also subanalytic. The existence of a finite C^1 Whitney stratification of $\mathbb{R}^m$ such that both $p \circ \mathrm{Pers} \circ \Phi$ and $\Psi \circ \mathrm{Pers} \circ \Phi$ are differentiable on the top-dimensional strata and that $\Psi \circ q$ is differentiable on the images of these strata via $p \circ \mathrm{Pers} \circ \Phi$ follows from the Whitney stratifiability of subanalytic sets and maps [28, 29]. The chain rule in top-dimensional strata is then the usual chain rule for differentiable maps between differentiable manifolds.

Remark 6.3. Corollary 6.2 extends to more general o-minimal structures for which Whitney stratifiability holds, such as for instance analytic-geometric categories [53].

6.3. Examples

We now explore two concrete examples of maps Φ (Example 6.4) and Ψ (Example 6.5) that are popular in applications. We provide explicit formulas for their derivatives, which can be combined via the chain rule of Corollary 6.2. For the sake of simplicity, our examples deal only with the 1-parameter version ($n = 1$) of pipeline (6.1). Other examples involving more parameters can be found in the literature; see, e.g., [46, 51].

Example 6.4 (Rips filtration). Let $r \in \mathbb{N}^*$ be fixed, and let K be the full simplex on $[\![1, r]\!]$, whose faces are in one-to-one correspondence with the power set $2^{[\![1,r]\!]} \setminus \{\emptyset\}$. As before, X is the underlying space of K. For any finite set $A = \{a_1, \ldots, a_r\}$ of labeled points in Euclidean space $\mathbb{R}^d$, we consider the geometric realization in X of the following monotonic function $f_A \colon K \to \mathbb{R}$:

$$\begin{cases} f_A(\{i_0\}) = 0 & \forall 1 \leq i_0 \leq r, \\ f_A(\{i_0, \ldots, i_s\}) = \max\limits_{0 \leq j,l \leq s} \|a_{i_j} - a_{i_l}\|_2 & \forall s \geq 1 \text{ and } \forall 1 \leq i_0 < \cdots < i_s \leq r. \end{cases}$$

This function is called the *Rips filtration* of the labeled point set A. It serves as a combinatorial approximation to the distance to A in the ambient space $\mathbb{R}^d$ (see, e.g., [48, Chapters 4–5] for the precise connection). Its parameters are the coordinates of the labeled points of A in $\mathbb{R}^d$. We consider the subspace of monotonic functions $f_A \colon K \to \mathbb{R}$ for A ranging over all the sets of r labeled points in $\mathbb{R}^d$ (with possible repetitions of the points). Referring to pipeline (6.1), we thus have $n = 1$, $m = dr$, and $\Phi \colon A \mapsto f_A$. This map Φ is known to be subanalytic [26], and it admits a C^1 Whitney stratification where each top-dimensional stratum is the locus of the labeled point clouds that have nonzero and distinct pairwise distances inducing the same total order on the pairs of labels, while the union of the lower-dimensional strata is composed of those labeled point clouds with at least one trivial or two equal pairwise distance(s). The derivatives of Φ at $A \in \mathbb{R}^{dr}$ in some top-dimensional stratum have

the following elementary expression: for any simplex $\sigma \in K$,

$$\left(\frac{\partial \Phi(A)}{\partial a_i}\right)_\sigma = \begin{cases} \dfrac{a_i - a_j}{\|a_i - a_j\|_2} & \text{if } \exists j \neq i \in \sigma \text{ s.t. } \|a_i - a_j\|_2 = \max\limits_{i',j' \in \sigma} \|a_{i'} - a_{j'}\|_2, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, even if A itself is in a top-dimensional stratum S_A, its image f_A may be (in fact, is generally) located in a stratum T_A of the affine Whitney stratification of $\mathbb{R}^{\angle K}$ that is not top dimensional. This is because the appearance of an edge in the Rips filtration is concomitant to the appearance of all its cofaces that have it as their longest edge. Therefore, the map $p \circ \mathrm{Pers}$ per se may not be differentiable at $f_A \in \mathbb{R}^{\angle K}$.

The composite map $p \circ \mathrm{Pers} \circ \Phi$, however, is differentiable at A. More precisely, the pairwise distances in each point cloud $A' \in S_A$ yield the same total order on the pairs of indices $1 \leq i < j \leq r$. In turn, this total order yields a unique pre-order $\preceq_A$ on the simplices of K by the (increasing) values of f_A; in other words, the point clouds in S_A are mapped to the stratum T_A. Then fixing a priori a total order $\leq_K$ on the simplices of K, the point clouds in S_A induce the same total lexicographic order $\leq_A := (\preceq_A, \leq_K)$. In turn, $\leq_A$ yields a unique assignment of the simplices of K to the finite endpoints of the bars in the barcode of $H_*(f_{A'})$ for any $A' \in S_A$, hence a unique assignment ζ of the simplices of K to coordinates in the lifting space $\mathbb{R}^{3k}$, where k is the number of bars in the barcode of $H_*(f_A)$. The derivatives of $p \circ \mathrm{Pers} \circ \Phi$ at A then have the following simple expression: for any coordinate $1 \leq l \leq 3k$,

$$\left(\frac{\partial (p \circ \mathrm{Pers} \circ \Phi)(A)}{\partial a_i}\right)_l = \begin{cases} \dfrac{a_i - a_j}{\|a_i - a_j\|_2} & \text{if } l = \zeta(\sigma) \text{ for some } \sigma \in K \text{ and } \exists j \neq i \in \sigma \\ & \text{s.t. } \|a_i - a_j\|_2 = \max\limits_{i',j' \in \sigma} \|a_{i'} - a_{j'}\|_2, \\ \left(\dfrac{\partial (p \circ \mathrm{Pers} \circ \Phi)(A)}{\partial a_i}\right)_{l-1} & \text{if } l \equiv 2 \bmod 3 \text{ and } (p \circ \mathrm{Pers} \circ \Phi)(A)_l = \\ & (p \circ \mathrm{Pers} \circ \Phi)(A)_{l-1} \text{ (bar endpoint at infinity),} \\ 0 & \text{otherwise.} \end{cases}$$

Recall from our setup (Section 6.1) that p sends infinite bars $[t, \infty)$ to vectors $(t, t, 1)$, hence the second case in the above expression for the derivatives of $p \circ \mathrm{Pers} \circ \Phi$.

Example 6.5 (Total persistence). For $n = 1$, let Ψ: Barcodes $\to \mathbb{R}$ map each finite barcode to the sum of the lengths of its finite bars, also known as its *total persistence*:

$$\forall \mathrm{Bar} \in \mathrm{Barcodes}, \quad \Psi(\mathrm{Bar}) := \sum_{\substack{I \in \mathrm{Bar} \\ \sup I < \infty}} (\sup I - \inf I).$$

This sum is finite and can be rearranged as follows:

$$\forall \text{Bar} \in \text{Barcodes}, \quad \Psi(\text{Bar}) = \sum_{\substack{I \in \text{Bar} \\ \sup I < \infty}} \sup I - \sum_{\substack{I \in \text{Bar} \\ \sup I < \infty}} \inf I.$$

Pre-composing Ψ by q gives the following linear map $\mathbb{R}^{3k} \to \mathbb{R}$ for every barcode size k:

$$x \mapsto \sum_{i=0}^{k-1} x_{3i+2} - \sum_{i=0}^{k-1} x_{3i+1}.$$

One can indeed check that $\Psi \circ q$ as defined above yields $\Psi \circ q \circ p = \Psi$, the crux of the matter being that the contribution of every infinite bar I to $(\Psi \circ q \circ p)(\text{Bar})$ is zero as desired because I is mapped to $(\inf I, \inf I, 1)$ by p then to $\inf I - \inf I = 0$ by $\Psi \circ q$.

The derivatives of $\Psi \circ q$ are then constant for every barcode size k:

$$\frac{\partial(\Psi \circ q)}{\partial x_l} = \begin{cases} -1 & \text{if } l \equiv 1 \bmod 3, \\ 1 & \text{if } l \equiv 2 \bmod 3, \\ 0 & \text{otherwise.} \end{cases}$$

7. Optimization

Referring again to pipeline (6.1), we assume $\mathscr{L}$ is some loss function on persistence modules that one seeks to minimize over the choice of parameter vector $x \in \mathbb{R}^m$ from which the modules originate. We also assume the maps Φ and $\Psi \circ q$ have been chosen to be subanalytic and locally Lipschitz continuous – as the most common choices in practice are. Then, by Corollary 6.2, the composite map $F := \Psi \circ \text{Pers} \circ \Phi$ itself is subanalytic, so it is differentiable almost everywhere in $\mathbb{R}^m$ and has a well-defined Clarke subdifferential $\partial F(x)$ at every point $x \in \mathbb{R}^m$:

$$\partial F(x) := \text{CH}\Big\{ \lim_{i \to \infty} \nabla F(y_i) \mid (y_i)_{i \in \mathbb{N}} \to x \text{ and } F \text{ differentiable at every } y_i \Big\},$$

where CH is the notation for the convex hull in $\mathbb{R}^m$. We consider the following differential inclusion:

$$\frac{dx}{dt} \in -\partial F(x(t)) \quad \text{for almost every } t \geq 0,$$

whose solutions are the trajectories of the subgradient of F. In order to approximate these trajectories, we use the stochastic subgradient descent algorithm, whose descent step is written as follows:

$$x_{i+1} := x_i - \alpha_i(g_i + \xi_i),$$

where $(x_i)_{i\in\mathbb{N}}$ is the sequence of iterates, $(g_i \in \partial F(x_i))_{i\in\mathbb{N}}$ is any sequence of subgradients of F at the iterates, $(\alpha_i)_{i\in\mathbb{N}}$ is a sequence of non-negative learning rates, and $(\xi_i)_{i\in\mathbb{N}}$ is a sequence of random variables (the stochastic part of the descent). We make the following standard assumptions on these quantities – see [23, Assumption C]:

(i) $\sum_{i\in\mathbb{N}} \alpha_i = +\infty$, while $\sum_{i\in\mathbb{N}} \alpha_i^2 < +\infty$,

(ii) $\sup_{i\in\mathbb{N}} \|x_i\|_2 < +\infty$ almost surely,

(iii) almost surely, conditioned on the past, each ξ_i has zero mean, and its second moment is bounded from above by $\mu(x_i)$ for some fixed function $\mu\colon \mathbb{R}^m \to \mathbb{R}_+$ that is bounded on bounded sets.

Under these assumptions, one can show that stochastic subgradient descent converges almost surely to critical points of F.

Theorem 7.1 ([23, Corollary 5.9]). *Given that F is subanalytic and locally Lipschitz continuous, under assumptions (i)–(iii), almost surely every limit point x of the sequence of iterates $(x_i)_{i\in\mathbb{N}}$ is critical for F (i.e., $0 \in \partial F(x)$), and the sequence of function values $(F(x_i))_{i\in\mathbb{N}}$ converges.*

We now give two illustrative examples of applications. For the sake of simplicity, we show basic applications that rely on the choices of maps Φ and Ψ from Examples 6.4 and 6.5. In particular, the persistence modules involved in our examples have only one parameter. Other, more elaborate applications involving persistence modules with more parameters and/or relying on different choices of maps Φ and Ψ have been considered in the literature, notably in deep learning contexts; see [15, 46, 51] for some examples.

Example 7.2 (Topological optimization). Let A be a finite set of samples drawn independently from the uniform distribution on the square $[-1, 1]^2 \subset \mathbb{R}^2$, as illustrated in Figure 9 (left). The goal is to move the points around in order to create as large 'holes' as possible in the distribution. For this, we use the setup of Examples 6.4 and 6.5, letting $d = 2$, $n = 1$, $m = 2\#A$, Φ be the Rips filtration for labeled point clouds in $\mathbb{R}^2$, and Ψ be the opposite of the total persistence. Since we are interested in creating 1-dimensional holes, we consider homology in degree 1, so in the following, Pers stands for persistent homology in degree 1. Our approach is to minimize the functional $\Psi \circ \mathrm{Pers} \circ \Phi$, that is, to maximize the total persistence in homology degree 1 of the Rips filtration of A – recall that the Rips filtration of A serves as a combinatorial approximation of the distance to A in $\mathbb{R}^2$, as mentioned in Example 6.4.

Applying stochastic subgradient descent with standard parameter values to the functional $\Psi \circ \mathrm{Pers} \circ \Phi$, starting from A, gives the result shown in Figure 9 (center) after 100 steps. As expected, 1-dimensional holes have started appearing in the

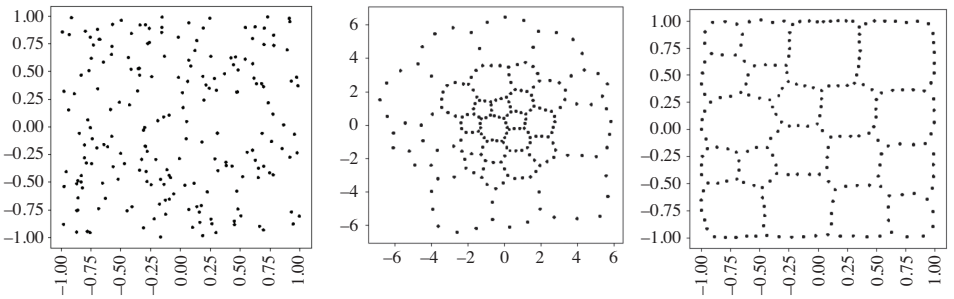


Figure 9. Left: The input point cloud A. Center: The output after 100 steps of stochastic subgradient descent on the functional $\Psi \circ \mathrm{Pers} \circ \Phi$, starting from A, where persistent homology is taken in degree 1. Right: Output after 100 steps of stochastic subgradient descent on the regularized functional $(\Psi \circ \mathrm{Pers} \circ \Phi)(A) + \lambda \sum_{a \in A} \max\{0,\ \|a\|_\infty - 1\}$.

distribution. However, the functional does not converge, as the optimization process keeps dispersing the points indefinitely. This is because scaling the data points by a factor $\nu > 0$ results in scaling their pairwise distances by ν, and therefore scaling the functional by ν as well. And indeed, our initial objective of creating as large holes as possible in the distribution without any further requirements is ill-defined.

We therefore modify our objective by adding the constraint that the points of A should remain in the square $[-1, 1]^2$. We treat this as a soft constraint and add the following regularization term to our functional, where λ is a user-defined mixing parameter:

$$\lambda \sum_{a \in A} \max\{0,\ \|a\|_\infty - 1\}.$$

Stochastic subgradient descent applied to the regularized functional eventually converges, thanks to the regularization term, and the output, depicted in Figure 9 (right), qualitatively meets our new objective.

Example 7.3 (Topological regularization). Let A be a finite set of data points in Euclidean space $\mathbb{R}^3$, sampled from two entangled thickened circles, as illustrated in Figure 10 (left). The goal is to project the data down to 2 dimensions, while preserving the geometry and underlying topological structure of the data as much as possible.

In order to do this, we use an autoencoder architecture, which implements the composition of two maps:

- a projection map $\pi_\theta \colon \mathbb{R}^3 \to \mathbb{R}^2$, called the *encoder*, parametrized by some vector $\theta \in \Theta$,
- an embedding map $\iota_\omega \colon \mathbb{R}^2 \to \mathbb{R}^3$, called the *decoder*, parametrized by some vector $\omega \in \Omega$.

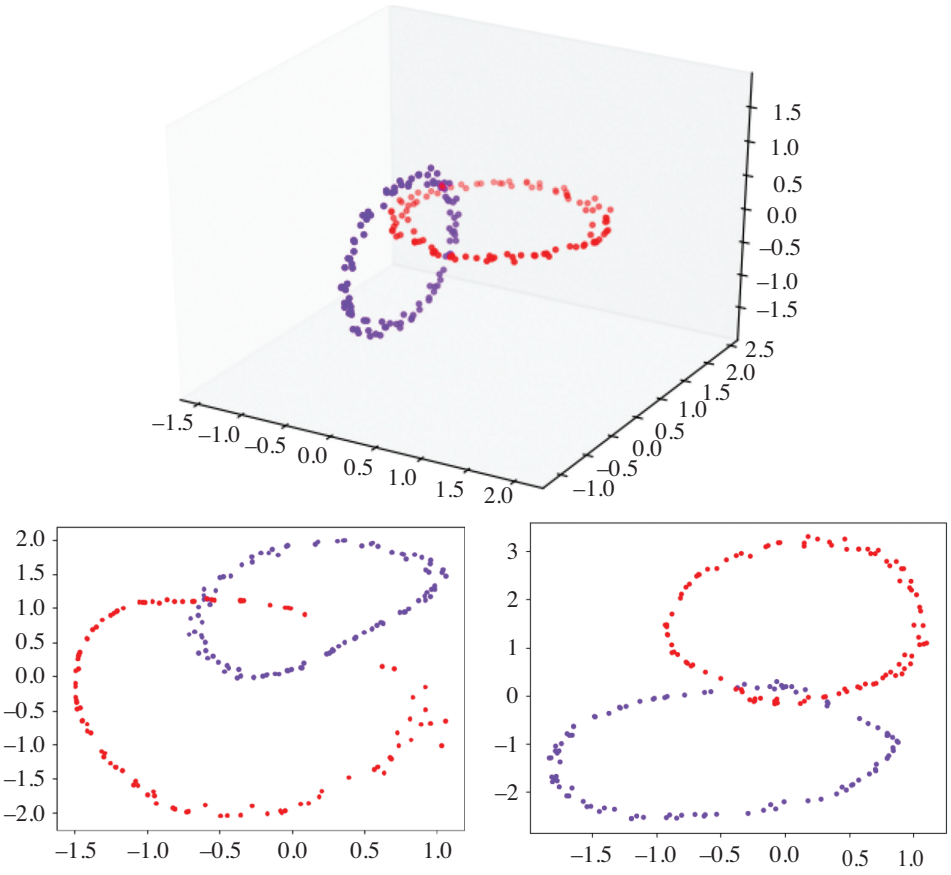


Figure 10. (From [15]) Left: The input point cloud A in 3 dimensions, with data coming from the two thickened circles marked in different colors. Center: Projection of A to 2 dimensions by the autoencoder trained for 100 epochs with the standard functional (7.1). Right: Projection of A to 2 dimensions by the autoencoder trained for 100 epochs with the regularized functional.

The idea is then to make the composite map $\iota_\omega \circ \pi_\theta : \mathbb{R}^3 \to \mathbb{R}^3$ as close to the identity of $\mathbb{R}^3$ as possible. For this, we set the parameters θ, ω in such a way as to minimize the quadratic error between the points of A and their image through $\iota_\omega \circ \pi_\theta$:

$$\min_{\theta,\omega} \sum_{a \in A} \|a - (\iota_\omega \circ \pi_\theta)(a)\|_2^2. \tag{7.1}$$

The resulting projection of the point cloud A to 2 dimensions after this training phase is shown in Figure 10 (center). Besides the unavoidable metric distortion of the data, we see a break in the topological structure of its underlying two circles. In order to repair it, we add the following regularization term to our functional, which measures

the discrepancy between the persistent homology in degree 1 of the Rips filtration of A and that of its projection $\pi_\theta(A)$:

$$\lambda\, \mathrm{d}_1((\mathrm{Pers} \circ \Phi_3)(A),\ (\mathrm{Pers} \circ \Phi_2)(\pi_\theta(A))),$$

where $\lambda > 0$ is a user-defined mixing parameter, Φ_d is the Rips filtration for labeled point clouds in $\mathbb{R}^d$, Pers is the persistent homology map for homology in degree 1, and d_1 is a variant of the bottleneck distance between persistence modules (Definition 5.3) where the maximum of the interleaving distances between their interval summands is replaced by the sum of these interleaving distances. As can be seen from Figure 10 (right), adding this regularization term to the functional helps the trained projection map π_θ better preserve the topological structure of the two circles underlying the data.

8. Perspectives

Many of the results presented in this text are preliminary and call for further investigations in order to consolidate and extend the parts of the theory of persistence modules they contribute to. Here, we report on a few concrete directions for future developments related to the differentiation, optimization, and application of the TDA pipeline.

Theorem 7.1 provides convergence guarantees for stochastic subgradient descent on functionals that factor through the persistence module category $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$, but the result says nothing about the convergence rate. Theoretical bounds on the convergence rate were obtained in [39] for the case $n = 1$, using a stratified variant of gradient sampling. Whether similar bounds can be obtained when $n > 1$ remains open.

Strictly speaking, the framework described in Sections 6 and 7 does not equip the persistence module category $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ itself with a differential structure but rather takes a detour through a certain lift of the space of (signed) barcodes. An important follow-up question is whether the differential structure on the lifting space can be pushed down to the space of (signed) barcodes, and from there, to the persistence module category $\mathsf{vect}_{\Bbbk,\mathrm{fp}}^{\mathbb{R}^n}$ itself. This was shown in [40] to be possible in the case $n = 1$, using the theory of diffeological spaces [32] and leveraging the Lipschitz continuity of the left inverse q to the lifting map p as well as the isometry theorem, Theorem 5.4. Whether this is also possible when $n > 1$ remains open, and the main bottleneck is that, in this setting, (signed) barcodes are no longer a complete algebraic nor metric invariant of persistence modules.

The question of computing (signed) barcodes efficiently is central for applications. While it is possible to use standard computer algebra systems like GAP [52] for

this purpose, in terms of algorithmic complexity, these libraries do not scale up well with the size of the simplicial complexes on which filtrations are built – these simplicial complexes may contain millions of simplices in practice. Dedicated software libraries have been developed over the years by the TDA community, which include highly efficient routines to compute barcodes in the 1-parameter setting [5, 31, 42]. In the multi-parameter setting, effective algorithms to compute minimal projective resolutions of 2-parameter persistence modules were recently proposed [34, 38], opening an avenue for the efficient computation of their signed barcodes relative to principal upsets. The computation can be made even more efficient in the special case of persistence modules induced in homology of degree 0 [45]. However, computing minimal projective resolutions relative to other classes of projectives remains largely open despite some promising developments [16].

Barcodes and their signed version are but one of many existing invariants for persistence modules. Indeed, given that barcodes are not a complete invariant in the multi-parameter setting, a wealth of other invariants have been proposed, among which are the generalized persistence diagrams [35], elder-rule staircodes [14], multi-persistence module approximations [41], multigraded Hilbert series [30], generalized rank invariant landscapes (GRIL) [54], meta-diagrams [19], or birth-death functions [43], to name a few. While some of them like GRIL are provably stable and differentiable [46, 54], many still have unknown theoretical behavior in terms of stability and differentiability. In order to prevent the parallel development of multiple specialized frameworks for differential calculus and optimization with each one of these invariants, there is a need for the emergence of a unified theory that would encompass most of them. A first attempt in this direction was made recently [51], encouraging further investigation.

Acknowledgments. Part of this exposition is based on the author's own work with his collaborators, whom he wishes to thank for their invaluable contributions. He also wishes to thank the anonymous referee for their useful feedback.

References

[1] C. Amiot, T. Brüstle, and E. J. Hanson, Invariants of persistence modules defined by order-embeddings. [v1] 2024, [v2] 2025, arXiv:2402.09190v2

[2] H. Asashiba, E. G. Escolar, K. Nakashima, and M. Yoshiwaki, Approximation by interval-decomposables and interval resolutions of persistence modules. *J. Pure Appl. Algebra* **227** (2023), no. 10, article no. 107397, 20 pp. Zbl 1545.16014 MR 4576885

[3] M. Auslander and Ø. Solberg, Relative homology and representation theory 1. Relative homology and homologically finite subcategories. *Comm. Algebra* **21** (1993), no. 9, 2995–3031 Zbl 0792.16017 MR 1228751

[4] H. Bakke Bjerkevik, On the stability of interval decomposable persistence modules. *Discrete Comput. Geom.* **66** (2021), no. 1, 92–121 Zbl 1471.55006 MR 4270636

[5] U. Bauer, Ripser: efficient computation of Vietoris–Rips persistence barcodes. *J. Appl. Comput. Topol.* **5** (2021), no. 3, 391–423 Zbl 1476.55012 MR 4298669

[6] U. Bauer and M. Lesnick, Induced matchings and the algebraic stability of persistence barcodes. *J. Comput. Geom.* **6** (2015), no. 2, 162–191 Zbl 1405.68398 MR 3333456

[7] N. Berkouk and F. Petit, Ephemeral persistence modules and distance comparison. *Algebr. Geom. Topol.* **21** (2021), no. 1, 247–277 Zbl 1472.55003 MR 4224741

[8] H. B. Bjerkevik, M. B. Botnan, and M. Kerber, Computing the interleaving distance is NP-hard. *Found. Comput. Math.* **20** (2020), no. 5, 1237–1271 Zbl 1460.55006 MR 4156997

[9] B. Blanchette, T. Brüstle, and E. J. Hanson, Homological approximations in persistence theory. *Canad. J. Math.* **76** (2024), no. 1, 66–103 Zbl 1551.55004 MR 4687766

[10] B. Blanchette, T. Brüstle, and E. J. Hanson, Exact structures for persistence modules. In *Representations of algebras and related topics*, pp. 121–160, EMS Ser. Congr. Rep., EMS Press, Berlin, 2025 MR 4993173

[11] M. B. Botnan and W. Crawley-Boevey, Decomposition of persistence modules. *Proc. Amer. Math. Soc.* **148** (2020), no. 11, 4581–4596 Zbl 1461.16013 MR 4143378

[12] M. B. Botnan, S. Oppermann, and S. Oudot, Signed barcodes for multi-parameter persistence via rank decompositions and rank-exact resolutions. *Found. Comput. Math.* **25** (2025), no. 5, 1815–1874 Zbl 08127335 MR 4983902

[13] M. B. Botnan, S. Oppermann, S. Oudot, and L. Scoccola, On the bottleneck stability of rank decompositions of multi-parameter persistence modules. *Adv. Math.* **451** (2024), article no. 109780, 53 pp. Zbl 1552.55006 MR 4761951

[14] C. Cai, W. Kim, F. Mémoli, and Y. Wang, Elder-rule-staircodes for augmented metric spaces. *SIAM J. Appl. Algebra Geom.* **5** (2021), no. 3, 417–454 Zbl 1477.55005 MR 4297819

[15] M. Carriere, F. Chazal, M. Glisse, Y. Ike, H. Kannan, and Y. Umeda, Optimizing persistent homology based functions. In M. Meila and T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, pp. 1294–1303, Proc. Mach. Learn. Res. (PMLR) 139, PMLR, 2021

[16] W. Chachólski, A. Guidolin, I. Ren, M. Scolamiero, and F. Tombari, Koszul complexes and relative homological algebra of functors over posets. *Found. Comput. Math.* **25** (2025), no. 4, 1121–1165 Zbl 08087069 MR 4951524

[17] F. Chazal, W. Crawley-Boevey, and V. de Silva, The observable structure of persistence modules. *Homology Homotopy Appl.* **18** (2016), no. 2, 247–265 Zbl 1377.55015 MR 3575998

[18] F. Chazal, V. de Silva, M. Glisse, and S. Oudot, *The structure and stability of persistence modules*. SpringerBriefs Math., Springer, Cham, 2016, 120 pp. Zbl 1362.55002 MR 3524869

[19] N. Clause, T. K. Dey, F. Mémoli, and B. Wang, Meta-diagrams for 2-parameter persistence. *Discrete Comput. Geom.* **74** (2025), no. 4, 872–898 Zbl 08141744 MR 4993963

[20] Ş. Cobzaş, R. Miculescu, and A. Nicolae, *Lipschitz functions*. Lecture Notes in Math. 2241, Springer, Cham, 2019, 591 pp. Zbl 1431.26002 MR 3931701

[21] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, Stability of persistence diagrams. *Discrete Comput. Geom.* **37** (2007), no. 1, 103–120 Zbl 1117.54027 MR 2279866

[22] W. Crawley-Boevey, Decomposition of pointwise finite-dimensional persistence modules. *J. Algebra Appl.* **14** (2015), no. 5, article no. 1550066, 8 pp. Zbl 1345.16015 MR 3323327

[23] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* **20** (2020), no. 1, 119–154 Zbl 1433.65141 MR 4056927

[24] T. K. Dey and Y. Wang, *Computational topology for data analysis*. Cambridge University Press, Cambridge, 2022, 434 pp. Zbl 1498.57001 MR 4381505

[25] T. K. Dey and C. Xin, Computing bottleneck distance for multi-parameter interval decomposable persistence modules. [v1] 2018, [v3] 2019, arXiv:1803.02869v3

[26] V. Divol and F. Chazal, The density of expected persistence diagrams and its kernel based estimation. *J. Comput. Geom.* **10** (2020), no. 2, 127–153 Zbl 1473.55003 MR 4088069

[27] P. Gabriel, Unzerlegbare Darstellungen I. *Manuscripta Math.* **6** (1972), no. 1, 71–103 Zbl 0232.08001 MR 0332887

[28] R. M. Hardt, Stratification of real analytic mappings and images. *Invent. Math.* **28** (1975), no. 2, 193–208 Zbl 0298.32003 MR 0372237

[29] R. M. Hardt, Topological properties of subanalytic sets. *Trans. Amer. Math. Soc.* **211** (1975), 57–70 Zbl 0303.32008 MR 0379882

[30] H. A. Harrington, N. Otter, H. Schenck, and U. Tillmann, Stratifying multiparameter persistent homology. *SIAM J. Appl. Algebra Geom.* **3** (2019), no. 3, 439–471 Zbl 1450.55002 MR 4000203

[31] A. Hylton, G. Henselman-Petrusek, J. Sang, and R. Short, Tuning the performance of a computational persistent homology package. *Softw. Pract. Exp.* **49** (2019), no. 5, 885–905

[32] P. Iglesias-Zemmour, *Diffeology*. Math. Surveys Monogr. 185, American Mathematical Society, Providence, RI, 2013, 439 pp. Zbl 1269.53003 MR 3025051

[33] M. Kerber, M. Lesnick, and S. Oudot, Exact computation of the matching distance on 2-parameter persistence modules. *J. Comput. Geom.* **11** (2020), no. 2, 4–25 Zbl 1476.55016 MR 4200191

[34] M. Kerber and A. Rolle, Fast minimal presentations of bi-graded persistence modules. In *Proceedings of the Symposium on Algorithm Engineering and Experiments (Alenex)*, pp. 207–220, Society for Industrial and Applied Mathematics, Philadelphia, 2021 Zbl 07302448

[35] W. Kim and F. Mémoli, Generalized persistence diagrams for persistence modules over posets. *J. Appl. Comput. Topol.* **5** (2021), no. 4, 533–581 Zbl 1500.55004 MR 4334502

[36] C. Landi, The rank invariant stability via interleavings. In E. Chambers, B. Fasy, and L. Ziegelmeier (Eds.), *Research in computational topology*, pp. 1–10, Assoc. Women Math. Ser. 13, Springer, Cham, 2018 Zbl 1432.55012 MR 3904998

[37] M. Lesnick, The theory of the interleaving distance on multidimensional persistence modules. *Found. Comput. Math.* **15** (2015), no. 3, 613–650 Zbl 1335.55006 MR 3348168

[38] M. Lesnick and M. Wright, Computing minimal presentations and bigraded Betti numbers of 2-parameter persistent homology. *SIAM J. Appl. Algebra Geom.* **6** (2022), no. 2, 267–298 Zbl 1498.55006 MR 4429408

[39] J. Leygonie, M. Carrière, T. Lacombe, and S. Oudot, A gradient sampling algorithm for stratified maps with applications to topological data analysis. *Math. Program.* **202** (2023), no. 1–2, 199–239 Zbl 1525.49015 MR 4653450

[40] J. Leygonie, S. Oudot, and U. Tillmann, A framework for differential calculus on persistence barcodes. *Found. Comput. Math.* **22** (2022), no. 4, 1069–1131 Zbl 1507.55008 MR 4462407

[41] D. Loiseaux, M. Carrière, and A. J. Blumberg, Multi-parameter Module Approximation: an efficient and interpretable invariant for multi-parameter persistence modules with guarantees. *J. Appl. Comput. Topol.* **9** (2025), no. 4, article no. 26, 60 pp. Zbl 08130858 MR 4977351

[42] C. Maria, J.-D. Boissonnat, M. Glisse, and M. Yvinec, The Gudhi library: simplicial complexes and persistent homology. In H. Hong and C. Yap (Eds.), *Proceedings of the 4th International Congress on Mathematical Software–ICMS 2014, Seoul, South Korea, August 5–9, 2014*, pp. 167–174, Springer, Berlin, 2014 Zbl 1402.57001

[43] A. McCleary and A. Patel, Edit distance and persistence diagrams over lattices. *SIAM J. Appl. Algebra Geom.* **6** (2022), no. 2, 134–155 Zbl 1547.55005 MR 4405184

[44] E. Miller, Homological algebra of modules over posets. *SIAM J. Appl. Algebra Geom.* **9** (2025), no. 3, 483–524 Zbl 08087502 MR 4941899

[45] D. Morozov and L. Scoccola, Computing Betti tables and minimal presentations of zero-dimensional persistent homology. In *41st International Symposium on Computational Geometry*, pp. 69:1–69:15, LIPIcs. Leibniz Int. Proc. Inform. 332, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2025 MR 4934422

[46] S. Mukherjee, S. N. Samaga, C. Xin, S. Oudot, and T. K. Dey, D-GRIL: end-to-end topological learning with 2-parameter persistence. *In Proceedings of the 42nd International Symposium on Computational Geometry* (SoCG), 2026

[47] S. Oudot and L. Scoccola, On the stability of multigraded Betti numbers and Hilbert functions. *SIAM J. Appl. Algebra Geom.* **8** (2024), no. 1, 54–88 Zbl 1536.55007 MR 4695718

[48] S. Y. Oudot, *Persistence theory: from quiver representations to data analysis*. Math. Surveys Monogr. 209, American Mathematical Society, Providence, RI, 2015, 218 pp. Zbl 1335.55001 MR 3408277

[49] L. Polterovich, D. Rosen, K. Samvelyan, and J. Zhang, *Topological persistence in geometry and analysis*. Univ. Lecture Ser. 74, American Mathematical Society, Providence, RI, 2020, 128 pp. Zbl 1464.55003 MR 4249570

[50] H. Schenck, *Algebraic foundations for applied topology and data analysis*. Math. Data 1, Springer, Cham, 2022, 224 pp. Zbl 1521.55001 MR 4559594

[51] L. Scoccola, S. Setlur, D. Loiseaux, M. Carrière, and S. Oudot, Differentiability and optimization of multiparameter persistent homology. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 43986–44011, Proc. Mach. Learn. Res. (PMLR) 235, PMLR, 2024

[52] The GAP Group, GAP – Groups, Algorithms, and Programming, version 4.13.1. 2024, https://www.gap-system.org visited on 18 March 2026

[53] L. van den Dries and C. Miller, Geometric categories and o-minimal structures. *Duke Math. J.* **84** (1996), no. 2, 497–540 Zbl 0889.03025 MR 1404337

[54] C. Xin, S. Mukherjee, S. N. Samaga, and T. K. Dey, GRIL: A 2-parameter persistence based vectorization for machine learning. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 313–333, Proc. Mach. Learn. Res. (PMLR), 2023

[55] A. Zomorodian and G. Carlsson, Computing persistent homology. *Discrete Comput. Geom.* **33** (2005), no. 2, 249–274 Zbl 1069.55003 MR 2121296

Steve Oudot
Inria and École polytechnique, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France;
steve.oudot@inria.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/28

Spectral gaps in quantum computing

David Pérez-García

Abstract. How does the spectral gap of a quantum many-body system scale with the system size? This turns out to be a very relevant question both in condensed matter physics and in quantum computing. However, there are very few available techniques to give bounds on that scaling. In the first part of this paper, we will review recent results which explain why it is undecidable to know if the spectral gap will vanish or not with the system size. In the second part, we will review complementary results which show that, despite this general impossibility result, one can still prove spectral gap estimates in a rather wide family of systems. As a consequence, one obtains new bounds on the lifetime of quantum memories.

1. Introduction

In this paper, we will review some recent results obtained in collaboration with different authors (J. Bausch, T. S. Cubitt, M. Kastoryano, A. Lucia, A. Pérez-Hernández and M. M. Wolf) regarding spectral gaps in lattice quantum spin systems [6, 15, 16, 22, 28, 38]. These results are steps in the larger program of understanding spectral gaps in interacting quantum systems. Both by the techniques and motivations, the focus will be located in the interconnection of the spectral gap problem with quantum computing.

The mathematical setup is the following. We consider "quantum particles" in a finite-size lattice, which for simplicity in the presentation we assume to be the square lattice in one or two dimensions. That is, we will consider the lattice $\Lambda_L = (\mathbb{Z} \cap [1, L])^\eta$ with $\eta = 1, 2$. We will usually consider periodic boundary conditions (i.e., a circle or a torus), though sometimes we will need to assume open boundary conditions.

We will be interested in the case $L \to \infty$, but we will study only the case of finite L and see how the relevant quantities behave asymptotically when $L \to \infty$. Note that in the mathematical physics literature, one usually considers directly the

Mathematics Subject Classification 2020: 81P73 (primary); 81P68 (secondary).
Keywords: quantum computing, quantum many-body systems, quantum memories, undecidability.

thermodynamic limit $L = \infty$ using operator algebraic techniques [31]. We will however consider the quantum information perspective of just looking at the asymptotic behavior of finite-size systems.

To make this presentation accessible to non-experts in quantum theory, we will start by quickly reviewing the main objects of interest.

Associated with each point v in the lattice Λ_L, there will be a complex Hilbert space $\mathcal{H}_v$, which we will take finite dimensional, and hence isomorphic to $\mathbb{C}^d$. The total Hilbert space of the system $\mathcal{H}_L$ will be $\bigotimes_{v\in\Lambda_L} \mathcal{H}_v$. Note that the dimension of the Hilbert space of the system grows exponentially with the system size L.

States on the system are semidefinite-positive operators ρ acting on the Hilbert space $\mathcal{H}_L$ and having $\mathrm{tr}(\rho) = 1$. They encode all measurable properties of the system via Born's rule $\mathrm{tr}(A\rho)$, which specifies the expected value of an observable A, when the system is in state ρ. Observables are modeled by self-adjoint operators on $\mathcal{H}_L$.

Pure states are the extreme points of the convex set of states and are therefore rank-one projectors associated with normalized vectors $|v\rangle \in \mathcal{H}_L$. In a slight abuse of notation, one usually identifies a pure state with the vector $|v\rangle$. Note that this is not a one-to-one correspondence since $\alpha|v\rangle$ ($|\alpha| = 1$) defines the same state as $|v\rangle$. We will use the standard *ket* notation $|\cdot\rangle$ for pure states. Non-pure states are called mixed states.

If the system is in state ρ but one is only interested in a subsystem, i.e., a subset $S \subset \Lambda_L$, one considers the *reduced state* ρ_S, which encodes all measurable properties of that subsystem and is defined as

$$\rho_S = (\mathbb{1}_S \otimes \mathrm{tr}_{S^c})(\rho).$$

The operator $(\mathbb{1}_S \otimes \mathrm{tr}_{S^c})$ is called partial trace and denoted by tr_{S^c}.

The behavior of the system is described by its *Hamiltonian* H_L, which is nothing but a self-adjoint operator that encodes the interactions between the particles and that is interpreted physically as the observable *energy*. Its eigenvalues $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$ are the *energy levels* of the system. The smallest eigenvalue λ_0 is called *ground energy*, and the associated normalized eigenvectors are called *ground states*. The ground states encode all the observable properties of the system when the temperature of the system goes to 0. The difference between the two smallest eigenvalues $\gamma = \lambda_1 - \lambda_0$ is called the *spectral gap*. It can be interpreted as the energy one needs to pump into the system to change its properties from the ground state to some different *excited state*.

We will restrict our study to finite-range Hamiltonians with translation invariance. The motivation behind this is the fact that the interactions between particles usually do not depend on their position on the lattice and decay in many cases very fast with distance. Therefore, it is a sensible approximation to consider that those interactions are just 0 when the distance between the particles is larger than a given threshold R,

which we call the interaction range. At the technical level, this hypothesis allows us to define easily the Hamiltonian H_L for all system sizes L (both with open and periodic boundary conditions) and hence to talk meaningfully about the case $L \to \infty$.

Formally, for any subset S of the lattice with diameter less than R, we will consider an interaction h_S, which is a self-adjoint operator on $\mathcal{H}_S = \bigotimes_{v \in S} \mathcal{H}_v$. In an abuse of notation, we will also call h_S its extension, tensoring with the identity operator, to the Hilbert space $\mathcal{H}_L$ of the full system. Any pair of sets S, S' which are related by a translation will have the same interaction ($h_S = h_{S'}$) under the identification by translation of the spaces $\mathcal{H}_S$ and $\mathcal{H}_{S'}$. The total Hamiltonian of the system will then be

$$H_L = \sum_{S,\, \mathrm{diam}(S) \leq R} h_S.$$

1.1. The spectral gap problem and its importance

The spectral gap problem can be stated as follows. Given a family of short-range translational invariant Hamiltonians $\{H_L\}_L$, where each H_L is defined on the lattice Λ_L, how does the spectral gap γ_L behave as $L \to \infty$? In particular, we are interested in the case in which the system is *gapped*, i.e., there exists $\gamma > 0$ so that $\gamma_L \geq \gamma$ for all L.

The main historical motivation to study the spectral gap is due to the fact that it defines the notion of "quantum phase of matter" in condensed matter physics via the following informal definition: "two interactions are in the same phase if there is a smooth path of interactions connecting them so that the spectral gap remains uniformly bounded throughout the path" (see, e.g., [5] for a rigorous definition). Therefore, phase transition points are detected via a closure of the spectral gap with the system size. The intuitive idea is that the spectral gap *protects* the system from jumping into excited states with different properties, and hence one needs to *close the gap* in order to get relevant changes in the properties of the ground state. However, formalizing this intuition has turned out to be a very challenging task, particularly the crucial question of the stability of the spectral gap under small perturbations of the interactions [8, 29, 34].

The spectral gap also plays a key role in adiabatic quantum computation. In this model of quantum computation, the solution of a given problem is encoded in the ability to engineer the ground state of a target Hamiltonian. The system is initialized in the ground state of a trivial Hamiltonian, then the interactions are smoothly changed from the initial to the target one. The so-called "adiabatic theorems" show that if the parameters are changed slowly enough, then the system is kept close to the ground state throughout the path, and hence one obtains the desired solution to the computation. The total time required for the adiabatic algorithm scales as $\mathrm{poly}\left(\frac{1}{\gamma_*}\right)$,

where γ_* is the infimum of the spectral gaps of all Hamiltonians considered throughout the path. Therefore, the complexity of the algorithm is precisely characterized by the spectral gap.

We refer the reader to the review [3] for detailed information about adiabatic quantum computation, including references to several adiabatic theorems.

1.2. Content and structure

The manuscript will review two types of complementary results. In the first part, we will show that the spectral gap problem is as hard as one can imagine: it is undecidable in the sense of Turing. It will be based on [6, 15]. The proof of this negative result will strongly rely on quantum computing ideas. In the second part, we will turn to positive results. Despite being undecidable in the worst case, there are still interesting families of Hamiltonians for which the spectral gap can be decided. We will focus on a particularly interesting and wide family of models motivated by quantum information theory, the so-called *tensor networks* [13]. We will describe a sufficient condition which guarantees that the system is gapped [22] based on a holographic approach which dates back to [14]. A notable application of that result in the context of quantum computing is the proof that all Kitaev's quantum double models, despite being non-trivial topological phases of matter, are not self-correcting quantum memories in the presence of thermal noise, no matter how small the temperature is [28]. This constitutes an important step toward proving the conjecture that there cannot exist self-correcting quantum memories in 2D [9, 10, 26, 44].

Section 2 will be devoted to the undecidable results, Section 3 to the case of tensor network Hamiltonians, and finally, Section 4 to the problem of thermalization in topological quantum memories. We will introduce at the beginning of each section the necessary concepts and definitions to understand the statements of the results. We will also try to give some of the main ideas behind the proofs, but keep the presentation at a sufficiently non-technical level to be as accessible as possible for non-experts. We refer the reader to the corresponding papers for a technical exposition, including detailed proofs.

2. Undecidability of the spectral gap

2.1. Basics on computability theory

Let us recall the needed basic definitions of computability theory: Turing machines, undecidability and the halting problem.

Definition 2.1. A (deterministic) Turing machine (TM) is defined by a triple (Σ, Q, δ), where Σ is a finite alphabet with an identified blank symbol, Q is a finite

set of internal control states with an identified initial state q_0 and final (or halting) state q_f, and δ is a transition function $\delta : Q \times \Sigma \to \Sigma \times Q \times \{L, R\}$.

The TM reads and writes in a two-way infinite tape of cells (identified with $\mathbb{Z}$) via a head which moves along the tape. The evolution of the machine is specified by its transition function δ. If the head is in a given position of the tape, δ takes as input the symbol read at that position and the internal state of the machine and specifies (i) which symbol from the alphabet Σ must be written on the tape in the current head position, (ii) which internal state from Q the machine must transition to, and (iii) where the head must move next, left (L) or right (R).

The machine keeps running until it enters the halting state. Then it stops, and the output is the set of non-blank symbols written on the tape.

We say that the TM halts on input x if, when initialized on input $x \in \Sigma^*$ written in positions $0, 1, 2, \ldots$, with the head in cell 0 and internal state q_0, it eventually enters the halting state.

It is a remarkable fact, already proven by [45], that there exist *universal* Turing machines (UTM). UTMs can simulate the behavior of any other TM T as follows. Given input x and a description of T, it halts if and only if T halts on input x, and, in that case, it produces as output the output of T when run on x.

An *undecidable problem* is a YES/NO decision problem which cannot be solved by any TM in the sense that there does not exist any TM which, given as input an instance of the problem, always halts and outputs the right answer YES or NO (identified with particular alphabet symbols).

The first and most important undecidable problem, to which most other undecidable problems reduce, is the *halting problem*, i.e., deciding whether a given TM halts or not. The undecidability of the halting problem was already proven by Turing in [45]. By the comments just made on UTMs, it is also undecidable to determine if a fixed UTM will halt on a given input x. In the following, we will fix a UTM and refer to it as *the* UTM.

2.2. Statements of results

We will state the undecidability of the spectral gap for 1D systems. In that case, we will consider the case of open boundary conditions for single-site and nearest neighbor interactions. That is, the overall Hamiltonian H_L will be given by two interaction terms $h^{(1)}$ and $h^{(2)}$, which are, respectively, operators on $\mathbb{C}^d$ and $\mathbb{C}^d \otimes \mathbb{C}^d$. The total Hamiltonian will be the sum of those local terms by translation:

$$H_L = \sum_{i=1}^{L-1} h^{(2)}_{i,i+1} + \sum_{i=1}^{L} h^{(1)}_i.$$

Note that we are in the open boundary condition case since we are not including the interaction $h^{(2)}_{L,1}$, which must be added for periodic boundary conditions.

The main result of [6] is a constructive proof of the following.

Theorem 2.2. *There exist (explicitly constructible) $d \times d$ matrices a, a', a'' and $d^2 \times d^2$ matrices b, b', b'', b''', b'''' with the following properties:*

(1) *a and b are diagonal with entries in $\mathbb{Z}$, i.e., they correspond to a purely classical interaction.*

(2) *a', a'', b', b'' are self-adjoint with entries in $\mathbb{Q}[\sqrt{2}]$, i.e., they are of the form $x + y\sqrt{2}$ with x and y being rational numbers.*

(3) *b''', b'''' have entries in $\mathbb{Q}$.*

(4) *For any $\eta \in \mathbb{N}$ and any rational number $0 < \beta \leq 1$ (which can be chosen arbitrarily small), define the following interaction terms with parameter η:*

$$\begin{aligned} h^{(1)}(\eta) &= a + \beta[2^{-2|\eta|}a' + a''], \\ h^{(2)}(\eta) &= b + \beta[2^{-2|\eta|}b' + b'' + \mathrm{e}^{\mathrm{i}\pi\phi(\eta)}b''' \\ &\qquad + \mathrm{e}^{-\mathrm{i}\pi\phi(\eta)}b'''^{\dagger} + \mathrm{e}^{\mathrm{i}\pi 2^{-2|\eta|}}b'''' + \mathrm{e}^{-\mathrm{i}\pi 2^{-2|\eta|}}b''''^{\dagger}]. \end{aligned}$$

Here, we denote by $|\eta|$ the number of digits in the binary expansion of η, $\eta = \eta_1\eta_2\cdots\eta_{|\eta|}$, and we define $\phi(\eta)$ to be its binary fraction with interleaved 1's, i.e., $\phi(\eta) = 0.\eta_1 1\eta_2 1\cdots\eta_{|\eta|-1}1\eta_{|\eta|}$. Then $\|h^{(1)}(\eta)\| \leq 2$, $\|h^{(2)}(\eta)\| \leq 1$ and the family of Hamiltonians $\{H_L(\eta)\}_L$ defined on Λ_L by

$$H_L(\eta) = \sum_{i=1}^{L-1} h^{(2)}_{i,i+1}(\eta) + \sum_{i=1}^{L} h^{(1)}_i(\eta)$$

satisfies the following:

(a) *If the UTM does not halt on input η, then $\{H_L(\eta)\}_L$ is gapped. Moreover, the spectral gap $\gamma_L(H_L(\eta)) \geq 1$ for all $L \in \mathbb{N}$.*

(b) *If the UTM halts on input η, then $\{H_L(\eta)\}_L$ is gapless in the following strong sense. There exists a finite interval of size c above its ground energy such that the spectrum of $H_L(\eta)$ becomes dense therein as $L \to \infty$.*

By the undecidability of the halting problem, this immediately implies that the existence of a spectral gap is undecidable for 1D Hamiltonians.

There are some remarkable comments to be made. First of all, since the interactions given by a, b are classical, and β is arbitrarily small, the result holds for

arbitrarily small quantum perturbations of classical interactions. Moreover, the physical dimension and the interactions are fixed and explicitly constructed. The only dependency in η, the input of the UTM, is in the coupling constants. Finally, there is a completely analogous result for 2D, proven in [15] (the result was announced beforehand in [16]). The main difference is that in the 2D case, one can cover the case of periodic boundary conditions, and the roles of gap and gapless are interchanged with respect to the halting property of the UTM. That is, if the UTM does not halt on input η, then $\{H_L(\eta)\}_L$ is gapless; whereas, if the UTM halts, then $\{H_L(\eta)\}_L$ is gapped.

2.3. Idea of the proof

The proof has four main steps which are shared by the 1D and 2D cases, though the details of those ingredients change substantially between the two cases, and we refer the readers to the papers [6, 15] for full details.

Step 1: Writing η on the tape. In order to feed an arbitrary input $\eta \in \mathbb{N}$ to the UTM, we need to write η on the tape. A counting argument shows that this is impossible with a *classical* TM while keeping its size (alphabet size and number of internal states) constant. This is why it is the key *quantum* step in the construction, since this will be possible with a *quantum* Turing machine (QTM), concretely with a QTM associated with the *quantum phase estimation algorithm*. To construct such a QTM, we rely strongly on the seminal results and ideas of [7]. Formally, we construct a QTM with a fixed number of internal states such that, when fed with an input given by any sequence of 1's larger than the size of η, it writes η deterministically on the tape and halts. Then, we dovetail this QTM with the UTM, which then runs on the desired input η.

The quantum phase estimation algorithm is one of the most important basic subroutines in quantum computing [35]. Given a black-box oracle access to a unitary of the form

$$U = \begin{pmatrix} 1 & 0 \\ 0 & e^{2\pi\phi\, i} \end{pmatrix},$$

the goal is to obtain the binary expansion of the *phase* $\phi \in [0, 1)$. As in many other quantum algorithms, the key step is the quantum Fourier transform (QFT), which has only polynomial complexity [35]. This means that the QFT is a unitary operation acting on the space of n *qubits* $(\mathbb{C}^2)^{\otimes n}$ which can be decomposed into poly(n) elementary unitaries which act non-trivially only in one or two of the qubits.

In order to define the QFT, we consider the canonical basis in the space of n qubits $(\mathbb{C}^2)^{\otimes n}$, given by $e_{i_1} \otimes \cdots \otimes e_{i_n}$, which we identify with $\widetilde{e}_k$. Here, k is the natural number with binary expansion $i_n \cdots i_1$, i.e., $k = \sum_{j=1}^{n} i_j 2^{j-1}$, and $\widetilde{e}_k$ is seen

as the canonical basis of the space $\mathbb{C}^N$, with $N = 2^n$. Note that, in the basis $\widetilde{e}_k$, k runs from 0 to $N-1$. The QFT is defined by its action on $\widetilde{e}_k$ as

$$\mathrm{QFT}(\widetilde{e}_k) = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} e^{\frac{2\pi jk}{N} i} \widetilde{e}_j .$$

Let us assume for simplicity that ϕ has a binary rational expansion or size n, i.e., $\phi = \sum_{j=1}^{n} \phi_j 2^{-j}$ (otherwise the algorithm is not exact and has some error). Let us define $\widetilde{\phi} = 2^n \phi = \sum_{j=1}^{n} \phi_j 2^{n-j}$, which is a natural number between 0 and $N-1$.

The idea of the quantum phase estimation algorithm is to have n ancillary qubit systems which interact with the qubit on which U acts via so-called controlled U unitaries. After an appropriate initialization, this allows us to end up, in the ancillary qubits, precisely in the state $\mathrm{QFT}(\widetilde{e}_{\widetilde{\phi}})$. Therefore, after applying the inverse QFT, one recovers the state $\widetilde{e}_{\widetilde{\phi}}$ in the ancillary qubits. One can obtain $\widetilde{\phi}$ just by measuring the qubits in the canonical basis.

Step 2: A 1D Hamiltonian whose ground state energy depends on the behavior of the UTM on input η. For this step, we need a mapping from QTM to Hamiltonians so that the ground state reflects the evolution of the QTM. In this way, by applying that mapping to the QTM that first writes η on the tape and then runs the UTM on it, we can encode the halting problem in the ground state energy. This mapping was precisely done in a remarkable paper of Gottesman and Irani [20], though some non-trivial modifications are required to make it work here. The idea of constructing Hamiltonians whose ground state reflects quantum evolutions has a long history [2,23,25,36], dating back to Feynman [19], and it is one of the most important techniques in quantum complexity theory.

After this step, we have, for each $\eta \in \mathbb{N}$, a 1D translational invariant Hamiltonian whose ground state energy for a chain of size L and open boundary conditions is different depending on the behavior of the UTM on η. Depending on whether, in the construction, one penalizes halting or non-halting in the energy (both things are possible), the energy is larger or smaller in the halting and non-halting cases. In the 2D construction, the ground state energy is larger for the halting case, and in the 1D construction, it is larger in the non-halting case. The problem is that the difference in energy between the two cases is exponentially small in L. This is why we need the next step.

Step 3: Amplifying the ground state energy difference between halting and non-halting cases. This step is implemented in completely different ways in 1D and 2D. The key ingredient is to have infinitely many well-identified segments of sufficiently large size so that the UTM has enough tape and time to halt (in the Hamiltonian

encoding of Step 2). Since the halting time is an uncomputable function, the easiest solution is to get segments of arbitrarily large lengths. This is the route taken in the 2D case, where we rely on an aperiodic tiling construction due to Robinson [40] to guarantee the existence of those segments.

In 1D, aperiodic tilings do not exist, and hence one needs a different construction. The key technical lemma, which has interest on its own, is the construction of a 1D translational invariant Hamiltonian whose ground state has a periodic structure whose period is exactly the halting time of the UTM on input η. This is remarkable since the given 1D Hamiltonian is explicitly constructed, but the periodic structure in its ground state is uncomputable.

After this step, we manage to make the energy difference between halting and non-halting diverging to ∞ as $L \to \infty$.

Step 4: From ground state energy difference to spectral gap. The final step combines the *undecidable* Hamiltonian constructed in Step 3 with (i) a trivial Hamiltonian having product ground state and constant gap and (ii) a critical Hamiltonian having spectrum that becomes dense in $[0, \infty)$. It is not difficult to find a way to couple these Hamiltonians so that the resulting spectrum has the properties required by Theorem 2.2.

3. Spectral gap for tensor network parent Hamiltonians

The theory of tensor networks (TN) lies at the intersection of quantum information theory and condensed matter physics as an ansatz to describe accurately low-energy states of short-range quantum Hamiltonians. The origin of TN can be traced back to three seminal papers in the late 1980s and early 1990s: the definition of the *AKLT model* [1], its generalization via *finitely correlated states* [18], currently known as matrix product states (MPS), and the proposal of the *DMRG algorithm* [48], which turned out to be a variational optimization of the energy in the family of MPS.

Since then, TN theory has shown its power on two fronts. Numerically, it has motivated several algorithms to simulate quantum many-body systems that are now the state of the art in the field. Analytically, it has provided mathematically rigorous results in the study and classification of quantum phases of matter. Very recently, TN have started to find applications in a wide variety of other fields, ranging from machine learning to quantum gravity or quantum computing.

We refer the reader to the reviews [13,37,46] for more information about, respectively, the numerical aspects of TN, the application of TN in a wide number of different areas, and the mathematically rigorous results on TN, both regarding their

ability to approximate low-energy states of quantum Hamiltonians and their use to classify quantum phases of matter.

3.1. Basics on tensor networks

Here, we will restrict our attention to the family of tensor networks known as projected entangled pair states (PEPS), whose 1D version is precisely the family of MPS. There are several equivalent ways to define PEPS, and we refer to [13] for a deeper introduction to them. Here, we will mainly follow the presentation of [43]. Let us start with the 1D case, which is easier to describe.

Definition 3.1. Given a set of d matrices $A_1, \dots, A_d$ of dimension $D \times D$, we define its associated matrix product state (MPS) of size L, $|\Psi_A^L\rangle$, as the normalized pure state proportional to

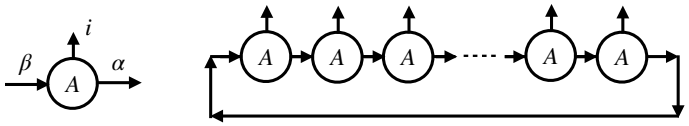
$$\sum_{i_1,\dots,i_L} \operatorname{tr}(A_{i_1} A_{i_2} \cdots A_{i_L})\, e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_L}$$

(where one is implicitly assuming that the last expression is non-zero).

The subindex A in the notation $|\Psi_A^L\rangle$ refers to the fact that the MPS is determined by the tensor

$$A = \sum_{i,\alpha,\beta} (A_i)_{\alpha,\beta}\, e_i \otimes e_\alpha \otimes e_\beta^*.$$

We will represent tensor A with the following graphical notation, where the arrow represents whether we are considering a Hilbert space or its dual, and different legs correspond to different tensor components. That is, we are looking at the tensor as a linear map from the ingoing systems to the outgoing ones. Indices α, β are called *virtual*, and D is called virtual or bond dimension. Index i is the *physical* index, and d is called the physical dimension of tensor A; i.e., for each vertex v, $\mathcal{H}_v = \mathbb{C}^d$.



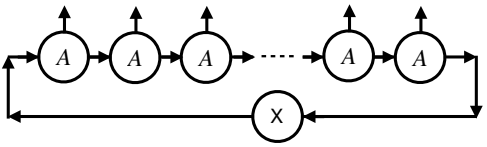
Given an MPS, one can define an associated *parent Hamiltonian* as follows. Take a region of R consecutive sites and define the space

$$\mathcal{G}_R = \left\{ \sum_{i_1,\dots,i_R} \operatorname{tr}(X A_{i_1} A_{i_2} \cdots A_{i_R})\, e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_R} : X \in \mathcal{M}_{D\times D} \right\}.$$

This is nothing but the image of $\Gamma_R : \mathcal{M}_{D\times D} \to \bigotimes_{v\in R} \mathcal{H}_v$ given by

$$X \mapsto \sum_{i_1,\ldots,i_R} \operatorname{tr}(XA_{i_1}A_{i_2}\cdots A_{i_R})\, e_{i_1}\otimes e_{i_2}\otimes\cdots\otimes e_{i_R}.$$

One can understand Γ_R as a map between Hilbert spaces by considering in $\mathcal{M}_{D\times D}$ the Hilbert–Schmidt scalar product $\operatorname{tr}(A^*B)$.

$\Gamma_R(X) =$ 

The interaction h_R is defined as the projection onto the orthogonal complement of $\mathcal{G}_R$. The parent Hamiltonian H_L is just the sum for all sets of R consecutive sites of those interactions h_R. It is clear from the definition that

(1) each $h_R \geq 0$ (meaning semidefinite positive), and hence $H_L \geq 0$,

(2) $h_R|\Psi_A^L\rangle = 0$ and hence $H|\Psi_A^L\rangle = 0$.

This implies immediately that $|\Psi_A^L\rangle$ is the ground state of H_L. It also implies that the Hamiltonian is frustration-free. That is, $|\Psi_A^L\rangle$ is the ground state of each one of the local interactions h_R defining the global Hamiltonian H_L.

It is a non-trivial question, related to quantum analogues of the classical Wielandt inequality [41] and the so-called *canonical form* of the tensor A, to determine the smallest R so that the MPS $|\Psi_A^L\rangle$ is the unique (or almost unique) ground state of H_L. We refer the reader to [12,13] for further information. Here, we just state that if R_0 is the smallest size which makes the map Γ_{R_0} injective (assuming it exists), then taking $R > R_0$ is enough to guarantee that the given MPS is the unique ground state of its parent Hamiltonian. Such R_0 is usually called the injectivity length. In this case, one can also prove that the parent Hamiltonian is always gapped [18].

Given an MPS associated with tensor A and a region R, one can also consider the important concept of *boundary state*, as $\rho_{\partial R} = \Gamma_R^*\Gamma_R$. By its very definition, ρ_R is a semidefinite-positive operator on the Hilbert space $\mathcal{M}_{D\times D}$ associated with the boundary of region R. Note that $\rho_{\partial R}$ lives in the virtual degrees of freedom of the MPS.

The definition of PEPS in two (and higher) dimensions is analogous. For simplicity, we will restrict just to the 2D case.

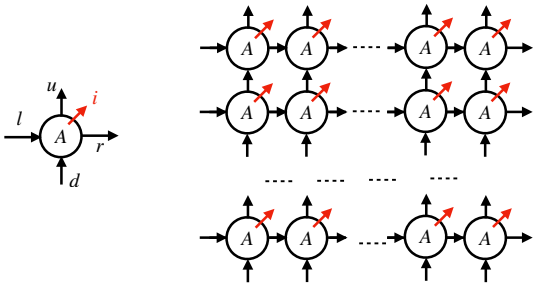
Definition 3.2. Given a set of d tensors $A_1,\ldots,A_d$ of degree 4 and dimension $D\times D\times D\times D$, we define its associated projected entangled pair state (PEPS) of

size L as the normalized state $|\Psi_A^L\rangle$ of $\mathcal{H}_L = \bigotimes_{v\in\Lambda_L} \mathcal{H}_v$ proportional to the vector obtained by contracting the tensor

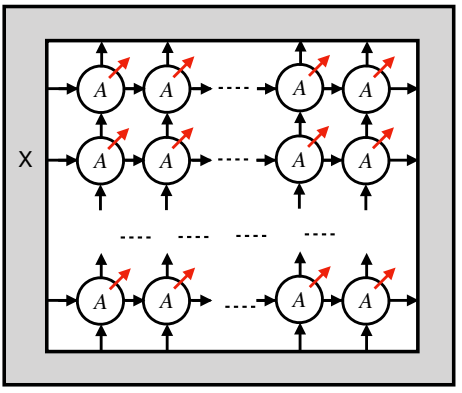
$$A = \sum_{i,u,l,d,r} (A_i)_{\alpha,\beta}\, e_i \otimes e_u \otimes e_r \otimes e_d^* \otimes e_l^* \tag{3.1}$$

on the square lattice Λ_L, where one takes periodic boundary conditions (assuming such a vector is non-zero).

The indices u, r, d, l stand for up, right, down, left, indicating the leg they refer to in the following diagrammatic notation, where, as in the 1D case, joining legs indicate the implementation of the duality mapping between one space and its dual.



Given a region $R \in \Lambda_L$, which we take to be a square for simplicity, we define, as in the 1D case above, the map $\Gamma_R : \bigotimes_{e\in\partial R}(\widehat{\mathcal{H}}_e) \to \bigotimes_{v\in R} \mathcal{H}_v$ assigning to a boundary condition X in the virtual boundary of R the corresponding vector (or unnormalized state) in R. (Here, $\widehat{\mathcal{H}}_e$ denotes the Hilbert space isomorphic to $\mathbb{C}^D$ corresponding to edge e.) Diagrammatically, $\Gamma_R(X)$ is given as follows.



Following the MPS case, we define $\mathcal{G}_R$ as the image of Γ_R and h_R as the interaction given by the projection onto the orthogonal complement of $\mathcal{G}_R$. The associated,

frustration-free, parent Hamiltonian is obtained by translating h_R throughout the lattice Λ_L, where we are taking periodic boundary conditions.

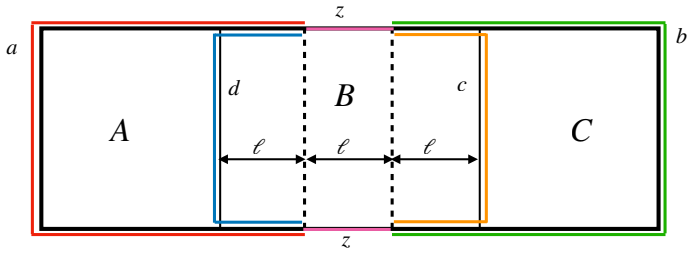
As in the 1D case, if R_0 is a square region of size $L_0 \times L_0$, where Γ_{R_0} is injective (assuming it exists), and if one defines the parent Hamiltonian interaction h_R by taking R of size $(L_0+1)\times(L_0+1)$, then the given PEPS $|\Psi_A^L\rangle$ will be the unique ground state [39]. We will refer to this situation as the "injective case".

One can also define in this case the boundary state of a region R by $\rho_{\partial R} = \Gamma_R^* \Gamma_R$, which is an unnormalized mixed state in the virtual Hilbert space of the boundary of R, $\bigotimes_{e\in\partial R} \widehat{\mathcal{H}}_e$.

3.2. Sufficient condition for gap in PEPS parent Hamiltonians

Unfortunately, even if one restricts to PEPS parent Hamiltonians, the spectral gap problem remains undecidable [42]. This forces one to restrict to particular subfamilies of PEPS and/or to aim at just the necessary or sufficient conditions for the existence of gap.

Among the most important necessary conditions for gap is the property that correlations in the ground state must decay exponentially with the distance [21, 33]. This already shows that, in the 2D case, parent Hamiltonians of PEPS are not always gapped, even in the injective case, since there exist PEPS with power-law decaying correlations [47]. Here, we will focus on a sufficient condition for gap at the level of the boundary state $\rho_{\partial R}$. The holographic perspective in PEPS, where global properties, like the gap, are characterized by properties of the boundary, which is 1D and hence simpler, has been very successful, as illustrated extensively in the recent review [13]. For the particular case of the gap, it was conjectured by numerical evidence in [14] that gap should be related to some form of *locality* of the boundary state. This was later formalized in [22] via the notion of *approximate factorization*. For that, we consider regions of the form given by the following figure.



Definition 3.3. A given PEPS has factorizable boundary if there exists a function $\xi(\ell)$ which decays to 0 at least polynomially fast (i.e., $\xi(\ell) \leq C\ell^{-\alpha}$ for constants $C, \alpha > 0$) so that for regions ABC as in the figure, there exist states $\sigma_{\partial AB}$, $\sigma_{\partial B}$,

$\sigma_{\partial BC}, \sigma_{\partial ABC}$ which factorize as follows:

$$\sigma_{\partial AB} = \Omega_{zc}\Delta_{az}, \quad \sigma_{\partial B} = \Omega_{zc}\Omega_{dz}, \quad \sigma_{\partial BC} = \Delta_{zb}\Omega_{dz}, \quad \sigma_{\partial ABC} = \Delta_{zb}\Delta_{az},$$

with $\Delta_{az}, \Delta_{zb}, \Omega_{dz}, \Omega_{zc}$ invertible matrices, and such that they approximate the true boundary states $\rho_{\partial AB}, \rho_{\partial B}, \rho_{\partial BC}, \rho_{\partial ABC}$ as follows:

$$\left\| \rho_{\partial R}^{\frac{1}{2}} \sigma_{\partial R}^{-1} \rho_{\partial R}^{\frac{1}{2}} - \mathbb{1} \right\| \leq \xi(\ell) \quad \text{for } R \in \{AB, BC, ABC\},$$
$$\left\| \rho_{\partial B}^{-\frac{1}{2}} \sigma_{\partial B} \rho_{\partial B}^{-\frac{1}{2}} - \mathbb{1} \right\| \leq \xi(\ell).$$

One of the main theorems of [22] is as follows.

Theorem 3.4. *In the injective case, the parent Hamiltonian of a PEPS has gap if the PEPS has factorizable boundary.*

Versions of this theorem also apply beyond the injective case; examples include the so-called MPO-injective PEPS (proven in [22]), which are generalizations of the G-isometric PEPS considered in Section 4.1, as well as the general case where the boundary states have a stronger form of approximate factorization: the operators $\sigma_{\partial R}$ appearing in Definition 3.3 have tensor product form and are full rank in the support of the corresponding boundary states $\rho_{\partial R}$ (see [28] for details). This latter case is the one that will be used in Section 4; there, we will refer to it simply as Theorem 3.4.

The proof of Theorem 3.4 is done by reducing to the so-called *martingale condition*, which in turn implies gap as proven in [32]. The reason that the boundary state is so closely related to the Hamiltonian can be easily understood from the fact that, in the injective case, the Hamiltonian is defined precisely via the boundary state:

$$h_R = \mathbb{1} - k_R, \quad \text{with } k_R = \Gamma_R \rho_{\partial R}^{-1} \Gamma_R^*.$$

To see it, just observe that, by definition, k_R, and hence h_R, is self-adjoint. Moreover,

$$k_R^2 = \Gamma_R \rho_{\partial R}^{-1} \underbrace{\Gamma_R^* \Gamma_R}_{\rho_{\partial R}} \rho_{\partial R}^{-1} \Gamma_R^* = k_R.$$

So k_R, and hence h_R, is a projector. Finally, it is clear from the definition that the support of k_R (and hence the kernel of h_R) is exactly the image of Γ_R.

4. Thermalization in topological quantum memories

4.1. Topological order in PEPS

One of the main advantages of PEPS as a way to represent low-energy quantum states is that they allow the characterization of global properties of interest in the

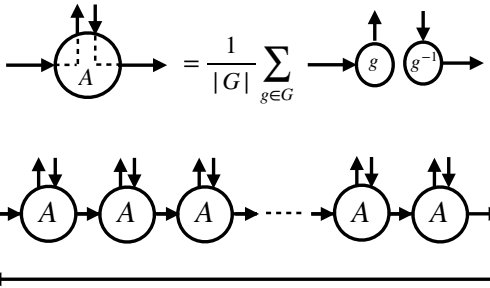
local tensor A (3.1) which defines the state. This has been particularly successful for topological order, which turns out to emerge from symmetries of the tensor A. In the most general case, those symmetries come from finite-dimensional weak Hopf algebras [11, 30]. Here, we will restrict to the particular, and simpler, case of finite groups G, leading to the notion of G-isometric PEPS defined in [43].

In topologically ordered systems, the dimension of the subspace of ground states (called ground space) in the torus is strictly larger than one. On top of that, the key feature is that all ground states are locally indistinguishable, meaning that they have exactly the same reduced state ρ_S in all contractible regions S of the torus Λ_L.

This makes those systems good candidates for quantum memories, where the encoded Hilbert space is the ground space. Topological order implies that only global operations, acting non-trivially in non-contractible regions of the torus, can affect the information encoded in the memory.

A good proxy to understand G-isometric PEPS and why they have topological order is their 1D analogue, the family of G-isometric MPS.

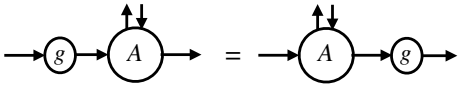
Definition 4.1. Given a finite group G, define the tensor $A = \frac{1}{|G|}\sum_{g\in G} L_g \otimes L_{g^{-1}}$, where L_g stands for the left regular representation. Note that now both the virtual and physical systems are seen as matrix spaces endowed with the Hilbert–Schmidt scalar product as in the figure (for simplicity, we use the notation g to represent the operator L_g). The associated MPS is called the G-isometric MPS.



For each $g \in G$, we define the (unnormalized) MPS with g-boundary condition as

$$|\Psi_A^g\rangle = \sum_{i_1,\dots,i_L} \operatorname{tr}(L_g A_{i_1} A_{i_2} \cdots A_{i_L})\, e_{i_1} \otimes e_{i_2} \otimes \cdots \otimes e_{i_L}.$$

Those states are also translational invariant since the operator L_g can be pulled through the tensor A.



(4.1)

Proposition 4.2. *The family of G-isometric MPS fulfills the following properties:*

(i) *The parent Hamiltonian is nearest neighbor, and the interaction between consecutive sites $i, i+1$ is given by $\mathbb{1}_i \otimes \mathbb{1}_{i+1} - k_{i,i+1}$, where, for a given region R, $k_R = \Gamma_R \Gamma_R^*$.*

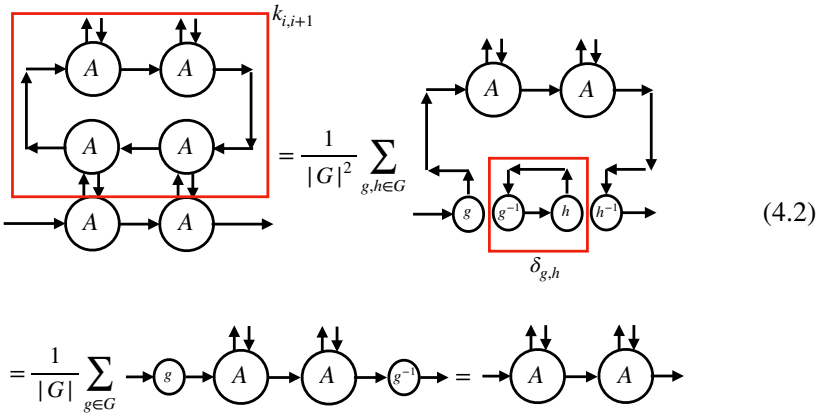
(ii) *The terms of the parent Hamiltonian commute. Moreover,*

$$k_{i,i+1}k_{i+1,i+2} = k_{i+1,i+2}k_{i,i+1} = k_{i,i+1,i+2}.$$

(iii) *The set of ground states of the global Hamiltonian H_L is spanned by the set $\{|\Psi_A^g\rangle : g \in G\}$, where $|\Psi_A^g\rangle$ and $|\Psi_A^h\rangle$ are equal if g and h are in the same conjugacy class and are orthogonal otherwise.*

It is instructive to prove this proposition, since the more interesting 2D case follows along similar ideas, but proofs are easier to illustrate in 1D. The proof also shows the power of the graphical notation.

Proof. From the definition, it is clear that $k_{i,i+1}$ is self-adjoint. That it is a projector and leaves the G-isometric MPS invariant is a trivial consequence of the following graphical equation:



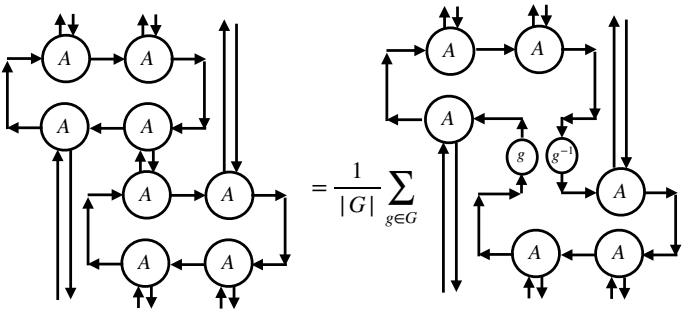
(4.2)

where in the last equality, we have used the pulling through equation (4.1), and in the first equality, we are using the orthogonality properties of the left regular representation

$$\frac{1}{|G|} \operatorname{tr}(L_{g^{-1}} L_h) = \delta_{g,h}.$$

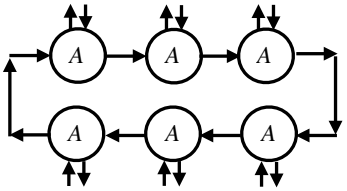
Note that, in particular, this argument shows that the boundary state of a region R of consecutive spins is exactly given by $\rho_{\partial R} = \frac{1}{|G|} \sum_{g\in G} L_g \otimes L_{g^{-1}}$, where the tensor product structure now refers to the left and right boundaries of the region.

That $k_{1,2}k_{2,3} = k_{1,2,3}$ is proven as follows. Diagramatically, the operator $k_{1,2}k_{2,3}$ can be written as follows.



$$= \frac{1}{|G|} \sum_{g \in G}$$

By using the pulling through equation (4.1), this equals



which is nothing but $k_{1,2,3}$. A similar argument shows that $k_{2,3}k_{1,2} = k_{1,2,3}$.

A similar argument proves that $K_L = \prod_{i=1}^{L} k_{i,i+1}$ equals $\sum_{g \in G} |\Psi_A^g\rangle\langle\Psi_A^g|$. Since the set of states stabilized by K_L is exactly the set of ground states of H_L, one concludes the first part of (iii). The pulling through equation (4.1) enforces that if g and h are in the same conjugacy class, $|\Psi_A^g\rangle$ and $|\Psi_A^h\rangle$ define the same state. That they are orthogonal otherwise can be seen as follows. By employing similar arguments with the graphical notation, it is not difficult to see that

$$\langle \Psi_A^h \mid \Psi_A^k \rangle = \frac{1}{|G|} \sum_g \mathrm{tr}(L_g L_{h^{-1}} L_{g^{-1}} L_k)$$

which is 0 if h and k are not in the same conjugacy class. ∎

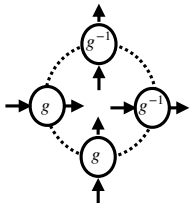
We can now turn to the definition of G- isometric PEPS.

Definition 4.3. Given a finite group G, define the tensor as follows.

$$A = \frac{1}{|G|} \sum_{g \in G} L_g \otimes L_g \otimes L_{g^{-1}} \otimes L_{g^{-1}} =$$

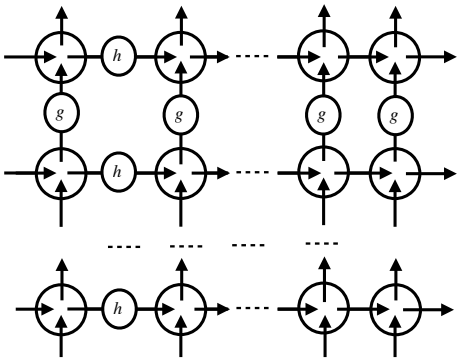


$$= \frac{1}{|G|} \sum_{g \in G}$$



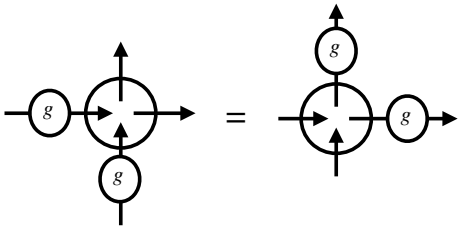
As in the 1D case, both the virtual and physical systems are seen as matrix spaces endowed with the Hilbert–Schmidt scalar product. In this case, the arrows in the interior of the circle correspond to the physical degrees of freedom. The associated PEPS are called the G-isometric PEPS.

The following theorem, proven in [43], summarizes some of the properties of G-isometric PEPS. Note the extra property (iv) with respect to Proposition 4.2, which is the one associated with topological order. In the theorem, $|\Psi_A^{(g,h)}\rangle$ refers to the state associated with the following diagram with periodic boundary conditions.



As shown in [43], states $|\Psi_A^{(g,h)}\rangle$, with $gh = hg$, are precisely the ground states of Kitaev's quantum double model associated with group G [24].

Similarly to the 1D case, the pulling through equation associated with tensor A



together with the fact that g commutes with h implies that the positions of strings associated with g and h in the definition of $|\Psi_A^{(g,h)}\rangle$ can be moved freely.

Theorem 4.4. *The set of G-isometric PEPS enjoys the following properties:*

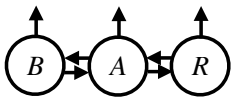
(i) *The parent Hamiltonian acts on plaquettes, and the interaction on a plaquette p is given by $\mathbb{1} - k_p$, where, for a region R, we define $k_R = \Gamma_R \Gamma_R^*$.*

(ii) *The terms of the parent Hamiltonian commute with each other. Indeed, for overlapping plaquettes p, q, $k_p k_q = k_q k_p = k_{p \cup q}$.*

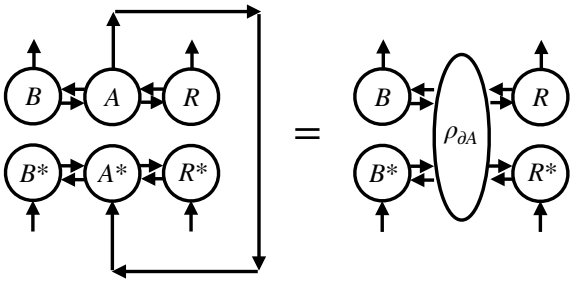
(iii) *The set of ground states of the global Hamiltonian H_L is spanned by the set $\{|\Psi_A^{(g,h)}\rangle : (g,h) \in G, gh = hg\}$, where $|\Psi_A^{(g,h)}\rangle$ and $|\Psi_A^{(g',h')}\rangle$ are equal if (g,h) and (g',h') are in the same pair-conjugacy class, meaning that there exists $k \in G$ with $g' = kgk^{-1}, h' = khk^{-1}$, and they are orthogonal otherwise.*

(iv) *All states $|\Psi_A^{(g,h)}\rangle$ are indistinguishable in the bulk, in the sense that they have the same reduced state in any contractible region of the torus Λ_L.*

The proof of (i)–(iii) is identical to the corresponding 1D cases of Proposition 4.2. Item (iv) can be proven by the following argument. Take a contractible region R and assume it to be a square for simplicity. One can pull the strings of g and h out of the region R. Let us separate the lattice Λ_L into three regions, R, the sites in the annulus next to R, call it A, and the rest of the lattice B. By blocking the tensors in each region of the PEPS, we have the following tripartite tensor network.



The reduced density matrix associated with region $R \cup B$ is given by the following diagram:



where the middle tensor corresponds precisely to the boundary state of region A, which is easily seen to be

$$\frac{1}{|G|}\sum_g \widetilde{L}_g^{\text{out}} \otimes \widetilde{L}_g^{\text{in}}.$$

Here, $\widetilde{L}_g^{\text{out}}$ is a tensor product over all edges connecting A with B of elements of the form L_g (or $L_g{}^{-1}$ depending on the orientation of the edge). The operator $\widetilde{L}_g^{\text{in}}$ is defined analogously for the edges connecting A with R. The operator $\widetilde{L}_g^{\text{in}}$ can be pulled through region R and disappears, exactly as it happened in figure (4.2). Therefore, the reduced density matrix for region $R \cup B$ is of the form $\rho_B \otimes \rho_R$, where ρ_R is exactly proportional to $k_R = \Gamma_R \Gamma_R^*$ and therefore the same for all ground states.

4.2. Lifetime of quantum memories

Motivated by the possibility of having a self-correcting quantum memory, i.e., one for which there is no need to implement active error correction, there has been an intensive search for topologically ordered quantum systems which could keep the quantum information stored in the ground state subspace for times which scale exponentially with the system size. One usually considers the case of thermal noise, where the temperature is low enough, but finite and independent of the system size.

It has been conjectured, based on arguments about energy barriers, that in 2D, self-correcting quantum memories cannot exist [9, 27]. However, this has only been rigorously proven for particular models [4, 26, 44]. We refer to [10] for a review on the topic.

The standard model for thermal noise is given by Davies dynamics [17]. This is a semigroup of completely positive trace-preserving maps $e^{\mathcal{L}_\beta t}$ defined on the set of states of the system. As usual, β corresponds to the inverse temperature. The specific form of the infinitesimal generator $\mathcal{L}_\beta$, called Lindbladian, is irrelevant for this paper. Its explicit form can be found in [28], where one considers, instead of the parent Hamiltonian of the G-isometric PEPS, the original Kitaev's Hamiltonian [24] which has the same ground space and similar properties.

The Lindbladian $\mathcal{L}_\beta$ is in our case finite range and translational invariant, so one can define the semigroup $e^{\mathcal{L}_\beta t}$ for all Λ_L. The semigroup $e^{\mathcal{L}_\beta t}$ has as a unique fixed point the thermal state of Kitaev's Hamiltonian H at inverse temperature β, i.e.,

$$\sigma_\beta = \frac{e^{-\beta H}}{\mathrm{tr}[e^{-\beta H}]}.$$

In the thermal state, all information encoded in the ground space is washed out. Therefore, understanding the lifetime of the quantum memory given by G-isometric PEPS (when considering Kitaev's Hamiltonian on them) boils down to estimating the mixing time of the associated Davies semigroup. By standard arguments [28], it is enough to give a bound on the spectral gap of $-\mathcal{L}_\beta$, independently of the system size, to ensure that the system is not a self-correcting quantum memory, since this would imply that, for any fixed ε, the time needed to guarantee that

$$\sup_{\rho_0} \| e^{\mathcal{L}_\beta t}(\rho_0) - \sigma_\beta \|_1 \leq \varepsilon \tag{4.3}$$

scales only polynomially with the size of the system L. The norm considered in (4.3) is the trace norm $\|A\|_1 = \mathrm{tr}[(AA^*)^{\frac{1}{2}}]$, which is the natural norm to measure distances between quantum states. Note that if two states are ε-close in trace norm, by the duality between trace and operator norms and Born's rule, both give ε-close expectation values for every observable which is normalized in operator norm.

The spectral gap for $-\mathcal{L}_\beta$ was proven so far only for the case of abelian groups [4, 26, 44], but the case of non-abelian ones remained open. Note that only non-abelian groups allow for universal topological quantum computation. The main contribution of [28] is precisely to extend the result to all finite groups.

Theorem 4.5. *For any finite group G, and any inverse temperature β, there exists a constant γ so that for any Λ_L, the spectral gap of $-\mathcal{L}_\beta$ is lower bounded by γ.*

The proof goes as follows. If we consider σ_β as a vector in $\bigotimes_{v\in\Lambda_L} \mathcal{M}_{d\times d}$ with the Hilbert–Schmidt scalar product in each $\mathcal{M}_{d\times d}$, the fact that the terms in the Hamiltonian commute with each other implies that σ_β is a PEPS, which we call $|\sigma_\beta\rangle$. One can then consider its parent Hamiltonian H_β. Note that this is very different from the parent or Kitaev Hamiltonians for the original G-isometric PEPS. In particular, the interaction terms of this Hamiltonian H_β do not commute with each other. The key insight, which somehow traces back to [47], is the following.

Lemma 4.6. *The operator $-\mathcal{L}_\beta$ has gap if H_β has gap.*

One can then use Theorem 3.4 to reduce the problem to proving that the boundary states of the PEPS $|\sigma_\beta\rangle$ have the appropriate approximate factorization property (see comments below Theorem 3.4). This is the technical part of the proof of Theorem 4.5.

Acknowledgments. The author deeply thanks the coauthors of all the works presented in this manuscript: J. Bausch, T. S. Cubitt, M. Kastoryano, A. Lucia, A. Pérez-Hernández and M. M. Wolf.

Funding. The author acknowledges support from the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 (grants CEX2023-001347-S and PID2020-113523GB-I00) and from the Universidad Complutense de Madrid (grant FEI-EU-22-06). This work has been financially supported by the Ministry for Digital Transformation and of the Civil Service of the Spanish Government through the QUANTUM ENIA project call – Quantum Spain project, and by the European Union through the Recovery, Transformation and Resilience Plan – NextGenerationEU within the framework of the Digital Spain 2026 Agenda.

References

[1] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki, Rigorous results on valence-bond ground states in antiferromagnets. *Phys. Rev. Lett.* **59** (1987), no. 7, 799–802

[2] D. Aharonov, D. Gottesman, S. Irani, and J. Kempe, The power of quantum systems on a line. *Comm. Math. Phys.* **287** (2009), no. 1, 41–65 Zbl 1180.68162 MR 2480741

[3] T. Albash and D. A. Lidar, Adiabatic quantum computation. *Rev. Modern Phys.* **90** (2018), no. 1, article no. 015002, 64 pp. MR 3788424

[4] R. Alicki, M. Fannes, and M. Horodecki, On thermalization in Kitaev's 2D model. *J. Phys. A* **42** (2009), no. 6, article no. 065303, 18 pp. Zbl 1159.81011 MR 2525435

[5] S. Bachmann, S. Michalakis, B. Nachtergaele, and R. Sims, Automorphic equivalence within gapped phases of quantum lattice systems. *Comm. Math. Phys.* **309** (2012), no. 3, 835–871 Zbl 1233.82006 MR 2885611

[6] J. Bausch, T. S. Cubitt, A. Lucia, and D. Perez-Garcia, Undecidability of the spectral gap in one dimension. *Phys. Rev. X* **10** (2020), no. 3, article no. 031038, 24 pp.

[7] E. Bernstein and U. Vazirani, Quantum complexity theory. *SIAM J. Comput.* **26** (1997), no. 5, 1411–1473 Zbl 0895.68042 MR 1471988

[8] S. Bravyi, M. B. Hastings, and S. Michalakis, Topological quantum order: stability under local perturbations. *J. Math. Phys.* **51** (2010), no. 9, article no. 093512, 33 pp. Zbl 1309.81067 MR 2742836

[9] S. Bravyi and B. Terhal, A no-go theorem for a two-dimensional self-correcting quantum memory based on stabilizer codes. *New J. Phys.* **11** (2009), no. 4, article no. 043029, 20 pp.

[10] B. J. Brown, D. Loss, J. K. Pachos, C. N. Self, and J. R. Wootton, Quantum memories at finite temperature. *Rev. Modern Phys.* **88** (2016), no. 4, article no. 045005, 51 pp. MR 3603187

[11] N. Bultinck, M. Mariën, D. J. Williamson, M. B. Şahinoğlu, J. Haegeman, and F. Verstraete, Anyons and matrix product operator algebras. *Ann. Physics* **378** (2017), 183–233 Zbl 1367.81084 MR 3614057

[12] J. I. Cirac, J. Garre-Rubio, and D. Pérez-García, Mathematical open problems in projected entangled pair states. *Rev. Mat. Complut.* **32** (2019), no. 3, 579–599 Zbl 1426.81060 MR 3995422

[13] J. I. Cirac, D. Pérez-García, N. Schuch, and F. Verstraete, Matrix product states and projected entangled pair states: concepts, symmetries, theorems. *Rev. Modern Phys.* **93** (2021), no. 4, article no. 045003, 65 pp. MR 4369910

[14] J. I. Cirac, D. Poilblanc, N. Schuch, and F. Verstraete, Entanglement spectrum and boundary theories with projected entangled-pair states. *Phys. Rev. B* **83** (2011), no. 24, article no. 245134, 12 pp.

[15] T. Cubitt, D. Perez-Garcia, and M. M. Wolf, Undecidability of the spectral gap. *Forum Math. Pi* **10** (2022), article no. e14, 102 pp. Zbl 1513.81051 MR 4438691

[16] T. S. Cubitt, D. Perez-Garcia, and M. M. Wolf, Undecidability of the spectral gap. *Nature* **528** (2015), no. 7581, 207–211

[17] E. B. Davies, Markovian master equations. *Comm. Math. Phys.* **39** (1974), no. 2, 91–110 Zbl 0294.60080 MR 0359633

[18] M. Fannes, B. Nachtergaele, and R. F. Werner, Finitely correlated states on quantum spin chains. *Comm. Math. Phys.* **144** (1992), no. 3, 443–490 Zbl 0755.46039 MR 1158756

[19] R. P. Feynman, Quantum mechanical computers. *Optics News* **11** (1985), no. 2, 11–20

[20] D. Gottesman and S. Irani, The quantum and classical complexity of translationally invariant tiling and Hamiltonian problems. In *2009 50th annual IEEE symposium on foundations of computer science*, pp. 95–104, IEEE, 2009 Zbl 1292.68067 MR 2648392

[21] M. B. Hastings and T. Koma, Spectral gap and exponential decay of correlations. *Comm. Math. Phys.* **265** (2006), no. 3, 781–804 Zbl 1104.82010 MR 2231689

[22] M. J. Kastoryano, A. Lucia, and D. Perez-Garcia, Locality at the boundary implies gap in the bulk for 2D PEPS. *Comm. Math. Phys.* **366** (2019), no. 3, 895–926 Zbl 1411.81066 MR 3927082

[23] J. Kempe, A. Kitaev, and O. Regev, The complexity of the local Hamiltonian problem. *SIAM J. Comput.* **35** (2006), no. 5, 1070–1097 Zbl 1102.81032 MR 2217138

[24] A. Yu. Kitaev, Fault-tolerant quantum computation by anyons. *Ann. Physics* **303** (2003), no. 1, 2–30 Zbl 1012.81006 MR 1951039

[25] A. Yu. Kitaev, A. H. Shen, and M. N. Vyalyi, *Classical and quantum computation*. Grad. Stud. Math. 47, American Mathematical Society, Providence, RI, 2002, 257 pp. Zbl 1022.68001 MR 1907291

[26] A. Kómár, O. Landon-Cardinal, and K. Temme, Necessity of an energy barrier for self-correction of abelian quantum doubles. *Phys. Rev. A* **93** (2016), no. 5, article no. 052337, 17 pp.

[27] O. Landon-Cardinal and D. Poulin, Local topological order inhibits thermal stability in 2D. *Phys. Rev. Lett.* **110** (2013), no. 9, article no. 090502, 5 pp.

[28] A. Lucia, D. Pérez-García, and A. Pérez-Hernández, Thermalization in Kitaev's quantum double models via tensor network techniques. *Forum Math. Sigma* **11** (2023), article no. e107, 71 pp. Zbl 1535.81100 MR 4672046

[29] S. Michalakis and J. P. Zwolak, Stability of frustration-free Hamiltonians. *Comm. Math. Phys.* **322** (2013), no. 2, 277–302 Zbl 1270.81089 MR 3077916

[30] A. Molnar, A. Ruiz de Alarcón, J. Garre-Rubio, N. Schuch, J. I. Cirac, and D. Pérez-García, Matrix product operator algebras I: representations of weak Hopf algebras and projected entangled pair states. 2022, arXiv:2204.05940v2, to appear in *Comm. Math. Phys*.

[31] P. Naaijkens, *Quantum spin systems on infinite lattices. A concise introduction*. Lecture Notes in Phys. 933, Springer, Cham, 2017, 177 pp. Zbl 1375.82002 MR 3617688

[32] B. Nachtergaele, The spectral gap for some spin chains with discrete symmetry breaking. *Comm. Math. Phys.* **175** (1996), no. 3, 565–606 Zbl 0838.60091 MR 1372810

[33] B. Nachtergaele and R. Sims, Lieb–Robinson bounds and the exponential clustering theorem. *Comm. Math. Phys.* **265** (2006), no. 1, 119–130 Zbl 1104.82013 MR 2217299

[34] B. Nachtergaele, R. Sims, and A. Young, Quasi-locality bounds for quantum lattice systems. Part II. Perturbations of frustration-free spin models with gapped ground states. *Ann. Henri Poincaré* **23** (2022), no. 2, 393–511 Zbl 1539.82046 MR 4386441

[35] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000, 676 pp. Zbl 1049.81015 MR 1796805

[36] R. Oliveira and B. M. Terhal, The complexity of quantum spin systems on a two-dimensional square lattice. *Quantum Inf. Comput.* **8** (2008), no. 10, 900–924 Zbl 1161.81007 MR 2522687

[37] R. Orús, Tensor networks for complex quantum systems. *Nat. Rev. Phys.* **1** (2019), no. 9, 538–550

[38] D. Pérez-García and A. Pérez-Hernández, Locality estimates for complex time evolution in 1D. *Comm. Math. Phys.* **399** (2023), no. 2, 929–970 Zbl 07678861 MR 4576761

[39] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, PEPS as unique ground states of local Hamiltonians. *Quantum Inf. Comput.* **8** (2008), no. 6–7, 650–663 Zbl 1154.81323 MR 2484304

[40] R. M. Robinson, Undecidability and nonperiodicity for tilings of the plane. *Invent. Math.* **12** (1971), no. 3, 177–209 Zbl 0197.46801 MR 0297572

[41] M. Sanz, D. Pérez-García, M. M. Wolf, and J. I. Cirac, A quantum version of Wielandt's inequality. *IEEE Trans. Inform. Theory* **56** (2010), no. 9, 4668–4673 Zbl 1366.81131 MR 2807352

[42] G. Scarpa, A. Molnár, Y. Ge, J. J. García-Ripoll, N. Schuch, D. Pérez-García, and S. Iblisdir, Projected entangled pair states: fundamental analytical and numerical limitations. *Phys. Rev. Lett.* **125** (2020), no. 21, article no. 210504, 6 pp. MR 4180430

[43] N. Schuch, I. Cirac, and D. Pérez-García, PEPS as ground states: degeneracy and topology. *Ann. Physics* **325** (2010), no. 10, 2153–2192 Zbl 1198.81063 MR 2718544

[44] K. Temme, Thermalization time bounds for Pauli stabilizer Hamiltonians. *Comm. Math. Phys.* **350** (2017), no. 2, 603–637 Zbl 1359.81087 MR 3607458

[45] A. M. Turing, On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc. (2)* **42** (1937), no. 1, 230–265 Zbl 0016.09701 MR 1577030

[46] F. Verstraete, V. Murg, and J. I. Cirac, Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems. *Adv. Phys.* **57** (2008), no. 2, 143–224

[47] F. Verstraete, M. M. Wolf, D. Perez-Garcia, and J. I. Cirac, Criticality, the area law, and the computational power of projected entangled pair states. *Phys. Rev. Lett.* **96** (2006), no. 22, article no. 220601, 4 pp. Zbl 1228.81096 MR 2231431

[48] S. R. White, Density matrix formulation for quantum renormalization groups. *Phys. Rev. Lett.* **69** (1992), no. 19, 2863–2866

David Pérez-García
Department of Applied Mathematics and Mathematical Analysis, Universidad Complutense de Madrid, Plaza de Ciencias 3, 28040 Madrid; ICMAT, C/ Nicolas Cabrera 13-15, 28049 Madrid, Spain; dperezga@ucm.es

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/29

The matrix A_2 conjecture

Stefanie Petermichl

Abstract. This survey begins with the history of modern weighted theory and its consequences for the Hilbert transform with scalar weight. It proceeds through further development in the area that leads to the counterexample to the famous matrix $\mathbf{A}_2$ conjecture.

1. The Hilbert transform

Defined as a singular integral operator on the real line, the Hilbert transform $\mathcal{H}$ is described as convolution against $1/\pi t$ and takes the form

$$\mathcal{H} f(x) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{f(t)}{x-t} \mathrm{d}t,$$

understood as a principal value integral. It has a particularly simple form as a Fourier multiplier

$$\widehat{\mathcal{H} f}(\xi) = -i \operatorname{sign}(\xi) \hat{f}(\xi).$$

There exists a related version on the circle $\mathbb{T}$, where the singular convolution kernel becomes

$$\frac{1}{2\pi} \cot\Big(\frac{t}{2}\Big)$$

and the Fourier transform is defined on integers. It follows immediately that $\mathcal{H}$ gives access to the boundary values of complex conjugate functions and that the analytic projection $\mathcal{P}_+$ can be written as $\mathcal{P}_+ = 1/2(\mathcal{I} + i\mathcal{H})$, where $\mathcal{I}$ is the identity.

The Hilbert transform itself is central to numerous branches of mathematics and certainly lies at the core of real and complex analysis as well as operator theory, but it has also long been applied as a frequency filter in AM/FM antennas. Through its multiplier form on the Fourier transform side, its formal application to $e^{imt} e^{int}$ with $|m| \gg |n|$ only sees the sign of m as $\operatorname{sign}(m+n) = \operatorname{sign}(m)$. This makes the

Mathematics Subject Classification 2020: 42B20 (primary); 42B35, 47A30 (secondary).
Keywords: matrix weight, Hilbert transform.

Hilbert transform a frequency filter, one of several crucial features when trying to filter a signal from its carrier wave. The operation of $\mathcal{H}$ is achieved by basic electrical components.

The Hilbert transform $\mathcal{H}$ is trivially bounded in L^2 by the Plancherel theorem, and using some standard methods, a weak bound at the endpoint $p = 1$, interpolation and duality, one obtains L^p boundedness for $1 < p < \infty$ (in addition to the weak-type bound in L^1 and the trivial L^2 bound). There are also many direct proofs of this fact, not using interpolation, one of which gives the sharp estimate for $\mathcal{H}$ in L^p. The proof using the weak-type bound and interpolation is the one that generalizes most easily and is usually the one found in textbooks on harmonic analysis. If a non-negative measure μ enters the picture, the boundedness in L^2_μ is no longer obvious and is false in many cases. One can show easily that it is necessary that μ be absolutely continuous with respect to Lebesgue measure and is thus described via a non-negative density w called a weight. In addition, there is a simple testing condition on the weight w that characterizes boundedness (see below).

The wish for a weighted theory of the Hilbert transform is well motivated within mathematics; it stems from certain PDE problems, stationary stochastic processes and the Riemann Hilbert problem in conjunction with Toeplitz operators, to name a few.

In order to achieve boundedness in $L^2_w(\mathbb{R})$ (in this text this means that f takes values in $\mathbb{R}$; we do not care much if our functions are defined on $\mathbb{T}$ or $\mathbb{R}$) with norm

$$\|f\|_{L^2_w(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(t)|^2 w(t) \mathrm{d}t \right)^{1/2},$$

one needs the necessary and sufficient Muckenhoupt condition [14]

$$[w]_{A_2} = \sup_I \langle w \rangle_I \langle w^{-1} \rangle_I < \infty.$$

Here, the supremum runs over all intervals I, and $\langle w \rangle_I = 1/|I| \int_I w(t) \mathrm{d}t$ denotes the average of w over I. It is easy to check that $\langle w \rangle_I \langle w^{-1} \rangle_I \geq 1$ always holds. Examples of A_2 weights are power weights such as $w(x) = |x|^\alpha$ with $|\alpha| < 1$.

Theorem 1.1 (Hunt–Muckenhaupt–Wheeden). *There exists a constant Φ that only depends on $[w]_{A_2}$ such that*

$$\|\mathcal{H}\|_{L^2_w(\mathbb{R}) \to L^2_w(\mathbb{R})} \leq \Phi([w]_{A_2}).$$

The condition $[w]_{A_2} < \infty$ is necessary for the boundedness.

When the Hilbert transform acts on vector functions $f : \mathbb{R} \to \mathbb{R}^d$ in an $L^2_W(\mathbb{R}^d)$ (functions with values in $\mathbb{R}^d$) with norms written as

$$\|f\|_{L^2_W(\mathbb{R}^d)} = \left(\int_{\mathbb{R}} \langle W(t) f(t), f(t) \rangle_{\mathbb{R}^d} \mathrm{d}t \right)^{1/2},$$

the weight is a positive semidefinite symmetric matrix function W, and the corresponding Muckenhoupt condition [34] becomes

$$[W]_{\mathbf{A}_2} = \sup_I \|\langle W\rangle_I^{1/2}\langle W^{-1}\rangle_I^{1/2}\|^2 < \infty,$$

where one can again show that the operator $\langle W\rangle_I^{1/2}\langle W^{-1}\rangle_I\langle W\rangle_I^{1/2}$ is expanding so that all terms in the above supremum are at least 1. The condition $[W]_{\mathbf{A}_2} < \infty$ is again necessary and sufficient for the boundedness of $\mathcal{H} : L^2_W(\mathbb{R}^d) \to L^2_W(\mathbb{R}^d)$, where $\mathcal{H}$ acts component wise on the vector function f.

Theorem 1.2 (Treil–Volberg). *There exists a constant $\mathbf{\Phi}$ only depending on $[W]_{\mathbf{A}_2}$ and the dimension d such that*

$$\|\mathcal{H}\|_{L^2_W(\mathbb{R}^d)\to L^2_W(\mathbb{R}^d)} \le \mathbf{\Phi}([W]_{\mathbf{A}_2}).$$

The condition $[W]_{\mathbf{A}_2} < \infty$ is necessary for the boundedness.

In this text, we are concerned with sharp weighted estimates, which means we are interested in bounds such as

$$\|\mathcal{H}\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \lesssim \Phi([w]_{A_2}), \quad \|\mathcal{H}\|_{L^2_W(\mathbb{R}^d)\to L^2_W(\mathbb{R}^d)} \lesssim \mathbf{\Phi}([W]_{\mathbf{A}_2}),$$

where we care about the exact growth near infinity of the positive functions Φ and $\mathbf{\Phi}$ that are defined on $[1,\infty)$. We say that an estimate is sharp if for any $\widetilde{\Phi}$ or $\widetilde{\mathbf{\Phi}}$ that satisfy the inequalities above there holds $\Phi = O(\widetilde{\Phi})$ and $\mathbf{\Phi} = O(\widetilde{\mathbf{\Phi}})$. This text discusses on the one hand that one can choose $\Phi(t) = t$ for scalar weights and no better than this, a well-known theorem by the author [27]. On the other hand, we highlight our recent breakthrough, Domelevo–Petermichl–Treil–Volberg [10], showing that the estimate with $\mathbf{\Phi}(t) = t^{3/2}$, previously obtained in Nazarov–Petermichl–Treil–Volberg [23], is sharp.

Theorem 1.3 (Petermichl). *There exists $C > 0$ such that*

$$\|\mathcal{H}\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \le C[w]_{A_2} \quad \forall w \in A_2.$$

There exists $c > 0$ such that for all sufficiently large Q, there exists a weight w with $[w]_{A_2} \le Q$ and a function $0 \neq g \in L^2_w(\mathbb{R})$ such that

$$\|\mathcal{H}g\|_{L^2_w(\mathbb{R})} \ge c[w]_{A_2}\|g\|_{L^2_w(\mathbb{R})}.$$

Theorem 1.4 (Nazarov–Petermichl–Treil–Volberg). *There exists $C > 0$ only depending on d such that*

$$\|\mathcal{H}\|_{L^2_W(\mathbb{R}^d)\to L^2_W(\mathbb{R}^d)} \le C[W]^{3/2}_{\mathbf{A}_2} \quad \forall W \in \mathbf{A}_2.$$

Theorem 1.5 (Domelevo–Petermichl–Treil–Volberg). *There exists $c > 0$ such that for all sufficiently large Q, there exists a 2×2 matrix weight W with $[W]_{\mathbf{A}_2} \leq Q$ and a function $0 \neq g \in L^2_W(\mathbb{R}^2)$ such that*

$$\|\mathcal{H} g\|_{L^2_W(\mathbb{R}^2)} \geq c[W]^{3/2}_{\mathbf{A}_2} \|g\|_{L^2_W(\mathbb{R}^2)}.$$

One should mention that it had been widely believed that $\Phi(t) = t^{3/2}$ even before such an estimate with $\Phi(t) = t^{3/2}$ was found by Petermichl–Pott [28] and of course before it was then again improved to the linear and sharp bound $\Phi(t) = t$ [27]. For the matrix counterpart, it had been shown by Nazarov–Petermichl–Treil–Volberg that one could choose $\mathbf{\Phi}(t) = t^{3/2}$ [23], but it was in turn believed that the growth would again be linear. Surprisingly, in the matrix case, the growth exponent remains $3/2$, as demonstrated by Domelevo–Petermichl–Treil–Volberg [10] with an involved example. The reason for this is, of course, the non-commutativity of the matrix multiplication. In the scalar case, we have an estimate for $\langle w \rangle_I \langle w^{-1} \rangle_I$, but in the matrix case for $\langle W^{-1} \rangle_I^{1/2} \langle W \rangle_I \langle W^{-1} \rangle_I^{1/2}$ and not $\langle W \rangle_I \langle W^{-1} \rangle_I$.

We give the part of the scalar weighted history that we deem important for the estimate of the Hilbert transform as well as those results and techniques that influenced the matrix weighted case. The recent scalar weighted history is much broader, with many more interesting results. Likewise, there have been a number of advances in the matrix weighted case that we do not mention. Again, we focus on the results targeting our recent counterexample to the matrix $\mathbf{A}_2$ conjecture. To keep the flow simple and concise, we often do not state theorems in full generality or even state the results qualitatively. The reader is invited to find the precise results in the provided references.

2. First significant results

The Fourier system is the best system to describe the action of the Hilbert transform, but it is not well adapted to working with weights – this is a central difficulty of these estimates. Analysts use local systems for such tasks, and a dyadic model approach in this context was picked up by Nazarov–Treil–Volberg through the use of the Haar system in conjunction with the model operator $\mathcal{T}_\sigma$. To this end, we define the dyadic grid $\mathcal{D}$ as the following collection of dyadic intervals:

$$\mathcal{D} = \{2^k([0,1) + n) : k, n \in \mathbb{Z}\}.$$

The Haar system is a basic wavelet system

$$\{h_I : I \in \mathcal{D}\}, h_I = \frac{1}{\sqrt{|I|}}(\chi_{I_+} - \chi_{I_-}).$$

The h_I are an orthonormal system of L^2. It is remarkable that step functions were used – they are not smooth, and the Hilbert transform produces mild singularities and does not have good behavior on these basis functions – but they give clarity in other ways, have a direct link to powerful tools and inherit a probabilistic structure. But to achieve anything, the use of a model operator was often needed. It replaces the Hilbert transform in part of the proof or at least models how to proceed. In the early proofs, a type of transference would then give the result on $\mathcal{H}$. In modern proofs, the model is all but a guideline. One model operator is densely defined via

$$\mathcal{T}_\sigma : h_I \mapsto \sigma_I h_I,$$

where the multiplier is a sign toss $\sigma_I = \pm 1 \forall I \in \mathcal{D}$. The model operator $\mathcal{T}_\sigma$ is called the martingale transform for the obvious reasons (but for the sake of this survey, we suppress the valuable connection of these thematics to probability theory). Much like the Hilbert transform, $\mathcal{T}_\sigma$ has L^p bounds, uniform in σ, for $1 < p < \infty$ and a uniform weak estimate at the endpoint $p = 1$. The sharp uniform norm estimate for $\mathcal{T}_\sigma$ is also known but differs from the one for $\mathcal{H}$.

In the view of the author, the two papers by Nazarov–Treil–Volberg on the two-weight problem for $\mathcal{T}_\sigma$ [25] and Treil–Volberg on the boundedness of the Hilbert transform in a matrix weighted space [34] significantly helped jump-start the first sharp weighted estimates in the second part of the 1990s.

2.1. The two-weight problem for $\mathcal{T}_\sigma$

Nazarov–Treil–Volberg [25] characterize the boundedness of $\mathcal{T}_\sigma : L^2_w(\mathbb{R}) \to L^2_v(\mathbb{R})$ and develop the Bellman function technique. This means they get several necessary conditions on w, v that are also sufficient for boundedness of $\mathcal{T}_\sigma$ for all σ. To get this characterization, the authors expand test functions f and g in disbalanced Haar systems, adapted to the weights w and v. The disbalanced Haar system can easily be derived for $u = w$ or $u = v$ as $h_I^u = \alpha_I^u \chi_I + \beta_I^u h_I$ with some constants α_I^u, β_I^u only depending upon I and the weight u so as to ensure orthogonality and normalization in $L^2_u(\mathbb{R})$. The dualized form

$$|\langle \mathcal{T}_\sigma f, g\rangle_{L^2(\mathbb{R})}| \le \sum_{I \in \mathcal{D}} |\langle f, h_I\rangle_{L^2(\mathbb{R})} \langle g, h_I\rangle_{L^2(\mathbb{R})}| \tag{2.1}$$

thus turns into an estimate of four terms when switching to the disbalanced Haar systems. The term stemming from h_I^w and h_I^v is governed by a dyadic condition similar to the A_2 condition for two different weights, while mixed terms with h_I^w and χ_I^v or h_I^v and χ_I^w require square function-type conditions with embedding theorems. The term from χ_I^w and χ_I^v is often referred to as the bilinear term; it required in this case a Sawyer-type condition on positive objects.

2.2. The angle between past and future

Treil–Volberg [34] characterize the boundedness of the Hilbert transform on vector functions in $L^2_W(\mathbb{R}^d)$ through the use of the dyadic model $\mathcal{T}_\sigma$. That is, they show that the Muckenhoupt condition for matrix weights

$$[W]_{\mathbf{A}_2} = \sup_I \|\langle W\rangle_I^{1/2}\langle W^{-1}\rangle_I^{1/2}\|^2 < \infty$$

is necessary and sufficient for the boundedness of $\mathcal{H} : L^2_W(\mathbb{R}^d) \to L^2_W(\mathbb{R}^d)$. Part of their argument uses a beautiful matrix version of the so-called Fefferman–Kenig–Pipher A_∞ inequality [12] and contains quantitative estimates that have not even been improved today. This paper also derives many of the basic properties of matrix $\mathbf{A}_2$ weights. The constants in the proof were not tracked by the authors as the paper pre-dated the spiked interest in quantitative estimates, but by shortening the proof via tricks in Petermichl–Pott [28] that gave a $3/2$ bound in the scalar case, Bickel–Petermichl–Wick [3] showed that one could get an estimate of the form $\mathbf{\Phi}(t) = t^{3/2}\log^2(t)$ for $\mathcal{T}_\sigma$. When adding the so-called Haar shift $\mathcal{S}_{\mathrm{cl}}$ (see later) to the toolbox, one could also obtain this estimate for $\mathcal{H}$.

3. First sharp estimates

These advances in the use of the Haar system and development of novel techniques were followed by the first sharp estimates for dyadic operators, for example, by Hukovic–Treil–Volberg for the square function [13] and by Wittwer for the martingale transform [35]. Later, the first continuous classical singular integral operator was estimated by Petermichl–Volberg with its sharp dependence, the Beurling–Ahlfors operator on the complex plane [30]. The latter estimate solved a famous regularity problem on the Beltrami equation. Several years later, the author proved the A_2 conjecture, namely the linear bound for the Hilbert transform.

3.1. The square function

The dyadic square function is

$$\mathcal{S} f(x) = \left(\sum_{I\in\mathcal{D}} \langle f, h_I\rangle^2_{L^2(\mathbb{R})} \frac{\chi_I(x)}{|I|}\right)^{1/2}.$$

It was shown by Hukovic–Treil–Volberg [13] that

$$\|\mathcal{S}\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \lesssim [w]_{A_2^{\mathrm{dy}}},$$

where the dependence is sharp and the supremum when determining the A_2 characteristic runs over dyadic intervals only (indicated by the superscript 'dy'). The authors use the reduction to disbalanced systems. The core estimate is proved via a Carleson embedding and the Bellman proof of the embedding condition.

Theorem 3.1 (Hukovic–Treil–Volberg). *There exists $C > 0$ such that*

$$\|S\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \le C[w]_{A_2^{\mathrm{dy}}},$$

and this estimate is sharp.

3.2. The linear estimate of $\mathcal{T}_\sigma$

The following celebrated theorem of Wittwer [35] proves through a mix of techniques in [25, 34] the estimate

$$\sup_\sigma \|\mathcal{T}_\sigma\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \lesssim [w]_{A_2^{\mathrm{dy}}}.$$

Since square functions are expectations of martingale multipliers, this estimate is stronger than that in [13] and also more involved as it has to address the bilinear term in the four-sum proof coming from the indicators in the disbalanced systems. One example of a core estimate is the following:

$$\sum_{I\in\mathcal{D}} |\langle w, h_I\rangle_{L^2(\mathbb{R})}\langle w^{-1}, h_I\rangle_{L^2(\mathbb{R})}| \lesssim [w]_{A_2}. \tag{3.1}$$

The estimate can be accomplished by a number of Bellman functions. One example is the function

$$B_Q(x,y) = \sqrt{Q}\sqrt{xy} - \frac{1}{4}xy$$

on the domain $\Omega = \{x, y \ge 0, 1 \le xy \le Q\}$. There holds on Ω: $0 \le B_Q(x,y) \le Q$, and for its Hessian, we have

$$-\mathrm{d}^2 B_Q = \begin{pmatrix} \sqrt{Q}\frac{\sqrt{y}}{4\sqrt{x^3}} & \frac{1}{4} - \sqrt{Q}\frac{1}{4\sqrt{xy}} \\ \frac{1}{4} - \sqrt{Q}\frac{1}{4\sqrt{xy}} & \sqrt{Q}\frac{\sqrt{x}}{4\sqrt{y^3}} \end{pmatrix} \ge \begin{pmatrix} 0 & \pm\frac{1}{4} \\ \pm\frac{1}{4} & 0 \end{pmatrix}.$$

To see the inequality for $-\mathrm{d}^2 B_Q$, one replaces the $1/4$ by c and estimates the determinant

$$Q\frac{1}{16xy} - \left(c^2 + Q\frac{1}{16xy} - c\sqrt{Q}\frac{1}{2\sqrt{xy}}\right) \ge -c^2 + \frac{c}{2},$$

which is non-negative when $0 \le c \le \frac{1}{2}$. The choice $1/4 \pm 1/4$ locates c in this range. Via a standard Bellman function argument, we obtain the desired estimate. The terms $\pm 1/4$ on the $\mathrm{d}x\mathrm{d}y$ slots will arrange to collect terms of the form $\sim |\langle w, h_I\rangle_{L^2(\mathbb{R})}\langle w^{-1}, h_I\rangle_{L^2(\mathbb{R})}|$ when plugging averages of w and v into the Bellman function.

Theorem 3.2 (Wittwer). *There exists $C > 0$ such that for all σ,*

$$\|\mathcal{T}_\sigma\|_{L^2_w(\mathbb{R})\to L^2_w(\mathbb{R})} \le C\,[w]_{A_2^{\mathrm{dy}}},$$

and this estimate is sharp.

3.3. The Beurling–Ahlfors operator

Petermichl–Volberg [30] solve a famous regularity problem via a sharp weighted $L^2_w(\mathbb{C})$ estimate for the Beurling–Ahlfors operator, defined through convolution in the complex plane with the kernel $1/\pi z^2$:

$$\mathcal{B} f(z) = \frac{1}{\pi}\int_{\mathbb{C}} \frac{f(\zeta)}{(z-\zeta)^2}\mathrm{d}A(\zeta),$$

again understood in the principal value sense. Strictly speaking, an exponent different from 2 was required, but for this type of operator, the exponent $p = 2$ is critical, and via the extrapolation theorem by Rubio de Francia, one obtains the required sharp $L^p_w(\mathbb{C})$ bound with dependence on the A_p characteristic of the weight. The operator $\mathcal{B}$ exchanges ∂ and $\overline{\partial}$ and thus plays a central role in the Beltrami equation, which is given by $\overline{\partial} f - \mu\partial f = 0$ for the Beltrami coefficient $\mu \in L^\infty$ with $\|\mu\|_{L^\infty} = k < 1$. Astala–Iwaniec–Saksman [1] proved that $\mathcal{I} - \mu\mathcal{B}$ and $\mathcal{I} - \mathcal{B}\mu$ are invertible in L^p for $p \in (1 + k, 1 + 1/k)$ using any weighted estimate for $\mathcal{B}$. The sharp estimate for $\mathcal{B}$ proved the critical exponent case, namely that these maps have dense ranges in $L^{1+1/k}$ and are injective in L^{1+k}. Injectivity at the critical exponent shows that weakly quasiregular maps are quasiregular.

The proof of the sharp weighted estimate for $\mathcal{B}$ builds on the estimate by Wittwer by a novel type of Bellman argument using the heat extension instead of dyadic extensions via dyadic averages. Indeed, Wittwer proved the dualized estimate

$$\sum_{I\in\mathcal{D}} |\langle f, h_I\rangle_{L^2(\mathbb{R})}\langle g, h_I\rangle_{L^2(\mathbb{R})}| \lesssim [w]^{\mathrm{dy}}_{A_2}\|f\|_{L^2_w(\mathbb{R})}\|g\|_{L^2_{w^{-1}}(\mathbb{R})}. \tag{3.2}$$

Via the Fourier transform, one can see that if $\mathcal{R}_1$ is the first Riesz transform in $\mathbb{R}^2$, then for test functions $f, g \in C_0^\infty$, there holds

$$\int_{\mathbb{R}^2} \mathcal{R}_1^2 f g\,\mathrm{d}x_1\mathrm{d}x_2 = -2\int_0^\infty \int_{\mathbb{R}^2} \frac{\partial f}{\partial x_1}\frac{\partial g}{\partial x_1}\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}t,$$

where the letters f, g stand for the heat extensions of the functions f, g originally defined on $\mathbb{R}^2$. The absolute value of the right-hand side is estimated by a constant multiple of

$$\int_0^\infty \int_{\mathbb{R}^2} \left| \frac{\partial f}{\partial x_1} \frac{\partial g}{\partial x_1} \right| \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}t,$$

which bears resemblance to the right-hand side of inequality (2.1) or left-hand side of (3.2). In turn, we have for the Beurling–Ahlfors operator $\mathcal{B} = \mathcal{R}_1^2 - \mathcal{R}_2^2 - 2i\mathcal{R}_1\mathcal{R}_2$, where we see that the imaginary part is a rotation of the real part. Estimating $\mathcal{R}_i^2$, where without loss of generality $i = 1$, is sufficient.

If one assumes a Bellman function solving (3.2), we plug heat averages into this Bellman function to obtain an estimate linear in an A_2 heat characteristic. Thanks to the rapid decay of the kernel, this is equivalent to the classical characteristic.

Theorem 3.3 (Petermichl–Volberg). *There exists $C > 0$ such that*

$$\|\mathcal{B}\|_{L^2_w(\mathbb{C}) \to L^2_w(\mathbb{C})} \leq C[w]_{A_2},$$

and this estimate is sharp.

3.4. The sharp estimate for the Hilbert transform

Through

$$\langle \mathcal{H} f, g \rangle_{L^2(\mathbb{R})} \leq \frac{1}{2\pi} \int_{\mathbb{D}} |\nabla f| |\nabla g| \log \frac{1}{|z|} \mathrm{d}A(z),$$

Petermichl–Wittwer [31] approached the estimate for the Hilbert transform. In this estimate, f, g on the right-hand side are the Poisson extensions of the functions f, g originally defined on $\mathbb{T}$. A similar method had previously been applied in [24] with a Bellman function that was less powerful, getting therefore a worse estimate. Plugging Poisson extensions into the Bellman function for (3.2) gives a nice estimate with respect to the Poisson A_2 characteristic. But due to the slow decay of the Poisson kernel, this only gives a quadratic bound in terms of the desired classical characteristic. Indeed, the Poisson A_2 can be much larger, and to show sharpness at the critical exponent $p = 2$ is quite involved and was accomplished by Kakaroumpas–Treil [18], a paper that had a great impact on finalizing the blowup of the norm of the Hilbert transform in the matrix weighted case.

To the situation's rescue and in order to prove Theorem 1.3, another tool had become available: a representation formula for the Hilbert transform using the dyadic shift

$$\mathcal{S}_{\mathrm{cl}} : h_I \mapsto h_{I_-} - h_{I_+}.$$

Petermichl [26] showed that when considering dyadic grids of the form

$$\mathcal{D}^{\alpha,r} = \{([\alpha, \alpha + r2^k) + nr2^k) : k, n \in \mathbb{Z}\}$$

for $\alpha \in \mathbb{R}$ and $r > 0$, one can consider the shift operators densely defined on the Haar functions built on intervals in $\mathcal{D}^{\alpha,r}$ as

$$\mathcal{S}_{\mathrm{cl}}^{\alpha,r} : h_I^{\alpha,r} \mapsto h_{I_-}^{\alpha,r} - h_{I_+}^{\alpha,r}.$$

It can be shown that an average in translations α followed by a logarithmic average in r produces a non-zero multiple of the Hilbert transform. This reduces weighted bounds for $\mathcal{H}$ to estimates for $\mathcal{S}_{\mathrm{cl}}$, a much simpler task if one uses the Haar base. Originally used only for upper estimates, we will see that this operator also plays a role in the blowup, or the lower estimate, in the matrix weighted case.

Staying with dyadic methods, the main tool at that time was Bellman functions. The core estimate of this work is motivated by the action of the dyadic shift:

$$\sum_{I \in \mathcal{D}} |\langle w, h_I \rangle_{L^2(\mathbb{R})} \langle w^{-1}, h_{I_-} \rangle_{L^2(\mathbb{R})}| \lesssim [w]_{A_2}.$$

There is a subtle but substantial difference to inequality (3.1): the index '$-$' on the second Haar function makes the term hard to estimate with Bellman functions as it requires a pairing of differences at a different time of the dyadic martingale. The solution consisted of the observation that the Bellman function used for this problem has a Hessian comparable to a diagonal matrix. This allows us to collect $(\mathrm{d}x)^2 + (\mathrm{d}y)^2$ instead of the seemingly smaller $|\mathrm{d}x\mathrm{d}y|$. With the increments separated, one could manage to insert an increment from the 'wrong' time. The sharp weighted estimate could thus be completed for the Haar shift.

Theorem 3.4 (Petermichl). *There exists $C > 0$ such that*

$$\|\mathcal{S}_{\mathrm{cl}}\|_{L^2_w(\mathbb{R}) \to L^2_w(\mathbb{R})} \leq C [w]_{A_2}^{\mathrm{dy}},$$

and this estimate is sharp and independent of the choice of the dyadic grid.

Theorem 1.3 followed as a corollary. This idea of 'separating increments' was developed later by Dragičević–Treil–Volberg [11] in the form of the so-called ellipse lemma. Finally, it was shown by Treil [32] using an abstract argument that one can always do such operations, even if one does not know the Bellman function explicitly. Treil's estimate also has advantages when the dyadic shift operators have higher complexity (for example, $h_I \mapsto h_{I\_\_\_}$ has complexity 3). The original argument of Petermichl [27] can be extended to shift operators of higher complexity, but the estimate one obtains is exponential in complexity, while that of Treil is linear. The author thinks the method of Treil has important implications in the area of convexity, but its features also became relevant, for example, for the representation formula of Hytönen [16] for all Calderón–Zygmund operators via generalized (high-complexity) dyadic shifts and general A_2 conjecture, the extension of the linear estimates to all Calderón–Zygmund operators.

4. Sparse revolution

In parallel to the Bellman method, more classical approaches via a form of sparse operator existed, and they became very popular due to developments around 2015. A sparse collection $\mathcal{R}$ of dyadic intervals has the following property. Defining

$$E_I = I \setminus \bigcup_{J \in \mathcal{R}, J \subsetneq I} J, \tag{4.1}$$

then $\mathcal{R}$ is sparse if, say, $|E_I| \geq 1/2|I|$. A sparse operator is an operator of the form

$$\mathcal{L} f(x) = \sum_{I \in \mathcal{R}} \langle |f| \rangle_I \chi_I(x),$$

where the $\mathcal{L}$ reminds us of Lerner. It had been known that sparse operators allow a simple weighted estimate through the use of maximal functions.

In 2015, there were several groups in parallel achieving a pointwise domination formula in terms of sparse operators for the Hilbert transform and other Calderón–Zygmund operators and in particular also for dyadic operators. We highlight the beautiful top-down proof by Lacey [20] that was extremely simple for the martingale multiplier or the dyadic shift but extended to Calderón–Zygmund operators. We state it for $\mathcal{H}$.

Theorem 4.1 (Lacey). *For any compactly supported function $f \in L^1(\mathbb{R})$, we have*

$$|\mathcal{H} f| \lesssim \mathcal{L}(f),$$

where $\mathcal{L}$ depends upon f and is a sum of at most three operators, each of which is sparse relative to a choice of dyadic grid on $\mathbb{R}$.

Building on an unweighted weak-type bound, a stopping time argument gives a pointwise domination of the dyadic version below, where the reader is again invited to pay close attention to the quantifiers. For all $f \in L^1$, there exists a sparse collection $\mathcal{R}$ so that almost surely

$$|\mathcal{T} f(x)| \lesssim \sum_{I \in \mathcal{R}} \langle |f| \rangle_I \chi_I(x)$$

for dyadic operators $\mathcal{T}$ such as $\mathcal{S}_{\mathrm{cl}}$ or $\mathcal{T}_\sigma$. The sparse intervals are the stopping intervals of hitting the double threshold of $|f|$ or $|\mathcal{T} f|$. A weak-type bound is required to observe that one stops rarely, thus getting property (4.1). There also holds such an estimate for the Hilbert transform (as stated above) and other Calderón–Zygmund operators. In the case of the Hilbert transform, only three different dyadic grids are needed; in general, the number of dyadic grids required depends upon the dimension.

Singular integral operators such as $\mathcal{H}$ rely heavily on cancellation of their kernel, so a domination by a positive object is strange – it is achieved through a choice of a collection of sparse intervals that in turn depends upon f.

With the availability of these ideas, many interesting sharp estimates followed that were out of reach with the methods that had been used prior. The first application was the extension beyond Calderón–Zygmund theory by Bernicot–Frey–Petermichl [2] that also set the notion of weak sparse forms. Shortly after, a paper on the bilinear Hilbert transform by Culiuc–Di Plinio–Ou [6] followed, along with many more. The interest in sharp matrix weighted bounds had resurfaced around 2013. But this sudden speed motivated us to revisit the matrix weighted questions, hoping that with all these new techniques, the problem would be ready to be solved.

5. Matrix weights: Positive results

We give a few recent notable positive results. Some of them are the equivalent in the scalar weighted theory, while some are not. These results with an identical formulation as in the scalar case fortified the belief that the matrix $\mathbf{A}_2$ conjecture was true. Results with worse estimates were believed to be not sharp.

5.1. Matrix weighted Carleson lemma

Culiuc–Treil [7] were able to show the weighted Carleson lemma without the use of the $\mathbf{A}_2$ condition, so precisely the same as in the scalar case. While in the scalar case, the weighted and unweighted versions are indistinguishable, there was a notable effort to get the matrix weighted version. Here is the statement of the matrix version, as proved in [7].

Theorem 5.1 (Culiuc–Treil). *Let $I_0 = [0, 1)$. Let (A_I) be a sequence of positive semidefinite matrices. There exists a constant C only depending on d such that there holds*

$$\begin{aligned} &\frac{1}{|K|} \sum_{J \in \mathcal{D}(K)} \langle W \rangle_J A_J \langle W \rangle_J \leq \langle W \rangle_K \qquad \forall K \in \mathcal{D} \\ \Rightarrow \quad &\sum_{J \in \mathcal{D}(I_0)} \| A_I^{1/2} \langle Wf \rangle_I \|^2 \leq C \| f \|^2_{L^2_W(\mathbb{R}^d)} \quad \forall f \in L^2_W(\mathbb{R}^d). \end{aligned}$$

The condition on the left-hand side is trivially necessary for the embedding as it represents a special case. The proof is a 'Bellman function with a parameter' argument. In the scalar case, the function

$$B(F, x, w, M) = F - \frac{x^2}{w + M}$$

represents the martingale variables $F = \langle f^2 w\rangle_J$, $x = \langle wf\rangle_J$, $w = \langle w\rangle_J$ and the super martingale $M = 1/|J| \sum_{I\in\mathcal{D}(J)} \alpha_I \langle w\rangle_I^2$. It is concave in $\{F \geq f^2, 0 \leq M \leq w\}$, and on this set, there holds $0 \leq B(F, x, M) \leq F$. Another crucial estimate is

$$\frac{\partial B}{\partial M} = \frac{f^2}{(w+M)^2} \geq \frac{f^2}{4w^2}. \tag{5.1}$$

This function gave the (sharp) estimate for the scalar (weighted) Carleson embedding. For its matrix analog $\mathbf{B}(\mathbf{F}, \mathbf{f}, \mathbf{W}, \mathbf{M}) = \mathbf{F} - \langle (\mathbf{W}+\mathbf{M})^{-1}\mathbf{f}, \mathbf{f}\rangle$, similar estimates hold, except that the simple derivative estimate (5.1) fails. Indeed,

$$\frac{\partial \mathbf{B}}{\partial \mathbf{M}} \Delta \mathbf{M} = \langle (\mathbf{W}+\mathbf{M})^{-1} \Delta \mathbf{M} (\mathbf{W}+\mathbf{M})^{-1}\mathbf{f}, \mathbf{f}\rangle,$$

and there is no way to estimate $(\mathbf{W}+\mathbf{M})^{-1}\Delta\mathbf{M}(\mathbf{W}+\mathbf{M})^{-1}$ from below. The increment $\Delta\mathbf{M}$ corresponds to some $\langle W\rangle_J A_J \langle W\rangle_J$, and in the factors $(\mathbf{W}+\mathbf{M})^{-1}$, the $\mathbf{M}$ has to disappear to have the resulting $\langle W\rangle_J^{-1}\langle W\rangle_J A_J \langle W\rangle_J \langle W\rangle_J^{-1} = A_J$. The idea is thus to multiply $\mathbf{M}$ by a scalar t and let $t \to 0$, certainly making the $\mathbf{M}$ disappear. Though this argument produces an inner derivative and a factor t. Simply letting $t \to 0$ does not produce the estimate; however, a bit more work in complex analysis allows us to do this with a penalty of an expected dimensional constant, here d^2, where d is the side length of the matrices. The scalar case has a numeric estimate 4, and it is known that this is optimal. It is interesting that Kraus–Moucha–Roth [19] found a complex analysis argument for the part as $t \to 0$ that gave the estimate $4d^2$. It remains an open question if the dimensional growth is optimal.

The result of Culiuc–Treil [7] readily implied with standard methods an estimate of the 'matrix-weighted unweighted maximal function', that is, Doob's inequality with matrix measure $W(x)\mathrm{d}x$. Namely, $M_{\mathrm{D}}^W : L^2_W(\mathbb{R}^d) \to L^2(\mathbb{R})$,

$$M_{\mathrm{D}}^W f(x) = \sup_{I\in\mathcal{D}, x\in I} \|W^{1/2}(x)\langle W\rangle_I^{-1}\langle Wf\rangle_I\|_{\mathbb{R}^d},$$

is bounded. This estimate was obtained through linearization by Petermichl–Pott–Reguera [29] as a consequence of [7].

Theorem 5.2 (Petermichl–Pott–Reguera). *Let W be any matrix weight. Then there exists C only depending on d such that*

$$\|M_{\mathrm{D}}^W\|_{L^2_W(\mathbb{R}^d)\to L^2(\mathbb{R})} \leq C.$$

Compare with the scalar case, where the supremum is usually taken before the multiplication with $w(x)$ and $M_w : L^2_w(\mathbb{R}) \to L^2_w(\mathbb{R})$,

$$M_w f(x) = \sup_{I: x\in I\in\mathcal{D}} |\langle f\rangle_{I,w}| = \sup_{I: x\in I\in\mathcal{D}} w(I)^{-1}\left|\int_I fw\right| = \sup_{I: x\in I\in\mathcal{D}} \langle w\rangle_I^{-1}|\langle fw\rangle_I|.$$

But $\|M^w f\|_{L^2(\mathbb{R})} = \|M_w f\|_{L^2_w(\mathbb{R})}$ since

$$w^{1/2}(x) M_w f(x) = \sup_{I : x \in I \in \mathcal{D}} |w^{1/2}(x) \langle f \rangle_{I,w}| = \sup_{I : x \in I \in \mathcal{D}} |w^{1/2}(x) \langle w \rangle^{-1} \langle f w \rangle_I|.$$

Here, it does not matter if we take the supremum before or after we multiply pointwise by $w^{1/2}(x)$, but the matrix case leaves no choice.

5.2. Linear bound for the square function

Hytönen–Petermichl–Volberg [15] proved the linear bound for the matrix weighted square function, later extended by Treil [33] to the nonhomogeneous case. The square function in this setting again has the weight W incorporated in a natural way. It is

$$\mathcal{S}_W : L^2_W(\mathbb{R}^d) \to L^2(\mathbb{R}), \quad \mathcal{S}_W f(x) = (\mathbb{E} \|W(x)^{1/2} \mathcal{T}_\sigma f(x)\|^2_{\mathbb{R}^d})^{1/2}.$$

Theorem 5.3 (Hytönen–Petermichl–Volberg). *There exists C only depending on d such that*

$$\|\mathcal{S}_W\|_{L^2_W(\mathbb{R}^d) \to L^2(\mathbb{R})} \leq C [W]_{\mathbf{A}_2}.$$

This is the same sharp bound as in the scalar case as $\mathcal{S} : L^2_w(\mathbb{R}) \to L^2_w(\mathbb{R})$ is equivalent to $\mathcal{S}_w : L^2_w(\mathbb{R}) \to L^2(\mathbb{R})$. From far, it looked as if this was the natural progression – in the scalar case, the square function was also quite a bit easier than, for example, the estimate for $\mathcal{H}$, which eventually followed. However, the authors noticed that a typical core difficulty does not arise if one has the integrated form of the square function at hand:

$$\|\mathcal{S}_W f\|^2_{L^2(\mathbb{R})} = \sum_{I \in \mathcal{D}} \langle \langle W \rangle_I \langle f, h_I \rangle_{L^2(\mathbb{R}^d)}, \langle f, h_I \rangle_{L^2(\mathbb{R}^d)} \rangle_{\mathbb{R}^d},$$

where there is a slight abuse of notation as h_I are real valued. When trying to estimate $\|W(x)^{1/2} \mathcal{T}_\sigma f(x)\|_{\mathbb{R}^2}$ (without the expectation) in $L^2(\mathbb{R})$, there are various options on how to dominate by a version of a sparse operator that in turn will be estimated by some appropriate maximal functions. There are a number of ways to do this. A common theme is the pointwise multiplier $W(x)^{1/2}$ that persists throughout. The suprema of the maximal operators typically act only after this pointwise multiplication. This makes these maximal operators difficult to stop, and every attempt cashes in an extra half power of $[W]_{\mathbf{A}_2}$, either by splitting off terms or employing awkward definitions of maximal operators. So there were good reasons why the square function could be estimated with the linear bound, while other operators could not.

5.3. The 3/2 bound

The authors Nazarov–Petermichl–Treil–Volberg [23] chose one of their numerous proofs for the publication of the estimate

$$\|\mathcal{H}\|_{L^2_W(\mathbb{R}^d)\to L^2_W(\mathbb{R}^d)} \lesssim [W]^{3/2}_{\mathbf{A}_2},$$

using the so-called convex body domination followed by an estimate involving certain sparse square functions. These sparse square function estimates were interesting in their own right and helped in turn to estimate the matrix weighted square function $\mathcal{S}_W$ linearly. The proof in [23] is by sparse domination and as such also works for all Calderón–Zygmund operators.

The convex body average of a vector function f over the interval I is defined as

$$\langle\!\langle f \rangle\!\rangle_I = \{\langle \varphi_I f\rangle_I : \varphi_I : I \to [-1,1]\}.$$

When $f \in L^1_{\chi_I}(\mathbb{R}^d)$ (L^1 on I with values in $\mathbb{R}^d$) is non-trivial, then there exists a unique subspace $E \in \mathbb{R}^d$ such that $\langle\!\langle f \rangle\!\rangle_I \subset E$ such that $\langle\!\langle f \rangle\!\rangle_I$ has non-empty interior in E, hence the name convex body. The convex body sparse operators for a sparse collection $\mathcal{R}$ are then

$$\mathcal{L} f(x) = \sum_{I\in\mathcal{R}} \langle\!\langle f \rangle\!\rangle_I \chi_I(x).$$

It is thus shown in [23] that there holds the domination

$$\mathcal{H} f(x) \in C \sum_i \sum_{I\in\mathcal{R}_i} \langle\!\langle f \rangle\!\rangle_I \chi_I(x).$$

Using simple functions, it is easy to see that if $g(x) \in \langle\!\langle f \rangle\!\rangle_I$ almost surely on I, then there exists a measurable function $K : I \times I \to \mathbb{R}$ with $\|K\|_{L^\infty} \le |I|^{-1}$ such that $g(x) = \int_I K(x,y) f(y)\mathrm{d}y$. This approach was one of many – all giving $\mathbf{\Phi}(t) = t^{3/2}$. The theorem was stated in the introduction, Theorem 1.4. The choice to write down the estimate in terms of square functions turned out natural in retrospect, as it generalized beautifully to the case beyond the Calderón–Zygmund theory in [17] by Kakaroumpas–Nguyen–Vardakis.

6. First negative results in matrix weights

It had been observed very early that matrix weights lacked certain basic features we were used to from the scalar setting, for example, the openness condition of the weight classes [5]. While we had further observed that many techniques failed completely or at least presented with difficulty that ruined constants, it remained a very difficult task to show that there is a definite failure of certain quantitative results in the matrix case. In this section, we present some of these negative results.

6.1. The convex body maximal operator is not bounded

The use of convex bodies helped to mimic what would be absolute values inside the average, in the scalar case, expressions like $\langle |f| \rangle_I$, which arise in the definition of the sparse operators. While the maximal operators below (D for Doob and LP for Littlewood–Paley),

$$M_{\mathrm{D}}^w f(x) = \sup_{I : x \in I \in \mathcal{D}} |w(x)^{1/2} \langle w \rangle_I^{-1} \langle f w \rangle_I|,$$
$$M_{\mathrm{LP}}^w f(x) = \sup_{I : x \in I \in \mathcal{D}} w(x)^{1/2} \langle w \rangle_I^{-1} \langle |f| w \rangle_I,$$

are essentially the same, both bounded by the constant 2, their matrix analogs have very different behavior. For instance,

$$M_{\mathrm{D}}^W f(x) = \sup_{I \in \mathcal{D}, x \in I} \| W^{1/2}(x) \langle W \rangle_I^{-1} \langle W f \rangle_I \|_{\mathbb{R}^d}$$

is bounded from $L^2_W(\mathbb{R}^d) \to L^2(\mathbb{R})$. To mimic the absolute value inside the average, one uses convex bodies:

$$M_{\mathrm{LP}}^W f(x) = M^W f(x) = \sup_{I \in \mathcal{D}, x \in I} \sup_{\varphi_I : I \to [-1,1]} \| W^{1/2}(x) \langle W \rangle_I^{-1} \langle \varphi_I W f \rangle_I \|_{\mathbb{R}^d}.$$

As was shown by Nazarov–Petermichl–Škreb–Treil [22], this convex body maximal function is in fact unbounded. The unboundedness is derived from the failure of the convex body Carleson lemma with matrix weight, which is also part of the work in [22]. If (A_I) is again a sequence of positive definite 2×2 matrices and if there holds the Carleson condition, then it does not follow that we may insert any φ_I into the conclusion of the Culiuc–Treil matrix weighted Carleson lemma.

Theorem 6.1 (Nazarov–Petermichl-Škreb–Treil). *There exist A_I, W_n, f and admissible φ_I so that $\| f \|^2_{L^2_{W_n}(\mathbb{R}^2)}$ is uniformly bounded but*

$$\frac{1}{|K|} \sum_{J \in \mathcal{D}(K)} \langle W_n \rangle_J A_J \langle W_n \rangle_J \leq \langle W_n \rangle_K \quad \forall K \in \mathcal{D},$$
$$\sum_{J \in \mathcal{D}(I_0)} \| A_I^{1/2} \langle \varphi_I W_n f \rangle_I \|^2_{\mathbb{R}^2} \to \infty \qquad \text{as } n \to \infty.$$

The counterexample uses a weight that only varies via a rotation in its eigenbasis and increasing eccentricity by a consecutive and patient procedure that rotates in every step the matrix weight while maintaining the martingale property. A constant $f \equiv e$ (almost any vector) will work to show the blowup. Access to the function φ_I allowed us to choose $\varphi_I = \chi_{I_+}$ inducing a shift in time on the term

$\langle \varphi_I W_n \rangle_I \sim \langle W_n \rangle_{I_+}$ as compared to just $\langle W_n \rangle_I$. While W_n and (A_I) were chosen to be in harmony so that operators A_I never see the large eigenvalues of W_n, this was brought out of balance by the time shift induced by φ_I, collecting a constant term on each one of the weight's moves. The index n describes how many moves were done. Letting $n \to \infty$ gives the blowup.

We were able to pass the example to the convex body maximal function.

Theorem 6.2 (Nazarov–Petermichl–Škreb–Treil). *There exists a matrix weight W such that M_{LP}^W does not map $L^2_W(\mathbb{R}^d) \to L^2(\mathbb{R})$.*

6.2. The bilinear Carleson lemma fails

A much simpler example gives the failure of a crucial ingredient in the original proof of the linear bound for $\mathcal{H}$, a bilinear weighted Carleson lemma. In the scalar case, it is the following.

Theorem 6.3 (Petermichl–Wittwer). *If (α_I) is a non-negative sequence of scalars, then*

$$\frac{1}{K} \sum_{I \in \mathcal{D}(K)} \alpha_I \le 1 \quad \forall K \in \mathcal{D}$$

$$\Rightarrow \quad \sum_{I \in \mathcal{D}} \frac{\alpha_I}{\langle w \rangle_I \langle w^{-1} \rangle_I} \langle w | f | \rangle_I \langle w^{-1} | g | \rangle_I \lesssim \| f \|_{L^2_w(\mathbb{R})} \| g \|_{L^2_{w^{-1}}(\mathbb{R})}$$

for all $f \in L^2_w(\mathbb{R})$ and $g \in L^2_{w^{-1}}(\mathbb{R})$.

The dyadic version is found in [27], but the original version is by Petermichl–Wittwer [31], written with Poisson averages instead of dyadic averages, and it was stated with two more conditions that were later observed to be implied by the one condition above. These were

$$\begin{aligned} \frac{1}{|K|} \sum_{I \in \mathcal{D}(K)} \frac{\alpha_I}{\langle w \rangle_I} &\le \langle w^{-1} \rangle_K \quad \forall K \in \mathcal{D}, \\ \frac{1}{|K|} \sum_{I \in \mathcal{D}(K)} \frac{\alpha_I}{\langle w^{-1} \rangle_I} &\le \langle w \rangle_K \quad \forall K \in \mathcal{D}. \end{aligned} \tag{6.1}$$

The statement of Theorem 6.3 was new, but the proof was, up to a small additional observation, implicit in [25]. Notice that other than in the linear case, the Carleson condition is not necessary for the conclusion.

A matrix version by Petermichl–Pott–Reguera [29] seemed to be very weak, and it did not help at all with the $\mathbf{A}_2$ conjecture.

Theorem 6.4 (Petermichl–Pott–Reguera). *If (α_I) is a non-negative scalar sequence, then*

$$\frac{1}{K}\sum_{I\in\mathcal{D}(K)}\alpha_I \le 1 \quad \forall K\in\mathcal{D}$$

$$\Rightarrow \quad \sum_{I\in\mathcal{D}}\alpha_I |\langle \langle W\rangle_I^{-1}\langle Wf\rangle_I, \langle W^{-1}\rangle_I^{-1}\langle W^{-1}g\rangle_I\rangle_{\mathbb{R}^d}| \lesssim \|f\|_{L^2_W(\mathbb{R}^d)}\|g\|_{L^2_{W^{-1}}(\mathbb{R}^d)}$$

for all $f\in L^2_W(\mathbb{R}^d), g\in L^2_{W^{-1}}(\mathbb{R}^d)$.

The proof was no longer by Bellman function but a simpler argument using level sets.

It is however interesting that the redundancy of the extra conditions (6.1) works in the matrix case. Petermichl–Pott–Reguera [29] showed that for non-negative (α_I), there holds

$$\frac{1}{K}\sum_{I\in\mathcal{D}(K)}\alpha_I \le 1 \quad \forall K\in\mathcal{D} \Rightarrow \frac{1}{K}\sum_{I\in\mathcal{D}(K)}\alpha_I\langle W^{-1}\rangle^{-1} \le 4\langle W\rangle_K \,\forall K\in\mathcal{D}$$

and the symmetric statement with W and W^{-1} exchanged.

Domelevo–Petermichl–Škreb [9] showed that one cannot in any way improve Theorem 6.4. Indeed, replacing (α_I) by a matrix sequence (A_I) fails, and replacing the inner product by a norm fails even in the presence of the scalar sequence. Not even assuming $W\in\mathbf{A}_2$ is helpful as one needs instead an actual barrier on the conditioning numbers of W. To use this theorem in a four-sum proof to estimate, say, $\mathcal{T}_\sigma$ in the matrix case, the inner product would have to be exchanged by a product of norms and the sequence has to be matrix valued.

7. The failure of the matrix $\mathbf{A_2}$ conjecture

It was eventually shown by Domelevo–Petermichl–Treil–Volberg [10] that in the matrix weighted case, the $3/2$ exponent obtained in [23] cannot be improved. In the scalar case, one can choose a simple weight with one singularity to see the optimality of the linear estimate. A power weight $w_\alpha(x)=|x|^\alpha$ belongs to A_2 if and only if $-1<\alpha<1$ with $[w_\alpha]_{A_2}=1/(1-\alpha^2)$. Let us choose α close to 1. Through the choice of an appropriate related test function $f_\alpha = x^{-\alpha}\chi_{(0,1)}$, one can compute directly that the weighted norm of $\mathcal{H}$ must grow at least linearly as seen by the following two

computations:

$$\begin{aligned}\|f_\alpha\|^2_{L^2_{w_\alpha}(\mathbb{R})} &= \int_0^1 x^{-2\alpha} x^\alpha \mathrm{d}x = \frac{1}{1-\alpha},\\ \|\mathcal{H} f_\alpha\|^2_{L^2_{w_\alpha}} &\geq \int_2^\infty \Big(\int_0^1 \frac{1}{x-t} f_\alpha(t)\mathrm{d}t\Big)^2 w_\alpha(x)\mathrm{d}x\\ &\gtrsim \int_2^\infty \frac{1}{x^2}\frac{1}{(1-\alpha)^2} w_\alpha(x)\mathrm{d}x \gtrsim \frac{1}{(1-\alpha)^3}.\end{aligned}$$

It is curious that the proof of the matrix bound in [23] returns a linear bound for specific situations, such as when one tries to do anything similar to the scalar example, that is, to choose a weight with one singularity. To be more clear, simple sparse operators, where the collection $\mathcal{R}$ has the property that each interval in $\mathcal{R}$ has at most one direct descendant, enjoy linear matrix weighted bounds. Leaning on this obstacle, Domelevo–Petermichl–Treil–Volberg use a sparse collection that just fails to be simple and construct on top of it a dyadic sparse paraproduct that has the feared 3/2 exponent in its estimate. Such a weight has arbitrarily many singularities.

We first show the desired lower bound for some dyadic operators in terms of the dyadic $\mathbf{A}_2^{\mathrm{dy}}$ condition. The weight we construct is necessarily quite complicated, constructed by a top-down approach: we apply subtle rotations of the eigenvectors of the averages in one step (which is only possible in the vector case), and in the next step, we disbalance the weight, which is in essence the action also giving the sharp blowup in the scalar case if one looks at the power weights. The rotation, which is only possible in the vector case, adds some subtle extra blowup: for the constructed weight, we then prove the estimate from below for several dyadic operators relevant to our task. We start with a special sparse paraproduct, which will be the main building block for all the dyadic estimates. We deduce that the classical Haar shift $\mathcal{S}_{\mathrm{cl}}$ has the desired blowup as well. We then prove the blowup for a special dyadic operator $\mathcal{H}^{\mathrm{dy}}$ which in our situation serves as a dyadic model for the Hilbert transform. While doing so, a number of quite recent estimates come into play – to allow us to not worry about all the terms, we utilize many times estimates that we know are linear – among them some sparse square functions from [23] or the square function estimate itself [15].

As explained earlier, the average of the Haar shifts $\mathcal{S}_{\mathrm{cl}}^{\alpha,r}$ gives the Hilbert transform – as such, getting upper estimates for $\mathcal{H}$ from the estimates of the Haar shift is more immediate, intuitive and obvious; this idea has been used for years. The case of a blowup is more involved: in our passage from the dyadic operators $\mathcal{H}^{\mathrm{dy}}$ to $\mathcal{H}$, we take inspiration from Bourgain [4] in the use of separating frequencies (the periodizations). A modification of his construction by considering second-order martingale differences allowed us to obtain the blowup for $\mathcal{H}$ (with a modified weight) from the blowup for $\mathcal{H}^{\mathrm{dy}}$. The choice of the dyadic operator $\mathcal{H}^{\mathrm{dy}}$ is forced upon us by the construction; we do not have any freedom here.

The construction does not stop here. The weight we constructed for the dyadic operator and its periodization are in the dyadic class $\mathbf{A}_2^{\mathrm{dy}}$; however, they both have large jumps and do not satisfy the classical matrix $\mathbf{A}_2$ condition.

To overcome this difficulty, we borrow from the technique of quasi-periodization, developed by Nazarov [21] in his counterexample to the Sarason conjecture. This technique allows us to get the classical $\mathbf{A}_2$ condition, but only in a two-weight situation. However, in the present paper, we are in the one-weight situation and thus have even less freedom. So, in the final step, we use a modification of the so-called iterated remodeling technique, developed by Kakaroumpas–Treil [18] for the scalar case.

While the passage from the dyadic blow-up estimate to $\mathcal{H}$ is complicated and quite rigid in that the Hilbert transform seems to make specific demands as to which dyadic operator it will accept, the construction of the matrix weight and the blowup of the sparse paraproduct are key. We try to explain qualitatively the construction of the weight and follow up on why the 3/2 comes into play.

For the construction of W, we start on I_0 and set the initial values W_0, V_0 of our martingales W, V. By the end of the construction, at time ∞, they will be our weight W and its inverse W^{-1},

$$W_0 = \alpha_0 a_0 a_0^\star + \beta_0 b_0 b_0^\star, \quad V_0 = \beta_0^\sharp a_0 a_0^\star + \alpha_0^\sharp b_0 b_0^\star,$$

so that they are extremal with respect to the $\mathbf{A}_2^{\mathrm{dy}}$ characteristic, say $\mathbf{Q}$, so that $\alpha_0\beta_0^\sharp = \alpha_0^\sharp\beta_0 = \mathbf{Q}$. We have the initial eccentricity $E_0 = \alpha_0/\beta_0 = \alpha_0^\sharp/\beta_0^\sharp$ and set our starting angle for the first rotation $\delta_0 = 0.1E_0^{-1/2}$. Set $\mathcal{R}_0 = \{I_0\}$. Assume $\mathcal{R}_n = \{I \in \mathcal{R}(I_0) : |I| = 2^{-2n}\}$ is constructed with

$$W_n = \alpha_n a_n a_n^\star + \beta_n b_n b_n^\star, \quad V_n = \beta_n^\sharp a_n a_n^\star + \alpha_n^\sharp b_n b_n^\star,$$

$\alpha_n\beta_n^\sharp = \alpha_n^\sharp\beta_n = \mathbf{Q}$, $E_n = \alpha_n/\beta_n = \alpha_n^\sharp/\beta_n^\sharp$, and we set $\delta_n = 0.1E_n^{-1/2}$. We then rotate the eigenvectors by $\pm\delta_n$ while maintaining the martingale property, receiving two new sets of main axes $(a_n^\pm, b_n^\pm)$ with modified coefficients. In the next step, we stretch the large eigenvalue by the right amount $r = 2 - 1/\mathbf{Q}$ and put the result on the interval to the right with its coefficients α_{n+1} and β_{n+1}. We put on the left side what we have to in order to preserve the martingale property. These right intervals are then the set $\mathcal{R}_{n+1}$, and the construction is such that again $\alpha_{n+1}\beta_{n+1}^\sharp = \alpha_{n+1}^\sharp\beta_{n+1} = \mathbf{Q}$.

We will eventually terminate this construction with one more step, after enough steps have been made so that enough damage has been done to the estimate. This last step is simple but essential so as to end up with an actual matrix weight function and its inverse.

In the following, the function $f : I_0 \to \mathbb{R}^2$, and we denote by $h_I^\infty = \sqrt{|I|}h_I$ the L^∞ normalized Haar functions. From above, $\mathcal{R}$ is the sparse collection and

$\mathcal{R}_<(I_0) = \mathcal{R}\backslash\{I_0\}$. The sparse paraproduct is

$$\Pi f(x) = \sum_{I\in\mathcal{R}_<(I_0)} \langle f\rangle_I h_I^\infty(x).$$

Then f will take the form $\mathcal{X}_{I_0}W^{-1}e$ for some appropriate e, and we will compute $\|\Pi f\|_{L^2_W(\mathbb{R}^2)}$ for a specially constructed $W : I_0 \to \mathbb{R}^{2\times 2}$. Denote $\mathcal{R}_<(J) = \{I \subsetneq J : I \in \mathcal{R}\}$. Then

$$\begin{aligned}
&\|\Pi(W^{-1}\chi_{I_0}e)\|^2_{L^2_W(\mathbb{R}^2)}\\
&\quad= \sum_{I\in\mathcal{R}_<(I_0)}\sum_{J\in\mathcal{R}_<(I_0)} \langle\langle W^{-1}\rangle_I h_I^\infty e, \langle W^{-1}\rangle_J h_J^\infty e\rangle_{L^2_W(\mathbb{R}^2)}\\
&\quad= \sum_{I\in\mathcal{R}_<(I_0)} \|\langle W\rangle_I^{1/2}\langle W^{-1}\rangle_I e\|^2_{\mathbb{R}^2}|I|\\
&\qquad\quad + 2\sum_{I\in\mathcal{R}_<(I_0)}\sum_{J\in\mathcal{R}_<(I)}\int \langle\langle W^{-1}\rangle_I e, W(s)h_I^\infty(s)h_j^\infty(s)\langle W^{-1}\rangle_J e\rangle_{\mathbb{R}^2}\mathrm{d}s.
\end{aligned}$$

The blowup we need will be in the second sum $2\mathbf{\Sigma}$, and this term will be positive. Denote $\mathcal{R}_\pm(I) = \{J \in \mathcal{R}_<(I) : J \subset I_\pm\}$. When $J \in \mathcal{R}_\pm(I)$, then $\langle Wh_I^\infty h_J^\infty\rangle_J = \pm\langle Wh_J^\infty\rangle_J$, and therefore,

$$\begin{aligned}
\mathbf{\Sigma} = &\sum_{I\in\mathcal{R}_<(I_0)}\sum_{J\in\mathcal{R}_+(I)} \langle\langle W^{-1}\rangle_I e, \langle Wh_J^\infty\rangle_J\langle W^{-1}\rangle_J e\rangle_{\mathbb{R}^2}|J|\\
&- \sum_{I\in\mathcal{R}_<(I_0)}\sum_{J\in\mathcal{R}_-(I)} \langle\langle W^{-1}\rangle_I e, \langle Wh_J^\infty\rangle_J\langle W^{-1}\rangle_J e\rangle_{\mathbb{R}^2}|J|.
\end{aligned}$$

In the scalar case, the summands would be estimated by $\langle w^{-1}\rangle_I[w]_{A_2}|J|$. If $I \in \mathcal{R}_<(I_0)$ with $|I| = 2^{-2n}$ and we sum over $J \in \mathcal{R}_<(I)$ with $|J| = 2^{-2k}$ for $k > n$, then

$$\begin{aligned}
\sum_{J\in\mathcal{R}_{k<}(I)} \langle w^{-1}\rangle_I\langle w\rangle_J\langle w^{-1}\rangle_J|J| &\le [w]_{A_2}\langle w^{-1}\rangle_I \sum_{J\in\mathcal{R}_{k<}(I)}|J|\\
&= [w]_{A_2}\langle w^{-1}\rangle_I 2^{n-k}|I|.
\end{aligned}$$

We would then sum over $k > n$:

$$\sum_{J\in\mathcal{R}_{k<}(I)} \langle w^{-1}\rangle_I\langle w\rangle_J\langle w^{-1}\rangle_J|J| \le \sum_{k>n}[w]_{A_2}\langle w^{-1}\rangle_I 2^{n-k}|I| \le [w]_{A_2}\langle w^{-1}\rangle_I|I|.$$

For scalar weights, $\sum_{I\in\mathcal{R}(I_0)}\langle w^{-1}\rangle_I|I| \lesssim [w]_{A_2}\langle w^{-1}\rangle_{I_0}$, giving the quadratic bound for $\mathbf{\Sigma}$ and thus linear bound for Π.

In the matrix case, there is no estimate for $\langle W^{-1}\rangle_J \langle W\rangle_J$. Our example is constructed so that most terms have an estimate from below of the form

$$\mathbf{Q} r^{k-n} \langle \langle W^{-1}\rangle e, e\rangle_{\mathbb{R}^2} |J|,$$

where $r = 2 - 1/\mathbf{Q}$. Thus, instead of summing the geometric series $\sum_{l=1}^{\infty} 2^{-l} = 1$, we sum $\sum_{l=1}^{\infty} \big(r/2\big)^k = 2\mathbf{Q} - 1$ which gives us an extra factor $\mathbf{Q}$ in the sum.

Thanks to the stretching in every second step, the weight is very disbalanced so that the paraproduct Π is comparable to a slice of the classical Haar shift. By a 'slice' we mean the dyadic sum of the Haar shift restricted to every second generation of dyadic intervals. As it turns out, this is not enough to get a lower estimate for $\mathcal{H}$. The dyadic operator $\mathcal{H}^{\mathrm{dy}}$ involves the (sliced) odd operator $\mathcal{S}_{\mathrm{cl}} - \mathcal{S}_{\mathrm{cl}}^{\star} + c\mathcal{S}_0$, where $\mathcal{S}_0 : h_{I_\pm} \mapsto \pm h_{I_\mp}$ is a new type of odd dyadic shift, recently used by Domelevo–Petermichl [8] in a question on Banach space geometry. For our situation, the $\mathcal{S}_0$ did not stress us as the slice we used has small norm, but the demand on $\mathcal{S}_{\mathrm{cl}} - \mathcal{S}_{\mathrm{cl}}^{\star}$ was a cliffhanger: both these operators $\mathcal{S}_{\mathrm{cl}}$ and $\mathcal{S}_{\mathrm{cl}}^{\star}$ have the same contributions, but luckily with opposing signs.

Funding. This work was partially supported by the Alexander von Humboldt Foundation.

References

[1] K. Astala, T. Iwaniec, and E. Saksman, Beltrami operators in the plane. *Duke Math. J.* **107** (2001), no. 1, 27–56 Zbl 1009.30015 MR 1815249

[2] F. Bernicot, D. Frey, and S. Petermichl, Sharp weighted norm estimates beyond Calderón–Zygmund theory. *Anal. PDE* **9** (2016), no. 5, 1079–1113 Zbl 1344.42009 MR 3531367

[3] K. Bickel, S. Petermichl, and B. D. Wick, Bounds for the Hilbert transform with matrix A_2 weights. *J. Funct. Anal.* **270** (2016), no. 5, 1719–1743 Zbl 1335.44001 MR 3452715

[4] J. Bourgain, Some remarks on Banach spaces in which martingale difference sequences are unconditional. *Ark. Mat.* **21** (1983), no. 1–2, 163–168 Zbl 0533.46008 MR 0727340

[5] M. Bownik, Inverse volume inequalities for matrix weights. *Indiana Univ. Math. J.* **50** (2001), no. 1, 383–410 Zbl 0992.42006 MR 1857041

[6] A. Culiuc, F. Di Plinio, and Y. Ou, Domination of multilinear singular integrals by positive sparse forms. *J. Lond. Math. Soc. (2)* **98** (2018), no. 2, 369–392 Zbl 1402.42013 MR 3873113

[7] A. Culiuc and S. Treil, The Carleson embedding theorem with matrix weights. *Int. Math. Res. Not. IMRN* **2019** (2019), no. 11, 3301–3312 Zbl 1428.42018 MR 3961702

[8] K. Domelevo and S. Petermichl, The dyadic and the continuous Hilbert transforms with values in Banach spaces. [v1] 2022, [v2] 2023, arXiv:2212.00090v2

[9] K. Domelevo, S. Petermichl, and K. A. Škreb, Failure of the matrix weighted bilinear Carleson embedding theorem. *Linear Algebra Appl.* **582** (2019), 452–466 Zbl 1426.42013 MR 3995793

[10] K. Domelevo, S. Petermichl, S. Treil, and A. Volberg, The matrix A_2 conjecture fails, i.e. $3/2 > 1$. 2024, arXiv:2402.06961v1

[11] O. Dragičević, S. Treil, and A. Volberg, A theorem about three quadratic forms. *Int. Math. Res. Not. IMRN* **2008** (2008), article no. rnn072, 9 pp. Zbl 1245.42012 MR 2439569

[12] R. A. Fefferman, C. E. Kenig, and J. Pipher, The theory of weights and the Dirichlet problem for elliptic equations. *Ann. of Math. (2)* **134** (1991), no. 1, 65–124 Zbl 0770.35014 MR 1114608

[13] S. Hukovic, S. Treil, and A. Volberg, The Bellman functions and sharp weighted inequalities for square functions. In V. P. Havin and N. K. Nikolski (Eds.), *Complex analysis, operators, and related topics*, pp. 97–113, Oper. Theory Adv. Appl. 113, Birkhäuser, Basel, 2000 Zbl 0972.42011 MR 1771755

[14] R. Hunt, B. Muckenhoupt, and R. Wheeden, Weighted norm inequalities for the conjugate function and Hilbert transform. *Trans. Amer. Math. Soc.* **176** (1973), 227–251 Zbl 0262.44004 MR 0312139

[15] T. Hytönen, S. Petermichl, and A. Volberg, The sharp square function estimate with matrix weight. *Discrete Anal.* (2019), article no. 2, 8 pp. Zbl 1431.42019 MR 3936542

[16] T. P. Hytönen, The sharp weighted bound for general Calderón–Zygmund operators. *Ann. of Math. (2)* **175** (2012), no. 3, 1473–1506 Zbl 1250.42036 MR 2912709

[17] S. Kakaroumpas, T. H. Nguyen, and D. Vardakis, Matrix-weighted estimates beyond Calderón–Zygmund theory. *Math. Ann.* **394** (2026), no. 3, article no. 57, 56 pp. Zbl 08165566 MR 5033744

[18] S. Kakaroumpas and S. Treil, "Small step" remodeling and counterexamples for weighted estimates with arbitrarily "smooth" weights. *Adv. Math.* **376** (2021), article no. 107450, 52 pp. Zbl 1453.42011 MR 4178922

[19] D. Kraus, A. Moucha, and O. Roth, A sharp Bernstein-type inequality and application to the Carleson embedding theorem with matrix weights. *Anal. Math. Phys.* **12** (2022), no. 1, article no. 40, 6 pp. Zbl 1482.42063 MR 4376241

[20] M. T. Lacey, An elementary proof of the A_2 bound. *Israel J. Math.* **217** (2017), no. 1, 181–195 Zbl 1368.42016 MR 3625108

[21] F. Nazarov, A counterexample to Sarason's conjecture. Unpublished manuscript, https://users.math.msu.edu/users/fedja/prepr.html visited on 20 March 2026

[22] F. Nazarov, S. Petermichl, K. A. Škreb, and S. Treil, The matrix-weighted dyadic convex body maximal operator is not bounded. *Adv. Math.* **410** (2022), no. part A, article no. 108711, 20 pp. Zbl 1504.42065 MR 4491249

[23] F. Nazarov, S. Petermichl, S. Treil, and A. Volberg, Convex body domination and weighted estimates with matrix weights. *Adv. Math.* **318** (2017), 279–306 Zbl 1382.42010 MR 3689742

[24] F. Nazarov and S. Treil, The weighted norm inequalities for Hilbert transform are now trivial. *C. R. Acad. Sci. Paris Sér. I Math.* **323** (1996), no. 7, 717–722 Zbl 0858.44004 MR 1416164

[25] F. Nazarov, S. Treil, and A. Volberg, The Bellman functions and two-weight inequalities for Haar multipliers. *J. Amer. Math. Soc.* **12** (1999), no. 4, 909–928 Zbl 0951.42007 MR 1685781

[26] S. Petermichl, Dyadic shifts and a logarithmic estimate for Hankel operators with matrix symbol. *C. R. Acad. Sci. Paris Sér. I Math.* **330** (2000), no. 6, 455–460 Zbl 0991.42003 MR 1756958

[27] S. Petermichl, The sharp bound for the Hilbert transform on weighted Lebesgue spaces in terms of the classical A_p characteristic. *Amer. J. Math.* **129** (2007), no. 5, 1355–1375 Zbl 1139.44002 MR 2354322

[28] S. Petermichl and S. Pott, An estimate for weighted Hilbert transform via square functions. *Trans. Amer. Math. Soc.* **354** (2002), no. 4, 1699–1703 Zbl 1005.42002 MR 1873024

[29] S. Petermichl, S. Pott, and M. C. Reguera, A matrix weighted bilinear Carleson lemma and maximal function. *Anal. Math. Phys.* **9** (2019), no. 3, 1163–1180 Zbl 1428.42021 MR 4014862

[30] S. Petermichl and A. Volberg, Heating of the Ahlfors–Beurling operator: weakly quasiregular maps on the plane are quasiregular. *Duke Math. J.* **112** (2002), no. 2, 281–305 Zbl 1025.30018 MR 1894362

[31] S. Petermichl and J. Wittwer, A sharp estimate for the weighted Hilbert transform via Bellman functions. *Michigan Math. J.* **50** (2002), no. 1, 71–88 Zbl 1040.42008 MR 1897034

[32] S. Treil, Sharp A_2 estimates of Haar shifts via Bellman function. In *Recent trends in analysis*, pp. 187–208, Theta Ser. Adv. Math. 16, Theta, Bucharest, 2013 Zbl 1299.42046 MR 3411052

[33] S. Treil, Mixed A_2-A_∞ estimates of the non-homogeneous vector square function with matrix weights. *Proc. Amer. Math. Soc.* **151** (2023), no. 8, 3381–3389 Zbl 1514.42019 MR 4591773

[34] S. Treil and A. Volberg, Wavelets and the angle between past and future. *J. Funct. Anal.* **143** (1997), no. 2, 269–308 Zbl 0876.42027 MR 1428818

[35] J. Wittwer, A sharp estimate on the norm of the martingale transform. *Math. Res. Lett.* **7** (2000), no. 1, 1–12 Zbl 0951.42008 MR 1748283

Stefanie Petermichl
Institut für Mathematik, Universität Würzburg, Emil-Fischer-Straße 41, 97074 Würzburg, Germany; stefanie.petermichl@uni-wuerzburg.de

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/30

Proofs, computational complexity, and games

Pavel Pudlák

Abstract. This paper discusses the interplay between proof complexity, computational complexity, and combinatorial games. Classical proof theory typically concerns itself with the provability of sentences without considering the lengths of proofs. One of the basic results is that, for sufficiently rich theories, it is impossible to distinguish provable sentences from false sentences using an algorithm. In proof complexity, the lengths of proofs play a central role; hence the question is how difficult it is to distinguish propositions with polynomial-size proofs from false propositions, in particular, whether this is possible using an algorithm running in polynomial time. This problem is represented using the concept of canonical pairs of proof systems, which is closely connected with the concept of interpolation pairs of proof systems. Interpolation pairs allow us to connect proof systems to computational complexity problems and, for some proof systems, even to prove exponential lower bounds on the lengths of proofs. Our goal is to extend this approach to lower bounds to stronger systems, in particular, to constant-depth Frege proof systems. To this end, we have characterized the interpolation pairs of constant-depth Frege systems using certain combinatorial games. In this paper, we will define these games and explain their connection to some other concepts studied in proof complexity, specifically, the communication complexity versions of TFNP classes. Our aim is to show that the three concepts studied in logic and computational complexity theory, *playing combinatorial games, computing,* and *solving problems using communication between two players,* are different facets of the same thing.

1. Introduction

Classical *computability theory* studies questions asking which functions are computable, which problems are decidable, etc. Here, computability means that some algorithm produces an answer in a finite amount of time. Knowing that a problem can be solved in a finite time certainly provides an essential piece of information, but more precise information is needed in practice. *Computational complexity* therefore considers resources required to perform computations. These are mostly time and

Mathematics Subject Classification 2020: 03D15 (primary); 68Q15 (secondary).
Keywords: proofs, complexity, games.

space (by which we mean memory), but also others, such as randomness, nondeterminism, etc. While this may look like a part of applied science, many mathematicians are attracted to this field by deep theoretical problems, such as the P vs. NP question, which is considered by many as one of the most difficult mathematical problems.

Similarly, classical *proof theory* studies the provability of sentences in principle, i.e., disregarding the lengths of proofs, while *proof complexity* is concerned with the lengths of proofs. There are many parallels between these research areas; one is that several fundamental problems in proof complexity resist all attempts to be solved, like the fundamental problems in computational complexity.

There are problems that belong to both proof theory and computability theory. The most important of these is: for a given formal system, say a theory with a computable set of axioms, how difficult is it to distinguish provable sentences from false sentences? In this paper, I will deal with a similar problem in proof and computational complexity: for a given propositional proof system, how difficult is it to distinguish propositions with polynomial-length proofs from false propositions? This question is formalized by the concept of the *canonical pair* of a proof system. Canonical pairs are specific pairs of disjoint NP sets, and the question is how complex a set we need to separate such a pair.[1] The related concept of the *interpolation pair* of a proof system is connected with the method of *feasible interpolation* by which exponential lower bounds were proved for some proof systems. By studying these pairs, we want to determine limits of this method.

In [4], we characterized the canonical pair of Resolution, which is the depth 0 Frege system. This also characterizes the interpolation pair of the depth 1 Frege system. In [26], I gave a combinatorial characterization of interpolation pairs for all levels of the bounded-depth Frege hierarchy, which is also a characterization of canonical pairs because the canonical pair of the depth d Frege system is equivalent to the interpolation pair of the depth $(d + 1)$ Frege system. These characterizations are based on certain combinatorial games[2] introduced in that paper. This paper will introduce these games and, for depths 0 and 1, show how they naturally emerge from the communication complexity TFNP classes associated with the proof systems. Moreover, I will briefly mention games with more than two players and their connection to multiparty communication complexity and a couple of combinatorial games that have been studied in connection with constant-depth Frege systems.

[1] Even if both sets are NP-complete, it may be possible to separate them using a set in P.

[2] One can also say *one* game with the parameter d.

2. Propositional proof systems

When I started with logic, I considered propositional logic absolutely uninteresting. I wondered why some logicians pay so much attention to the laws and rules of propositional logic when there is a simple procedure to test whether a propositional formula is a tautology. This view conformed with the widespread opinion that mathematics starts with infinity because finite problems are trivial. We can indeed check a tautology simply by testing all truth assignments to its variables, but when we consider *the complexity of testing*, things become interesting (or if we consider non-classical logics, which I will not do in this paper). Many tautologies have simple proofs, while others, for example, random formulas in a suitable probability distribution, seem hard.

But why should a mathematician study such problems? Is it not just a problem for practical computations? The answer is both yes and no. Yes, because propositional proofs are used in automated reasoning (including the now popular AI systems). No, because the lengths of propositional proofs are connected with provability in first-order theories. Lower bounds on the lengths of propositional proofs can, in principle, be used to show the unprovability of Π_1-sentences in first-order theories axiomatizing fragments of arithmetic, a fact discovered by S. A. Cook [10]. We are able to prove superpolynomial lower bounds on propositional proofs only in very special instances, yet there are several interesting applications of this method.

Another reason for studying propositional proof complexity is that in this area, there are deep problems similar to the fundamental problems in computational complexity. For instance, the problem of whether it is possible to prove every tautology in a Frege proof system using a polynomial number of formulas seems as hard as the problem of proving superpolynomial lower bounds on circuits computing explicitly defined Boolean functions. (The latter is still a widely open problem after 75 years of research.)

When discussing the lengths of propositional proofs, we have to specify what we mean by proofs. This is done by defining the important concept of *propositional proof systems*. In the rest of the paper, I will only say "proof system" because I will not consider proof systems for other logics.

The general concept of a proof system P only requires three properties:

(1) P is sound (proves only tautologies).

(2) P is complete (proves all tautologies).

(3) It is decidable in polynomial time whether a given string is a proof of a given formula.

Here, I will only consider proof systems that are fragments of a standard proof system based on axiom schemas and deduction rules. There are many such formalizations

that differ in the choice of the language, axioms, and rules, but they are all equivalent in a well-defined sense. The official name of this class of proof systems is *Frege systems*. In order to define bounded-depth fragments of Frege systems, we restrict the language to the connectives $\wedge, \vee, \neg$ and allow negations only at variables. Then we say that a formula ϕ has depth d if the number of alternations of ANDs and ORs is d. The *depth d Frege system,* F_d, is the subsystem in which only depth d formulas are allowed.

I will define more precisely the two systems at the bottom of this hierarchy, F_0 and F_1. The first one is usually presented as a *refutation system* and is called *Resolution*. In a refutation system, one starts with the negation of the assumptions and tries to derive a contradiction. Resolution is used to refute CNF formulas, which are conjunctions of disjunctions. The disjunctions are presented as sets of literals and called *clauses*, where *literals* are variables and negated variables. The CNF to be refuted is presented as a set of clauses. A refutation is a sequence of clauses derived from the clauses of the CNF formula and previously derived clauses. There is only one derivation rule, called the *resolution rule*, which is usually stated as follows:

$$\frac{C \cup \{x\} \quad D \cup \{\neg x\}}{C \cup D},$$

which means that from the two assumptions $C \cup \{x\}$, $D \cup \{\neg x\}$, conclusion $C \cup D$ follows. The aim is to derive the empty clause $\emptyset$ that represents a contradiction.

The system F_1 is an extension in which clauses may contain not only literals but also conjunctions of literals. It has an axiom $\{x, \neg x\}$ for every variable x and rules

$$\frac{C \cup \{\alpha \wedge z\} \quad D \cup \{\neg z\}}{C \cup D}, \qquad \frac{C \cup \{\alpha\} \quad D \cup \{z\}}{C \cup D \cup \{\alpha \wedge z\}}, \qquad \frac{C}{C \cup D},$$

where α is a (possibly empty) conjunction of literals, z is a literal, and $\neg z$ is the complementary literal.

3. Canonical and interpolation pairs of proof systems

Definition 3.1 (Razborov [28]). The *canonical pair* of a proof system P is the pair of two disjoint NP sets:

- $C_P^A = \{(\phi, 1^n)$; there a P-refutation of ϕ of length $\leq n\}$,
- $C_P^B = \{(\phi, 1^n)$; ϕ is satisfiable$\}$,

where ϕ is a propositional formula (a CNF without loss of generality) and 1^n, a string of 1s of length n, is a limit on the lengths of proofs given in unary.

I will denote $C_P := (C_P^A, C_P^B)$.

Informally, think of C_p^A as the set of formulas with short P-refutations and C_p^B as the set of satisfiable formulas. The "padding" 1^n is a convenient way to avoid saying what short means. It must also be in C_P^B because otherwise, it would be easy to distinguish elements of C_P^A from elements of C_P^B.

The concept of canonical pairs is a way to connect a proof system with a computational problem – instead of asking about the lengths of proofs, we can ask what the computational complexity of separating these disjoint NP-pairs is. The stronger the proof system is, the more difficult it is to separate the pairs. The intuition is that it is easier to separate two things that are far apart than those that are close to each other. In the canonical pairs, the second set is always the same, while the size of the first increases with the increasing strength of the proof system because more tautologies have short proofs.

We do not know whether the canonical pairs of F_d can be separated by a set in P. Building on our previous result [22], Bonet et al. [5] showed that for large depth d, the canonical pairs of F_d cannot be separated by a set in P if one assumes some strong conjecture from cryptography. There are no results of this kind based on standard conjectures about complexity classes such as $\mathsf{P} \neq \mathsf{NP}$.[3] The most interesting problem is the complexity of the canonical pair of F_0 (Resolution). Essentially, it is a win-win question: a polynomial separation of this pair would have application in SAT-solving and automated theorem proving, where Resolution is used; on the other hand, a superpolynomial lower bound may lead to new lower bounds in proof complexity.

Since it is unclear whether cryptographic hypotheses could help prove inseparability for low levels of Frege hierarchy, we can try the opposite: to find applications of the hypotheses that some canonical pairs are not polynomially separable in cryptography. The problem is, however, that cryptography needs sets in $\mathsf{NP} \cap \mathsf{coNP}$, not just a pair of disjoint NP sets.

The connection of canonical pairs with lower bounds on proofs is more general. To this end, I will introduce another concept.

Definition 3.2 ([25]). The *interpolation pair* of a proof system P is the pair of two disjoint NP sets:

- $I_P^A = \{(\phi_1, \phi_2, \Pi)$; $\mathrm{Var}(\phi_1) \cap \mathrm{Var}(\phi_2) = \emptyset$, Π is a P-refutation of $\phi_1 \wedge \phi_2$, and ϕ_1 is satisfiable$\}$,
- $I_P^B = \{(\phi_1, \phi_2, \Pi)$; $\mathrm{Var}(\phi_1) \cap \mathrm{Var}(\phi_2) = \emptyset$, Π is a P-refutation of $\phi_1 \wedge \phi_2$, and ϕ_2 is satisfiable$\}$,

[3] The famous result about nonautomatability of Resolution [2] implies that C_{Res}^A is NP-complete. The fact that both sets in a disjoint NP-pair are NP-complete, however, does not imply that they cannot be separated by a set in P.

where $\mathrm{Var}(\phi)$ denotes the set of variables present in ϕ.

I will denote $I_P := (I_P^A, I_P^B)$.

Informally, this concept captures the complexity of the problem of deciding which of the formulas ϕ_1 or ϕ_2 is satisfiable, with the promise that one of them is. It is intended to formalize the question of whether the method of feasible interpolation applies to a given proof system. This method works as follows. Let X_1 and X_2 be two disjoint NP sets. We can formalize the sets by a sequence of propositional formulas ϕ_1^n and ϕ_2^n, $n = 1, 2, \ldots$. Then their disjointness means that $\phi_1^n \wedge \phi_2^n$ is unsatisfiable. If I_P^A and I_P^B can be separated by a set in P, but X_1 and X_2 cannot, then there cannot be polynomial-size P-refutations of $\phi_1^n \wedge \phi_2^n$.

Of course, we do not know how to prove that some disjoint NP-pair is not polynomially separable, but often one shows that refutations of suitable formulas imply separation by *monotone* Boolean circuits and then uses known lower bounds on monotone Boolean circuits [27]. Furthermore, there are stronger monotone computation models which can be used to prove lower bounds using this method, one of which is monotone circuits that compute with real numbers [13, 19, 24]. My ultimate goal is to use combinatorial games, by which some interpolation pairs are characterized, as a sort of generalized monotone Boolean circuits to prove lower bounds for slightly stronger systems.

The two concepts defined above are closely related. For the Frege system, for example, the pairs are equivalent with respect to polynomial time reductions. For bounded-depth Frege systems, we have: $I_{\mathsf{F}_{d+1}}$ is equivalent to C_{F_d} for $d \geq 0$; see [4].

4. Games and circuits

The interpolation pairs, hence also canonical pairs, of F_d have been characterized using certain combinatorial games. To show how games can be used to characterize interpolation pairs, I will start with the simplest case, which is F_0 formalized as the Resolution refutation system.

A *basic game* is defined by

(1) a directed acyclic graph with a single source;

(2) the non-leaf nodes labeled by A and B;

(3) leaves labeled by a and b.[4]

[4] I call the nodes with no outgoing arrows *leaves*, although the graphs are not trees in general.

The game starts at the root, and at a node labeled A, Alice decides which arrow they will follow, and at a node labeled B, it is Bob who makes the decision. We define that the leaves labeled by a are winning positions for Alice and those labeled by b are winning positions for Bob.

The interpolation pair of F_1 is characterized by basic games in the following sense. Let G^A (resp. G^B) be the games in which Alice (Bob) has a winning strategy. Then the pair I_{F_1} is polynomially equivalent to the pair (G^A, G^B).[5]

In basic games, always one of the players has a winning strategy according to what is usually referred to as Zermelo's theorem. This simple fact is proved by showing by induction that at each node v, one of the players has a winning strategy when the game starts at v. The induction starts at the leaves and goes in the reverse order of the arrows. Note that this gives us a polynomial time algorithm to compute which player has a winning strategy. In fact, this algorithm can be presented as a circuit that computes with values a and b. To connect it with Boolean circuits, let $a := 0$ and $b := 1$. Then the circuit becomes a *monotone Boolean circuit* in which Alice's nodes are conjunctions and Bob's nodes are disjunctions. Clearly, the converse is also true: given a monotone circuit (with gates AND and OR) and an assignment to its inputs, we can interpret it as a basic game.

Game schemas. In order to identify basic games with monotone Boolean circuits, we should consider games in which leaves are left without labels and use strings of leaf labels as inputs. Let us call such structures *game schemas*. One can use this paradigm also for other games in which the winner is determined by a label on the terminal node. Thus, we can view game schemas as *generalized circuits*. In all these cases, the monotonicity of the computed functions is an intrinsic property – if Alice has a winning strategy and we enlarge the set of leaves labeled with a, then Alice obviously still has a winning strategy.

Positional strategies. Let us furthermore observe that if a player has a winning strategy, then they also have a *positional* winning strategy, which means that the instruction on how to play at a node v only depends on v, i.e., it is independent of the history of the play so far. The concept of a positional strategy is needed in characterizing the interpolation pairs of proof systems higher in the bounded-depth Frege system hierarchy. These characterizations use more complex games, in which the above fact is not true – there are games in which no player has a positional winning strategy.

[5] This is essentially the main result in [18], where it is stated for Boolean circuits instead of games (see also [24] for a simpler proof); furthermore, it is a special case of the main theorem in [26].

5. Circuits and PLS

The class TFNP is a class of relations $R(x, y)$ that satisfy the condition that for some polynomial p, for every x, there exists y such that $|y| \leq p(|x|)$ and $R(x, y)$ holds true. The best way to imagine a TFNP problem is as a *search problem* where a solution is guaranteed and can be checked in polynomial time. Unlike the typical complexity classes such as NP, it seems that there are no complete problems in TFNP. In the seminal paper [15], several natural subclasses have been identified, which do have complete problems, and more classes have been defined later.

One of the classes introduced in [15] is PLS, polynomial local search, which is connected with Boolean circuits. A problem in PLS is given by a directed acyclic graph with bounded outdegree and a single source node. The nodes of the graph are numbers, and if there is an arrow from i to j, then $i < j$. Furthermore, there is a *feasibility predicate*, a property of the nodes. The source is feasible, and for every feasible node which is not a leaf, one of the successors is feasible. A solution is a feasible leaf. If the graph is given explicitly and the feasibility condition is computable in polynomial time, then the solution can be found in polynomial time. Therefore, to make the problem nontrivial, we assume that the graph is given *succinctly*. This means that the graph is of exponential size, but the successors and the feasibility condition of nodes are computable in polynomial time.

There are two more ways to make the problem nontrivial: using *communication complexity* or using *query complexity*, also known as *black box model*. These two computational models are connected (see, e.g., [12]), but here, I will focus only on the communication complexity versions of the TFNP classes.

In the communication complexity version, instead of assuming that the graph is defined succinctly, we restrict the complexity of computing the primitives using bounds on communication complexity. So we will assume that there are two players, Alice and Bob, such that they are able to compute the feasibility of a node and the successors of the node by exchanging only a logarithmic number of bits. Thus, we get the class *communication complexity*-PLS, which I will abbreviate by ccPLS. Another name for this kind of communication complexity is *DAG-like* communication complexity, which is used to differentiate it from the standard communication complexity model in which the possible communication paths form a tree.[6]

6. From ccPLS to games

In communication complexity, we have two players, like in basic games, but otherwise the setup seems to be very different from basic games. In particular,

[6]DAG is an abbreviation for *directed acyclic graph*.

- in games, the players compete, while in communication complexity, they cooperate to find a feasible leaf;
- in communication complexity, the players have their own strings that they use, while there is nothing like that in games;
- and what should the criterion for winning be in a game corresponding to the communication protocol?

The key is to focus on one player, say, Alice. Alice has an input string x, and this completely determines her actions. Let S_x be the protocol that she follows, and let us view it as a *strategy*. So x is a code of a strategy.

Concerning the winning positions, note that we can always make S_x a *winning* strategy. The trivial way is to label all leaves with a. More interestingly, there is a minimum assignment of a's that makes S_x a winning strategy. Such an assignment u_x is defined by putting a on those leaves that can be reached when Alice plays her strategy S_x and Bob tries all possible ways to play the game.

We can, certainly, do the same for Bob. Thus, for each input y for Bob, we get his strategy T_y and the minimum assignment v_y to leaves that makes T_y winning. When Alice and Bob both play according to the communication protocol on the inputs x and y, they end up on a leaf that is labeled a in u_x and b in v_y.

Thus, we naturally arrived at the problem that was used to characterize monotone circuit depth by Karchmer and Wigderson in [17] and circuit size by Razborov in [28]. In the setup of these papers, a monotone Boolean function f and two inputs x and y such that $f(x) = 0$ and $f(y) = 1$ are given. The goal of the players is to find an index such that $x_i = 0$ and $y_i = 1$. Let us call the problem they are solving the *Karchmer–Wigderson problem*. In the theory of Boolean functions, the minimal weight strings for which a monotone function is zero are called *minterms*; the dual concept is called *maxterms*. Our u_x correspond to the minterms and v_y to the maxterms.

To get games or circuits, we still need to extend the communication graph. In the DAG-like communication protocol, the players decide which node to proceed to from the current one by communicating $O(\log n)$ bits. From a node U, they can get to polynomially many nodes V_i. The problem is not that this is more than two, but the fact that the decision is not made by one player. So we represent this communication protocol by a tree T_U and connect its root to U and leaves to the nodes V_i. The tree represents communication of the players in which one player sends one bit, hence decides to which successor in T_U the computation proceeds. So it suffices to insert such a tree for every node of the communication graph to get a basic game.

In this way, we get some game schema G, but it may not relate to the problem the communication protocol solved. Suppose the protocol solved the Karchmer–Wigderson problem for a monotone function f. Let $X := \{x \ ; \ f(x) = 1\}$ and $Y := \{y \ ; \ f(y) = 0\}$. Assume X are inputs given to Alice, and Y are inputs to

Bob. Let α_x, for $x \in X$, be Alice's strategies for playing with given x and β_y, for $y \in Y$, be Bob's strategies for playing with y. The fact that the communication game solved the Karchmer–Wigderson problem ensures that for $x \in X$, Alice's strategy α_x beats all Bob's strategies β_y, $y \in Y$, in game (G, x) (G with the winning positions determined by 1's in the vector x). But this does not imply that α_x is a winning strategy in game (G, x) because Bob may play a strategy that is not among the strategies β_y. Nevertheless, it suffices to omit some arrows and nodes from G to get a game schema G' such that α_x is a winning strategy in (G', x). This may look surprising because in G', Bob is not forced to play one of the strategies β_y. The reason why it works is a property of the feasibility predicates. When we extend the communication protocol by trees, the players only need to send one bit to each other in order to check the feasibility predicate $F_v(x, y)$ at a given node v. This implies that the predicate is a conjunction of two predicates $F_v^A(x) \wedge F_v^B(y)$, which can be computed separately by Alice and Bob. Consequently, if x, x' are two inputs for Alice and y, y' for Bob, we have

$$F_v(x, y) \wedge F_v(x', y') \Rightarrow F_v(x, y') \wedge F_v(x', y). \tag{6.1}$$

In terms of strategies, $F_v(x, y)$ is equivalent to the condition that the path determined by strategies α_x and β_y leads to a leaf u that is labeled a in x and b in y.

Lemma 6.1. *Let G' be the game schema that results from G by omitting all arrows that do not occur on any path defined by a pair of strategies α_x, β_y for $x \in X, y \in Y$.*[7] *Then α_x is a winning strategy in (G', x) for every $x \in X$, and β_y is a winning strategy in (G, y) for every $y \in Y$.*

Proof. Because of symmetry, it suffices to consider $x \in X$. I will show that every node u in G' satisfies the following condition:

- ⋆ If u is feasible for (x, y), then u is a winning position for Alice in (G, x) with α_x being a winning strategy.

The proof is by induction, starting from the leaves and going toward the root.

Let u be a (x, y)-feasible leaf for some $y \in Y$. Since the communication protocol solves the Karchmer–Wigderson problem for f, the value of x on u must be a. Hence u is a winning position in (G, x).

Let u be a non-leaf. Suppose u belongs to Alice, and let v be the node to which Alice will go playing α_x. If u is (x, y)-feasible, then so is v. By the induction assumption, v is a winning position for Alice with the winning strategy α_x, so the same holds for u.

[7] To formally satisfy the definition of a basic game, we should also delete all nodes that become not reachable from the root.

Let u belong to Bob and v, w be its successors in G'. Suppose u is (x, y)-feasible. Then either v or w is (x, y)-feasible; let it be v. Since the arrow $u \to w$ has not been deleted, there is a path defined by some pair $\alpha_{x'}, \beta_{y'}$ that has the arrow $u \to w$. By property (6.1), u is also (x, y')-feasible. Since strategy $\beta_{y'}$ points to w, w is (x, y')-feasible. Hence, by the induction assumption, both v and w are winning positions for Alice, and, consequently, v is also a winning position.

I leave the case when u belongs to Bob and has only one successor to the reader. ∎

What I have done is a proof of Razborov's result, mentioned before, presented in terms of games.

Theorem 6.2 (Razborov [28]). *The minimum size of a monotone circuit computing a monotone Boolean function f is equal, up to a polynomial factor, to the minimum size of a graph in a DAG-like communication protocol solving the Karchmer–Wigderson problem for f.*

7. More players

I will now digress to show a potential application of a generalization of basic games to more players. In communication complexity, protocols with more players are considered, called *multiparty communication complexity*. There are two main versions, called *number in hand* and *number on the forehead*. In both versions, the input is split into k parts, where k is the number of players. In the number-in-hand model, each player sees only one part that is assigned to them. In the number-on-the-forehead model, each player sees all parts except their own.

We can study both versions using games. For the sake of simplicity, I will consider the number-in-hand model with three players, Alice, Bob, and Cindy. The generalization of the basic game to three players is straightforward. The inner nodes are now labeled by A, B, and C, the leaves by a, b, and c. The corresponding game schema is without labels on the leaves.

With more than two players, Zermelo's theorem does not hold true. Nevertheless, we can still compute who has a winning strategy, if there is a player who has such a strategy. To this end, we will need four values a, b, c, and 0, where 0 denotes that at the node nobody has a winning strategy. Thus, we can define circuits corresponding to these game schemas. The following proposition characterizes the problems that can be studied using this concept.

Proposition 7.1. *Let $W_A, W_B, W_C \subseteq \{a, b, c\}^n$. Then there exists a game schema G such that Alice has a winning strategy when we label the leaves with W_A, Bob has a*

winning strategy when we label the leaves with W_B, and Cindy has a winning strategy when we label the leaves with W_C if and only if the following condition holds true:

- *For all $\alpha \in W_A, \beta \in W_B, \gamma \in W_C$, there exists $i \in [n]$ such that*

$$\alpha(i) = a \wedge \beta(i) = b \wedge \gamma(i) = c.$$

Clearly, if $\alpha \in W_A$ (similarly if $\beta \in W_B$ or $\gamma \in W_C$), then only the values on indices i such that $\alpha(i) = a$ matter. Therefore, it is more natural to use *sets* of indices instead of vectors in $\{a, b, c\}^n$.

Intuitively, one should get larger lower bounds when there are more players, because the players know less about the input and hence must communicate more. So far, this idea has been used only in some special cases.

Problem 7.2. Prove, for explicitly given sets W_A, W_B, W_C, a larger lower bound on the size of these circuits than we have for monotone Boolean circuits!

Example. In this example, we will have k players. Let $n := 2^k + 1$. The input vectors will be edge colorings by k colors of the complete graph on n vertices. For $j = 1, \ldots, k$, the winning set of the jth player W_j is the set of colorings such that the graph defined by the edges with color j contains two cliques that cover all vertices. One can easily see that these sets satisfy the condition from Proposition 7.1 above, namely, for any string of colorings $(\kappa_1, \ldots, \kappa_k), \kappa_1 \in W_1, \ldots, \kappa_k \in W_k$, there exists an edge e such that $\kappa_1(e) = 1, \ldots, \kappa_k(e) = k$. By the generalization of Proposition 7.1, there exists a game in which sets W_i are the winning sets of players i for all i. Proving a lower bound on the size of such games maybe a way to get better lower bounds for the well-known tautology called *BitPHP*. For tree-like cutting plane proofs, such an improvement was shown by P. Beame and M. Whitmeyer in [3].[8]

Lower bounds on the number-on-forehead multiparty communication complexity are notoriously hard but could resolve important open problems in both circuit complexity and proof complexity. Therefore, it would be more interesting to prove lower bounds on the size of the corresponding games. In game theory, the number-on-forehead model corresponds to the same games except that instead of asking about winning strategies of single players, we ask about winning strategies of coalitions of all players except one.[9]

[8]The cutting planes proof system is stronger than Resolution but incomparable with higher fragments of constant-depth Frege systems.

[9]Note that a player i does not have a winning strategy if and only if the coalition of all players j, $j \neq i$, has a winning strategy. So these problems are, in a sense, dual to each other.

8. The interpolation pair of F_1

Recall that F_1 is a generalization of Resolution in which we allow conjunctions in the clauses. Furthermore, the interpolation pair of F_1 is equivalent to the canonical pair of Resolution. The first combinatorial characterization of this pair appeared in [4] and was based on the *point-line game* introduced in that paper.

The setup of a point-line game consists of

- the board, which is a directed acyclic graph with one source, and all nodes of outdegree 2, except for leaves;
- the nodes of the graph are containers with slots – *the points*;
- the source node is empty, each leaf node has one point;
- if $U \to V$ are nodes connected by an arrow, then there is a partial mapping from the points in V to the points in U – *the lines*;
- there are two players, Red and Blue, and each node is owned by one player and marked by the color of the owner.

The game is played as follows:

- starting at the source, the players follow the arrows, and the color of the node decides who will choose the arrow;
- they move red and blue pebbles along the lines;
- when they move from a red node U to some node V, and there are points in V that are not connected to points in U, they will be covered by blue pebbles; similarly for a blue node U, the free points in V will be covered by red pebbles; thus, all points in the currently used nodes are filled with pebbles;
- the player whose pebble appears in the leaf wins.

See the example in Figure 1.

A *positional strategy* of the Red player is a set of arrows, one for each red node; similarly for the Blue player.

Lemma 8.1. *Given a set of arrows S, it is decidable in polynomial time whether S is a winning positional strategy of the Red player. Hence the set of games with winning positional strategies for the Red player is in* NP. *The same for the Blue player.*

Theorem 8.2 (Beckmann, Pudlák, Thapen [4]). *Let S^R (respectively, S^B) be the set of games in which the Red (Blue) player has positional winning strategies. Then the interpolation pair of F_1 and (S^R, S^B) are polynomially reducible to each other.*

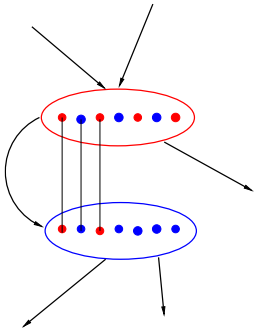


Figure 1. Red decided to move to the blue node. The first three pebbles are moved along the lines; the rest of the bottom node is covered by blue pebbles.

The condition that the strategies must be positional seems to be ad hoc in order to get NP-pairs; however, it naturally appears in the proof of the theorem, and we will see another reason for it in a moment.

The point-line game without pebbles. It is important to realize that the point-line game can be played without pebbles. The difference is that such a game does not end at a leaf that the players reach, because they still have to determine the winner. To this end, the players go back along the path by which they arrived at the leaf and follow the predecessors of the unique point on the leaf. Eventually, they reach a node such that the point does not have a predecessor in the preceding node. This is the place where the color of the pebble would be decided if they played with pebbles. If the preceding node is red, the color of the pebble would be blue, and vice versa. This pebble would go all the way to the leaf.

Note that in such a game, the players have to record the path by which they arrived at a leaf. If they use pebbles, they do not have to do that because the information they need is encoded by the color of the pebbles.

The point-line game schema. Using colors of certain nodes as inputs in order to compute a partial Boolean function is rather awkward, and it would be difficult to analyze such a computational model. A more natural way is to allow points in the source node and use them to place different strings of colored pebbles. The simulation of the original games by this modification is easy, and I leave it to the reader. So *the point-line game schema* is like the point-line game except that we allow points in the source node, and these points are used as placeholders for input bits represented by pebbles of two colors.

9. The point-line game and the class CPLS

The class *colored polynomial local search,* CPLS, is an extension of PLS introduced by Krajíček, Skelley, and Thapen in [23]. This class is connected with the point-line game in the same way as PLS is connected with the basic game. To describe this connection, I will present CPLS as a point-line game played by *one player* on an *exponentially large board* and *without pebbles*. The goal of the game is to find a point without a predecessor in a predecessor node. In more detail, as follows.

The setup is

- the nodes and the points are given by numbers $\leq 2^n$ (n is the input size parameter); arrows go from a smaller number to a larger one; node 0 has no point; each leaf has a unique point;
- the arrows, lines, and points are given by polynomial time algorithms as follows:
  - (1) for a node U, an algorithm computes the successors of U, or outputs U *is a leaf*;
  - (2) for two nodes $U \to V$ and a point $a \in V$, an algorithm computes the predecessor of a in U, if there is any, otherwise outputs *no predecessor in* U;
  - (3) for a leaf of the graph S, an algorithm computes the unique point in S.

The goal is

- to find a triple $U \to V$, $a \in V$ such that a has no predecessor in U.

These triples are, moreover, associated with some elements, which are outputs. The outputs are not relevant to what I want to do because I only want to transform the problem into a game schema.

To transform CPLS into a communication complexity class, we proceed in the same way as in the case of PLS; i.e., we use a polynomial domain instead of the exponential one, and let the two players compute the arrows, lines, and points using small communication protocols.

The basic idea of the transformation of ccCPLS into a game is the same as for ccPLS; i.e., we interpret the protocols for players as their winning strategies. The transformation of the graphs of the communication protocol into a graph of the game is, however, more complicated because we have, in fact, two graphs: one whose vertices are, what we call, nodes, and another whose vertices are points and edges are lines.

The graph of nodes can be treated the same way as we did with ccPLS. Thus, for every node U, we replace the outgoing arrows with a tree T_U that represents possible communication paths to determine the successor.

Concerning the points and lines, it is more difficult to simulate communication by playing a point-line game. The problem is that, for two nodes $U \to V$ and a point $a \in V$, we need to compute a neighbor of a in U, and this is in the opposite direction. In the point-line game without pebbles, players also go in the opposite direction, but they just follow the line and do not make any decisions. Therefore, I will consider a generalization of the game in which, for every $a \in V$, there is a basic game G_a with the root at a and leaves being some points of U. Then the simulation is easy: use the communication trees as these basic games.

Such a generalization differs from point-line games considerably. We can interpret the nodes of G_a as points in V, but then we will have lines inside V. To make it more like the original game, we can do the following. Let us call the nodes of the basic games G_a points and the arrows lines. Insert a chain of nodes between U and V, $V_0 = U, V_1, \ldots, V_{k-1}, V_k = V$, where k is the maximal depth of the trees that are the graphs of G_a, and let the arrows only be $U \to V_1 \to \cdots \to V_{k-1} \to V$. Then put the points of all G_a into the nodes of the chain so that the lines are only between the consecutive nodes. One may need to add additional points and lines when the branches are shorter than k. Then it remains to deal with the markings that associate players with points in G_a. We can assume, without loss of generality, that in all G_a, the same player starts, and they alternate regularly. So we can assign colors to nodes V_i so that they correspond to the colors of the points in these games. As a result, we can remove the markings of the points.

To complete the full circle from point-line games to CPLS, to communication complexity CPLS, and back to point-line games, one would need to prove a type of result similar to Razborov's theorem, Theorem 6.2. I believe it is possible, but for now, I leave it for future research.

Remark 1. One can check that such games are essentially the same as the Turing machine games of depth 2 that I will define in the next section. Since Turing machine games of depth 2 characterize the same interpolation pair as point-line games, they are polynomially reducible to each other. Thus, we get point-line games from ccCPLS.

Remark 2. One thing to note is that in this way, we necessarily get *positional* strategies, which justifies the condition that we imposed on strategies when defining the point-line game. Being positional in general means that the strategy only uses "local" information. The locality is an intrinsic property of the communication complexity versions of TFNP problems. Therefore, we always get positional strategies from TFNP problems.

10. Games for all levels of the bounded-depth Frege hierarchy

I will present a characterization of the interpolation pairs of F_d for all d from [26]. It is based on games that I call simply depth d games. Here, I will call them *Turing machine games of depth d*.

Definition 10.1. A *Turing machine (TM) game of depth d* is given by (M, n, A, R, B), where

- M is a nondeterministic Turing machine;
- n is the length of a finite tape;
- A is a finite, but possibly large, alphabet used by the machine;
- (R, B) is a partition of A into winning positions for Red and Blue players, respectively.

The game is played as follows:

- initially, the tape is blank;
- starting at the leftmost square, the machine moves forth and back from one end to the other d times while reading and rewriting the symbols on the tape;[10]
- two players, Red and Blue, alternate in deciding which action the nondeterministic machine takes (i.e., what it prints on the tape and which state the automaton takes on);
- the last printed symbol decides who the winner is.

A possible situation is depicted in Figure 2.

Admittedly, the machine in the definition is not what we usually call a Turing machine because in the standard definition, the size of the alphabet and the number of states of the automaton are bounded by a constant. It is possible to slightly change the definition to achieve the constant bound, but the resulting concept would not be natural.

Definition 10.2. The *size of a TM game* is the number of states of $M + |A| + n$. A *positional strategy* for a TM game is a strategy that only uses the number of the round, the state of M, and the symbol on the tape read by M.

Lemma 10.3 ([26]). *For every $d \geq 1$, there exists a polynomial time algorithm (polynomial in the size of M) to decide if a given positional strategy is a winning strategy.*

[10] A machine which moves its head independently of the content of the tape and its state is called *oblivious*, so the machine in the definition is oblivious.

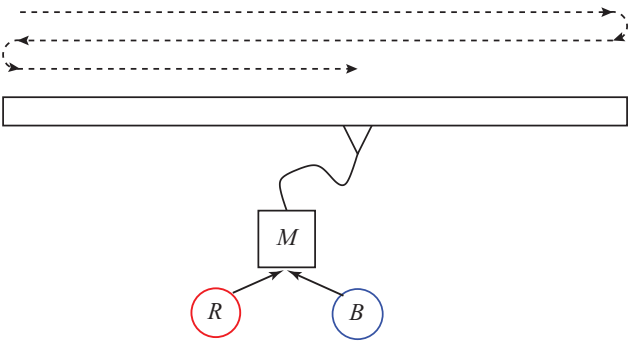


Figure 2. Turing machine M, controlled by players R and B, in the middle of the third round moves to the right.

Hence the existence of a positional strategy for a given player is an NP predicate. Unfortunately, my proof of this lemma only gives algorithms in which the exponent of the polynomial bound grows with increasing d. Thus, I am not able to use these games to characterize the interpolation pair of the unbounded-depth Frege system.

Let W_d^R (W_d^B) denote the set of depth d TM games with winning positional strategies for the Red (respectively, Blue) player. The characterization of the interpolation pairs is given by the following theorem.

Theorem 10.4 ([26]). *For all $d \geq 1$, the interpolation pairs of F_d (the depth d Frege system) and (W_d^R, W_d^B) are polynomially reducible to each other.*

It is not difficult to see that the depth 1 TM game is the basic game: given the machine, the nodes of the basic game are triples (x, s, i), where i is the position of the head, x is the symbol on the ith square, and s is the state. To reduce the point-line game to the depth 2 TM game is also easy if we take the version with all pebbles black. The machine going in one direction follows the arrows of the graph; going in the opposite direction follows the lines. The converse reduction can be done using the generalized version of Section 9.

The paradigm used in the proof of Theorem 10.4 is the search for a falsified clause in a Resolution refutation when the variables are split between two players. In such a situation, always one of the players can control the step in the search procedure because the assumptions of the resolution rule contain only one variable which is not in the conclusion. So the player who knows the value of this variable can decide which of the assumptions is false, assuming the conclusion is false. For proofs of depth $d > 0$, I use the key idea of Skelley and Thapen [29] that one can define a calculus which is equivalent to F_d and in which the only essential rule is the resolution rule. Such a calculus cannot be the standard sequent calculus with cuts restricted to propositional variables because this restriction would make it very weak. What they

actually did was to use *deep inference rules*. These rules allow one to derive a new formula by changing a given formula inside; in other words, the rules can be applied *inside formulas*.

In my proof, I use a different calculus with two deep inference rules that I call resolution, which is the usual resolution rule applied inside formulas, and *dual resolution*. Both rules can be applied inside formulas:

$$\frac{(A \vee p) \wedge (B \vee \neg p)}{A \vee B}, \qquad \frac{A \wedge B}{(A \wedge p) \vee (B \wedge \neg p)}.$$

The Turing machine, roughly speaking, searches for unsatisfied formulas of the highest depth d, which is the depth of the proof. Then it reverses the order of search and searches their unsatisfied subformulas of depth $d - 1$ and so on until it finds an unsatisfied initial clause. When the TM goes in one direction, it uses instances of the resolution rule to search for an unsatisfied formula, the same way as it is done when we use interpolation for Resolution; when it goes in the opposite direction, it uses dual resolution.

11. Some related combinatorial games

Various games have been studied in game theory. Here, I want to mention only two that are related to the games presented in this article. In contrast to our games, these games are played on directed graphs that are in general not acyclic; thus, the number of rounds is infinite.

Parity games. A parity game is given by a directed graph whose set of nodes is $\{0, 1, 2, \dots, n\}$, and for each node, there is at least one outgoing edge. The set of nodes is divided into two parts, V_0 and V_1, which we associate with two players, Player 0 and Player 1. Without loss of generality, we can assume that V_0 are even numbers and V_1 are odd. A play starts with a pebble on 0, and then players move the pebble along the arrows of the graph. When the pebble is on a node in V_i, then Player i moves the pebble. After they play infinitely many rounds, the winner is decided by the least node v that was visited infinitely many times: if $v \in V_i$, then Player i wins.

E. A. Emerson showed that in a parity game, always one of the players has a *positional* winning strategy [11]. In such a strategy, for every node, the player always uses the same arrow. This can be used to define the games in which Player i has a winning strategy by an NP condition. In 2017, C. S. Calude, S. Jain, B. Khoussainov, W. Li, and F. Stephan found a quasi-polynomial algorithm to decide who has a winning strategy in parity games [7].

Simple stochastic games. These games were introduced by A. Condon [8] as a special case of stochastic games studied before. The game is given by a directed graph whose nodes are divided into three sets $V_{\min}, V_{\max}, V_{\text{ave}}$ and with three distinguished nodes s, v_0, v_1. Starting at node s, players Min and Max move the pebble if it is at a node assigned to them. If the pebble is on a node $v \in V_{\text{ave}}$, the pebble is moved to a child of v randomly. We assume that the outdegree of every node is at least 1, except for the sinks v_0, v_1. The aim of Min (resp. Max) is to reach v_0 (resp. v_1). Given a pair of positional strategies σ for Min and τ for Max, we can compute the probability that they reach v_1. The *value of the game* is inf sup of these probabilities over all σ, τ. The decision problem for simple stochastic games is to determine whether the value is $> 1/2$.

No truly subexponential algorithms are known for the decision problem for simple stochastic games. Condon [8] proved that the problem is in $\mathsf{NP} \cap \mathsf{coNP}$. In 2011, K. Chatterjee and N. Fijalkow strengthened this result to $\mathsf{UP} \cap \mathsf{coUP}$ [9]. Let us note that for parity games, this was shown by M. Jurdziński [16], who did it by reducing the problem to finding the unique fixed points of certain contractive maps, which in turn was first done by U. Zwick and M. Paterson [30].

In [4], we showed that the decision problem for parity games can be reduced to the interpolation pair of F_1, hence also to point-line games and Turing machine games of depth 2. Strengthening the results of A. Atserias and E. Maneva [1] for mean payoff games,[11] L. Huang and T. Pitassi [14] proved that the decision problem for simple stochastic games is reducible to the interpolation pair of F_2. In [4], we gave a more direct proof of this result. We observed that our proof does not use the full power of F_2, but we were not able to improve it to F_1. Below, we restate these reductions in terms of Turing machine games.

Theorem 11.1. *The decision problem for parity games is polynomially reducible to deciding who has a* positional *winning strategy in depth* 2 *Turing machine games. The decision problem for simple stochastic games is polynomially reducible to deciding who has a* positional *winning strategy in depth* 3 *Turing machine games.*

12. Conclusions and open problems

In this paper, I have focused on the relation of games and communication complexity versions of two TFNP classes. These games characterize the interpolation pairs of F_0 and F_1 and the TFNP classes that correspond to theories T_2^1 and T_2^2. It seems

[11] I will not define these games here.

that this connection can be extended to the interpolation pairs of all F_i and the game induction principles GI_i introduced in [29] that are associated with theories T^i_2. This raises the natural question of whether one can get in such a way a game from any TFNP class, for which the communication complexity version makes good sense. The answer is that one always gets *some* game, but it may not be what we expect. In the two examples above, ccPLS and ccCPLS, we were able to incorporate the small trees representing the communications at nodes into the original graph. Adding these trees may essentially distort the principle for TFNP classes that are not based on DAGs.

There are more connections between proof systems and various computational models. In particular, the black box (a.k.a. query complexity) versions of TFNP have been associated with all important proof systems studied in proof complexity, and for several proof systems also classes of generalized monotone circuits have been found [6, 12, 21]. An approach to feasible interpolation for the bounded-depth Frege system, different from the one here, was proposed in [20]. This suggests a lot of questions, but I confine myself to a few relevant to this paper's topic.

(1) What is the computational complexity of deciding who has a positional winning strategy in a given depth d TM game, assuming that one player has a positional winning strategy?

(2) For $d > 1$, prove a lower bound on the complexity of depth d TM games computing an explicitly given partial *monotone* Boolean function.

(3) In particular, for some $d > 1$ (ideally for all), prove that (W^R_{d+1}, W^B_{d+1}) cannot be polynomially reduced to (W^R_d, W^B_d) by a *monotone* projection.

(4) Characterize the canonical pair of the unbounded-depth Frege system.

Furthermore, I presented connections between games and communication complexity rather informally. My goal was to show that games offer an alternative view of open problems in proof complexity, but more precise definitions and proofs are needed, which I will do in subsequent publications.

Acknowledgments. I would like to thank Paul Beame, Emil Jeřábek, Jan Krajíček, Neil Thapen, and an anonymous referee for their comments on a draft of this paper.

Funding. This work was supported by the Czech Academy of Science (RVO 67985840) and GAČR grant 23-04825S.

References

[1] A. Atserias and E. Maneva, Mean-payoff games and propositional proofs. *Inform. and Comput.* **209** (2011), no. 4, 664–691 Zbl 1237.91043 MR 2790836

[2] A. Atserias and M. Müller, Automating resolution is NP-hard. In *2019 IEEE 60th annual symposium on foundations of computer science (FOCS)*, pp. 498–509, IEEE, Los Alamitos, CA, 2019 Zbl 08078792 MR 4228185

[3] P. Beame and M. Whitmeyer, Multiparty communication complexity of collision-finding and cutting planes proofs of concise pigeonhole principles. In K. Censor-Hillel, F. Grandoni, J. Ouaknine, and G. Puppis (eds.), *Proceedings of the 52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, pp. 21:1–21:20, LIPIcs. Leibniz Int. Proc. Inform. 334, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, 2025 MR 4934173

[4] A. Beckmann, P. Pudlák, and N. Thapen, Parity games and propositional proofs. *ACM Trans. Comput. Log.* **15** (2014), no. 2, article no. 17, 30 pp. Zbl 1291.03111 MR 3205545

[5] M. L. Bonet, T. Pitassi, and R. Raz, On interpolation and automatization for Frege systems. *SIAM J. Comput.* **29** (2000), no. 6, 1939–1967 Zbl 0959.03044 MR 1756400

[6] S. Buss, N. Fleming, and R. Impagliazzo, TFNP characterizations of proof systems and monotone circuits. In Y. Tauman Kalai (ed.), *14th innovations in theoretical computer science conference (ITCS 2023)*, pp. 30:1–30:40, LIPIcs. Leibniz Int. Proc. Inform. 251, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, 2023 Zbl 07918351 MR 4587044

[7] C. S. Calude, S. Jain, B. Khoussainov, W. Li, and F. Stephan, Deciding parity games in quasi-polynomial time. *SIAM J. Comput.* **51** (2022), no. 2, STOC17-152–STOC17-188 Zbl 1542.68075 MR 4413072

[8] A. Condon, On algorithms for simple stochastic games. In J.-Y. Cai (ed.), *Advances in computational complexity theory*, pp. 51–71, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 13, American Mathematical Society, Providence, RI, 1993 Zbl 0808.90141 MR 1255328

[9] K. Chatterjee and N. Fijalkow, A reduction from parity games to simple stochastic games. In G. D'Agostino and S. La Torre (eds.), *Proceedings of the 2nd international symposium on "Games, Automata, Logics and Formal Verification" (GandALF 2011)*, pp. 74–86, Electron. Proc. Theor. Comput. Sci. (EPTCS) 54, Open Publishing Association, Waterloo, 2011 Zbl 1460.91027

[10] S. A. Cook, Feasibly constructive proofs and the propositional calculus (preliminary version). In *Proceedings of the 7th annual ACM symposium on theory of computing (Albuquerque, NM, 1975)*, pp. 83–97, Association for Computing Machinery, New York, 1975 Zbl 0357.68061 MR 0502226

[11] E. A. Emerson, Automata, tableaux, and temporal logics. In R. Parikh (ed.), *Logics of programs (Brooklyn, NY, 1985)*, pp. 79–88, Lecture Notes in Comput. Sci. 193, Springer, Berlin, 1985 Zbl 0603.03005 MR 0808790

[12] M. Göös, P. Kamath, R. Robere, and D. Sokolov, Adventures in monotone complexity and TFNP. In A. Blum (ed.), *10th innovations in theoretical computer science conference (ITCS 2019)*, pp. 38:1–38:19, LIPIcs. Leibniz Int. Proc. Inform. 124, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern, 2019 Zbl 1547.68237 MR 3899832

[13] P. Hrubeš and P. Pudlák, A note on monotone real circuits. *Inform. Process. Lett.* **131** (2018), 15–19 Zbl 1422.68115 MR 3734966

[14] L. Huang and T. Pitassi, Automatizability and simple stochastic games. In L. Aceto, M. Henzinger, and J. Sgall (eds.), *Automata, languages and programming. Part I*, pp. 605–617, Lecture Notes in Comput. Sci. 6755, Springer, Heidelberg, 2011 Zbl 1334.68100 MR 2874140

[15] D. S. Johnson, C. H. Papadimitriou, and M. Yannakakis, How easy is local search? *J. Comput. System Sci.* **37** (1988), no. 1, 79–100 Zbl 0655.68074 MR 0973658

[16] M. Jurdziński, Deciding the winner in parity games is in UP ∩ co-UP. *Inform. Process. Lett.* **68** (1998), no. 3, 119–124 Zbl 1338.68109 MR 1657581

[17] M. Karchmer and A. Wigderson, Monotone circuits for connectivity require super-logarithmic depth. *SIAM J. Discrete Math.* **3** (1990), no. 2, 255–265 Zbl 0695.94021 MR 1039296

[18] J. Krajíček, Interpolation theorems, lower bounds for proof systems, and independence results for bounded arithmetic. *J. Symb. Log.* **62** (1997), no. 2, 457–486 Zbl 0891.03029 MR 1464108

[19] J. Krajíček, Interpolation by a game. *MLQ Math. Log. Q.* **44** (1998), no. 4, 450–458 Zbl 0915.03047 MR 1654289

[20] J. Krajíček, A form of feasible interpolation for constant depth Frege systems. *J. Symb. Log.* **75** (2010), no. 2, 774–784 Zbl 1194.03052 MR 2648165

[21] J. Krajíček, Randomized feasible interpolation and monotone circuits with a local oracle. *J. Math. Log.* **18** (2018), no. 2, article no. 1850012, 27 pp. Zbl 1522.03321 MR 3878474

[22] J. Krajíček and P. Pudlák, Some consequences of cryptographical conjectures for S^1_2 and EF. In D. Leivant (ed.), *Logic and computational complexity (Indianapolis, IN, 1994)*, pp. 210–220, Lecture Notes in Comput. Sci. 960, Springer, Berlin, 1995 Zbl 1540.03105 MR 1449663

[23] J. Krajíček, A. Skelley, and N. Thapen, NP search problems in low fragments of bounded arithmetic. *J. Symb. Log.* **72** (2007), no. 2, 649–672 Zbl 1118.03051 MR 2320295

[24] P. Pudlák, Lower bounds for resolution and cutting plane proofs and monotone computations. *J. Symb. Log.* **62** (1997), no. 3, 981–998 Zbl 0945.03086 MR 1472134

[25] P. Pudlák, On reducibility and symmetry of disjoint NP pairs. *Theoret. Comput. Sci.* **295**, (2003), no. 1–3, 323–339 Zbl 1045.68058 MR 1964673

[26] P. Pudlák, The canonical pairs of bounded depth Frege systems. *Ann. Pure Appl. Logic* **172** (2021), no. 2, article no. 102892, 41 pp. Zbl 1498.03152 MR 4161762

[27] A. A. Razborov, Lower bounds on the monotone complexity of some Boolean functions (in Russian). *Dokl. Akad. Nauk SSSR* **281** (1985), no. 4, 798–801. English translation. *Dokl. Math.* **31** (1985), 354–357 Zbl 0621.94027 MR 0785629

[28] A. A. Razborov, Unprovability of lower bounds on circuit size in certain fragments of bounded arithmetic (in Russian). *Izv. Ross. Akad. Nauk Ser. Mat.* **59** (1995), no. 1, 201–224. English translation. *Izv. Math.* **59** (1995), no. 1, 205–227 Zbl 0838.03045 MR 1328561

[29] A. Skelley and N. Thapen, The provably total search problems of bounded arithmetic. *Proc. Lond. Math. Soc. (3)* **103** (2011), no. 1, 106–138 Zbl 1225.03081 MR 2812503

[30] U. Zwick and M. Paterson, The complexity of mean payoff games on graphs. *Theoret. Comput. Sci.* **158** (1996), no. 1–2, 343–359 Zbl 0871.68138 MR 1388974

Pavel Pudlák
Institute of Mathematics, Czech Academy of Sciences, Žitná 25, 11567 Prague,
Czech Republic; pudlak@math.cas.cz

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/31

The $K(\pi, 1)$-conjecture for Artin groups

Mario Salvetti

Abstract. An Artin group G_W is the fundamental group of the space Y_W, which consists of the regular orbits of a Coxeter group W acting on the complexified Tits cone. The best-known example occurs when W is the symmetric group, in which case G_W is the classical braid group, and Y_W is the *configuration space* of n distinct points in $\mathbb{C}$. A famous long-standing conjecture, attributed to Arnol'd, Brieskorn, Pham, and Thom, asserts that the orbit space Y_W is a $K(\pi, 1)$-space. For finite W, the conjecture was proved by Deligne (1972) in a landmark paper. Recently, in collaboration with Giovanni Paolini (2021), we proved the conjecture for the next significant class of Artin groups of *affine type*.

We discuss various aspects of our proof, providing an overview of the main methods while avoiding the numerous technical details. In particular, we focus on combinatorial methods applied to the so-called *dual approach* to Artin groups. The latter involves the development of *Garside theory* and its application to *non-crossing partition* intervals. Finally, we highlight some naturally related problems and conjectures that can be formulated for any Artin group. Delucchi, Paolini, and Salvetti (2024) have recently resolved these conjectures affirmatively for all rank-three groups.

1. Introduction

We consider Coxeter groups and their associated Artin groups (precise definitions are provided below). Since Deligne's landmark paper [14], which established (among other results) the $K(\pi, 1)$-property for the orbit configuration space of spherical-type Artin groups, the topological study of this important class of groups has seen only sporadic progress. However, in recent years, new significant results have rekindled strong interest in this area. In collaboration with Giovanni Paolini, we proved the $K(\pi, 1)$-conjecture for all the affine Artin groups [29]. More recently, further advancements have been achieved using different approaches, primarily based on Gromov's theory of hyperbolic groups [19, 21].

Mathematics Subject Classification 2020: 20F36 (primary); 32S22, 57Q70, 55P20 (secondary).
Keywords: Artin groups, Coxeter groups, Eilenberg–MacLane spaces.

In this paper, an extended exposition of my talk at the ECM, we focus on our approach to the $K(\pi, 1)$-conjecture. After providing all necessary definitions and stating the conjecture in its general form, we introduce the so-called *dual approach* to the theory. Within this framework, one considers all reflections R as generators of the Coxeter group W, which in turn defines a length function ℓ_R and induces a poset structure on W. We then consider the interval $[1, c]$ in this poset, consisting of all elements less than or equal to a fixed Coxeter element c. This interval is commonly referred to as the *non-crossing partition* interval, a term originating from the symmetric group case, as we explain later.

Starting from the non-crossing partition interval, we define a *dual Artin group* and an *interval complex*. The interval complex is a *CW*-complex whose fundamental group is the dual Artin group. A key property in Garside theory is when the associated poset forms a *lattice*. This lattice property ensures that the dual Artin group is a Garside group and that the interval complex is a $K(\pi, 1)$-space.

Next, we outline the main components of our proof, which partially extend the theory developed by McCammond and collaborators [26] while introducing combinatorial methods. The essential steps are as follows: first, we prove that the interval complex remains a $K(\pi, 1)$-space even when the poset is not a lattice; second, we construct a *dual* finite model for the orbit configuration space; and finally, we apply combinatorial techniques – such as *EL*-shelling of the poset and discrete Morse theory – to simplify the interval complex, reducing it to the dual model (see also [18]).

In the last part of the paper, we consider potential generalizations of these constructions to arbitrary Coxeter groups and pose a few related questions and conjectures. Our concluding theorem provides affirmative answers to all these questions for rank-three Artin groups (see [16]).

2. Coxeter groups and Artin groups

A (finitely generated) Coxeter group is a group defined by the presentation

$$W = \left\langle S \mid s^2 = 1, \ \underbrace{sts\cdots}_{m(s,t)\text{ factors}} = \underbrace{tst\cdots}_{m(s,t)\text{ factors}} \ \text{for } s \neq t,\ m(s,t) \neq \infty \right\rangle$$

(see [5,22]), where S is a finite set and $m : S \times S \to \mathbb{N} \cup \{\infty\}$ is a function such that $m(s, s) = 1$ for all $s \in S$ and $m(s, t) = m(t, s) \in \{2, 3, 4, \ldots\} \cup \{\infty\}$ for all $s \neq t$ in S.

Any Coxeter group can be represented as a group generated by *linear reflections* in the real space $V = \mathbb{R}^S$. It is convenient to consider the contragredient $\rho^* : W \to \mathrm{GL}(V^*)$ of the *Tits representation* $\rho : W \to \mathrm{GL}(V)$, which is given by

extending to all W the map

$$\rho : s \mapsto r_s : x \mapsto x - 2B(x, e_s)e_s, \quad s \in S.$$

Here, e_s denotes the canonical basis vector in V, and B is the symmetric bilinear form given by

$$B(e_s, e_{s'}) := -\cos(\pi/m(s, s'))$$

(with $B(e_s, e_{s'}) = -1$ if $m(s, s') = \infty$).

The representation ρ preserves the bilinear form B. It turns out that W is *spherical* (finite) if and only if B is positive definite; W is *affine* if and only if B is positive semidefinite (but not definite); W is *hyperbolic* if B is non-degenerate of signature $(n-1, 1)$.

The action of W on V^*, which has a canonical basis given by the dual basis $e_s^*, s \in S$, of V, is free on the *base chamber*

$$C_0 := \bigcap_{s \in S} H_s^+.$$

Here, $H_s = \mathrm{Span}\{e_{s'}^* \mid s' \in S,\ s' \neq s\}$ are the *walls* of C_0, and $H_s^+ = H_s + \mathbb{R}_+ e_s^*$ is the half-space with boundary H_s which contains C_0. Denote by $D = \overline{C_0}$ the closure of the base chamber. The stabilizer of a face $F \subset D \setminus C_0$,

$$F = \left(\bigcap_{s \in T \subset S} H_s \right) \cap \left(\bigcap_{s \in S \setminus T} H_s^+ \right), \tag{2.1}$$

is the *parabolic subgroup* W_T generated by all $s \in T$. Notice that for an affine W, all proper parabolic subgroups are finite.

Let $R := \{r = w^{-1} s w \mid s \in S,\ w \in W\}$ be the set of all *reflections* in W. Each $r \in R$ is mapped by ρ^* into a linear reflection in V^* with fixed point set $H_r = w^{-1}.H_s$. The collection $\mathcal{A} = \{H_r, r \in R\}$ is called the *reflection arrangement* of W. The connected components of the complement of $\bigcup_{r \in R} H_r$ are the *chambers* of the arrangement.

The *Tits cone* $I = \bigcup_{w \in W} w.D$ is the orbit of D. One has the following properties (see [22, 36]):

- $I = \mathbb{R}^S$ if and only if W is finite; when W is affine, I is a half-space (with boundary $\mathrm{Ann}(V^\perp)$).
- D is a fundamental domain for the action of W on I.
- I is a convex cone in R^S, which is generally neither open nor closed.
- The reflection arrangement $\mathcal{A}$ stratifies I into convex *faces*, each conjugated to a unique face of C_0. The group W acts simply transitively on the set of codimension-0 faces (the chambers).

- Face (2.1) of C_0 has stabilizer equal to the *parabolic subgroup* W_T which is generated by the reflections $s \in T$. This face belongs to the interior $\mathrm{int}(I)$ if and only if W_T is finite. More generally, any face F' conjugated to F (by elements in some left coset of W_T) belongs to $\mathrm{int}(I)$ if and only if W_T is finite.

In particular, all codimension-1 faces of C_0 (the walls of C_0) belong to $\mathrm{int}(I)$ and a codimension-2 face $F = (H_s \cap H_{s'}) \cap \left(\bigcap_{t \neq s,s'} H_t^+\right)$ is in $\mathrm{int}(I)$ if and only if $m(s,s') \neq \infty$.

We now introduce the *configuration space* associated with W:

$$Y = (I_{\mathbb{C}}) \setminus \bigcup_{r \in R} H_{r,\mathbb{C}},$$

where $I_{\mathbb{C}} \cong I \times \mathbb{R}^S$ is the complexification of the Tits cone and $H_{r,\mathbb{C}} \cong H_r \times H_r$ is the complexification of the hyperplane H_r. (Equivalently, one can take $I \times I$ instead of $I_{\mathbb{C}}$.) The group W acts diagonally on $I_{\mathbb{C}}$, and it acts freely and properly discontinuously on Y. Thus, we have the *orbit configuration space*

$$Y_W = Y/W.$$

Example 2.1. For W the symmetric group S_n, one has $Y \simeq \mathbb{C}^n \setminus \bigcup_{i \neq j} \{z_i = z_j\}$ as the space of ordered n-tuples of points in $\mathbb{C}$ and $Y_W \simeq Y/S_n$ the space of unordered n-tuples of points in $\mathbb{C}$.

We also recall the *Salvetti complexes* $X \subset Y$ and $X_W \subset Y_W$ which are deformation retracts of the configuration space and of the orbit configuration space of W, respectively [33, 34]. In particular, X_W contains a $|T|$-cell for each *finite* parabolic subgroup W_T. The construction of X closely follows the approach in the original paper [33], involving only the faces contained within the interior of I (so corresponding to cosets of finite parabolic subgroups; see [36, Thm. 2]). Consequently, X_W is a finite union of subcomplexes associated with the maximal finite parabolic subgroups of W.

Definition 2.2. The *Artin group* G_W, associated with a Coxeter group W, is the group presented by the generators and relations of W, except the relations $s^2 = 1$ are omitted. Specifically, it is defined by

$$G_W = \left\langle S \mid \underbrace{sts\cdots}_{m(s,t)\text{ factors}} = \underbrace{tst\cdots}_{m(s,t)\text{ factors}} \text{ for } s \neq t,\ m(s,t) \neq \infty \right\rangle.$$

In Example 2.1 with $W = S_n$, the Artin group G_W is the classical braid group Br_n.

Theorem 2.3 ([7] in the finite case, [35] in general). *The following holds:*

$$\pi_1(Y_W) \cong G_W. \qquad \blacksquare$$

(A simple proof of this result can be obtained by examining the 2-skeleton of the associated complex X_W.)

We can now state the main conjecture.

Conjecture 2.4 ($K(\pi, 1)$-conjecture). *The orbit configuration space Y_W is a classifying space for the corresponding Artin group G_W.*

This famous conjecture is usually attributed to Arnol'd, Thom, Brieskorn, Pham, dating back to the 1960s and 1970s in connection with topological questions in singularity theory. One consequence of the conjecture is the existence of a $K(\pi, 1)$-space given by an explicit finite CW-complex (e.g., the abovementioned X_W).

The conjecture was established by Deligne in the 1970s for spherical Coxeter groups [14]. Recently, we found [29].

Theorem 2.5 (G. Paolini, M. Salvetti). *The $K(\pi, 1)$-conjecture is true for all affine Artin groups.*

In previous years, the conjecture was proved in some scattered cases: types $\widetilde{A}_n$, $\widetilde{C}_n$ [28]; large-type Artin groups [20]; FC-type groups [11]; locally reducible type [10]; type $\widetilde{B}_n$ [9]. Several additional related results were established in the meantime: a comprehensive survey up to that point can be found in [30]. Following [29], additional cases were addressed in [19, 21], including cases where the Coxeter diagram is a tree (subjected to certain constraints on the edge labelings) as well as some finite-volume hyperbolic cases.

3. Quick overview of Garside theory

Deligne's proof of the spherical case was inspired by and extended the earlier work by Garside on the braid group [17]. His work, in turn, led to the development of what is now known as Garside theory (a valuable contribution toward understanding certain key properties in the finite case was also made by [8], which was published at the same time as [14]).

A *Garside structure* on a group G is defined as a pair (M, Δ), where M is a monoid and $\Delta \in M$ is a distinguished element called the *Garside element*. The group G is the group of fractions of M. This pair satisfies a list of properties (see [13] for a comprehensive overview and references up to that point).

The main property of interest is that M, with respect to the poset structure induced by the divisibility relation, forms a *lattice*; that is, any two elements of M have a least common multiple (*lcm*) and a greatest common divisor (*gcd*). One consequence is the

existence of a *normal form* in G, expressed as $g = \Delta^n a$, $n \in \mathbb{Z}, a \in M$. This normal form provides a solution to the word problem in G.

Now, consider a group G with generating set $R = R^{-1}, 1 \notin R$. Let $\ell_R(g)$ be the length of an element $g \in G$ with respect to R (we can take $\ell_R(g)$ as the minimal length of an R-factorization of g). We can introduce a partial ordering on G as follows:

$$g \leq g' \iff \ell_R(g) + \ell_R(g^{-1}g') = \ell_R(g')$$

(equivalently, there is a reduced R-factorization of g' which begins with a reduced R-factorization of g). The Hasse diagram of $(G, \leq)$, graded by the length function ℓ_R, can be viewed as a subgraph of the Cayley graph of G. Specifically, for $g \in G, r \in R$, there is an oriented edge (labeled by r) from g to gr if and only if $\ell_R(g) < \ell_R(\mathrm{gr})$ (so $\ell_R(\mathrm{gr}) = \ell_R(g) + 1$).

Definition 3.1 (See [12, 23, 26]). For $\gamma \in G$, let $[1, \gamma]$ be the interval below γ in the poset $(G, \leq)$.

(1) The *interval group* G_γ is the group with generators $R_\gamma = R \cap [1, \gamma]$ and relations as follows: for any pair of maximal chains c, c' of $[1, \gamma]$, with labels given by words w, w' in R, there is a relation $w = w'$. (Equivalently, there is a relation for any loop in the (undirected) Hasse diagram of $[1, \gamma]$.)

(2) The *interval complex* K_γ is the CW-complex (delta-complex) obtained from the order complex $\Delta([1, \gamma])$ by identifying two simplices if they have the same edge labels (in particular, all 0-simplices are identified).

Recall that the order complex of a poset $\mathcal{P}$ is the simplicial complex whose q-simplices are the $(q + 1)$-chains of $\mathcal{P}$.

By examining the 2-skeleton of K_γ, it is easy to see that $\pi_1(K_\gamma) = G_\gamma$.

We say that the element $\gamma \in G$ is *balanced* if its left and right divisors coincide.

Theorem 3.2 (See [2, 12, 26, 27]). *Let γ be balanced and assume that $[1, \gamma]$ is a lattice. Then:*

- *The interval group G_γ is a Garside group.*
- *The interval complex K_γ is a $K(G_\gamma, 1)$.*

The Garside structure on G is given by the monoid G_γ^+ which has the same presentation as G_γ. The Garside element is γ itself.

3.1. Standard and dual Garside structures

When W is finite, the *standard Garside structure* is defined by taking the set of simple reflections S as the generating set and using the *longest element* $\delta \in W$ (with

respect to length function ℓ_S). In this case, the interval $[1, \delta]$ coincides with the entire group W.

The main points, following Deligne's proof within the previously introduced framework, can be summarized as follows.

Theorem 3.3. *Let W be finite. Then:*

(1) *The interval group W_δ is the standard Artin group G_W (and the associated monoid is the standard Artin monoid as defined in* [8, 14]*).*

(2) *The poset $[1, \delta]$ is a lattice. Therefore, G_W is a Garside group and K_δ is a $K(G_W, 1)$-space.*

(3) *K_δ is homotopy equivalent to the orbit configuration space Y_W.*

Point (1) is not too difficult. Point (2) follows from the fact that the weak Bruhat order on a finite Coxeter group is a lattice (see, e.g., [4]). Point (3) is the most delicate: one can see that K_δ is homotopy equivalent to the complex X_W mentioned earlier (see, for instance, [15]).

To introduce the *dual Garside structure* on W, we take all the reflections R as a generating set and choose any *Coxeter element* $c \in W$. Recall that a Coxeter element of W is a product of all the simple reflections in some order or, equivalently, a product of the reflections associated with the walls of a given chamber, in some order. Thus, there are many Coxeter elements, although most of them are often conjugate in W. The interval $[1, c]$ is commonly referred to as the *non-crossing partition* poset; we denote it by $\mathrm{NC}(W, c)$. This terminology originates from the case when W is the symmetric group S_n and $c = (1, 2, \ldots, n)$. Let S_n act on the n points $P_k = \exp\left(\frac{2\pi i}{k}\right), k = 0, \ldots, n-1$, in the unit circle. A permutation σ belongs to $[1, c]$ if and only if the convex hulls of its cycles do not intersect each other (see Figure 1 for $n = 4$) (see, e.g., [13, Ch. IX]).

Definition 3.4. The interval group $G'_W := G_c$ is called the *dual Artin group* of type W.

We also set $K_W := K_c$. The following holds.

Theorem 3.5 (See [2], generalizing [3] for the case A_n; see also [6]). *When W is finite, one has the following:*

(1) *There is an isomorphism $G'_W \cong G_W$ between the dual and standard Artin groups.*

(2) $\mathrm{NC}(W, c)$ *is a lattice (therefore, G'_W is a Garside group and $K_W = K(G'_W, 1)$).*

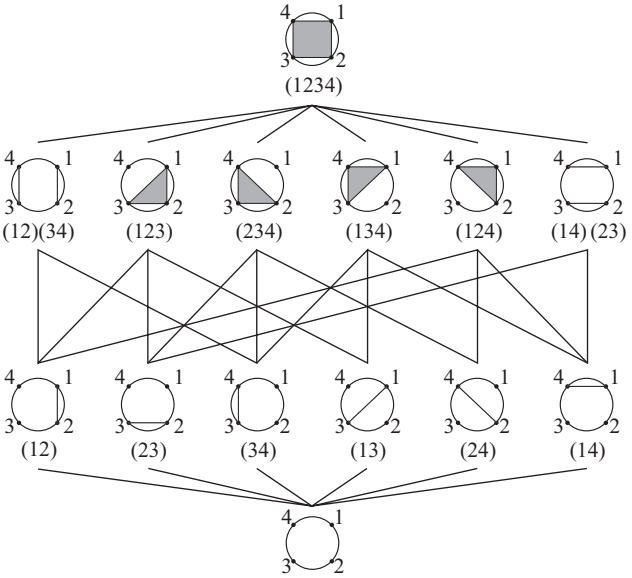


Figure 1. The non-crossing partition poset $\mathrm{NC}(S_4; c)$, where $c = (1, 2, 3, 4)$.

(3) $K_W \cong Y_W (\cong X_W)$ *(by the uniqueness of* $K(\pi, 1)$ *up to homotopy).*

Example 3.6. We illustrate the two structures in the case where W is of type A_2 (i.e., $W = S_3$). In this case, $W = \langle S = \{s_1, s_2\} \mid s_1^2 = s_2^2 = 1,\ s_1 s_2 s_1 = s_2 s_1 s_2 \rangle$. The standard structure is obtained by taking generators S and using the longest element $\delta = s_1 s_2 s_1$ as the Garside element. The dual structure uses as generators all reflections $R = \{s_1, s_2, s_3\}$, where $s_3 = s_1 s_2 s_1$. The Garside element is a Coxeter element, say $c = s_1 s_2$. We depict the associated posets and the complexes X_W, K_W in Figure 2.

4. The case W affine

Let now W be an affine Coxeter group. We want to discuss some main points from [29]. In the affine case, there is no standard Garside structure because there is no longest element in W. Attempts have been made to use the dual structure by choosing a Coxeter element $c \in W$. However, the strategy that works in the finite case does not generally apply here. Specifically, we have the following result.

Theorem 4.1 ([24]). *For W affine, the non-crossing partition poset* $\mathrm{NC}(W, c)$ *forms a lattice if and only if W is of type $\widetilde{A}_n$ (for certain Coxeter elements c), $\widetilde{C}_n$, or $\widetilde{G}_2$.*

Nonetheless, several interesting properties were discovered in [26], including the following.

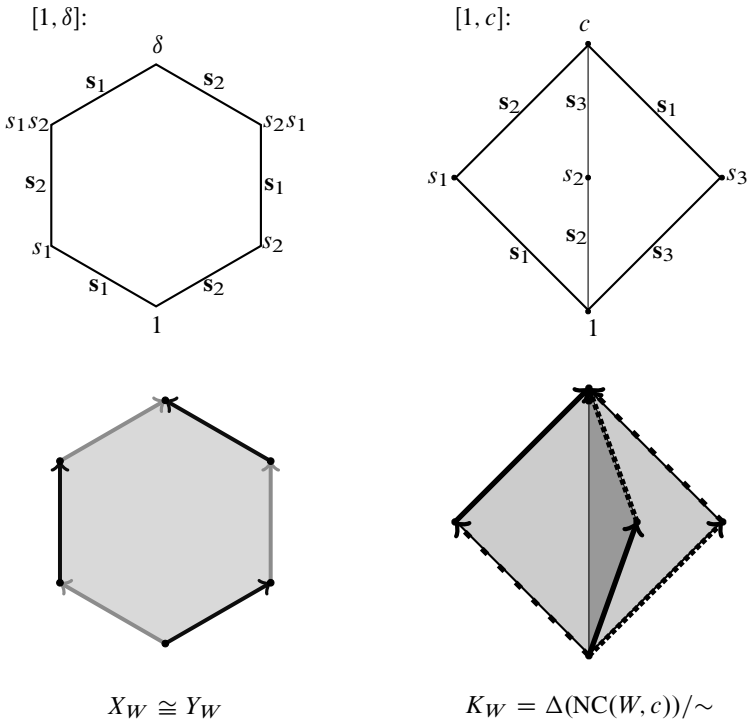


Figure 2. Standard (on the left) and dual (on the right) structures when $W = S_3$. Arrows of the same kind must be glued.

Theorem 4.2. *For W affine, there is an isomorphism between the dual and standard Artin groups: $G'_W \cong G_W$.*

We now summarize the main results proved in [29].

Theorem 4.3. *Let W be an affine Coxeter group. Then:*

(1) *The complex K_W is always a $K(\pi, 1)$-space, even when $\mathrm{NC}(W, c)$ is not a lattice.*

(2) *There exists a finite subcomplex $X'_W \subset K_W$ which is homotopy equivalent to Y_W.*

(3) *The complex K_W deformation retracts onto X'_W.* ∎

The $K(\pi, 1)$-conjecture stated in Conjecture 2.4 follows directly from Theorem 4.3.

4.1. $K(\pi, 1)$-property for K_W

We quickly outline the strategy to prove part (1) of Theorem 4.3.

To each element $w \in W$, acting as an isometry on the affine space, we associate the set $\mathrm{Min}(w)$ of points which are *minimally moved* by w with respect to the euclidean distance. This set forms an affine subspace. When w is hyperbolic (i.e., it has no finite fixed points), one has $\dim(\mathrm{Min}(w)) = \mathrm{rk}(W) - \ell_R(w) + 1$. In particular, for a Coxeter element c, the set $\mathrm{Min}(c)$ is an affine line, referred to as the *Coxeter axis* (relative to c).

We rely on the following facts established in [26]. A reflection $r \in R_0 = R \cap [1, c]$ is called *vertical* if $\mathrm{Fix}(r)$ intersects the Coxeter axis; otherwise, r is called *horizontal*. When W is affine, the set of horizontal reflections R_{hor} is finite. These reflections generate a subgroup H of W, which is a direct product of affine Coxeter groups of type $\widetilde{A}_i$. The poset $\mathrm{NC}(W, c)$ forms a lattice if and only if H is irreducible, which occurs in the cases covered by Theorem 4.1. Additionally, there is a finite set of translations $T \subset [1, c]$.

In cases where the poset $\mathrm{NC}(W, c) = [1, c]$ is not a lattice, a finite set of additional new translations $T_f \supset T$ is added, leading to a larger group $C \supseteq W$. This construction ensures that the new interval below c in C, denoted for clarity as $[1, c]^C$, becomes a lattice; furthermore, $\mathrm{NC}(W, c)$ is a subposet of $[1, c]^C$. Let then D and F be the subgroups of C generated by $R_{\mathrm{hor}} \cup T$ and $R_{\mathrm{hor}} \cup T_f$, respectively. The Coxeter element c belongs to both D and F, giving rise to intervals $[1, c]^D \subset D$, $[1, c]^F \subset F$. The latter interval is also a lattice.

Using these intervals, we construct additional interval complexes K_c^C, K_c^F, and K_c^D. Furthermore, we define the subcomplex $K_c^H \subset K_c^D$ consisting of all simplices $[x_1 \mid \cdots \mid x_k]$ such that the product $x_1 \cdots x_k$ belongs to H.

The situation is summarized in Figure 3. Here, the complexes K_c^C, K_c^F are $K(\pi, 1)$-spaces.

We prove the following in [29]:

- K_c^D is covered by $K_c^H \times \mathbb{R}$.
- K_c^H is a $K(\pi, 1)$-space, hence K_c^D is also a $K(\pi, 1)$-space.
- K_W is a $K(\pi, 1)$-space (which follows by a Mayer–Vietoris argument).

4.2. The dual model

When W is finite, we just set $X'_W = K_W$. Therefore, by Theorem 3.5 (3), we have $X'_W \cong X_W \cong Y_W$ in this case.

In general, let $T \subset S$ be such that the parabolic subgroup $W_T \subset W$ is finite. Let c be the Coxeter element expressed as a product of elements $s \in S$, in some order, and let c_T be the product of the elements in T, in the same order as they appear in the

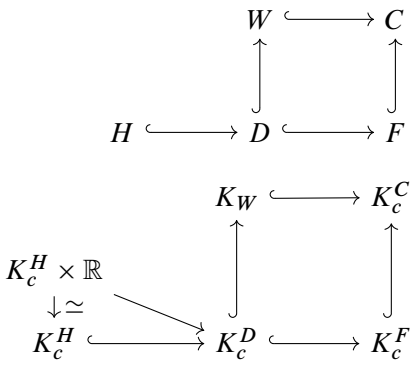


Figure 3. Diagrams of groups and complexes.

factorization of c. Then c_T is a Coxeter element of W_T, and it belongs to the interval $[1, c]$. Moreover, the interval complex K_{W_T} constructed from $[1, c_T] \subset W_T$ is a subcomplex of K_W (see [29, §7]).

We define X'_W in general as

$$X'_W = \bigcup_{W_T \subset W,\, W_T \text{ finite}} K_{W_T}.$$

To prove that $X'_W \simeq Y_W$, we construct a homotopy equivalence $\varphi : X_W \to X'_W$. Note that X_W can also be expressed as a union of subcomplexes X_{W_T}, corresponding to finite parabolic subgroups. The map φ is constructed by systematically gluing together homotopy equivalences between the corresponding subcomplexes.

Example 4.4. We show in Figure 4 the example of $W = \widetilde{A}_2$. Here, the Coxeter element is $w = abc$. The dotted line represents the Coxeter axis of W with respect to w. In this case, $X'_W = K_{W_{a,b}} \cup K_{W_{a,c}} \cup K_{W_{b,c}}$, and $X_W = X_{W_{a,b}} \cup X_{W_{a,c}} \cup X_{W_{b,c}}$. Each block K_{W_T} is homotopy equivalent to the corresponding block X_{W_T}.

4.3. The retraction $K_W \searrow X'_W$

To prove part (3) of Theorem 4.3, we apply *discrete Morse theory* to the complex K_W. We construct a discrete flow which contracts all the cells in $K_W \setminus X'_W$ (perfect matching). To control this matching, we make use of another result, which is of independent interest.

Theorem 4.5 ([29]). *The poset* $\mathrm{NC}(W, c)$ *is EL-shellable.*

(For the spherical case, an analogous result was proven in [1].) For the edge-labeled poset $\mathrm{NC}(W, c)$, *EL*-shellability is equivalent to the following:

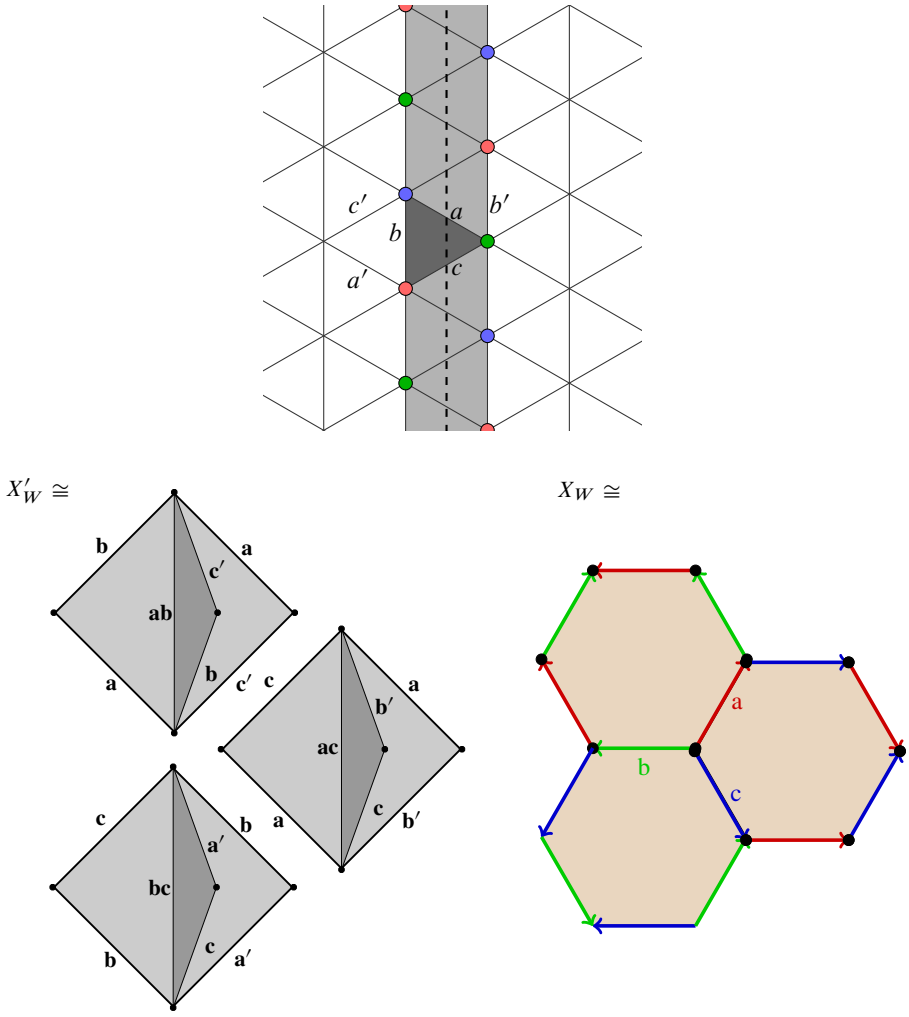


Figure 4. Case $\widetilde{A}_2$. Arrows with the same labels or colors are glued together.

- There is a *total ordering* $\prec$ of the reflections R_0 in $\mathrm{NC}(W, c)$ such that every $u \in \mathrm{NC}(W, c)$ has a *unique* increasing minimal factorization $u = r_1 \cdots r_k$, where $r_1 \prec \cdots \prec r_k$.
- This factorization is the *lexicographically smallest* minimal factorization of u.

To define an ordering of the reflections R_0, we use an *axial ordering* method. Let C_0 be the base chamber, and let ℓ_c be the Coxeter axis. The Coxeter element c acts as a translation along ℓ_c, and we orient ℓ_c accordingly. Fix a point $P_0 \in \ell_c \cap C_0$ and order R_0 as follows:

- Vertical reflections r such that $\mathrm{Fix}(r) \cap \ell_c$ lies after P_0 in the ℓ_c-orientation: these are ordered first, following the orientation of ℓ_c.
- Horizontal reflections: these are ordered according to a specific construction, which depends on the detailed structure of each irreducible component of the horizontal root system.
- Vertical reflections r such that $\mathrm{Fix}(r) \cap \ell_c$ lies before P_0 in the ℓ_c-orientation: these are ordered last, again respecting the orientation of ℓ_c.

Figure 5 illustrates the axial ordering for $W = \widetilde{A}_2$ and $W = \widetilde{G}_2$, where the Coxeter element is $w = abc$.

We use the axial ordering as the main tool to construct an explicit matching of the cells in $K_W \setminus X'_W$, in order to prove part (3) of Theorem 4.3.

Example 4.6. We illustrate the matching in the smallest affine case, $W = \widetilde{A}_1$. Here, $W \cong \mathbb{Z}_2 * \mathbb{Z}_2 = \langle s_0, s_1 \mid s_i^2 = 1 \rangle$ and $G_W \cong F[2] \cong \langle s_0, s_1 \rangle$. The reflections R correspond to words of odd length: $s_{-i} = s_0 s_1 s_0 \cdots$ ($2i + 1$ factors, $i \geq 0$), $s_i = s_1 s_0 s_1 \cdots$ ($2i - 1$ factors, $i \geq 1$). Thus, W is realized as the isometry group of the real line generated by reflections with respect to integer points. Let $c = s_1 s_0$ be the Coxeter element (acting on the right: c is the translation $x \mapsto x + 2$ of the real line). We have $c = s_{i+1} s_i$ for $i \in \mathbb{Z}$. The non-crossing partition lattice $\mathrm{NC}(W, c)$ consists of the elements $1, s_i, c$ with the relations $1 \prec s_i \prec c$ for $i \in \mathbb{Z}$. The edge labels are s_i for $(1, s_i)$ and s_{i-1} for (s_i, c) (see Figure 6). Therefore, the dual Artin group is presented as $G'_W = \langle s_i \mid s_{i+1} s_i = s_{j+1} s_j \ \forall i, j \rangle$, which can be easily seen to be isomorphic to G_W.

The axial ordering here is simply

$$s_1 \prec s_2 \prec \cdots (\prec s_{+\infty} \prec s_{-\infty} \prec) \cdots \prec s_{-1} \prec s_0.$$

Notice that the unique increasing 2-chain in $\mathrm{NC}(W, c)$ is $s_1 \prec s_0$.

The complex K_W is two-dimensional, with one 0-cell [ ], 1-cells $[s_i]$ for $i \in \mathbb{Z}$, and $[c]$. The 2-cells are $[s_{i+1} \mid s_i]$, $i \in \mathbb{Z}$, and their boundaries are represented by $\partial[s_{i+1} \mid s_i] = [s_i] \cup [c] \cup [s_{i+1}]$ (see Figure 7).

In Figure 8, we illustrate the Hasse diagram of the face poset $\mathcal{F}(K_W)$ of K_W. This poset consists of the simplices of K_W, ordered by face inclusion. The thick and dashed arrows indicate a proper matching on $\mathcal{F}(K_W)$. The unmatched (critical) cells are the 0-cell [ ] and two 1-cells $[s_0]$, $[s_1]$. These form the subcomplex $X'_W = K_{\{s_0\}} \cup K_{\{s_1\}} \cong S^1 \vee S^1 \cong X_W$. Consequently, by standard results in discrete Morse theory, K_W contracts onto X'_W.

The previous example considers the simplest affine group, $\widetilde{A}_1$, and sheds light on the general procedure used to contract K_W onto X'_W. Notice that the face poset

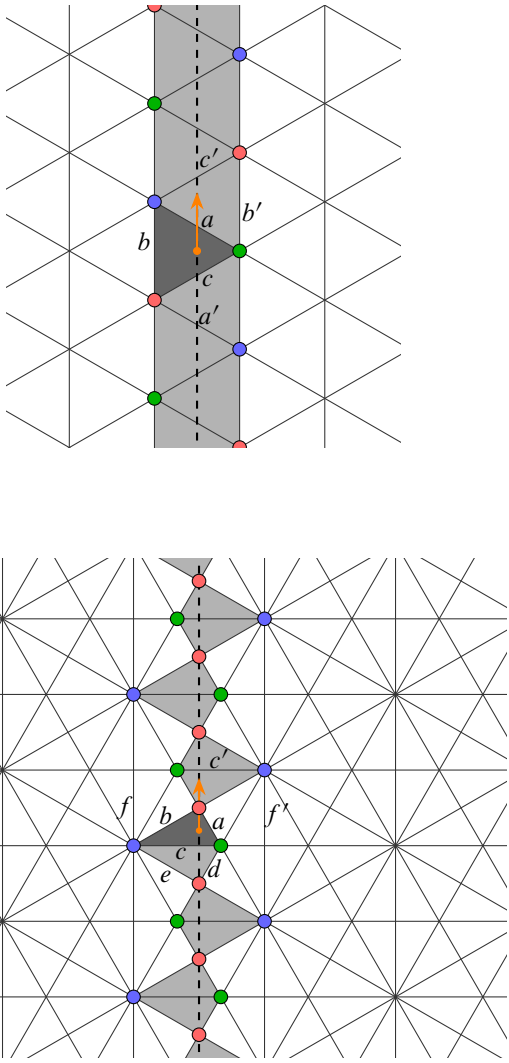


Figure 5. Type $\widetilde{A}_2$: axial ordering: $a \prec c' \prec \cdots \prec b \prec b' \prec \cdots \prec a' \prec c$. Type $\widetilde{G}_2$: axial ordering: $a \prec b \prec c' \prec \cdots \prec f \prec f' \prec \cdots d \prec e \prec c$.

$\mathcal{F}(K_W)$ (Figure 8) can be divided into two distinct parts: the first part, $\mathcal{F}_1$, consists of the cells [ ] and $[c]$, while the second part, $\mathcal{F}_2$, consists of the top two rows, excluding the cell $[c]$. These parts are preserved under conjugation by the Coxeter element c (with the relation $c^{-1}s_i c = s_{i-2}$).

The contraction $K_W \searrow X'_W$ can be carried out in two steps. First, one *shrinks* $\mathcal{F}_2$ onto a finite segment. In Figure 8, this is achieved using the matching indicated by

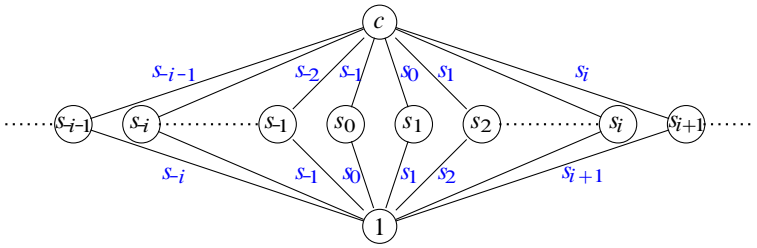


Figure 6. The non-crossing partition poset $\mathrm{NC}(W, c)$ for $W = \widetilde{A}_1$.

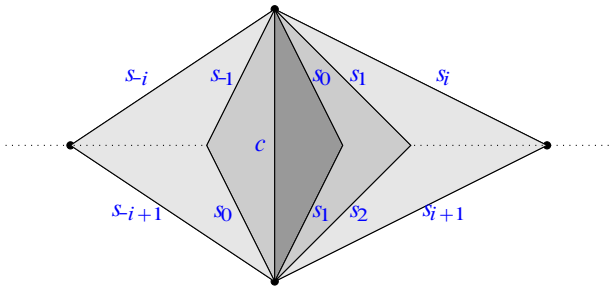


Figure 7. The interval complex K_W, $W = \widetilde{A}_1$: edges with same label are identified.

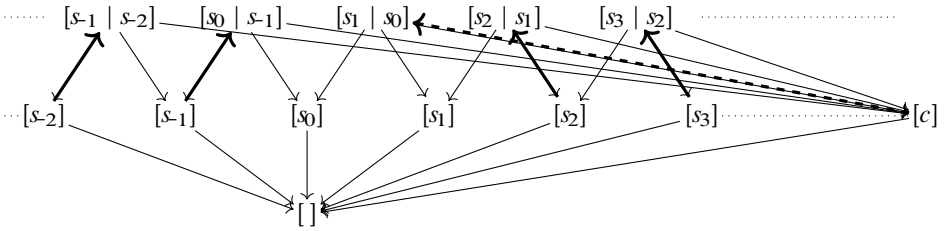


Figure 8. The face poset $\mathcal{F}(K_W)$: critical cells: $[\,]$, $[s_0]$, $[s_1]$, so $K_W \simeq S^1 \vee S^1$.

the thick arrows within $\mathcal{F}_2$. This results in a contraction of K_W onto a finite subcomplex K', which strictly contains X'_W. In this example, K' is the 2-simplex $[s_1 \mid s_0]$, together with its boundary cells (and with all vertices identified).

Next, a second matching – indicated by the dashed arrow – connects simplices in $\mathcal{F}_1$ and $\mathcal{F}_2$, further contracting K' onto X'_W. Specifically, the 1-cell $[c]$ is matched with the 2-cell $[s_1 \mid s_0]$, corresponding to the unique increasing factorization $c = s_1 s_0$.

A similar, though technically more complex, situation arises in the general case. First, the Hasse diagram of $\mathcal{F}(K_W)$ is decomposed into connected pieces, similar

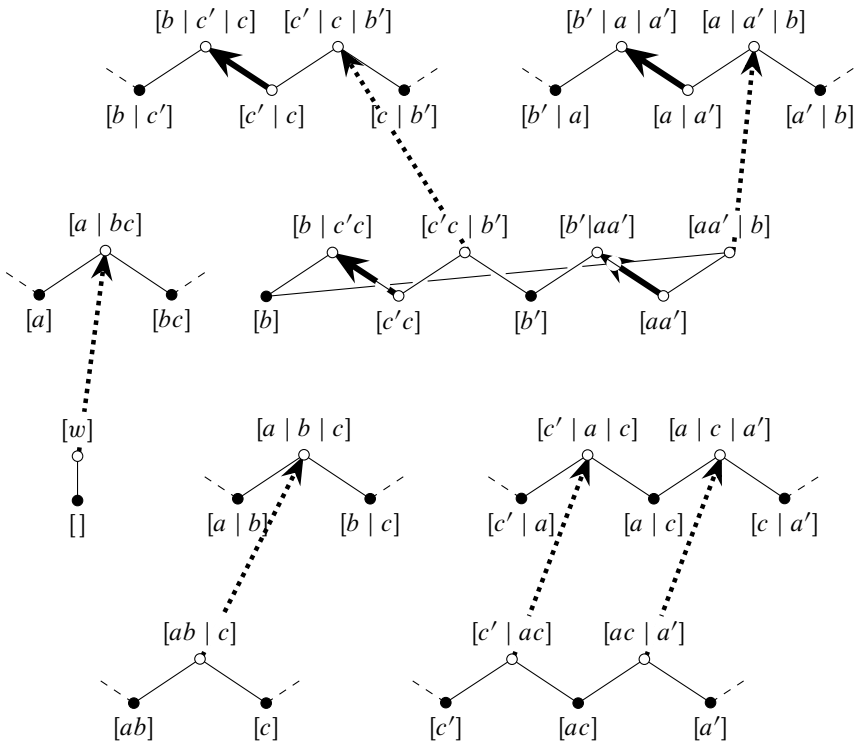


Figure 9. Components of K_W for $W = \widetilde{A}_2$, with Coxeter element $w = abc$. The perfect matching in $K' \setminus X'_W$ is indicated by the thick and dotted arrows.

to $\mathcal{F}_1$ or $\mathcal{F}_2$, appearing as zigzag sequences (which can be finite or infinite) connecting simplices of adjacent dimensions. Retraction of each infinite sequence onto a finite "interval" reduces K_W to a finite subcomplex K' containing X'_W. The cells in $K' \setminus X'_W$ are matched either within their own sequence or by connecting different sequences. The latter type of matching involves splitting certain elements according to their minimal factorization.

Example 4.7. These constructions are illustrated in Figure 9 for the case where W is of type $\widetilde{A}_2$.

There are nine sequences (two of which are finite). The black nodes represent the cells of X'_W. Each infinite sequence contracts onto the "convex envelope" of the cells of X'_W; together, these cells form the finite subcomplex K' containing X'_W (comprising both black and white nodes). The matching of the cells in $K' \setminus X'_W$ is represented by the thick black arrows (connecting pairs within the same sequence) and the dotted

arrows (indicating jumps between different sequences, obtained by "splitting" certain elements according to their minimal factorization).

4.4. Some general conjectures

The constructions described in the previous sections naturally lead to some interesting questions and conjectures. Below, we provide a list of these, starting with the $K(\pi, 1)$-conjecture as the first main conjecture.

(1) $K(\pi, 1)$-conjecture: for any Artin group, $Y_W \simeq K(G_W, 1)$.

(2) Isomorphism of dual and standard Artin groups: $G'_W \cong G_W$ for any Coxeter group W; known for spherical and affine Artin groups [2, 26, 29], and it is natural to conjecture that it holds for all Artin groups.

(3) Interval complex as a $K(\pi, 1)$-space: $K_W \simeq K(G'_W, 1)$. In the case where W is of spherical or affine type, the interval complex K_W is always a $K(\pi, 1)$-space, even when the lattice property of $\mathrm{NC}(W, c)$ does not hold (see Theorem 4.3 (1)). This observation suggests that Garside theory could potentially be extended. Therefore, the conjecture is that the interval complex remains a $K(\pi, 1)$-space under certain conditions, which have yet to be characterized, and are weaker than the lattice property.

(4) Equivalence of complexes: $K_W \simeq Y_W$ for all Artin groups. A stronger version of this conjecture is $K_W \searrow X'_W$. This follows directly from our construction in the affine case. The hope is that a similar construction can be generalized.

(5) Shellability of non-crossing partitions: $\mathrm{NC}(W, c)$ is always shellable. This has been proven for finite groups [1] and for affine-type groups (from Theorem 4.5). Recently, G. Paolini discovered (using computational methods) that, for certain hyperbolic groups, the axial ordering does not yield an *EL*-shelling. However, there might exist modifications of the axial ordering that work (we have identified such modifications in some cases).

It should be noted that some of these conjectures are not independent. For instance, (3) and (4) imply (1) (this is the strategy used in [29]). Moreover, (4) implies (2), as it can be deduced by examining the fundamental groups.

Some recent progress for conjecture (2) can be found in [31].

4.5. Rank-three case

We now ask whether the methods used in the affine case can be extended to other classes of Artin groups. A positive answer includes the class of all rank-three Artin groups. In Figure 10, we show an example of a rank-three Coxeter group with exponents m_{ij} given by $(3, 3, 4)$.

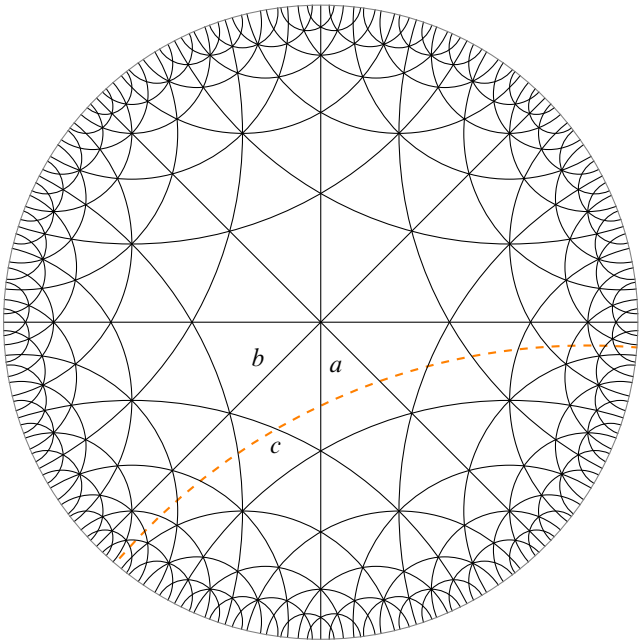


Figure 10. Example of a rank-three Coxeter group in the Poincaré model. The hyperbolic disk is tiled by triangles with angles $\frac{\pi}{3}, \frac{\pi}{3}, \frac{\pi}{4}$. The Coxeter element is $w = abc$, and the Coxeter axis is represented by the dotted line.

In particular, the following theorem implies that all the conjectures listed in Section 4.4 hold true for rank-three Coxeter groups.

Theorem 4.8 ([16]). *Let* **W** *be a Coxeter group of rank three and* $c \in W$ *any Coxeter element.*

(1) $\mathrm{NC}(W, c)$ *is a lattice, so* $K_W \simeq K(G'_W, 1)$ *(conjecture* (3)*).*

(2) $\mathrm{NC}(W, c)$ *is EL-shellable (conjecture* (5)*).*

(3) $K_W \searrow X'_W$ *(conjecture* (4)*).*

As a consequence, we obtain the following corollary.

Corollary 4.9. *Let* **W** *be a Coxeter group of rank three and* $c \in W$ *any Coxeter element. Then:*

(1) $Y_W \simeq K(G_W, 1)$ *(*$K(\pi, 1)$*-conjecture).*

(2) $G'_W \cong G_W$ *(conjecture* (2)*).*

(3) G_W *is a Garside group (in particular, the word problem is solvable, and the center of* G_W *is trivial unless* W *is finite).*

For the rank-three case, the $K(\pi, 1)$-conjecture was already known; all other properties presented here are new.

The lattice property (1) of Theorem 4.8 follows from the poset inclusion of $[1, c]$ into the poset of linear subspaces of $\mathbb{R}^3$ (ordered by inclusion; see [25]).

We remark that the *EL*-shelling of Theorem 4.8 (2) is also obtained by axial ordering of R_0 and involves several geometric considerations in the hyperbolic model.

The proof of part (3) of Theorem 4.8 is similar to that of part (3) of Theorem 4.3, but it is unexpectedly more challenging in certain aspects. In particular, it requires considering, in addition to the complexes $X'_W \subset K' \subset K_W$, two extra complexes:

$$X'_W \subseteq K_W^{\text{sparse}} \subseteq K_W^{\text{short}} (\cong K') \subseteq K_W^{\text{long}} \subseteq K_W.$$

Starting from the right, each complex in this chain contracts onto the one immediately to its left.

Some of these contractions can be performed in general under the presence of an axial ordering that induces an *EL*-shelling of $\mathrm{NC}(W, c)$. This suggests that conjecture (4) might hold at least when $\mathrm{NC}(W, c)$ admits an *EL*-shelling.

As discussed in Section 4.4, such a result would imply the isomorphism between the dual and standard Artin groups (conjecture (2)).

Remark 4.10. Several related problems are of particular interest. For example, in the finite case, Deligne proved in [14] that the complement of a simplicial arrangement of hyperplanes is a $K(\pi, 1)$-space. These families strictly include reflection arrangements. A natural question arises: is the complement of a simplicial affine arrangement (possibly not a reflection arrangement) always a $K(\pi, 1)$-space?

A related long-standing conjecture involves the $K(\pi, 1)$-property for the *complements of discriminants* of the universal deformation space of non-simple singularities, such as elliptic singularities (the case of A, D, and E singularities is covered by the Deligne's result). The fundamental groups of these spaces are generally not Artin groups, and their higher homotopy groups remain unknown (see [32]).

Acknowledgments. I would like to express my sincere thanks to the ECM Scientific Committee for inviting me to speak. I would also like to extend my appreciation to the ECM Organizers for the excellent coordination of such a large conference. I would also like to once again mention my coauthor Giovanni, who was my PhD student and is now a professor in Bologna. I am proud to have worked with such a talented mathematician.

Funding. The author acknowledges support from the MIUR Excellence Department Project awarded to the Department of Mathematics, University of Pisa, CUP I57G

22000700001, and from PRIN 2022 no. 2022S8SSW2 “Algebraic and geometric aspects of Lie theory.”

References

[1] C. A. Athanasiadis, T. Brady, and C. Watt, Shellability of noncrossing partition lattices. *Proc. Amer. Math. Soc.* **135** (2007), no. 4, 939–949 Zbl 1171.05053 MR 2262893

[2] D. Bessis, The dual braid monoid. *Ann. Sci. Éc. Norm. Super. (4)* **36** (2003), no. 5, 647–683 Zbl 1064.20039 MR 2032983

[3] J. Birman, K. H. Ko, and S. J. Lee, A new approach to the word and conjugacy problems in the braid groups. *Adv. Math.* **139** (1998), no. 2, 322–353 Zbl 0937.20016 MR 1654165

[4] A. Björner and F. Brenti, *Combinatorics of Coxeter groups*. Grad. Texts in Math. 231, Springer, Berlin, 2006, 366 pp. Zbl 1110.05001 MR 2133266

[5] N. Bourbaki, *Groupes et algèbres de Lie* (Chapter 4: Groupes de Coxeter et système de Tits; Chapter 5: Groupes engendrés par des réflexions; Chapter 6: Systémes de racines). Élém. Math. Fasc. XXXIV, Hermann, Paris, 1968, 288 pp. Zbl 0186.33001 MR 0240238

[6] T. Brady and C. Watt, Non-crossing partition lattices in finite real reflection groups. *Trans. Amer. Math. Soc.* **360** (2008), no. 4, 1983–2005 Zbl 1187.20051 MR 2366971

[7] E. Brieskorn, Sur les groupes de tresses [d’après V. I. Arnol’d]. In A. Dold and B. Eckmann (eds.), *Séminaire Bourbaki. Vol. 1971/72. Exposés 400–417*, pp. 21–44, Lecture Notes in Math. 317, Springer, Berlin, 1973 Zbl 0277.55003 MR 0422674

[8] E. Brieskorn and K. Saito, Artin-Gruppen und Coxeter-Gruppen. *Invent. Math.* **17** (1972), no. 4, 245–271 Zbl 0243.20037 MR 0323910

[9] F. Callegaro, D. Moroni, and M. Salvetti, Cohomology of Artin groups of type $\tilde{A}_n$, B_n and applications. *Geom. Topol. Monogr.* **13** (2008), 85–104 Zbl 1143.20031 MR 2508202

[10] R. Charney, The Deligne complex for the four-strand braid group. *Trans. Amer. Math. Soc.* **356** (2004), no. 10, 3881–3897 Zbl 1077.20054 MR 2058510

[11] R. Charney and M. W. Davis, The $K(\pi, 1)$-problem for hyperplane complements associated to infinite reflection groups. *J. Amer. Math. Soc.* **8** (1995), no. 3, 597–627 Zbl 0833.51006 MR 1303028

[12] R. Charney, J. Meier, and K. Whittlesey, Bestvina’s normal form complex and the homology of Garside groups. *Geom. Dedicata* **105** (2004), no. 1, 171–188 Zbl 1064.20044 MR 2057250

[13] P. Dehornoy, F. Digne, E. Godelle, D. Krammer, and J. Michel, *Foundations of Garside theory*. EMS Tracts Math. 22, European Mathematical Society, Zürich, 2015, 691 pp. Zbl 1370.20001 MR 3362691

[14] P. Deligne, Les immeubles des groupes de tresses généralisés. *Invent. Math.* **17** (1972), no. 4, 273–302 Zbl 0238.20034 MR 0422673

[15] E. Delucchi, Combinatorics of covers of complexified hyperplane arrangements. In F. El Zein, A. I. Suciu, M. Tosun, A. M. Uludağ, and S. Yuzvinsky (eds.), *Arrangements, local systems and singularities*, pp. 1–38, Progr. Math. 283, Birkhäuser, Basel, 2009 Zbl 1281.32002 MR 3025858

[16] E. Delucchi, G. Paolini, and M. Salvetti, Dual structures on Coxeter and Artin groups of rank three. *Geom. Topol.* **28** (2024), no. 9, 4295–4336 Zbl 1554.20070 MR 4845455

[17] F. A. Garside, The braid group and other groups. *Q. J. Math.* **20** (1969), no. 1, 235–254 Zbl 0194.03303 MR 0248801

[18] T. Haettel, Exposé Bourbaki 1195: La conjecture du $K(\pi, 1)$ pour les groupes d'Artin affines [*d'après* Giovanni Paolini et Mario Salvetti]. In *Séminaire Bourbaki. Vol. 2021/22*, pp. 495–546, Astérisque 438, Société Mathématique de France, Paris, 2022 Zbl 1521.20074 MR 4576027

[19] T. Haettel and J. Huang, Lattices, Garside structures and weakly modular graphs. *J. Algebra* **656** (2024), 226–258 Zbl 1543.20039 MR 4759503

[20] H. Hendriks, Hyperplane complements of large type. *Invent. Math.* **79** (1985), no. 2, 375–381 Zbl 0564.57016 MR 0778133

[21] J. Huang, Labeled four cycles and the $K(\pi, 1)$-conjecture for Artin groups. *Invent. Math.* **238** (2024), no. 3, 905–994 Zbl 1554.20073 MR 4824731

[22] J. E. Humphreys, *Reflection groups and Coxeter groups*. Cambridge Stud. Adv. Math. 29, Cambridge University Press, Cambridge, 1992, 204 pp. Zbl 0768.20016 MR 1066460

[23] J. McCammond, An introduction to Garside structures. 2005, preprint

[24] J. McCammond, Dual Euclidean Artin groups and the failure of the lattice property. *J. Algebra* **437** (2015), 308–343 Zbl 1343.20039 MR 3351966

[25] J. McCammond and G. Paolini, Factoring isometries of quadratic spaces into reflections. *J. Algebra* **605** (2022), 226–252 Zbl 1510.11098 MR 4421719

[26] J. McCammond and R. Sulway, Artin groups of Euclidean type. *Invent. Math.* **210** (2017), no. 1, 231–282 Zbl 1423.20032 MR 3698343

[27] J. Michel, A note on words in braid monoids. *J. Algebra* **215** (1999), no. 1, 366–377 Zbl 0937.20017 MR 1684142

[28] C. Okonek, Das $K(\pi, 1)$-Problem für die affinen Wurzelsysteme vom Typ A_n, C_n. *Math. Z.* **168** (1979), no. 2, 143–148 Zbl 0427.14001 MR 0544701

[29] G. Paolini and M. Salvetti, Proof of the $K(\pi, 1)$ conjecture for affine Artin groups. *Invent. Math.* **224** (2021), no. 2, 487–572 Zbl 1542.20155 MR 4243019

[30] L. Paris, $K(\pi, 1)$ conjecture for Artin groups. *Ann. Fac. Sci. Toulouse Math. (6)* **23** (2014), no. 2, 361–415 Zbl 1306.20040 MR 3205598

[31] S. Resteghini, Free products and the isomorphism between standard and dual Artin groups. 2024, arXiv:2410.17724v1

[32] K. Saito, Second homotopy classes associated with non-cancellative monoids. [v1] 2023, [v3] 2024, arXiv:2312.16951v3

[33] 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 0884802

[34] M. Salvetti, The homotopy type of Artin groups. *Math. Res. Lett.* **1** (1994), no. 5, 565–577 Zbl 0847.55011 MR 1295551

[35] H. van der Lek, *The homotopy type of complex hyperplane complements*. Ph.D. thesis, Katholieke Universiteit te Nijmegen, 1983

[36] È. B. Vinberg, Discrete linear groups generated by reflections (in Russian). *Izv. Akad. Nauk SSSR Ser. Mat.* **35** (1971), no. 5, 1072–1112. Zbl 0247.20054 English translation. *Math. USSR-Izv.* **5** (1971), no. 5, 1083–1119 Zbl 0256.20067 MR 0302779

Mario Salvetti
Department of Mathematics, University of Pisa, Largo B. Pontecorvo, 5, 56127 Pisa, Italy;
mario.salvetti@unipi.it

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/32

Universal symmetries: Global equivariant homotopy theory

Stefan Schwede

Abstract. Global equivariant homotopy theory is, informally speaking, the universal home for deformations of geometric structures with symmetry groups that vary coherently. It studies geometric and topological objects on which all compact Lie groups act at once and in a compatible and coherent way. In this survey article, we give a brief introduction to the subject with emphasis on recent applications to equivariant bordism.

1. Introduction

One important aim of algebraic topology is the classification of nice spaces, such as manifolds or cell complexes, up to deformation, or more precisely, up to homotopy equivalence. A major tool to distinguish different homotopy types is algebraic invariants, such as the fundamental group, higher homotopy groups, or cohomology theories. Some of the historically first and particularly prominent (co)homology theories are:

- singular cohomology $H^n(X;\mathbb{Z})$, made from 'singular simplices' in the space X, i.e., continuous maps $\Delta^n \to X$ from the standard simplices;
- the bordism groups $\mathcal{N}_n(X)$, where elements are represented by continuous maps $M^n \to X$ from smooth closed manifolds of dimension n, taken up to bordism;
- topological K-theory $K(X)$, where elements are formal differences of complex vector bundles over the space X.

A major milestone in algebraic topology was the insight in the late 1950s that homology theories and cohomology theories are represented by *spectra*. For example, the (co)homology theories just mentioned are represented:

- Singular (co)homology is represented by the *Eilenberg–MacLane spectrum*, often denoted $H\mathbb{Z}$.

Mathematics Subject Classification 2020: 55N91 (primary); 55P91, 57R85 (secondary).
Keywords: global homotopy theory, compact Lie group, bordism.

- The bordism homology theory is represented by the *Thom spectrum*, usually denoted MO. Popular variations of this theme endow the manifolds and bordisms with additional tangential or normal structure; these bordism homology theories are represented by spectra such as MSO (for oriented manifolds) or MU (for stably almost complex manifolds).
- Bott periodicity allows one to extend the Grothendieck rings of complex vector bundles to a multiplicative cohomology theory, represented by the *K-theory spectrum* KU.

The main advantage of spectra over (co)homology theories is that they are more flexible and easier to manipulate.

But: What are spectra? Well, the answer varies, depending on when the question is asked and who is being asked. A crude approximation to this concept is motivated by the feature that cohomology theories come with suspension isomorphisms, i.e., suspending a space shifts the reduced (co)homology groups. So one should obtain spectra from spaces by formally inverting the suspension operation. The crux lies in how 'inversion of suspension' is implemented. First passing to the homotopy category of based spaces and then formally inverting the suspension functor yields the *Spanier–Whitehead category* [44]. This category has certainly been useful, and when restricted to finite CW-complexes, it correctly models the homotopy category of compact spectra (sometimes also called 'small' or 'finite' spectra, the thick subcategory generated by the sphere spectrum).

In the early days of the subject, the only *raison d'être* of spectra was to represent (co)homology theories, and any definition that fit this purpose was fine. To represent homology and cohomology theories is still an important purpose of spectra to this day. Over time, the entirety of spectra was more and more studied for its own sake, and the collection of all spectra was organized into increasingly more refined mathematical structures. Until the 1990s, most stable homotopy theorists would probably have argued that spectra form a *triangulated category*. A spectrum would be some kind of sequence of based spaces connected by continuous maps from the suspension of one space to the next space; but getting the morphisms right is a bit tricky. The first fully functional category of spectra was constructed by Boardman in his thesis [5], and Adams' exposition [1, Part III] was particularly popular for a long time.

Then until the mid-2000s, *model categories* were the predominant concept for organizing spectra, and some (including the author) made a living by constructing and comparing model categories of spectra. In my opinion, the advent of higher category theory greatly clarified what spectra really *are* from a conceptual point of view: they are the free presentable stable ∞-category on one object (the sphere spectrum); see [33, Corollary 1.4.4.6]. In the realm of triangulated categories, there is no such

universal property for the stable homotopy category. And while the universal property can in principle be expressed in the language of model categories, it is much more clumsy to implement. For example, [42, Theorem 5.1] amounts to saying that a certain moduli space of left Quillen functors is nonempty and connected, but it falls short of showing that this space is actually contractible. Since the ∞-category of spaces (or, as some prefer, of 'homotopy types', '∞-groupoids', or 'animae') is the free presentable ∞-category on one object (the contractible anima), spectra are indeed the ∞-categorical stabilization of spaces. And since 'stable' essentially means that suspension is invertible, in this precise sense, spectra are obtained from spaces by inverting suspension.

2. Equivariant spectra

Interesting mathematical objects that occur 'in nature' tend to have symmetries; and who would not agree that embracing the natural symmetries is beneficial? For my story, the symmetry groups are either finite groups or, more generally, compact Lie groups. An important aim of equivariant algebraic topology is to classify certain kinds of equivariant spaces or manifolds up to symmetry-preserving deformation (equivariant homotopy equivalence); compare Figure 1.

Again, algebraic invariants such as equivariant cohomology theories come in handy to tell apart different equivariant homotopy types. It turns out that many classical (co)homology theories have equivariant refinements:

- Bredon (co)homology [9] is a G-equivariant extension of singular cohomology; its coefficients are not merely abelian groups, but rather 'coefficient systems', i.e., co- or contravariant functors from the G-orbit category to abelian groups. The group G can be very general here, such as an arbitrary topological group [25].
- The equivariant bordism group $\mathcal{N}_n^G(X)$ consists of G-bordism classes of continuous G-maps $M^n \to X$, where the source is now a smooth closed G-manifold of dimension n. To actually get a homology theory, restrictions on the group are

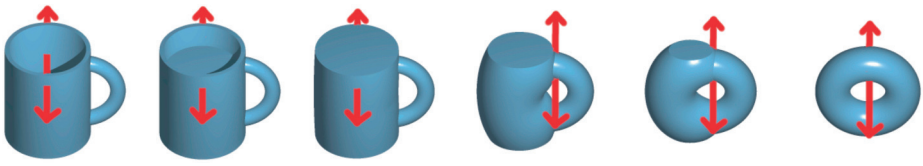


Figure 1. Equivariant deformation, preserving reflection symmetry.
Public domain, original gif file created by Lucas Vieira. https://en.wikipedia.org/wiki/File:Mug_and_Torus_morph.gif

necessary. Compact Lie groups are fine; and general Lie groups (not necessarily compact) work if one restricts to proper actions. I will return to equivariant bordism in more detail in Section 4.

- Equivariant K-theory $K_G(X)$ works nicely for compact Lie groups of equivariance; for compact G-spaces, elements are again formal differences of G-equivariant complex vector bundles over X.
- Every non-equivariant cohomology theory F^* can be turned into an equivariant cohomology theory F^*_G by applying F^* to the homotopy orbit construction. In other words, for a G-space X, one simply sets $F^*_G(X) = F^*(EG \times_G X)$, where EG is a universal free G-space, of the G-homotopy type of a G-CW-complex. The homotopy orbit construction is also known as the 'Borel construction', and hence F^*_G also goes under the name of *Borel cohomology*.

Now you can probably guess what is coming next: G-equivariant cohomology theories are represented by objects known as *G-spectra*, which can be obtained from G-spaces by 'inverting suspension'. However, in the equivariant situation, new subtleties arise, and the term 'inverting suspension' has even more possible interpretations. Depending on the flavor of suspensions to be inverted, the resulting objects are known as either *naive* or *genuine* G-spectra.

For our purposes, the 'good' class of G-spaces is the class of G-CW-complexes, which we study up to G-homotopy equivalence. For compact Lie groups, Illman's triangulation theorem [26] ensures that smooth G-manifolds belong to this class. One can get the same homotopy theory by working with general G-spaces, but then the correct notion of equivalence is that of an equivariant continuous map that induces a weak homotopy equivalence on all fixed points for all closed subgroups. Elmendorf's theorem [18] identifies this 'fine' homotopy theory of G-spaces with continuous functors from the G-orbit category to spaces. It became somewhat popular recently to even *define* G-spaces as functors from the orbit category to the ∞-category of spaces. While this is mathematically legitimate, degrading G-spaces to just another presheaf category makes me melancholic... I would like a G-space to be some kind of geometric or topological object with symmetries parameterized by G. But I might just be old-fashioned.

Let us return to G-spectra. For any topological group, there is always a good notion of G-spectra; and for some classes of groups such as compact Lie groups, there is an even better notion of G-spectra. The good version of G-spectra – often called 'naive' – is anything that models the stabilization, in the sense of higher categories, of the fine homotopy theory of G-spaces. So naive G-spectra are obtained from based G-spaces by inverting suspension; they represent (co)homology theories with a suspension isomorphism for ordinary suspensions. (There is also a 'doubly

naive' version of G-spectra, namely functors, in the ∞-categorical sense, from the classifying space BG to spectra. But let us not go into that.)

The even better version of G-spectra, called *genuine*, is only available for special classes of symmetry groups. For me, the class of compact Lie groups is the preferred setting; special things happen for the subclass of finite groups, which might be why some like to restrict to that context. I do believe that for phenomena that involve representations, equivariant vector bundles or equivariant bordism, it is a bad idea to restrict attention to finite groups, because the universal instances of important classes, such as Thom classes, Euler classes or Chern classes, tend to live at orthogonal or unitary groups.

So what is the difference between naive and genuine G-spectra? Succinctly, to get genuine spectra, we also invert all 'twisted suspensions', i.e., smashing with linear representation spheres. For this purpose, a *G-representation* is a finite-dimensional $\mathbb{R}$-vector space V on which G acts by linear isometries; and the *representation sphere* S^V is the one-point compactification, based at infinity. Figure 2 is a picture of the unit sphere of the natural Σ_3-representation on $\mathbb{R}^3$ by permuting the coordinates. It is simultaneously a picture of the one-point compactification of the reduced natural Σ_3-representation, permuting the coordinates of $\{(x, y, z) \in \mathbb{R}^3 : x + y + z = 0\}$.

Inverting all linear representation spheres has some benefits over inverting only spheres with trivial G-action. For example, cohomology theories represented by genuine G-spectra can be extended from integer-graded objects to $\mathrm{RO}(G)$-graded objects [30], where $\mathrm{RO}(G)$ is the real representation ring of G. I must issue the warning, however, that properly setting up an $\mathrm{RO}(G)$-grading is a slippery slope, in particular with respect to multiplicative structures. Readers unfamiliar with the subtleties and pitfalls of $\mathrm{RO}(G)$-gradings are encouraged to peruse Adams' sardonic and timelessly entertaining introduction to [2, Section 6]. Incidentally, this is the only paper I

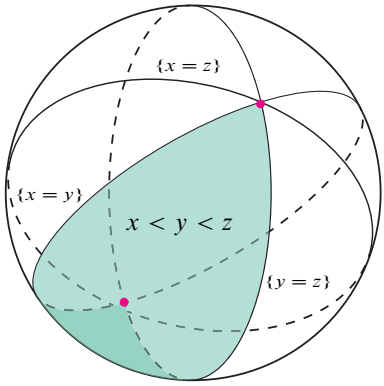


Figure 2. Unit sphere of the natural Σ_3-representation, with fundamental domain.

know of that is referenced in MathSciNet, and which references the Bible, specifically the book Genesis. Two other advantages of genuine G-spectra over naive G-spectra are better duality and transfer properties. In genuine G-spectra, finite G-CW-complexes become dualizable, and Atiyah duality holds for smooth compact G-manifolds. Transfers, also known as 'pushforward', 'wrong way' or 'Umkehr' maps, exist for equivariant fibrations whose fibers are finite CW-complexes. Special instances of the transfers provide the equivariant homotopy groups with the structure of a Mackey functor. For finite groups G, every genuine G-spectrum has a natural filtration whose subquotients 'are' G-Mackey functors. More formally, the heart of the most obvious t-structure on the genuine G-stable category is equivalent to the abelian category of G-Mackey functors.

Genuine G-spectra have been generalized beyond the class of compact Lie groups, most notably to profinite groups and to non-compact Lie groups [16, 19]; in these cases, the 'genuineness' is inherited from the finite quotients or the compact subgroups, respectively. One could contemplate inverting even more kinds of 'equivariant spheres', arising, for example, from homotopy representations. But this has no additional effect: every based finite G-CW-complex all of whose fixed point spaces, for all closed subgroups of G, are homology spheres, is already invertible in the genuine G-stable homotopy category. But there are variations that interpolate between 'doubly naive', 'naive' and 'genuine' by varying some of the available parameters: we could decide to test equivalences on fixed points for some, but not all, subgroups, and we could invert some, but not all, linear representation spheres. In my opinion, out of the zoo of different brands of G-spectra available, the genuine G-spectra are the richest and most interesting.

The cohomology theories mentioned above are representable by genuine G-spectra:

- For finite groups G, Bredon (co)homology with coefficients in a G-Mackey functor is represented by an associated *Eilenberg–MacLane G-spectrum*. Mind the fact that while Bredon homology and cohomology only need, respectively, a covariant or contravariant coefficient system, we need a G-Mackey functor to be representable by a genuine G-spectrum. This should not be surprising given that the equivariant homotopy groups of a genuine G-spectrum always come with this enhanced algebraic structure.
- The homology theory of G-equivariant bordism is represented by a *G-Thom spectrum* $\mathbf{mO}^G$. Well, not quite, but at least for a large class of groups. More precisely, for every compact Lie group G, the equivariant Thom–Pontryagin construction is a natural transformation from geometric bordism $\mathcal{N}_*^G(X)$ to the equivariant homology theory $\mathbf{mO}_*^G(X)$. And whenever G is isomorphic to a product of a finite group and a torus, then this map is an isomorphism. This result is usually

credited to Wasserman because it can be derived from his equivariant transversality theorem [47, Theorem 3.11]; a homotopy-theoretic proof is given in [37, Theorem 6.2.33]. The Thom–Pontryagin map is not bijective for more general compact Lie groups. For example, for $G = \mathrm{SU}(2)$, a homotopy-theoretic transfer in $\mathbf{mO}_0^{\mathrm{SU}(2)}$ from the maximal torus normalizer to $\mathrm{SU}(2)$ is not in the image; compare [37, Remark 6.2.34]. If we add additional tangential or normal structure to the picture, such as orientations or stably almost complex structures, it is not clear to me for which groups an appropriately adjusted Thom–Pontryagin map is bijective.

- *Stable equivariant bordism* $\mathfrak{N}_*^{G;S}(X)$ is a certain localization of equivariant bordism, obtained by formally inverting bordism classes associated with orthogonal G-representations. By a theorem of Bröcker and Hook [10], the homology theory $\mathfrak{N}_*^{G;S}(X)$ is represented, without any restriction on the compact Lie group, by another equivariant refinement $\mathbf{MO}^G$ of the classical bordism spectrum, a genuine G-spectrum known as *homotopical equivariant bordism*.
- Equivariant complex K-theory is represented by a genuine G-spectrum KU_G. This theory enjoys Bott periodicity for equivariant Spin^c-representations, and more generally Thom isomorphisms for equivariant Spin^c-vector bundles, the so-called *Atiyah–Bott–Shapiro* orientation.
- Every non-equivariant cohomology theory F^* is represented by a spectrum F. The associated Borel cohomology theory F_G^* is then represented by a genuine G-spectrum whose underlying spectrum is F, and on which G acts, in a certain precise sense, 'cofreely'.

There are always many different equivariant forms of any given non-equivariant spectrum. And as the two equivariant versions $\mathbf{mO}^G$ and $\mathbf{MO}^G$ of the classical Thom spectrum indicate, often several of the many available refinements are useful. Finding the 'best' equivariant form of a non-equivariant theory can be more of an art than a science.

Now that we discussed what genuine G-spectra do for us, what *are* they? Again, there are several legitimate answers that have evolved over time in much the same way as in the non-equivariant context, with 1-categorical models [31, 34] as the setup of choice for several decades, and with higher-categorical constructions becoming increasingly more popular in recent years. The idea of genuine G-spectra as the 'equivariant stabilization' of the fine homotopy theory of G-spaces has been implemented in different ways. One model-independent description works for all compact Lie groups: genuine G-spectra are the initial example of a presentably symmetric monoidal ∞-category equipped with a symmetric monoidal left adjoint from based G-spaces that inverts all linear representation spheres (see [21, Appendix C]); the universal functor is the equivariant suspension spectrum functor.

For finite groups, several other descriptions or characterizations of genuine G-spectra are available. If we work parameterized over the ∞-topos of genuine G-spaces, then there is an enhanced notion of 'stability' that includes the usual stability (i.e., suspension is invertible) and also a condition making precise that 'indexed sums are equivalent to indexed products'. Genuine G-spectra are then the free G-presentable G-stable G-∞-category on one object; compare [13, Theorem 10.7]. Moreover, the preferred t-structure on the genuine G-stable category with G-Mackey functors as its heart is a shadow of a much finer relationship: genuine G-spectra are equivalent to 'spectral Mackey functors' [4], i.e., additive functors to spectra from the ∞-category of spans of finite G-sets; see [11, Appendix A] for a streamlined proof.

3. Global spectra

All the examples of G-equivariant cohomology theories and G-equivariant spectra that I mentioned in the previous section occur 'uniformly for all groups'. This is always a clear sign that these theories are underlying a *globally equivariant* cohomology theory or a *global* spectrum. Of course, there are also many useful G-equivariant spectra that are specific to one group G, but since those do not support my narrative, I did not mention any examples.

What are global spectra? Well, this question again has more than one useful answer.

The orthogonal spectrum model. If the question 'What are global spectra?' refers to a mathematically rigorous definition, my answer is: global spectra are the localization of the category of orthogonal spectra at the class of global equivalences defined in [37, Definition 4.1.3]. Here, the reader may interpret 'localization' in terms of ∞-categories, or in the more classical sense of 1-categories. This definition of global spectra was proposed in January 2013, when I uploaded the first preprint version of the book [37] to my homepage. With the package came a model structure complementing the global equivalences and the proof that the global stable homotopy category is tensor triangulated and compactly generated by the suspension spectra of the global classifying spaces of all compact Lie groups. A somewhat differently looking definition of global spectra had been independently proposed by Anna Marie Bohmann in her thesis from 2011 [7, Chapter 4]. This part of Bohmann's thesis was later published as [6]; in Theorem 6.2 of that paper, Bohmann also shows that her 1-category of $\mathcal{I}$-spaces is equivalent to the category of orthogonal spectra. Bohmann does not discuss global equivalences.

The orthogonal spectrum model for global spectra is not only historically the first approach; there are still some features that, to my knowledge, and at the time of this

writing, can only be implemented in this setup. One example is the full theory of multiplicative ultra-commutativity, both stably and unstably, when taking all compact Lie groups into account. Another example is that without Joachim's orthogonal spectrum model for global equivariant K-theory [27], I would not know how to establish its rich structure as an ultra-commutative global ring spectrum.

But possibly the question 'What are global spectra?' really meant: How should one think of global spectra in conceptual and model-independent terms? Then, too, there are several possible answers. Before we explain what global spectra are conceptually, I hasten to clarify what they are not: global spectra are *not* the ∞-categorical stabilization of global spaces. In contemporary higher-categorical notation,

$$\mathrm{Sp}_{\mathrm{gl}} \not\simeq \mathrm{Sp} \otimes \mathrm{spc}_{\mathrm{gl}},$$

i.e., global spectra are not Lurie's tensor product, in presentable ∞-categories, of the ∞-category Sp of spectra and the ∞-topos $\mathrm{spc}_{\mathrm{gl}}$ of global spaces.

Global spectra are compatible families of equivariant spectra. I suspect that the idea of global spectra as 'coherently compatible families of equivariant spectra' predates all formal definitions by quite a bit, and might have already been around in the early days of equivariant stable homotopy theory. Practitioners of the subject were certainly well aware that several important families of equivariant cohomology theories, such as equivariant K-theory, Bredon homology, Borel homology, equivariant stable cohomology and equivariant bordism, naturally come in 'coherent families'.

The earliest published reference to a formal definition that I know of is by Lewis and May in [31, Chapter II, Definition 8.5]. Their version of global spectra consists of families $\{E_G\}_G$ of genuine G-spectra, in the sense of [31], for all compact Lie groups, equipped with transition morphisms $\xi_\alpha\colon \alpha^*(E_G) \to E_H$ in the homotopy category of H-spectra for all continuous homomorphisms $\alpha\colon H \to G$. These data must satisfy some conditions, including that ξ_α is an isomorphism whenever α is the inclusion of a closed subgroup. While these objects certainly qualify as 'coherent families of equivariant spectra', the coherence is rather crude in that the transition morphism ξ_α only lives *in the homotopy category*. This definition captures the natural examples, but today's higher-categorical perspective immediately craves a refined definition where the transition maps ξ_α are actual morphisms (and not just homotopy classes) and the equalities in the Lewis–May definition are witnessed by coherence homotopies.

It took some time before a satisfactory and fully higher-categorical definition of global spectra as coherent families of equivariant spectra was provided by Linskens, Nardin and Pol [32, Theorem 11.10]. In more technical terms, they identify the ∞-category of global spectra with a partially lax limit of a specific diagram of stable ∞-categories. That diagram, very roughly speaking, sends every compact Lie

group G to the ∞-category Sp_G of genuine G-spectra and every continuous homomorphism $\alpha\colon K \to G$ to the restriction functor $\alpha^*\colon \mathrm{Sp}_G \to \mathrm{Sp}_K$. The marked morphisms are the *injective* homomorphisms α, and the partial laxness implements the design feature that the transition morphisms $\xi_\alpha\colon \alpha^*(E_G) \to E_H$ of a global spectrum are equivalences whenever α is injective.

Global spectra are genuine cohomology theories on orbifolds and stacks. Various parts of pure mathematics have come up with notions of geometric objects that locally look like the quotient of a smooth object by a group action, in a way that remembers information about the isotropy groups of the action; examples are orbifolds in differential topology and stacks in algebraic geometry. Such 'stacky' objects can behave like smooth objects even if the underlying spaces have singularities. As for spaces, manifolds and schemes, cohomology theories are important invariants also for stacks and orbifolds, and examples such as ordinary cohomology or K-theory lend themselves to generalization. Orbispaces and global spaces are slightly different implementations of the same idea with spaces (or rather, homotopy types or animae) instead of 'smooth' objects. By now, there is a plethora of different approaches to defining global spaces. Historically, the first model is that of Gepner–Henriques [20]; it is given by contravariant continuous functors on a specific model for the global indexing category Glo, using homotopy orbits by the conjugation action on group homomorphisms. The author's orthogonal space model [37, Chapter 1] is a different setup that is easy to connect to orthogonal spectra. The closest thing to realizing the slogan that global objects have 'compatible actions of all groups' is the model as spaces with an action of the universal compact Lie group [39]; while the universal compact Lie group is neither compact nor a group, it contains all compact Lie groups as subgroups, in a specific way. For finite groups, even small 1-categories can be used to model global spaces; see [38]. Linskens, Nardin and Pol show in [32, Theorem 6.17] that global spaces are 'coherent families of equivariant spaces' in much the same way as global spectra are coherent families of equivariant spectra. Clough, Cnossen and Linskens [12] make precise in which way global spaces 'are' the homotopy theory of smooth stacks: they identify the ∞-category of global spaces with homotopy invariant sheaves of animae on the site of separated differentiable stacks.

Special cases of orbifolds are 'global quotients', sometimes denoted $G\backslash\backslash M$, for example, for a smooth action of a finite group G on a smooth manifold M. If $E = \{E_G\}$ is a global spectrum, the orbifold cohomology of $G\backslash\backslash M$ is supposed to be the G-equivariant E_G-cohomology of M. The partial-laxness condition in the coherent families picture of global spectra arises very naturally from the orbispace perspective: if H is a subgroup of G, and N is an H-manifold, then $H\backslash\backslash N$ and $G\backslash\backslash(G \times_H N)$ are different presentations of the same orbifold; so the E_H-cohomology of N better agree with the E_G-cohomology of $G \times_H N$, which forces

the underlying H-spectrum of E_G to be equivalent to E_H. In other words, the collection of equivariant spectra $\{E_G\}$ needs to be consistent under restriction to subgroups.

Global spectra are spectral global Mackey functors. The equivariant homotopy groups of global spectra come with a rich algebraic structure, namely that of a *global Mackey functor*. Restricted to finite groups, the notion of global Mackey functors has previously featured in various places in algebra and representation theory, sometimes under alternative names such as 'inflation functors' [48] or 'biset functors' [8].

Since I want to illustrate the computational power of global Mackey functors, I take the time to spell out a definition. My preferred definition is actually as additive functors from the *global Burnside category* of [37, Construction 4.2.1] to abelian groups. But there is an equivalent and more down-to-earth reformulation in terms of generating operations and relations that is more in tune with the narrative of this expository article.

Definition 3.1. A *global Mackey functor* M consists of:

(a) an *abelian group* $M(G)$ for every compact Lie group G;

(b) a *restriction homomorphism* $\alpha^*\colon M(G) \to M(K)$ for every continuous group homomorphism $\alpha\colon K \to G$; if α is the inclusion of a closed subgroup, we write res^G_K instead of α^*;

(c) a *transfer homomorphism* $\mathrm{tr}^G_H\colon M(H) \to M(G)$ for every closed subgroup H of G.

These data must satisfy the following relations:

(i) Restriction is contravariantly functorial: $(\alpha \circ \beta)^* = \beta^* \circ \alpha^*$.

(ii) Transfers are covariantly functorial: $\mathrm{tr}^G_H \circ \mathrm{tr}^H_K = \mathrm{tr}^G_K$.

(iii) Inner automorphisms act as the identity.

(iv) Transfers commute with restriction along epimorphisms.

(v) $\mathrm{tr}^G_H = 0$ whenever $\dim(N_G H) > \dim(H)$.

(vi) The *double coset formula* holds:

$$\mathrm{res}^G_K \circ \mathrm{tr}^G_H = \sum_{[M]\in K\backslash G/H} \chi^\sharp(M) \cdot \mathrm{tr}^K_{K\cap {}^g H} \circ c^*_g \circ \mathrm{res}^H_{K^g\cap H}\,. \tag{3.1}$$

Readers familiar with G-Mackey functors for finite groups will immediately realize that every global Mackey functor has underlying G-Mackey functors, by simply forgetting a lot of structure. Conversely, a generic G-Mackey functor does *not* extend to a global one. The simplest obstruction is that in a global functor, the presence of 'inflations' (i.e., restriction along surjective group homomorphisms) forces the

restriction map $\mathrm{res}^G_{\{1\}}\colon M(G) \to M(\{1\})$ to be a split epimorphism. For a general G-Mackey functor, this restriction map need not even be surjective.

Considering Mackey functors defined on compact Lie groups (as opposed to just finite groups) makes for a much richer story. While most of the structure behaves as for finite groups, the generalization exhibits some new features, most notably properties (v) and (vi) above. The double coset formula (vi) is the most complicated item in the definition, and for full details, I refer to [37, Section 3.4]. When H is a subgroup of G of the same dimension (and hence of finite index), the double coset formula (vi) simplifies a lot. In this case, the double coset space $K\backslash G/H$ is a discrete finite set, all the manifold components M are just points, all the internal Euler characteristics $\chi^\sharp(M)$ equal 1 and the double coset formula has the same form as in the realm of finite groups. But the double coset formula is particularly powerful when H and K have strictly smaller dimensions than G. Then the double coset space $K\backslash G/H$ is typically not discrete, but rather stratified by conjugacy classes of subgroups of K, and one needs to understand the stratification in concrete terms. I have gotten a lot of mileage out of special cases of the double coset formula where G is an orthogonal or unitary group, and K and H are smaller orthogonal or unitary groups or block subgroups.

In the 'coherent family' picture for global spectra, half of the algebraic structure of the associated global Mackey functors – the restriction homomorphisms – are the effect on homotopy groups of the transition maps $\xi_\alpha\colon \alpha^*(E_G) \to E_H$. The other half – the transfer maps – arise from the 'genuineness' of the equivariant constituent spectra in the coherent global family. But not only does a global spectrum E give rise to a global Mackey functor $\underline{\pi}_0(E) = \{\pi_0^G(E), \alpha^*, \mathrm{tr}_H^G\}$, this gadget is the full natural algebraic structure on the zeroth equivariant homotopy groups. Moreover, the global stable homotopy category comes with a preferred t-structure whose heart is the abelian category of global Mackey functors; see [37, Theorem 4.4.9].

For $\mathcal{F}in$-global spectra, i.e., global spectra indexed only on finite groups (as opposed to general compact Lie groups), there is an even tighter connection to Mackey functors: $\mathcal{F}in$-global spectra are equivalent to spectral global Mackey functors. The precise implementation of this slogan goes as follows. In [4], Barwick introduces a very general higher-categorical framework of spectral Mackey functors. Given an ∞-category with pullbacks and specified subcategories to parameterize 'forward' and 'backward' morphisms (subject, of course, to axioms), he defines spectral Mackey functors as additive functors from an ∞-category of spans to the ∞-category of spectra. When applied to finite G-sets for a finite group G, and allowing forward and backward functoriality for all morphisms, this yields a model for genuine G-spectra. Replacing finite G-sets by finite groupoids yields a model for $\mathcal{F}in$-global stable homotopy theory as spectral Mackey functors; here, arbitrary functors between finite groupoids are allowed for backward functoriality, but only faithful

functors give forward functoriality. Lenz [29, Theorem A] constructed an equivalence between this ∞-category of spectral Mackey functors on spans of finite groupoids and the ∞-category underlying Hausmann's symmetric spectrum model [22] for $\mathcal{F}in$-global spectra. Truncating this higher-categorical equivalence down to the level of 1-categories recovers the structure of global Mackey functors on the equivariant homotopy groups of global spectra.

After this discussion of the general nature of global spectra, I would like to mention some important global (co)homology theories, their representing global spectra and the associated global Mackey functors.

- Bredon equivariant (co)homology with integral coefficients is represented by the global Eilenberg–MacLane spectrum $H\underline{\mathbb{Z}}$. The associated global Mackey functor is 'constant', i.e., its value at every compact Lie group is $\mathbb{Z}$, and all restriction homomorphisms α^* are the identity. To satisfy the double coset formula, the transfers cannot generally be the identity; instead, tr_H^G is multiplication by the Euler characteristic of the homogeneous space G/H. In particular, if H has the same dimension as G, then the transfer is multiplication by the index $[G:H]$.
  More generally, every global Mackey functor has an associated global Eilenberg–MacLane spectrum. This construction implements an equivalence between the abelian category of global Mackey functors and the heart of the preferred t-structure on the global stable homotopy category; see [37, Theorem 4.4.9].
- Equivariant complex K-theory is represented by a global spectrum $\mathbf{KU}$. This spectrum comes with a highly structured commutative multiplication ('ultra-commutative') that not only yields products but also power operations and multiplicative norms on the equivariant cohomology theories. The Atiyah–Bott–Shapiro orientation is realized by a morphism of ultra-commutative ring spectra [27, Section 6]. The associated global Mackey functor sends a compact Lie group G to its representation ring $R(G)$, the Grothendieck group of finite-dimensional complex G-representations, with product induced by tensor product of representations. The restriction map $\alpha^*\colon R(G) \to R(K)$ is induced by restriction of representations along a continuous homomorphism $\alpha\colon K \to G$. The transfer map $\mathrm{tr}_H^G\colon R(H) \to R(G)$ along a closed subgroup inclusion $H \leq G$ is Segal's *smooth induction* [43, Section 2]. Whenever H has finite index in G, this sends an H-representation to the induced G-representation; in general, induction may send actual representations to virtual representations.
- The global sphere spectrum $\mathbb{S}$ is the coherent family of G-equivariant sphere spectra, with $\mathbb{S}_G$ representing G-equivariant cohomotopy. The associated Mackey functor sends G to the G-equivariant 0-stem $\pi_0^G(\mathbb{S})$, defined as the colimit over all G-representations V of homotopy classes of G-equivariant self-maps of S^V. When G is finite, then $\pi_0^G(\mathbb{S})$ is naturally isomorphic to the Burnside ring $A(G)$,

the Grothendieck group of finite G-sets. The unit morphism $\mathbb{S} \to \mathbf{KU}$ of the global K-theory spectrum refines the morphism of global Mackey functors whose value $A(G) \to R(G)$ sends a finite G-set to the associated permutation representation. In general, $\pi_0^G(\mathbb{S})$ is a free abelian group with basis the transfers $\mathrm{tr}_H^G(1)$ as H runs over all conjugacy classes of closed subgroups with finite Weyl group in G.

- The family of Borel equivariant spectra associated with a non-equivariant spectrum is global. The construction is implemented by the lax symmetric monoidal functor $b\colon \mathrm{Sp} \to \mathrm{Sp}_{\mathrm{gl}}$ from non-equivariant spectra to global spectra that is right adjoint to the forgetful functor; compare [37, Construction 4.5.21].

The two equivariant versions $\mathbf{mO}_G$ and $\mathbf{MO}_G$ of the classical Thom spectrum are both underlying global spectra, denoted **mO** and **MO**, respectively. I return to these and their complex cousins, **mU** and **MU**, in more detail in Section 4.

The extra global structure can be, and has been, put to good use. The following results are examples where globally equivariant structure was used in an essential way to resolve a problem in equivariant algebraic topology.

- In [36], I use the global structure to calculate the equivariant 0-stems of the G-equivariant spectra made from symmetric products of representation spheres.
- In [23], Hausmann proves a conjecture of Greenlees: for abelian compact Lie groups A, the homotopical equivariant bordism ring $\mathbf{MU}_A^*$ carries the universal A-equivariant formal group law.
- In [40], I prove a splitting theorem for the values of global Mackey functors at orthogonal, unitary and symplectic groups; an immediate consequence is the regularity of certain $U(n)$-equivariant Euler classes in homotopical equivariant bordism. This triggered the construction of $\mathbf{MU}_G$-Chern classes with enough regularity properties [41]. I will elaborate on these applications in the final section, Section 4.
- In [28], La Vecchia proves the Greenlees–May conjecture regarding a completion theorem for equivariant $\mathbf{MU}_G$-module spectra for general compact Lie groups, making essential use of the splitting from [40].
- In [24], Hausmann and I establish a universal characterization of geometric bordism for manifolds with commuting involutions, also known as equivariant bordism for elementary abelian 2-groups.

4. Case study: Equivariant bordism

In this final section, I want to illustrate by two examples how the global structure, specifically the restriction, inflation and transfer operations, and the double coset

formula, work in practice. Both examples involve *homotopical equivariant bordism* originally introduced by tom Dieck [46]. Before diving into the new results, I will provide some more context.

The non-equivariant bordism ring is well understood. In his celebrated paper on the subject, Thom [45] showed that the bordism ring $\mathcal{N}_*$ of smooth closed manifolds is a polynomial $\mathbb{F}_2$-algebra on infinitely many generators x_i of dimension $i \geq 2$ for $i+1$ not a power of 2. Thom also showed that for even i, the bordism classes of the real projective spaces $\mathbb{R}P^i$ can be taken as polynomial generators. Only slightly later, Dold [17] exhibited specific manifolds (now called the 'Dold manifolds') that can serve as the odd-dimensional polynomial generators. There is also a convenient necessary and sufficient criterion to decide when two smooth closed manifolds are bordant: all of their Stiefel–Whitney characteristic numbers need to coincide.

The serious study of equivariant bordism was initiated by the work [14] of Conner and Floyd. For a compact Lie group G, let $\mathcal{N}_n^G$ denote the group, under disjoint union, of n-dimensional smooth closed G-manifolds, taken up to equivariant bordism. Then for varying n, the graded abelian groups $\mathcal{N}_*^G$ have a ring structure by product of equivariant manifolds. Endowing non-equivariant manifolds with trivial G-actions makes $\mathcal{N}_*^G$ into an algebra over the non-equivariant bordism ring $\mathcal{N}_*$. Conner and Floyd carefully studied the bordism ring of manifolds with involution $\mathcal{N}_*^{C_2}$, i.e., the equivariant bordism ring for the group $G = C_2 = \{\pm 1\}$ with two elements. Among other things, Conner and Floyd showed that $\mathcal{N}_*^{C_2}$ is a free module of infinite rank (but of finite type) over the non-equivariant bordism ring $\mathcal{N}_*$. Alexander [3] described an explicit geometric basis constructed inductively from the $\mathbb{R}P^n$'s with involution $[x_0{:}x_1{:}\cdots{:}x_n] \mapsto [-x_0{:}x_1{:}\cdots{:}x_n]$, along with some partial information about the ring structure.

Not only do the bordism rings $\mathcal{N}_*^G$ exist 'uniformly' for all compact Lie groups, they also admit global functoriality. Indeed, a continuous homomorphism $\alpha\colon K \to G$ between compact Lie groups is automatically smooth, and hence restriction of actions along α yields a restriction homomorphism $\alpha^*\colon \mathcal{N}_n^G \to \mathcal{N}_n^K$. If H is a closed subgroup of G, and M a smooth closed H-manifold, then the induced G-space $G \times_H M$ has a unique smooth structure for which the projection $G \times M \to G \times_H M$ is a submersion. Moreover, this operation passes to a transfer homomorphism (or 'induction homomorphism')

$$\mathrm{tr}_H^G : \mathcal{N}_n^H \to \mathcal{N}_{n+\dim(G/H)}^G$$

between the bordism groups. If H has a smaller dimension than G, there is an increase in dimension by $\dim(G/H) = \dim(G) - \dim(H)$. This essentially (but not quite) makes the geometric bordism rings into a global Mackey functor.

As I already emphasized earlier, classical cohomology theories and spectra have many equivariant and global refinements. And sometimes, more than one such global

form is interesting; equivariant bordism is a prime example of this phenomenon. The complex (or unitary) bordism spectrum MU plays several important roles in stable homotopy theory:

(a) MU represents the bordism homology theory of smooth closed manifolds with stably almost complex structures.

(b) MU is the universal complex oriented cohomology theory, i.e., a multiplicative cohomology theory endowed with natural Thom isomorphisms.

(c) By Quillen's celebrated theorem [35], the coefficient ring MU^* together with the formal group law arising from the tautological complex orientation is an initial formal group law.

One might then desire, for every compact Lie group G, an equivariant theory that enjoys analogous properties. However, when G is nontrivial, properties (a) and (b) cannot simultaneously be satisfied. Indeed, manifolds have non-negative dimensions, so geometrically defined bordism theories are necessarily concentrated in non-negative degrees. However, equivariant complex orientations in particular provide Thom classes for G-representations; and for representations with trivial fixed points, the associated Euler classes are nontrivial homotopy classes in *negative* degrees.

In my opinion, the fact that equivariant theories cannot simultaneously be connective and complex oriented is a global feature, and not a bug. For example, there are two very interesting global ring spectra **mU** and **MU** that refine the non-equivariant theory, with **mU** being equivariantly connective and **MU** being equivariantly complex oriented. The fact that the geometric-to-homotopical morphism of global spectra $\mathbf{mU} \to \mathbf{MU}$ is *not* an equivalence at nontrivial groups can be seen as an incarnation of the failure of equivariant transversality.

Tom Dieck's theory $\mathbf{MU}^*_G$ is derived from equivariant Thom spaces of universal bundles over equivariant Grassmannians; it is the universal G-equivariantly complex oriented cohomology theory. The coefficient rings $\mathbf{MU}^*_G$ are reasonably well understood for *abelian* G, but, except possibly for $G = \{\pm 1\}$, not in the sense of a useful presentation by generators and relations. For abelian G, the graded ring $\mathbf{MU}^*_G$ is concentrated in even degrees and free as a module over the non-equivariant cobordism ring MU^*; the bundling homomorphism $\mathbf{MU}^*_G \to \mathrm{MU}^*(BG)$ is completion at the augmentation ideal; and the G-equivariant formal group law over $\mathbf{MU}^*_G$ arising from the tautological complex orientation is initial [23]. For nonabelian compact Lie groups, however, the equivariant bordism rings $\mathbf{MU}^*_G$ are still largely mysterious.

As already mentioned, the theories $\{\mathbf{MU}^*_G\}_G$ assemble into a global ring spectrum **MU**, and I now want to illustrate how the global structure has been used to elucidate the structure of the rings $\mathbf{MU}^*_G$ for unitary groups. A key player in this game is the *Euler class*

$$e_n \in \mathbf{MU}^{2n}_{U(n)}$$

of the tautological $U(n)$-representation on $\mathbb{C}^n$, already defined by tom Dieck [46, p. 347] when he introduced the theory. This Euler class participates in a long exact sequence (Gysin sequence):

$$\cdots \to \mathbf{MU}^{*-2n}_{U(n)} \xrightarrow{e_n \cdot} \mathbf{MU}^{*}_{U(n)} \xrightarrow{\mathrm{res}^{U(n)}_{U(n-1)}} \mathbf{MU}^{*}_{U(n-1)} \to \cdots .$$

The main result of [40] says that the restriction map $\mathrm{res}^{U(n)}_{U(n-1)}$ is surjective; thus, the long exact sequence decomposes into short exact sequences. Hence the Euler class e_n is *regular*, i.e., multiplication by e_n is injective.

The surjectivity of the restriction map $\mathrm{res}^{U(n)}_{U(n-1)}$ is not specific to **MU**, and it works in the same way for every global spectrum and even for every global Mackey functor. In fact, in [40], I use the global structure to exhibit a natural additive splitting. On the kernel of $\mathrm{res}^{U(n-1)}_{U(n-2)}$, the splitting is given by the composite

$$\mathbf{MU}^{*}_{U(n-1)} \xrightarrow{p^*} \mathbf{MU}^{*}_{U(n-1)\times U(1)} \xrightarrow{\mathrm{tr}^{U(n)}_{U(n-1)\times U(1)}} \mathbf{MU}^{*}_{U(n)} .$$

The first map is an inflation, i.e., restriction along the projection $p\colon U(n-1)\times U(1) \to U(n-1)$ to the first factor; the second map is the transfer from the block matrix subgroup $U(n-1)\times U(1)$ to $U(n)$. So this splitting has a similar flavor to parabolic induction in the representation theory of reductive algebraic groups. In verifying that the construction really provides a splitting, one needs to confront expressions like

$$\mathrm{res}^{U(n)}_{U(n-1)} \circ \mathrm{tr}^{U(n)}_{U(n-1)\times U(1)} ,$$

a specific composite of a transfer and a restriction, and thus work out an instance of the double coset formula (3.1) where the double coset space is not discrete. In this particular situation, the double coset space is a closed interval, with the minimal stratification by its two end points and the interior.

The final illustration of the global structure is the construction of Chern classes in homotopical equivariant bordism. Chern classes are important characteristic classes of complex vector bundles that were originally introduced in singular cohomology. Conner and Floyd [15, Corollary 8.3] constructed Chern classes for complex vector bundles in complex cobordism, nowadays referred to as Conner–Floyd–Chern classes.

Popular constructions of Chern classes derive them from the Euler classes of a complex orientation via either the projective bundle theorem or the 'splitting principle'. However, these tools are *not* available in the equivariant context. For example, the splitting theorem says that for complex oriented E, the homomorphism $E^*(BU(n)) \to E^*(BT)$ induced by the inclusion of the maximal torus

$T = U(1)^n$ into $U(n)$ is injective. In homotopical equivariant bordism, the restriction homomorphism

$$\mathrm{res}_T^{U(n)} : \mathbf{MU}^*_{U(n)} \to \mathbf{MU}^*_T$$

is however *not* injective for $n \geq 2$. So the traditional methods fail to produce Chern-like classes in homotopical equivariant bordism.

At this point, the global structure of **MU** comes to the rescue. We start from the Euler class $e_k \in \mathbf{MU}^{2k}_{U(k)}$ of the tautological $U(k)$-representation on $\mathbb{C}^k$, inflate along the projection $p\colon U(k) \times U(n-k) \to U(k)$ and then transfer up to $U(n)$: the k-th Chern class is

$$c_k = \mathrm{tr}^{U(n)}_{U(k)\times U(n-k)}(p^*(e_k)) \quad \text{in } \mathbf{MU}^{2k}_{U(n)}.$$

These Chern classes satisfy the analogous formal properties as their classical counterparts, including the equivariant refinement of the Whitney sum formula. Since c_k is defined as a transfer, verifying these properties again involves a double coset formula (3.1), this time with double coset space a simplex stratified by its faces. Despite the many formal similarities, there are crucial qualitative differences compared to Chern classes in complex oriented cohomology theories. The **MU**-Chern classes are not characterized by their restriction to the maximal torus, and some of them are zero divisors.

With the Whitney sum formula and the regularity of [40] at hand, it is only a short way to the main result of [41].

Theorem 4.1. *The Chern classes in homotopical equivariant bordism have the following properties:*

(i) *The Chern classes $c_n, c_{n-1}, \dots, c_1$ form a regular sequence in $\mathbf{MU}^*_{U(n)}$ that generates the augmentation ideal.*

(ii) *The completion of $\mathbf{MU}^*_{U(n)}$ at the augmentation ideal is a power series algebra over $\mathbf{MU}^*$ in the Chern classes, and tom Dieck's bundling map extends to an isomorphism*

$$(\mathbf{MU}^*_{U(n)})^\wedge_I \cong \mathbf{MU}^*(BU(n)).$$

I believe that the global techniques have further potential to illuminate the remaining mysteries surrounding homotopical equivariant bordism, such as the evenness conjecture or the quest for a notion of equivariant formal group law for nonabelian compact Lie groups.

This ends my tour of global homotopy theory. To wrap up, some of the main points I was trying to convey are:

- global homotopy theory is the home of equivariant phenomena with 'universal symmetry';

- many interesting equivariant cohomology theories are global;
- recognizing cohomology theories as global provides a rich algebraic structure;
- homotopical equivariant bordism witnesses the calculational impact of global equivariant structures.

Acknowledgments. I would like to thank Bastiaan Cnossen, Markus Hausmann, Tobias Lenz and Sil Linskens for numerous enlightening discussion on different aspects of global homotopy theory over many years. I am grateful to Catharina Stroppel and the anonymous referee for a host of helpful comments on this article.

Funding. This work was partially supported by the Hausdorff Center for Mathematics at Bonn University (DFG GZ 2047/1, project ID 390685813) and by the DFG SPP 1786 'Homotopy Theory and Algebraic Geometry' (SCHW860/1-1).

References

[1] J. F. Adams, *Stable homotopy and generalised homology*. Chicago Lect. Math., University of Chicago Press, Chicago, 1974, 373 pp. Zbl 0309.55016 MR 0402720

[2] J. F. Adams, Prerequisites (on equivariant stable homotopy) for Carlsson's lecture. In I. H. Madsen and R. A. Oliver (eds.), *Algebraic topology. Aarhus 1982 (Aarhus, 1982)*, pp. 483–532, Lecture Notes in Math. 1051, Springer, Berlin, 1984 Zbl 0553.55010 MR 0764596

[3] J. C. Alexander, The bordism ring of manifolds with involution. *Proc. Amer. Math. Soc.* **31** (1972), no. 2, 536–542 Zbl 0241.57015 MR 0290379

[4] C. Barwick, Spectral Mackey functors and equivariant algebraic K-theory (I). *Adv. Math.* **304** (2017), 646–727 Zbl 1348.18020 MR 3558219

[5] J. M. Boardman, *On stable homotopy theory and some applications*. Ph.D. thesis, University of Cambridge, 1964

[6] A. M. Bohmann, Global orthogonal spectra. *Homology Homotopy Appl.* **16** (2014), no. 1, 313–332 Zbl 1356.55009 MR 3217308

[7] A. M. R. Bohmann, *Topics in equivariant stable homotopy theory*. Ph.D. thesis, The University of Chicago, 2011 MR 2898610

[8] S. Bouc, Foncteurs d'ensembles munis d'une double action. *J. Algebra* **183** (1996), no. 3, 664–736 Zbl 0858.19001 MR 1401673

[9] G. E. Bredon, *Equivariant cohomology theories*. Lecture Notes in Math. 34, Springer, Berlin, 1967, 68 pp. Zbl 0162.27202 MR 0214062

[10] T. Bröcker and E. C. Hook, Stable equivariant bordism. *Math. Z.* **129** (1972), no. 3, 269–277 Zbl 0236.57020 MR 0321119

[11] D. Clausen, A. Mathew, N. Naumann, and J. Noel, Descent and vanishing in chromatic algebraic K-theory via group actions. *Ann. Sci. Éc. Norm. Supér. (4)* **57** (2024), no. 4, 1135–1190 Zbl 1550.19001 MR 4773302

[12] A. Clough, B. Cnossen, and S. Linskens, Global spaces and the homotopy theory of stacks. 2024, arXiv:2407.06877v2

[13] B. Cnossen, T. Lenz, and S. Linskens, Partial parametrized presentability and the universal property of equivariant spectra. *Trans. Amer. Math. Soc.* **378** (2025), no. 10, 6913–6974 Zbl 08092173 MR 4956364

[14] P. E. Conner and E. E. Floyd, *Differentiable periodic maps*. Ergeb. Math. Grenzgeb. (N.F.) 33, Springer, Heidelberg, 1964, 148 pp. Zbl 0125.40103 MR 0176478

[15] P. E. Conner and E. E. Floyd, *The relation of cobordism to K-theories*. Lecture Notes in Math. 28, Springer, Berlin, 1966, 116 pp. Zbl 0161.42802 MR 0216511

[16] D. Degrijse, M. Hausmann, W. Lück, I. Patchkoria, and S. Schwede, Proper equivariant stable homotopy theory. *Mem. Amer. Math. Soc.* **288** (2023), no. 1432, 142 pp. Zbl 1528.55001 MR 4627088

[17] A. Dold, Erzeugende der Thomschen Algebra $\mathfrak{N}$. *Math. Z.* **65** (1956), no. 1, 25–35 Zbl 0071.17601 MR 0079269

[18] A. D. Elmendorf, Systems of fixed point sets. *Trans. Amer. Math. Soc.* **277** (1983), no. 1, 275–284 Zbl 0521.57027 MR 0690052

[19] H. Fausk, Equivariant homotopy theory for pro-spectra. *Geom. Topol.* **12** (2008), no. 1, 103–176 Zbl 1135.55005 MR 2377247

[20] D. Gepner and A. Henriques, Homotopy theory of orbispaces. 2007, arXiv: math/0701916v1

[21] D. Gepner and L. Meier, On equivariant topological modular forms. *Compos. Math.* **159** (2023), no. 12, 2638–2693 Zbl 1532.55006 MR 4664814

[22] M. Hausmann, Symmetric spectra model global homotopy theory of finite groups. *Algebr. Geom. Topol.* **19** (2019), no. 3, 1413–1452 Zbl 1418.55006 MR 3954287

[23] M. Hausmann, Global group laws and equivariant bordism rings. *Ann. of Math. (2)* **195** (2022), no. 3, 841–910 Zbl 1503.57033 MR 4413745

[24] M. Hausmann and S. Schwede, The universal property of bordism of commuting involutions. 2024, arXiv:2406.00404v1

[25] S. Illman, Equivariant singular homology and cohomology. I. *Mem. Amer. Math. Soc.* **1** (1975), no. 2 (Number 156), 74 pp. Zbl 0297.55003 MR 0375286

[26] S. Illman, The equivariant triangulation theorem for actions of compact Lie groups. *Math. Ann.* **262** (1983), no. 4, 487–501 Zbl 0488.57014 MR 0696520

[27] M. Joachim, Higher coherences for equivariant K-theory. In A. Baker and B. Richter (eds.), *Structured ring spectra*, pp. 87–114, London Math. Soc. Lecture Note Ser. 315, Cambridge University Press, Cambridge, 2004 Zbl 1070.19007 MR 2122155

[28] M. La Vecchia, The local (co)homology theorems for equivariant bordism. *Geom. Topol.* **28** (2024), no. 2, 627–639 Zbl 1546.55012 MR 4718124

[29] T. Lenz, Global homotopy theory via spectral Mackey functors. [v1] 2022, [v2] 2025, arXiv:2202.07272v2

[30] G. Lewis, J. P. May, and J. McClure, Ordinary $RO(G)$-graded cohomology. *Bull. Amer. Math. Soc. (N.S.)* **4** (1981), no. 2, 208–212 Zbl 0477.55009 MR 0598689

[31] L. G. Lewis, Jr., J. P. May, and M. Steinberger, *Equivariant stable homotopy theory. With contributions by J. E. McClure*. Lecture Notes in Math. 1213, Springer, Berlin, 1986, 538 pp. Zbl 0611.55001 MR 0866482

[32] S. Linskens, D. Nardin, and L. Pol, Global homotopy theory via partially lax limits. *Geom. Topol.* **29** (2025), no. 3, 1345–1440 Zbl 08055644 MR 4918109

[33] J. Lurie, Higher algebra. Preprint dated 18 September 2017, available from the author's homepage

[34] M. A. Mandell and J. P. May, Equivariant orthogonal spectra and S-modules. *Mem. Amer. Math. Soc.* **159** (2002), no. 755, 108 pp. Zbl 1025.55002 MR 1922205

[35] D. Quillen, On the formal group laws of unoriented and complex cobordism theory. *Bull. Amer. Math. Soc.* **75** (1969), 1293–1298 Zbl 0199.26705 MR 0253350

[36] S. Schwede, Equivariant properties of symmetric products. *J. Amer. Math. Soc.* **30** (2017), no. 3, 673–711 Zbl 1367.55005 MR 3630085

[37] S. Schwede, *Global homotopy theory*. New Math. Monogr. 34, Cambridge University Press, Cambridge, 2018, 828 pp. Zbl 1451.55001 MR 3838307

[38] S. Schwede, Categories and orbispaces. *Algebr. Geom. Topol.* **19** (2019), no. 6, 3171–3215 Zbl 1435.55008 MR 4023338

[39] S. Schwede, Orbispaces, orthogonal spaces, and the universal compact Lie group. *Math. Z.* **294** (2020), no. 1–2, 71–107 Zbl 1439.55014 MR 4050064

[40] S. Schwede, Splittings of global Mackey functors and regularity of equivariant Euler classes. *Proc. Lond. Math. Soc. (3)* **125** (2022), no. 2, 258–276 Zbl 1529.55009 MR 4466235

[41] S. Schwede, Chern classes in equivariant bordism. *Forum Math. Sigma* **12** (2024), article no. e7, 11 pp. Zbl 1539.55003 MR 4685685

[42] S. Schwede and B. Shipley, A uniqueness theorem for stable homotopy theory. *Math. Z.* **239** (2002), no. 4, 803–828 Zbl 1009.55012 MR 1902062

[43] G. Segal, The representation ring of a compact Lie group. *Publ. Math. Inst. Hautes Études Sci.* **34** (1968), 113–128 Zbl 0209.06203 MR 0248277

[44] E. H. Spanier and J. H. C. Whitehead, A first approximation to homotopy theory. *Proc. Natl. Acad. Sci. USA* **39** (1953), no. 7, 655–660 Zbl 0051.14001 MR 0056290

[45] R. Thom, Quelques propriétés globales des variétés différentiables. *Comment. Math. Helv.* **28** (1954), no. 1, 17–86 Zbl 0057.15502 MR 0061823

[46] T. tom Dieck, Bordism of G-manifolds and integrality theorems. *Topology* **9** (1970), no. 4, 345–358 Zbl 0209.27504 MR 0266241

[47] A. G. Wasserman, Equivariant differential topology. *Topology* **8** (1969), no. 2, 127–150 Zbl 0215.24702 MR 0250324

[48] P. Webb, Two classifications of simple Mackey functors with applications to group cohomology and the decomposition of classifying spaces. *J. Pure Appl. Algebra* **88** (1993), no. 1–3, 265–304 Zbl 0787.55012 MR 1233331

Stefan Schwede
Mathematisches Institut, Universität Bonn, Endenicher Allee 60, 53115 Bonn, Germany;
schwede@math.uni-bonn.de

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/33

New tools in hierarchical hyperbolicity: A survey

Alessandro Sisto

Abstract. The aim of this short survey is to highlight various tools that have been developed to study hierarchically hyperbolic spaces (HHSs) in recent years, with particular emphasis on those that require little to no knowledge of the HHS machinery to be used.

1. Introduction

Hierarchically hyperbolic groups were defined in [10] as a generalisation of hyperbolic groups and a common framework for studying mapping class groups and (special) cube complexes. The definition is an axiomatic version of the machinery developed by Masur and Minsky to study mapping class groups [43, 44].

As it turns out, there are a lot more hierarchically hyperbolic groups and spaces (HHGs and HHSs) than those motivating the original definition. In fact, examples of a wide variety of different flavours are now known, from "most" types of 3-manifold groups to extra-large-type Artin groups; see among others [7–9, 12–14, 18, 20, 33–35, 37, 55]. Many new results have been obtained using the HHS machinery (see, for instance, [1, 3, 5, 9–11, 22, 25, 27, 30, 40, 47, 50]), and many of these are new even for mapping class groups or other very well-studied examples. Some of these results are mentioned below; just to mention a couple more here, these include the strong Tits alternative [22, 23], the quasiflat theorem [11], the existence of bicombings [30], uniform exponential growth for a large subclass [3], and mapping class groups being quasi-isometric to CAT(0) cube complexes [47].

The definition of hierarchical hyperbolicity is explained and commented on in detail in the survey [54]. This survey, instead, focuses on a variety of tools and techniques that have been developed in recent years to study HHSs, and which make the study of HHSs much more accessible to the "uninitiated", in various ways. Indeed, the content of some of them is that HHSs also carry other very rich structures, such as

Mathematics Subject Classification 2020: 20F65.
Keywords: hyperbolic groups, nonpositive curvature, mapping class groups.

injective metrics. The theory of combinatorial HHSs, on the other hand, makes it substantially easier to construct new examples of HHSs, as one no longer needs to check the definition directly. For the purposes of this survey, no particular knowledge of HHSs is required, and we only recall very briefly that the main data of an HHS structure on a metric space X is a collection of uniformly hyperbolic spaces $\{\mathcal{C}(Y)\}_{Y\in\mathcal{S}}$, each equipped with a coarsely Lipschitz map $\pi_Y : X \to \mathcal{C}(Y)$ (with uniform constants across all Y). The interested reader is referred to the survey [54] for more information, where we note that more "hierarchical" and basic tools are discussed already. Notably, the cubulation of hulls from [11] is discussed there, and interestingly, a few of the results mentioned below (including both re-metrisation results) heavily rely on it and its further refinements and improvements; see [15, 21, 25, 26]. I decided not to discuss this further here as I see this as more intrinsically hierarchical than the tools I want to focus on.

1.1. Outline

In Section 2, we describe combinatorial HHSs, which provide a very useful criterion to construct new HHSs. Indeed, after the development of combinatorial HHSs, there has been a substantial increase in activity in finding new HHSs.

In Section 3, we describe two ways of re-metrising an HHS. One yields injective metrics, while the other (which works under an additional but very general condition) yields asymptotically CAT(0) metrics. Both are metrics with nonpositive curvature characteristics and have been studied independently of HHSs (especially injective metrics).

In Section 4, we discuss various other tools being developed, including Dehn-filling-like quotients, curtains, $\mathbb{R}$-cubings, and higher-rank JSJ decompositions.

2. Combinatorial HHSs

In this section, we describe the criterion for hierarchical hyperbolicity devised in [7], which has now been used numerous times to construct new HHSs. It seems likely that more new examples will be constructed using this criterion. The discussion below covers the main idea and formulates the main statement, but more information is available in the "User's Guide" in [7, Section 1.5], and also in [32], where a more explicit description of the candidate combinatorial model for an HHS is given.

First, we begin with an informal discussion. Roughly speaking, the criterion says that if G acts "nicely" on a hyperbolic simplicial complex X where all links of simplices are also hyperbolic, then G is hierarchically hyperbolic. (By link of a simplex we mean the intersection of the links of its vertices.) Here, "nicely" includes the requirements that the action is cocompact and that stabilisers of maximal simplices

are finite. With these hypotheses, the group G is quasi-isometric to a graph W whose vertex set is the set of maximal simplices of X, and where one adds finitely many orbits of edges to make this set connected. (This is a simple generalisation of letting a finitely generated group G act on itself and adding finitely many orbits of edges to form the Cayley graph.) The idea is then that the hyperbolic spaces for an HHS structure on G are all links of simplices of X (where here we include the link of the empty simplex, which is X itself). This is not accurate, though, because X will have some discrete links, so it is basically never the case that a simplicial complex has all links hyperbolic. To remedy this, one considers "augmented links", adding edges suitably. This is similar to the procedure of adding edges to the set of maximal simplices to obtain a graph quasi-isometric to the group.

With this discussion in mind, here are the actual definitions. Throughout, let X be a finite-dimensional simplicial complex. First of all, we describe the setup of a pair (simplicial complex, graph whose vertices are maximal simplices), and also, we describe additional edges that will be needed later.

Definition 2.1 (X-graph, W-augmented graph). An *X-graph* is a graph W whose vertex set is the set of all maximal simplices of X.

For a flag simplicial complex X and an X-graph W, the *W-augmented graph* X^{+W} is the graph defined as follows:

- The 0-skeleton of X^{+W} is $X^{(0)}$.
- If $v, w \in X^{(0)}$ are adjacent in X, then they are adjacent in X^{+W}.
- If two vertices in W are adjacent, then we consider σ, ρ, the associated maximal simplices of X, and in X^{+W}, we connect each vertex of σ to each vertex of ρ.

Next, we describe augmented links, that is, links with additional edges. We also need an auxiliary space, denoted by Y_Δ below, which is perhaps the least intuitive feature of the whole setup.

Definition 2.2 (Augmented links). Let W be an X-graph. For each simplex Δ of X, let its *augmented link* $\mathcal{C}(\Delta)$ be the induced subgraph of X^{+W} spanned by $\mathrm{Link}(\Delta)^{(0)}$. Also, let Y_Δ be the subgraph of X^{+W} induced by the set of vertices

$$X^{(0)} - \bigcup_{\Sigma\colon \mathrm{Link}(\Sigma)=\mathrm{Link}(\Delta)} \Sigma^{(0)}.$$

Finally, we describe the axioms of combinatorial HHSs. A more general version of item (3) below suffices, but we state a simpler one which most often holds in practice. Note that all conditions except for (2) are combinatorial conditions on X and W, with (2) being the only geometric condition.

Definition 2.3 (Combinatorial HHS). A *combinatorial HHS* (X, W) consists of a flag simplicial complex X and an X-graph W satisfying the following conditions:

(1) There exists $n \in \mathbb{N}$ such that any chain $\mathrm{Link}(\Delta_1) \subsetneq \cdots \subsetneq \mathrm{Link}(\Delta_i)$, where each Δ_j is a simplex of X, has length at most n.

(2) There is a constant δ so that for each non-maximal simplex Δ, the subgraph $\mathcal{C}(\Delta)$ is δ-hyperbolic and (δ, δ)-quasi-isometrically embedded in Y_Δ.

(3) Whenever Δ and Σ are non-maximal simplices, there exist simplices Π, Π' such that $\mathrm{Link}(\Delta) \cap \mathrm{Link}(\Sigma) = \mathrm{Link}(\Delta \star \Pi) \star \Pi'$, where $\star$ denotes the join.

(4) If v, w are distinct non-adjacent vertices of $\mathrm{Link}(\Delta)$, for some simplex Δ of X, contained in W-adjacent maximal simplices, then they are contained in W-adjacent simplices of the form $\Delta \star \Sigma$.

We are finally ready to state the criterion from [7].

Theorem 2.4. *Let (X, W) be a combinatorial HHS. Then W is a hierarchically hyperbolic space.*

Note that [7] actually gives a concrete HHS structure for W, and the associated hyperbolic spaces are exactly the augmented links. When there is a group action of G on (X, W) which is compatible in a suitable sense, then one can conclude that G is a hierarchically hyperbolic group.

Very recently, a different version of combinatorial HHS has been used in [46].

2.1. Case study: Pants graphs

To conclude, we illustrate the criterion in a case where it works incredibly well, namely, the case of the pants graph. The pants graph $\mathcal{P}$ of a fixed surface Σ is the graph with vertex set all pants decompositions of Σ, and where two pants decompositions are connected if they differ by a "move". Here, a move consists of changing a single curve with one that intersects it once or twice (depending on whether the other curves cut out a one-holed torus or a four-holed sphere). Notice that the vertices of $\mathcal{P}$ are exactly the maximal simplices of the curve graph $\mathcal{C}$ of Σ so that, in the terminology above, $\mathcal{P}$ is a $\mathcal{C}$-graph. Checking the combinatorial conditions is left as an exercise to the reader. Each augmented link is the curve graph $\mathcal{C}(Y)$ of a subsurface Y, with additional edges connecting curves that intersect once or twice, but these do not change the quasi-isometry type. Therefore, augmented links are hyperbolic. Finally, the quasi-isometric embedding part of condition (2) can be checked using subsurface projections. As it turns out, for Y the subsurface associated with the simplex Δ, the curves that have well-defined subsurface projection to $\mathcal{C}(Y)$ correspond exactly to the vertex set of Y_Δ, and it is easy to see that subsurface projections

are coarsely Lipschitz (where they are defined). This provides a coarse retraction of Y_Δ onto $\mathcal{C}(\Delta)$, therefore showing that the former is quasi-isometrically embedded into the latter. To my knowledge, in all applications of combinatorial HHSs, the quasi-isometric embedding condition is checked by constructing a suitable coarse retraction.

3. Re-metrisation

Hierarchically hyperbolic groups display features of nonpositive curvature, but not all of them are nonpositively curved in the most classical sense, meaning that not all of them are CAT(0). Indeed, most mapping class groups are not CAT(0) [16], but it is also not known whether all hyperbolic groups are CAT(0). Despite this, hierarchically hyperbolic groups act properly and coboundedly on spaces that are nonpositively curved in other senses.

3.1. Injective metrics

The first result of this type is due to Haettel–Hoda–Petyt [30], who showed the following.

Theorem 3.1 ([30]). *All hierarchically hyperbolic groups act properly and coboundedly on some injective space.*

Injective spaces have different characterisations, one being that they are geodesic metric spaces where, given balls that pairwise intersect, they all have a common intersection. Often more useful than the definition is the explicit construction of the *injective hull* of a metric space X, which is the smallest injective space containing X (in particular, $E(X) = X$ if X is injective). Injective spaces have various features of nonpositive curvature, including barycentres for finite sets and bicombings with convexity properties. Both of these were exploited in [30] to prove new results about HHSs, namely that they contain finitely many conjugacy classes of finite subgroups and that they are semi-hyperbolic, and hence have solvable conjugacy problem.

Injective metric spaces have been brought to the attention of geometric group theorists by Lang, who showed that all hyperbolic groups act geometrically on injective spaces [39], despite the fact that we still do not know whether all hyperbolic groups are CAT(0). They have been an active area of study since, and the theorem above provided further motivation to pursue their study. The reader interested in injective spaces (perhaps to obtain new results about HHSs) is referred to [29, 39, 57]. As a word of warning, it is not known whether one can improve "coboundedly" to "cocompactly" in the theorem, and probably one cannot do this in fact. This sometimes causes problems when applying results on injective spaces to HHGs.

3.2. Asymptotically CAT(0) metrics

There is now another class of spaces for HHGs to act on, even though (at least for the moment) one needs the extra technical assumption of colourability, where there is essentially only one example of non-colourable HHG, constructed specifically to have this property in [31]. The relevant class of spaces is that of *asymptotically* CAT(0) *spaces*, which were defined by Kar [38] (at least in the case of geodesic spaces, we need a slight extension here). Roughly, a space is asymptotically CAT(0) if all triangles satisfy the CAT(0) inequality up to an error sublinear in their size. Note that injective spaces have fine nonpositive curvature features, while asymptotically CAT(0) spaces have coarse nonpositive curvature features. Here are all the relevant definitions.

Given three points x, y, z in a metric space, we denote by $\bar{\Delta}(x, y, z)$ a comparison triangle in the Euclidean plane $\mathbb{E}^2$, with vertices $\bar{x}, \bar{y}, \bar{z}$. Given a point p on a $(1, C)$-quasi-geodesic γ joining x, y for a fixed C, we call a point $\bar{p}$ on the geodesic $[\bar{x}, \bar{y}]$ a *comparison point* if $|d(p, x) - d(\bar{p}, \bar{x})| \leq C$. Also, we will denote $d(p) = \min\{d(x, p), d(p, y)\}$.

Definition 3.2 (Sublinear CAT(0)). Given a sublinear and non-decreasing function κ, we say that a triangle Δ of $(1, C)$-quasi-geodesics satisfies the CAT(0) *condition up to* κ if the following holds. Let x, y, z be the vertices of the triangle, and let p and q be points on the triangle. Fixing a comparison triangle and comparison points, we have

$$d(p, q) \leq d(\bar{p}, \bar{q}) + \kappa(d(p)) + \kappa(d(q)).$$

The following definition is a small variation on [38, Definition 6].

Definition 3.3. Given a constant C and a sublinear function κ, we say that a metric space X is (C, κ)-*asymptotically* CAT(0) if the following hold:

(1) Every pair of points of X is connected by a $(1, C)$-quasi-geodesic.

(2) Every triangle Δ of $(1, C)$-quasi-geodesics satisfies the CAT(0) condition up to κ.

One of the main theorems of [24] is as follows.

Theorem 3.4. *All colourable hierarchically hyperbolic groups act geometrically on some* (C, κ)*-asymptotically* CAT(0) *space.*

The main motivation to prove this theorem is that it is a key step in proving the Farrell–Jones conjecture for a large class of hierarchically hyperbolic groups, which is the main application in [24]. This is a bit more restrictive than colourable ones,

as we do not quite prove that a geometric action on an asymptotically CAT(0) space suffices for the Farrell–Jones conjecture to hold, since for one part of the argument, we use more of the HHS structure.

Asymptotically CAT(0) spaces have not been studied as much as injective spaces, but hopefully the theorem provides plenty of motivation to do so. We formulate this as a very open-ended problem.

Problem 3.5. Study asymptotically CAT(0) spaces.

4. Further tools

In this section, we discuss other tools more briefly (most of these are under development).

4.1. Dehn-filling-like quotients

While the theory I will describe in this subsection is still under development, it already has striking applications. In analogy with relatively hyperbolic Dehn filling [28, 45], there are now a few known ways of taking quotients of HHGs and obtaining new HHGs. The first instance of this was in [9], where, in the context of mapping class groups, it is shown that quotients of mapping class groups by suitable powers of pseudo-Anosovs are still HHG. (Relatedly, random quotients have been considered in [2].) In [7], instead, quotients by suitable powers of all Dehn twists are shown to be HHG, and the crucial feature of this construction is that it reduces a natural measure of complexity, where minimal complexity HHGs are simply hyperbolic groups. Therefore, this can be thought of as a first step in constructing hyperbolic quotients of mapping class groups, but this procedure requires certain hyperbolic groups encountered during the construction to be residually finite. Under the assumption that, say, all hyperbolic groups are residually finite, the inductive procedure is carried out in [7]. As it turns out, this allows us to relate residual finiteness of "enough" hyperbolic groups to various open questions on profinite properties of mapping class groups. Taking a further, striking step in this direction, Wilton recently showed, using the construction of [7] almost as a black box, that if "enough" hyperbolic groups are residually finite, then mapping class groups have the congruence subgroup property [56]. In a different direction, a related type of quotient of HHGs is used in [42] to show that many Artin groups are Hopfian. Roughly, the point for those Artin groups is that a single "complexity reduction" is enough to obtain a hyperbolic group. The general idea is then to exploit hyperbolic quotients to deduce algebraic properties of an HHG from those of its hyperbolic quotients.

We do not have a complete theory of "Dehn-filling-like" quotients of HHGs, but even partial versions of it have several potential applications, as illustrated by the examples above. I believe that a good question to drive research in this area is the following conjecture, stated in slightly simplified form for concreteness.

Conjecture 4.1 ([41]). *Let S be any surface of finite type, and let $g_1, \dots, g_l \in \mathrm{MCG}(S)$. Then there exists $N \in \mathbb{N} - \{0\}$ such that $\mathrm{MCG}(S)/\langle\langle\{g_i^N\}\rangle\rangle$ is hierarchically hyperbolic.*

The conjecture is proven for the five-holed sphere and almost all collections of elements g_i in [42].

4.2. Curtains

Curtains were originally introduced for CAT(0) spaces in [48] as generalisations of hyperplanes in CAT(0) cube complexes. The technology has been pushed considerably further in [49]. The construction is very general, and here we only focus on the HHS aspects of it. Roughly, a hyperbolic space can be split into two parts by fixing a geodesic, looking at the closest point projection onto it, and considering points that project before or after a certain point. In an HHS X, one can do something similar, first using one of the maps $\pi_Y : X \to \mathcal{C}(Y)$, where $\mathcal{C}(Y)$ is one of the hyperbolic spaces from the HHS structure. This feeds into the very general machinery of [49] to yield a metric space $X_{\mathcal{C}}$ which has many remarkable properties. First, it is injective, so this gives another construction of injective metric spaces for HHSs. Secondly, it admits a bicombing by rough geodesics which interact nicely with the HHS structure, namely hierarchy paths. It is not clear whether the previous construction of injective spaces for HHSs had this feature. There is also another property that $X_{\mathcal{C}}$ satisfies, namely almost convexity, a convexity property of balls. Finally, $X_{\mathcal{C}}$ is not only a metric space but also a median algebra. The HHS X is only coarse median, so $X_{\mathcal{C}}$ has better "local" properties, making it easier to work with; in some sense, $X_{\mathcal{C}}$ is at the same time the injective hull and the "median hull" of X.

Overall, besides the construction having many further applications beyond HHSs, the space $X_{\mathcal{C}}$ sketched above is a natural and convenient geometric model to use. We note that in forthcoming work of Hagen, Petyt, and Zalloum, curtains will be used to start with HHS data (hyperbolic spaces, relations, etc.) and construct an HHS out of this. A more categorical approach to this "inverse problem" is being developed by Tang.

4.3. $\mathbb{R}$-cubings

It is well known that asymptotic cones of hyperbolic groups are $\mathbb{R}$-trees (which can be defined as 0-hyperbolic spaces). Actions on $\mathbb{R}$-trees have several applications to

hyperbolic groups and beyond because of this. Casals-Ruiz, Hagen, and Kazachkov found a generalisation of $\mathbb{R}$-trees, which they call $\mathbb{R}$-cubings, and they show that asymptotic cones of HHSs are indeed $\mathbb{R}$-cubings and in turn use this to show that mapping class groups, among many other hierarchically hyperbolic groups, have unique asymptotic cones up to bilipschitz equivalence [17]. Roughly, an $\mathbb{R}$-cubing is a particular type of median space contained in a product of (typically infinitely many) $\mathbb{R}$-trees, consisting of all points satisfying a certain system of equations. These equations are essentially exact versions of HHS axioms, meaning that the constants involved in those (which can be thought of as margins of error) are set to 0. It is envisioned that one can use actions on $\mathbb{R}$-cubings to study certain HHGs in roughly the same way one uses $\mathbb{R}$-trees to study hyperbolic groups, by introducing non-discrete versions of ideas from special cube complexes in the sense of [36].

4.4. Higher-rank JSJ decompositions

Sela is working on a series of papers studying automorphisms of colourable HHGs satisfying a weak acylindricity condition [52, 53]. This is in analogy with the theory of JSJ decompositions of hyperbolic groups, and indeed, an object that can be viewed as a higher-rank JSJ decomposition is constructed. The higher-rank decomposition encodes the dynamics of individual automorphisms and the structure of the outer automorphism group of an HHG. For a one-ended hyperbolic group G, there exists an epimorphism from a finite index subgroup of $\mathrm{Out}(G)$ onto a direct product of the mapping class groups of 2-orbifolds (appearing in the canonical JSJ decomposition of G), where the kernel is a finitely generated virtually abelian group. The state of the art for HHGs, satisfying additional conditions, is not quite a direct analogue of this, but instead two groupoids get associated with G, and outer automorphisms are associated with finitely many morphisms rather than just one. There is still a certain homomorphism from a finite index subgroup of $\mathrm{Out}(G)$ into the direct sum of (finitely many) mapping class groups of 2-orbifolds and general linear groups. This data still suffices to show that certain natural conditions on G imply that $\mathrm{Out}(G)$ is locally virtually nilpotent, for instance, but many open questions on $\mathrm{Out}(G)$ still remain.

The reader is also referred to [4] for results on acylindricity conditions emerging from Sela's work.

5. Some open problems

There are many open questions on hierarchical hyperbolicity, and indeed a problem list is available at http://comet.lehman.cuny.edu/behrstock/HHS.html, together with further material on HHSs. Therefore, here I only mention the problems I discussed at my ECM talk. These arise in the context of trying to understand the topology of a

manifold from a triangulation, assuming that the fundamental group is hierarchically hyperbolic. Perhaps the tools advertised in this survey could be useful to tackle some of them.

The word and conjugacy problems are solvable for HHGs, but the proofs are rather indirect and give very slow running time, worse than exponential, in fact. It is then natural to ask the following questions.

Question 5.1. How quickly can one solve the word problem in an HHG?

Question 5.2. How quickly can one solve the conjugacy problem in an HHG?

It is a very deep and difficult theorem that the isomorphism problem is solvable for hyperbolic groups [19, 51]. One can then very ambitiously ask whether the same holds for all HHGs or, still ambitiously but less so, whether there are interesting subclasses where it can be solved.

Question 5.3. Is the isomorphism problem solvable for HHGs? Is it solvable at least in some interesting subclasses?

Finally, [24] proves the Farrell–Jones conjecture for "most" HHGs, the main ones left out being the non-colourable examples from [31]. These are however cubical groups, so they do satisfy the Farrell–Jones conjecture [6], and indeed I am not aware of any specific HHGs for which the Farrell–Jones conjecture is now not known to hold. Still, it is natural to ask the following, and in fact, it is very likely that removing the additional assumptions from [24] will require independently interesting mathematics.

Question 5.4. Does the Farrell–Jones conjecture hold for all HHGs?

Acknowledgements. I would like to thank Jason Behrstock, Mark Hagen, Giorgio Mangioni, Harry Petyt, and Abdul Zalloum for their very useful comments and suggestions.

References

[1] C. Abbott, J. Behrstock, and M. G. Durham, Largest acylindrical actions and stability in hierarchically hyperbolic groups. *Trans. Amer. Math. Soc. Ser. B* **8** (2021), no. 3, 66–104 Zbl 1498.20109 MR 4215647

[2] C. Abbott, D. Berlyne, G. Mangioni, T. Ng, and A. Rasmussen, Random quotients preserve acylindrical and hierarchical hyperbolicity. [v1] 2025, [v3] 2026, arXiv: 2507.16677v3

[3] C. R. Abbott, T. Ng, D. Spriano, R. Gupta, and H. Petyt, Hierarchically hyperbolic groups and uniform exponential growth. *Math. Z.* **306** (2024), no. 1, article no. 18, 33 pp. Zbl 1536.20052 MR 4680503

[4] S. Balasubramanya and T. Fernos, The semi-simple theory of higher rank acylindricity. [v1] 2024, [v3] 2025, arXiv:2407.04838v3

[5] O. Barak, On equational noetherianity of colorable hierarchically hyperbolic groups. 2024, arXiv:2410.00977v1

[6] A. Bartels and W. Lück, The Borel conjecture for hyperbolic and CAT(0)-groups. *Ann. of Math. (2)* **175** (2012), no. 2, 631–689 Zbl 1256.57021 MR 2993750

[7] J. Behrstock, M. Hagen, A. Martin, and A. Sisto, A combinatorial take on hierarchical hyperbolicity and applications to quotients of mapping class groups. *J. Topol.* **17** (2024), no. 3, article no. e12351, 94 pp. Zbl 1552.20181 MR 4822919

[8] J. Behrstock, M. Hagen, and A. Sisto, Hierarchically hyperbolic spaces II: combination theorems and the distance formula. *Pacific J. Math.* **299** (2019), no. 2, 257–338 Zbl 1515.20208 MR 3956144

[9] J. Behrstock, M. F. Hagen, and A. Sisto, Asymptotic dimension and small-cancellation for hierarchically hyperbolic spaces and groups. *Proc. Lond. Math. Soc. (3)* **114** (2017), no. 5, 890–926 Zbl 1431.20028 MR 3653249

[10] J. Behrstock, M. F. Hagen, and A. Sisto, Hierarchically hyperbolic spaces, I: curve complexes for cubical groups. *Geom. Topol.* **21** (2017), no. 3, 1731–1804 Zbl 1439.20043 MR 3650081

[11] J. Behrstock, M. F. Hagen, and A. Sisto, Quasiflats in hierarchically hyperbolic spaces. *Duke Math. J.* **170** (2021), no. 5, 909–996 Zbl 1542.20174 MR 4255047

[12] F. Berlai and B. Robbio, A refined combination theorem for hierarchically hyperbolic groups. *Groups Geom. Dyn.* **14** (2020), no. 4, 1127–1203 Zbl 1511.20170 MR 4186470

[13] D. Berlyne and J. Russell, Hierarchical hyperbolicity of graph products. *Groups Geom. Dyn.* **16** (2022), no. 2, 523–580 Zbl 1521.20085 MR 4502615

[14] E. Bongiovanni, Extensions of finitely generated Veech groups. *Algebr. Geom. Topol.* **26** (2026), no. 3, 989–1035 Zbl 08186079 MR 5056662

[15] B. H. Bowditch, Quasiflats in coarse median spaces, 2019, Preprint, https://bhbowditch.com/papers/quasiflats.pdf visited on 4 April 2026

[16] M. R. Bridson, Semisimple actions of mapping class groups on CAT(0) spaces. In F. P. Gardiner, G. González-Diez, and C. Kourouniotis (Eds.), *Geometry of Riemann surfaces*, pp. 1–14, London Math. Soc. Lecture Note Ser. 368, Cambridge University Press, Cambridge, 2010 Zbl 1204.30030 MR 2665003

[17] M. Casals-Ruiz, M. Hagen, and I. Kazachkov, Real cubings and asymptotic cones of hierarchically hyperbolic groups. 2024, Preprint, https://www.wescac.net/cones_july_2024-public.pdf visited on 4 April 2026

[18] M. Chesser, Stable subgroups of the genus 2 handlebody group. *Algebr. Geom. Topol.* **22** (2022), no. 2, 919–971 Zbl 1554.20088 MR 4464468

[19] F. Dahmani and V. Guirardel, The isomorphism problem for all hyperbolic groups. *Geom. Funct. Anal.* **21** (2011), no. 2, 223–300 Zbl 1258.20034 MR 2795509

[20] S. Dowdall, M. G. Durham, C. J. Leininger, and A. Sisto, Extensions of Veech groups II: hierarchical hyperbolicity and quasi-isometric rigidity. *Comment. Math. Helv.* **99** (2024), no. 1, 149–228 Zbl 1556.20069 MR 4709309

[21] M. G. Durham, Cubulating infinity in hierarchically hyperbolic spaces. 2023, arXiv: 2308.13689v1

[22] M. G. Durham, M. F. Hagen, and A. Sisto, Boundaries and automorphisms of hierarchically hyperbolic spaces. *Geom. Topol.* **21** (2017), no. 6, 3659–3758 Zbl 1439.20044 MR 3693574

[23] M. G. Durham, M. F. Hagen, and A. Sisto, Correction to the article "Boundaries and automorphisms of hierarchically hyperbolic spaces". *Geom. Topol.* **24** (2020), no. 2, 1051–1073 Zbl 1443.20071 MR 4153656

[24] M. G. Durham, Y. Minsky, and A. Sisto, Asymptotically CAT(0) metrics, Z-structures, and the Farrell–Jones Conjecture. 2025, arXiv:2504.17048v1

[25] M. G. Durham, Y. N. Minsky, and A. Sisto, Stable cubulations, bicombings, and barycenters. *Geom. Topol.* **27** (2023), no. 6, 2383–2478 Zbl 1535.20208 MR 4634751

[26] M. G. Durham and A. Zalloum, The geometry of genericity in mapping class groups and Teichmüller spaces via CAT(0) cube complexes. *Trans. Amer. Math. Soc.* **379** (2026), no. 6, 3763–3832 Zbl 08198613 MR 5067476

[27] A. Goldsborough, M. F. Hagen, H. Petyt, and A. Sisto, Induced quasi-isometries of hyperbolic spaces, Markov chains, and acylindrical hyperbolicity (with an appendix by J. Russell). *Groups Geom. Dyn.* (2025), DOI 10.4171/GGD/873

[28] D. Groves and J. F. Manning, Dehn filling in relatively hyperbolic groups. *Israel J. Math.* **168** (2008), no. 1, 317–429 Zbl 1211.20038 MR 2448064

[29] T. Haettel, Group actions on injective spaces and Helly graphs. 2023, arXiv:2307.00414v2

[30] T. Haettel, N. Hoda, and H. Petyt, Coarse injectivity, hierarchical hyperbolicity and semihyperbolicity. *Geom. Topol.* **27** (2023), no. 4, 1587–1633 Zbl 1516.20098 MR 4602421

[31] M. Hagen, Non-colorable hierarchically hyperbolic groups. *Internat. J. Algebra Comput.* **33** (2023), no. 2, 337–350 Zbl 1512.20150 MR 4581212

[32] M. Hagen, G. Mangioni, and A. Sisto, A combinatorial structure for many hierarchically hyperbolic spaces. *Pacific J. Math.* **341** (2026), no. 2, 305–377 Zbl 08176986 MR 5048996

[33] M. Hagen, A. Martin, and A. Sisto, Extra-large type Artin groups are hierarchically hyperbolic. *Math. Ann.* **388** (2024), no. 1, 867–938 Zbl 1536.20049 MR 4693949

[34] M. Hagen, J. Russell, A. Sisto, and D. Spriano, Equivariant hierarchically hyperbolic structures for 3-manifold groups via quasimorphisms. *Ann. Inst. Fourier (Grenoble)* **75** (2025), no. 2, 769–828 Zbl 08053903 MR 4921365

[35] M. F. Hagen and T. Susse, On hierarchical hyperbolicity of cubical groups. *Israel J. Math.* **236** (2020), no. 1, 45–89 Zbl 1453.20056 MR 4093881

[36] F. Haglund and D. T. Wise, Special cube complexes. *Geom. Funct. Anal.* **17** (2008), no. 5, 1551–1620 Zbl 1155.53025 MR 2377497

[37] S. Hughes and M. Valiunas, Commensurating HNN-extensions: hierarchical hyperbolicity and biautomaticity. *Comment. Math. Helv.* **99** (2024), no. 2, 397–436 Zbl 1540.20078 MR 4730350

[38] A. Kar, *Discrete groups and CAT(0) asymptotic cones*. Ph.D. thesis, The Ohio State University, 2008 MR 2712674

[39] U. Lang, Injective hulls of certain discrete metric spaces and groups. *J. Topol. Anal.* **5** (2013), no. 3, 297–331 Zbl 1292.20046 MR 3096307

[40] G. Mangioni, A combination theorem for hierarchically quasiconvex subgroups, and application to geometric subgroups of mapping class groups. [v1] 2024, [v2] 2026, arXiv:2409.03602v2, to appear in *Algebr. Geom. Topol.*

[41] G. Mangioni and A. Sisto, Rigidity of mapping class groups mod powers of twists. *Proc. Roy. Soc. Edinburgh Sect. A* (2025), 71 pp., DOI 10.1017/prm.2025.23

[42] G. Mangioni and A. Sisto, Short hierarchically hyperbolic groups II: quotients and the Hopf property for Artin groups. *Adv. Math.* **486** (2026), article no. 110736, 71 pp. Zbl 08152676 MR 5002582

[43] H. A. Masur and Y. N. Minsky, Geometry of the complex of curves I: hyperbolicity. *Invent. Math.* **138** (1999), no. 1, 103–149 Zbl 0941.32012 MR 1714338

[44] H. A. Masur and Y. N. Minsky, Geometry of the complex of curves II: hierarchical structure. *Geom. Funct. Anal.* **10** (2000), no. 4, 902–974 Zbl 0972.32011 MR 1791145

[45] D. V. Osin, Peripheral fillings of relatively hyperbolic groups. *Invent. Math.* **167** (2007), no. 2, 295–326 Zbl 1116.20031 MR 2270456

[46] J. Park, Factor system for graphs and combinatorial HHS. 2024, arXiv:2409.08663v1

[47] H. Petyt, Mapping class groups are quasicubical. [v1] 2021, [v3] 2024, arXiv: 2112.10681v3, to appear in *Amer. J. Math.*

[48] H. Petyt, D. Spriano, and A. Zalloum, Hyperbolic models for CAT(0) spaces. *Adv. Math.* **450** (2024), article no. 109742, 66 pp. Zbl 1542.20193 MR 4753310

[49] H. Petyt, A. Zalloum, and D. Spriano, Constructing metric spaces from systems of walls. 2024, arXiv:2404.12057v1

[50] J. Russell, D. Spriano, and H. C. Tran, Convexity in hierarchically hyperbolic spaces. *Algebr. Geom. Topol.* **23** (2023), no. 3, 1167–1248 Zbl 1517.20069 MR 4598806

[51] Z. Sela, The isomorphism problem for hyperbolic groups I. *Ann. of Math. (2)* **141** (1995), no. 2, 217–283 Zbl 0868.57005 MR 1324134

[52] Z. Sela, Automorphisms of groups and a higher rank JSJ decomposition I: RAAGs and a higher rank Makanin–Razborov diagram. *Geom. Funct. Anal.* **33** (2023), no. 3, 824–874 Zbl 1516.20095 MR 4597643

[53] Z. Sela, Automorphisms of groups and a higher rank JSJ decomposition II: the single ended case. [v1] 2022, [v2] 2023, arXiv:2209.12891v2

[54] A. Sisto, What is a hierarchically hyperbolic space? In M. Hagen, R. Webb, and H. Wilton (Eds.), *Beyond hyperbolicity*, pp. 117–148, London Math. Soc. Lecture Note Ser. 454, Cambridge University Press, Cambridge, 2019 Zbl 1514.20174 MR 3966608

[55] K. M. Vokes, Hierarchical hyperbolicity of graphs of multicurves. *Algebr. Geom. Topol.* **22** (2022), no. 1, 113–151 Zbl 1551.20085 MR 4413817

[56] H. Wilton and A. Sisto, The congruence subgroup property for mapping class groups and the residual finiteness of hyperbolic groups. 2024, arXiv:2410.00556v2

[57] A. Zalloum, Injectivity, cubical approximations and equivariant wall structures beyond CAT(0) cube complexes. 2023, arXiv:2305.02951v1

Alessandro Sisto
Maxwell Institute for Mathematical Sciences and Department of Mathematics, Heriot-Watt University, First Gait, Edinburgh EH14 4AS, UK; a.sisto@hw.ac.uk

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/34

Noncommutative singularity theory

Gavin Brown and Michael Wemyss

Abstract. This is an expository article on the noncommutative singularity theory of power series in noncommuting variables, its motivation from deformation theory, and its applications to contractibility of curves and the classification of smooth 3-fold flops.

1. Introduction

It is a fact of life that there are just too many functions, and to gain structural insight, some restriction is necessary. Taking that as given, what comes next is harder, since it involves making choices.

To first approximation, making the choice of a class of functions amounts to making the choice of a research area: differential geometry is often considered to be the study of C^∞ functions, algebraic geometry thought to be the exclusive realm of polynomial functions, and symplectic topology to be, well, the land where symplectomorphisms roam free.

Whilst this viewpoint provides helpful motivation, it is simply flat-out false from any reasonable technical perspective. There are many situations where, within an area 'defined' by a class of functions, proving certain theorems requires an enlargement of or extension to that class of functions. This may be frustrating. For example, if you begin with polynomials that look as nice as $u^2 + v^2 + x^2 + y^2$, it does initially take some convincing that formal power series really are objects that humans should consider and that analytic changes of coordinates really do make your life easier. If a problem can be stated entirely within one class of functions, there is understandably a reluctance to extend that class without a clear and compelling reason to do so.

The purpose of this article is to introduce noncommutative singularity theory as an object in its own right and to then explain why this extension from commutative to noncommutative polynomials (and power series) is needed to solve various contraction and classification problems within algebraic geometry.

Mathematics Subject Classification 2020: 14B05 (primary); 14B20, 14E30 (secondary).
Keywords: singularity theory, deformation theory, contraction algebras, birational geometry.

With this in mind, and attempting to accomplish multiple goals, the structure of this article is as follows. Section 2 outlines what noncommutative singularity theory is and states some of its initial results. This is all factual, but it does not address the why: that is partially contained in Section 3, which outlines the original intuition from deformation theory, but in fact is mostly explained in Sections 4 and 5, which cover in detail the applications to birational geometry. Throughout Sections 4 and 5, there is a particular focus on what the new noncommutative technology buys us, in the form of various theorems and geometric consequences. Two conjectures are briefly outlined in Section 5.4.

2. Noncommutative singularity theory

Polynomial functions are elements in the polynomial ring $\mathbb{C}[x_1, \dots, x_d]$ for some $d \geq 1$. To set language, noncommutative (polynomial) functions will simply be elements of the free algebra $\mathbb{C}\langle x_1, \dots, x_d \rangle$.

2.1. Classical singularity theory

To classify even smooth points in algebraic geometry, one must work suitably locally. For this article, this means working with elements of the commutative power series ring $\mathbb{C}[\![x_1, \dots, x_d]\!]$, although alternatives are available. The key point is that elements in the commutative power series ring are allowed to consist of infinite sums, and as such they can be represented in the form

$$f = \lambda_1 + \lambda_2 x + \lambda_3 y + \lambda_4 xy + \cdots .$$

The ring $\mathbb{C}[\![x_1, \dots, x_d]\!]$ is sometimes alternatively referred to as the formal power series ring, with the word 'formal' acting as a reminder that its elements do not need to satisfy any radius of convergence conditions.

At first glance, it is not so clear that moving from polynomials (finite sums) to power series (infinite sums) makes life easier, given that there seem to be many more elements. But it does, since there are now *many* more symmetries, in the form of analytic changes of coordinates. Given this huge symmetry group, it is then possible to ask about classifying functions. This is the domain of singularity theory.

Aim 2.1 (Singularity theory à la Arnold [1]). Up to specified isomorphism, classify all $f \in \mathbb{C}[\![x_1, \dots, x_d]\!]$ that satisfy some fixed numerical criteria and produce theory for when classification is not possible.

The following is perhaps the most fundamental class of examples. Although the condition below is a priori unmotivated, it is numerical, as it counts ideals satisfying a certain property.

Definition 2.2. An element $f \in \mathbb{C}[\![x_1, \ldots, x_d]\!] = \mathbb{C}[\![\mathrm{x}]\!]$ is said to be a simple singularity if

$$\#\{I \mid I \text{ proper ideal of } \mathbb{C}[\![\mathrm{x}]\!] \text{ with } f \in I^2\} < \infty.$$

This is a non-standard definition of a simple singularity. Like many numerical definitions within singularity theory, it is only really justified *after* classification, when it becomes more clear that it is equivalent to every other characterisation of simple singularities in the literature.

Theorem 2.3 (For example, [27, Section 8]). *If f is a simple singularity in the sense of Definition 2.2, then up to relabelling variables $z_1, \ldots, z_{d-2}, x, y$, and up to isomorphism (analytic or formal changes of coordinates in the domain), f is one of*

$$\begin{aligned}
&A_n \quad z^2 + x^2 + y^{n+1}, \quad && n \geq 1,\\
&D_n \quad z^2 + x^2 y + y^{n-1}, \quad && n \geq 4,\\
&E_6 \quad z^2 + x^3 + y^4,\\
&E_7 \quad z^2 + x^3 + xy^3,\\
&E_8 \quad z^2 + x^3 + y^5,
\end{aligned}$$

where $z^2 = z_1^2 + \cdots + z_{d-2}^2$.

From the viewpoint of later sections, several remarks are in order. First, in Definition 2.2, there is no mention of ADE, yet this still appears in the classification in Theorem 2.3. Second, Theorem 2.3 illustrates the very small number of simple singularities, namely just two (countable) infinite families and only three other cases. And third, with Theorem 2.3 in hand, working with simple singularities becomes much easier, since various problems may be reduced to checking statements against a short list of elegant forms.

2.2. Noncommutative singularity theory

To first approximation, the overall idea of noncommutative singularity theory is to replace the commutative object $\mathbb{C}[\![x_1, \ldots, x_d]\!]$ with the formal *noncommutative* power series ring $\mathbb{C}\langle\!\langle x_1, \ldots, x_d \rangle\!\rangle$ and then to play the same game as in Section 2.1. This is not quite true since in the noncommutative setting, there are many more 'free objects' given by quivers, and as in Remark 2.8 below, this makes the theory richer and widens applications.

In the first case, the ring $\mathbb{C}\langle\!\langle x_1, \ldots, x_d \rangle\!\rangle$ is similar to the free algebra in d variables, except now elements are possibly infinite sums

$$f = \lambda_1 + \lambda_2 x + \lambda_3 y + \lambda_4 xy + \lambda_5 yx + \cdots.$$

Whilst for $d \geq 2$, the ring $\mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ is not noetherian, has exponential growth, and is pathologically horrible from many perspectives, as in the commutative case, its automorphism group is huge, and so classifying elements satisfying certain numerical criteria still turns out to be possible. But, as in the classical case, doing this still requires choices.

Problem 2.4. When should $f, g \in \mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ be viewed as being the same? And what would be natural numerical criteria?

To explain first the notion of equivalence, recall that every $f \in \mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ can be differentiated with respect to any variable. The rule is simple: cycle, then score off the leftmost variable. For example, in humane coordinates x and y, to differentiate x^3y with respect to x,

$$x^3y = xxxy \xrightarrow{\text{cycle}} \begin{matrix} xxxy \\ xxyx \\ xyxx \\ yxxx \end{matrix} \xrightarrow{\text{score off}} \begin{matrix} xxy \\ xyx \\ yxx \\ 0 \end{matrix},$$

where 'score off' either scores off the leftmost x or returns zero if the monomial does not begin with x. By definition, it follows that $\delta_x(x^3y) = x^2y + xyx + yx^2$. Reassuringly, $\delta_y(y^3) = 3y^2$.

Definition 2.5. Given any f, the Jacobi algebra is

$$\mathcal{J}\mathrm{ac}\, f := \frac{\mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle}{(\!(\delta_z f \mid z \text{ is a variable})\!)},$$

where the double brackets $(\!(\ )\!)$ denote the closure of the two-sided ideal.

Example 2.6. The element $f = x^4 + xy^2 \in \mathbb{C}\langle\langle x, y \rangle\rangle$ has

$$\mathcal{J}\mathrm{ac}\, f = \frac{\mathbb{C}\langle\langle x, y \rangle\rangle}{(\!(4x^3 + y^2, xy + yx)\!)},$$

which as explained in Example 2.13 below is a nine-dimensional algebra.

Write $f \cong g$ provided that $\mathcal{J}\mathrm{ac}\, f \cong \mathcal{J}\mathrm{ac}\, g$.

Aim 2.7. With $\mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ fixed, and the notion of equivalence $\cong$ now defined, the aim of singularity theory remains: to classify all elements satisfying some numerical (or structural) criteria and produce theory for when classification is not possible.

Remark 2.8. An alternative description of the noncommutative power series ring $\mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ is the complete path algebra of the d-loop quiver. Other quivers are possible, and all of the constructions in this section and below naturally and easily generalise to that setting [11]. These other quivers are needed for multi-curve geometric applications; see Sections 3.4 and 4.1 and Remark 5.3 later.

2.3. The singularity theory toolkit

This subsection briefly outlines some of the singularity theory toolkit, adapted into the noncommutative setting.

The first and most fundamental problem is to establish the correct notion of dimension and the related numerical invariants. Using the Gelfand–Kirillov (GK) dimension, which is standard for noncommutative rings, is problematic since $\operatorname{GKdim} \mathbb{C}[\![x]\!] = \infty$. Fundamentally, GK dimension is not well adapted to complete rings. To overcome this, we return to the basic fact that $\mathcal{J}\mathrm{ac}\, f$ is local, with Jacobson radical, say, $\mathfrak{J}$. This gives an intrinsic filtration

$$\mathcal{J}\mathrm{ac}\, f \supseteq \mathfrak{J} \supseteq \mathfrak{J}^2 \supseteq \cdots.$$

Definition 2.9. The (Jacobson radical) $\mathfrak{J}$-dimension, $\operatorname{Jdim} \mathcal{J}\mathrm{ac}\, f$, is the growth rate of the steps in the chain above, namely

$$\inf\{r \in \mathbb{R} \mid \text{for some } c \in \mathbb{R},\ \dim \mathcal{J}\mathrm{ac}\, f/\mathfrak{J}^n \leq cn^r \text{ for every } n \in \mathbb{N}\}. \tag{2.1}$$

To calibrate, note that $\operatorname{Jdim} \mathbb{C}[\![x]\!] = 1$ and that $\operatorname{Jdim} \mathcal{J}\mathrm{ac}\, f = 0$ is equivalent to $\mathcal{J}\mathrm{ac}\, f$ being a finite-dimensional algebra.

Remark 2.10. There are several well-known results about GK dimension: for example, that it can take any real value ≥ 2 (or $0, 1$ or ∞) but no values in the 'Bergman gap' between 1 and 2. We do not know analogous results for $\mathfrak{J}$-dimension, even for the restricted class of Jacobi algebras. In the context of this survey, however, it is clear that if the $\mathfrak{J}$-dimension is ≤ 1, then it is either 0 or 1, just as for GK dimension.

Whilst the growth rate of the whole chain (2.1) gives the dimension, the numerical properties of its individual pieces also hold valuable information. In what follows, we will always assume that f contains no constant or linear terms, which we will abbreviate as $f \in \mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle_{\geq 2}$.

Definition 2.11. For $f \in \mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle_{\geq 2}$, and $m \geq 1$, the *mth corank of* f is defined to be

$$\mathcal{C}\mathrm{rk}_m(f) = \dim_{\mathbb{C}} \left(\frac{\mathfrak{J}^m}{\mathfrak{J}^{m+1}} \right),$$

where $\mathfrak{J}$ is the Jacobson radical of $\mathcal{J}\mathrm{ac}\, f$.

The first of these invariants $\mathcal{C}\mathrm{rk}(f) := \mathcal{C}\mathrm{rk}_1(f)$ is often referred to as simply the corank. The corank is determined by the linear conditions imposed by derivatives and is therefore uniquely determined by the degree two part of f.

Theorem 2.12 (Splitting lemma, for example, [11, Theorem 4.5]). *Suppose* $f \in \mathbb{C}\langle\langle x_1, \dots, x_d\rangle\rangle_{\geq 2}$. *Then* $f \cong x_1^2 + \cdots + x_r^2 + g$ *for some* $g \in \mathbb{C}\langle\langle x_{r+1}, \dots, x_d\rangle\rangle_{\geq 3}$, *where* $d - r = \mathcal{C}\mathrm{rk}(f)$. *In particular,*

$$\mathcal{J}\mathrm{ac}\, f \cong \frac{\mathbb{C}\langle\langle x_{r+1}, \dots, x_d\rangle\rangle}{(\!(\delta_{x_{r+1}} g, \dots, \delta_{x_d} g)\!)}.$$

The coranks are properties of the leading terms of elements of the closed Jacobian ideal of derivatives and can be calculated using a standard basis. In the commutative context, the standard basis (also known as a local Gröbner basis) is well known. Loosely speaking, one follows the usual Buchberger approach of cancelling leading terms, but regarding the leading terms as those of low degree. The theory and key results all hold in the setting of noncommutative power series [17]. From a user perspective, there are two points to note: noncommutativity means leading terms may have nontrivial cancellation with themselves, and, being non-noetherian, the resulting basis may be infinite (although not of course in nice settings such as $\mathrm{Jdim}\, \mathcal{J}\mathrm{ac}\, f = 0$).

Example 2.13. Continuing Example 2.6, consider $f = x^4 + xy^2$. The leading term of a non-zero power series is defined to be the term of least degree, with lexicographic order used to break ties. So, for example, the leading term of $xy^2 + yxy + y^3$ is xy^2.

Setting $A = y^2 + 4x^3$ and $B = xy + yx$ for the two derivatives, the leading term y^2 of A cancels itself in the expression $yA - Ay$ to give

$$\begin{aligned} yA - Ay &= 4yx^3 - 4x^3y \overset{B}{\equiv} 4yx^3 + 4x^2yx \overset{B}{\equiv} 4yx^3 - 4xyx^2 \\ &\overset{B}{\equiv} 4yx^3 + 4yx^3 = 8yx^3, \end{aligned}$$

where $\overset{B}{\equiv}$ denotes congruence modulo B. Thus, $C = yx^3 \in (A, B)$. Similarly, $Ax^3 - yC = 4x^6 \in (A, B)$. All other attempts to cancel leading terms reduce to zero, and so the (normalised) standard basis is

$$y^2 + 4x^3,\ xy + yx,\ yx^3,\ x^6.$$

The leading terms of these are y^2, xy, yx^3, x^6, and the set of monomials not divisible by any of these provides a basis of $\mathcal{J}\mathrm{ac}\, f$. In this case, we see at once therefore that

$$1,\ x,\ y,\ x^2,\ yx, x^3,\ yx^2,\ x^4,\ x^5$$

is a basis of $\mathcal{J}\mathrm{ac}\, f$, and so it has dimension 9, as claimed. Moreover, the degree m monomials in this monomial basis descend to a basis of $\mathfrak{J}^m/\mathfrak{J}^{m+1}$, which implies that

$$\mathcal{C}\mathrm{rk}_1(f) = \mathcal{C}\mathrm{rk}_2(f) = \mathcal{C}\mathrm{rk}_3(f) = 2 \quad \text{and} \quad \mathcal{C}\mathrm{rk}_4(f) = \mathcal{C}\mathrm{rk}_5(f) = 1.$$

Together with $\dim_{\mathbb{C}} \mathbb{C}\langle\!\langle x, y\rangle\!\rangle/\mathfrak{J} = 1$, these numbers add up to $9 = \dim_{\mathbb{C}} \mathcal{J}\mathrm{ac}\, f$.

Another key property of Jacobi algebras is that they are invariant under automorphisms, and so in particular under arbitrary analytic changes of coordinates.

Theorem 2.14 ([11, Proposition 3.7]). *If* $F \in \mathrm{Aut}\, \mathbb{C}\langle\!\langle x_1, \ldots, x_d\rangle\!\rangle$, *then* $\mathcal{J}\mathrm{ac}\, f \cong \mathcal{J}\mathrm{ac}\, F(f)$.

Other natural parts of the commutative theory, such as Saito's famous result [24] on quasi-homogeneous power series, also work in the noncommutative context, although not quite yet at the elementary entry level of this discussion; see [19].

Remark 2.15. Whilst any type of singularity theory is really 'just' the classification of Jacobi (or commutatively, Tjurina) algebras up to isomorphism, thinking in terms of functions allows us to access the toolkit of singularity theory. Even when the Jacobi algebras are finite dimensional, perhaps the main advantage of thinking in terms of functions is Theorem 2.14, as this enables the construction of many isomorphisms which seem completely out of reach of more classical approaches to the classification of noncommutative finite-dimensional algebras.

2.4. Proposal

The following proposal 2.16, which we view as the noncommutative version of simple singularities, is our main reason for developing noncommutative singularity theory.

Proposal 2.16. *Under the numerical criterion* $\mathrm{Jdim}\, \mathcal{J}\mathrm{ac}\, f \leq 1$, *we propose that*

(1) *a classification is in fact possible,*

(2) *the classification is ADE,*

(3) *this algebraic classification* is *(and implies) the classification of flops and of crepant divisorial contractions to curves.*

The proposal is still conjectural in its form stated above, as this is strongly related to Conjecture 5.6 below. The current state of the art, partially explained in Section 5, in part replaces the numerical condition $\mathrm{Jdim}\, \mathcal{J}\mathrm{ac}\, f \leq 1$ with a structural one and solves that: for all geometric applications including the classification of 3-fold flops,

this is enough. However, from the viewpoint of future ease of use (numerical properties are easier to verify than structural ones), and also to make the analogy with simple singularities in Theorem 2.3 more compelling, verifying Proposal 2.16 in the generality stated remains the founding question of noncommutative singularity theory.

3. Deformation theory

This section describes the links, via deformation theory, between noncommutative singularity theory on the one hand and both classification and contractibility problems in birational geometry on the other. Applications are given later, in Sections 4 and 5.

3.1. Deformation theory framework

One reason for deforming an object, say M, is the hope that doing so will attach an intrinsic invariant to M, and furthermore that this invariant will be useful. In good cases, it is possible to extract simple data from the invariant, for example, a number such as its dimension (if that even makes sense). Intrinsically attaching numbers to objects is hard, and the framework of deformation theory is generally a good way of doing this.

Alas, deformation theory involves choices. To ensure that the theory is intrinsic, it is natural to insist on the functorial approach. That is, given some object M, the first aim is to construct a functor

$$\mathcal{D}\mathrm{ef}_M : \mathrm{TestObjects} \to \mathrm{Sets}$$

which sends a test object R to a set $\{\text{deformations of } M \text{ over } R\}/\sim$. To make this rigorous, the following need to be defined, or rather declared:

(1) The category TestObjects.

(2) The definition of 'deformations of M over R'.

(3) The equivalence relation $\sim$.

There is a fourth choice, suppressed throughout this article, since we will always assume that the target category is the category of sets.

3.2. Commutative deformation theory

Let X be a variety and $\mathcal{F} \in \mathrm{Qcoh}\, X$ be a quasi-coherent sheaf. To deform $\mathcal{F}$ via the classical Grothendieck formulation, consider the functor

$$\mathrm{c}\mathcal{D}\mathrm{ef}\colon \mathfrak{cart} \to \mathrm{Sets},$$

where $\mathfrak{cart}$ is the category of commutative local artinian $\mathbb{C}$-algebras, sending

$$(R,\mathfrak{m}) \mapsto \left\{ (\mathcal{G},\eta) \;\middle|\; \begin{array}{l} \mathcal{G} \in \operatorname{Qcoh}(X \times \operatorname{Spec} R), \\ \mathcal{G} \text{ is flat over } \operatorname{Spec} R, \\ \eta\colon \mathcal{G}|_{X\times 0} \xrightarrow{\sim} \mathcal{F} \end{array} \right\} \Big/ \sim,$$

where $\sim$ is the natural equivalence relation outlined in a greater generality below. The point is that, roughly speaking, a deformation of $\mathcal{F}$ is a sheaf $\mathcal{G}$ which is infinitesimally thicker than $\mathcal{F}$ (namely, a sheaf on $X \times \operatorname{Spec} R$), roughly of the same size (flatness), which recovers the original $\mathcal{F}$ (restricted to the zero fibre). There are many choices of such isomorphisms on the zero fibre, so we carry those choices as part of the data.

3.3. Noncommutative deformation theory

It is very intuitive that more test objects will lead to more information. This subsection enlarges the category of test objects $\mathfrak{cart}$ from Section 3.2 by adding in (some) noncommutative artinian rings. The resulting deformation functor will be similar but requires some reinterpretation.

For simplicity, consider the same setup as in Section 3.2, namely X is a variety, and $\mathcal{F} \in \operatorname{Qcoh} X$. The noncommutative deformation functor is defined to be

$$\mathcal{D}\mathrm{ef}\colon \mathfrak{art}_1 \to \mathrm{Sets}$$

from finite-dimensional augmented $\mathbb{C}$-algebras to sets, sending

$$(\Gamma,\mathfrak{n}) \mapsto \left\{ (\mathcal{G},\vartheta,\eta) \;\middle|\; \begin{array}{l} \mathcal{G} \in \operatorname{Qcoh} X, \\ \vartheta\colon \Gamma \to \operatorname{End}_X(\mathcal{G}) \text{ homomorphism}, \\ - \otimes_\Gamma \mathcal{G}\colon \operatorname{mod}\Gamma \to \operatorname{Qcoh} X \text{ is exact}, \\ \eta\colon (\Gamma/\mathfrak{n}) \otimes_\Gamma \mathcal{G} \xrightarrow{\sim} \mathcal{F} \end{array} \right\} \Big/ \sim,$$

where, for any given Γ, $(\mathcal{G},\vartheta,\eta) \sim (\mathcal{G}',\vartheta',\eta')$ if and only if there exists $\tau\colon \mathcal{G} \xrightarrow{\sim} \mathcal{G}'$ such that for all $r \in \Gamma$,

$$\begin{array}{ccc} \mathcal{G} & \xrightarrow{\tau} & \mathcal{G}' \\ {\scriptstyle \vartheta(r)}\downarrow & & \downarrow{\scriptstyle \vartheta'(r)} \\ \mathcal{G} & \xrightarrow{\tau} & \mathcal{G}' \end{array} \qquad \text{and} \qquad \begin{array}{ccc} (\Gamma/\mathfrak{n}) \otimes_\Gamma \mathcal{G} & \xrightarrow{1\otimes\tau} & (\Gamma/\mathfrak{n}) \otimes_\Gamma \mathcal{G}' \\ & {\scriptstyle \eta}\searrow \quad \swarrow{\scriptstyle \eta'} & \\ & \mathcal{F} & \end{array}$$

commute. There are many equivalent characterisations of the category $\mathfrak{art}_1$, where we refer the reader to [16] for full details.

The commutative deformation functor is recovered via

$$\mathrm{c}\mathcal{D}\mathrm{ef} = \mathcal{D}\mathrm{ef}|_{\mathfrak{cart}}.$$

3.4. Deforming multiple objects

Perhaps the clearest way of observing the benefit of noncommutative deformation theory is when we try to deform more than one object, say $\mathcal{F}_1$ and $\mathcal{F}_2$. Classically, the commutative solution is to take the direct sum $\mathcal{F}_1 \oplus \mathcal{F}_2$ and apply the technology of Section 3.2. The commutative hull is then constructed as a factor of $\mathbb{C}[\![x_1, \dots, x_t]\!]$, where $t = \dim_{\mathbb{C}} \operatorname{Ext}^1_X(\mathcal{F}_1 \oplus \mathcal{F}_2, \mathcal{F}_1 \oplus \mathcal{F}_2)$.

But this clearly loses information, since Ext groups are directed: indeed, in general

$$\operatorname{Ext}^1_X(\mathcal{F}_1, \mathcal{F}_2) \ncong \operatorname{Ext}^1_X(\mathcal{F}_2, \mathcal{F}_1).$$

Since the order matters in the Ext groups, forcing the commutativity in $\mathbb{C}[\![x_1, \dots, x_t]\!]$ means that often almost all information is lost.

The solution to deforming n objects is to realise that, since Exts are directed, the prorepresenting hull should naturally be constructed as a factor of a quiver with n vertices, where the number of arrows from vertex i to vertex j should be the integer $\dim_{\mathbb{C}} \operatorname{Ext}^1_X(\mathcal{F}_i, \mathcal{F}_j)$. This naturally records the asymmetry in the Ext groups and also gives a more natural and compelling description of what the tangent space should be.

To do this on a technical level requires the introduction of the category of test objects $\mathfrak{art}_n$, which are, roughly speaking, finite-dimensional $\mathbb{C}$-algebras where all simple modules are one dimensional; this last condition rules out the appearance of matrix algebras. For many equivalent characterisations of the category $\mathfrak{art}_n$, we refer the reader to [16].

To deform the collection $\mathcal{F}_1, \dots, \mathcal{F}_n$, each $\mathcal{F}_i$ is replaced by an injective resolution $\mathcal{F}_i \hookrightarrow \mathcal{I}_i$, and from that the endomorphism DG-algebra $\mathcal{E}\mathrm{nd}_X(\oplus \mathcal{I}_i)$ is constructed. The Maurer–Cartan formulation then gives a functor

$$\mathcal{D}\mathrm{ef}_{\mathcal{F}_1,\dots,\mathcal{F}_n} \colon \mathfrak{art}_n \to \mathrm{Sets} \tag{3.1}$$

called the multi-pointed noncommutative deformation functor. Full details can be found in, for example, [13, 25]. As always, there are choices; it is also possible to consider

$$\mathcal{D}\mathrm{ef}_{\mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_n} \colon \mathfrak{art}_1 \to \mathrm{Sets}$$

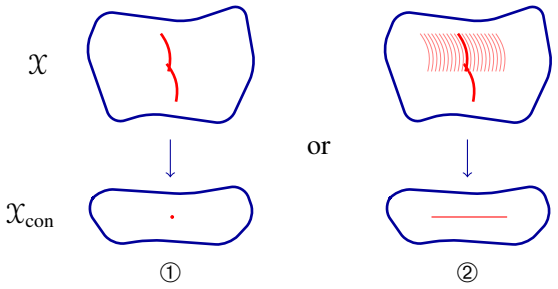
from Section 3.3. We describe in Section 4 below the behaviour of these functors in the situation of curve contractions and outline the reasons why, from almost all geometric viewpoints, (3.1) is the key.

4. The setting of curves

Our main interest is in formal or analytic neighbourhoods of rational curves on 3-folds.

4.1. Contractions

Pictorially, we will consider contractions of 3-folds, of two types:



(4.1)

where the contraction map f is projective birational, satisfying $\mathrm{R}f_*\mathcal{O}_{\mathcal{X}} = \mathcal{O}_{\mathcal{X}_{\text{con}}}$. In the left-hand case ①, the morphism is an isomorphism away from a point, whereas in the right-hand case ②, the morphism contracts a divisor to a curve and is an isomorphism away from the one-dimensional locus on the base. Although the above pictures are drawn in the algebraic category, we will often implicitly pick a closed point on the base within the non-isomorphism locus, complete at that point, and thus assume that the base $\mathcal{X}_{\text{con}} = \operatorname{Spec} \mathcal{R}$ is complete local.

Above the chosen point p, the fibre (given reduced scheme structure) is a collection of curves $C_1, \ldots, C_n$, all isomorphic to $\mathbb{P}^1$. The number of curves depends on which closed point on the base is chosen.

Regardless, the application of Section 3.4 to the collection of structure sheaves $\mathcal{O}_{C_1}, \ldots, \mathcal{O}_{C_n} \in \operatorname{Qcoh} \mathcal{X}$ gives rise to the noncommutative deformation functor

$$\mathcal{D}\mathrm{ef}_{C_1,\ldots,C_n} \colon \mathfrak{art}_n \to \mathrm{Sets}\,.$$

Theorem 4.1 ([12, 13]). *There is an algebra* A_{con} *and a functorial isomorphism*

$$\mathcal{D}\mathrm{ef}_{C_1,\ldots,C_n}(\Gamma) \cong \operatorname{Hom}(\mathrm{A}_{\text{con}}, \Gamma)/\sim,$$

where $\sim$ *is the action of inner automorphisms in* Γ.

The algebra A_{con} is called the contraction algebra. In fact, there is a second algebra A_{def} called the noncommutative deformation algebra, but it is shown in [12] that the two are isomorphic; from the viewpoint of Section 5.4 below, this latter algebra A_{def} is important, since it can be defined *without* needing to assume the existence of a contraction. Regardless, in general, A_{con} is very challenging to compute. At least with some extra smoothness assumptions, Section 4.3 explains the link to noncommutative singularity theory.

We briefly note that there is a small technical issue in [12, 14] since unlike [25, Proposition 2.11], those papers forgot to account for the action of inner automorphisms in the statement of Theorem 4.1. In practice, this does not affect anything; as explained in [25, Proposition 2.14], even after accounting for inner automorphisms, the prorepresenting object is still unique up to non-unique isomorphism. An alternative approach to this is to redefine the morphisms in the category $\mathfrak{art}_n$ to be conjugacy classes, as in [4].

4.2. Which deformation theory?

Given the abundance of deformation functors in Section 3, it seems natural to wonder which solves the geometric problems of interest.

Around the chosen point p on the base, above which there are curves $C_1, \dots, C_n$, we first ask which deformation functor detects the difference between situations ① and ②. In all cases, without assumptions on the singularities of $\mathcal{X}$, and without crepant assumptions, this is answered by the noncommutative deformation functor.

Theorem 4.2 ([13]). *Given a contraction $\mathcal{X} \to \mathcal{X}_{\mathrm{con}}$ as in (4.1), pick a point p on the base and consider the corresponding contraction algebra $\mathrm{A}_{\mathrm{con}}$ above that point. Then, locally around p,*

$$\begin{aligned} \operatorname{Jdim} \mathrm{A}_{\mathrm{con}} = 0 \quad &\Longleftrightarrow \quad \textit{Situation ① (flopping)}, \\ \operatorname{Jdim} \mathrm{A}_{\mathrm{con}} = 1 \quad &\Longleftrightarrow \quad \textit{Situation ② (div}\to\textit{curve)}. \end{aligned}$$

With extra smoothness assumptions, $\mathrm{A}_{\mathrm{con}}$ is a Jacobi algebra (see Section 4.3), and so in addition to detecting the nature of the geometry locally around p, the above theorem strongly motivates the study of potentials f satisfying the numerical condition $\operatorname{Jdim} \mathcal{J}\mathrm{ac}\, f \leq 1$.

The larger question is which deformation functor, if any, is sufficient to *classify* the curve neighbourhood, since this would be the strongest property of all. The following has become known as the Donovan–Wemyss conjecture, where for ease of exposition, we restrict to the case of a single curve. Classification requires some form of smoothness, and the rough slogan is that contraction algebras are always the classifying objects for smooth crepant contractible neighbourhoods.

Conjecture 4.3 (Donovan–Wemyss). *Let $\mathcal{X}_1 \to \operatorname{Spec} \mathcal{R}_1$ and $\mathcal{X}_2 \to \operatorname{Spec} \mathcal{R}_2$ be as in (4.1), where both are crepant, both $\mathcal{X}_i$ are smooth, and both $\mathcal{R}_i$ are complete local cDV. Write $\mathrm{A}_{\mathrm{con}}$ and $\mathrm{B}_{\mathrm{con}}$ for their corresponding contraction algebras, and in each case, assume that there is only one curve. Then*

$$\mathcal{R}_1 \cong \mathcal{R}_2 \iff \mathrm{A}_{\mathrm{con}} \cong \mathrm{B}_{\mathrm{con}}.$$

In the setting of flops ①, Conjecture 4.3 is now a theorem due to [3, 12, 18, 20], where after a series of reduction steps [20] apply their derived Auslander–Iyama correspondence. The divisor-to-curve case of Conjecture 4.3 remains open.

Regardless, it is useful to regard the word conjecture as a challenge to extend the scope of Conjecture 4.3 to the maximum extent possible; it may even be true that contractibility is not required and that the noncommutative deformation algebra classifies smooth crepant curve neighbourhoods in full generality.

4.3. Link to noncommutative singularity theory

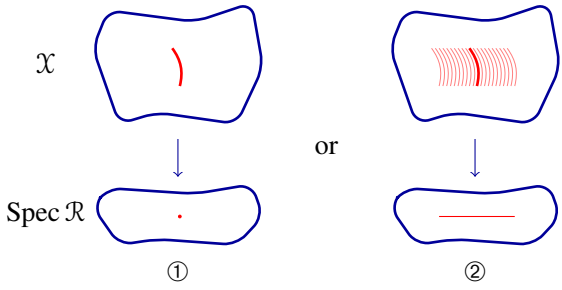
As already alluded to, when $\mathcal{X}$ is smooth and the morphism is crepant, it is automatic that $\mathrm{A}_{\mathrm{con}}$ is the Jacobi algebra of a quiver with potential. The quiver can be described easily from the data of the configuration of the curves contracted and their normal bundles [26, Theorem 2.15]. Describing the potential is significantly more challenging.

This is precisely the domain of noncommutative singularity theory: with the combinatorial data of the quiver fixed, it asks what functions (=potentials) can arise. To calibrate, in the case of a single crepant curve, which is often the hardest case, $\mathrm{A}_{\mathrm{con}}$ is a factor of $\mathbb{C}\langle\langle x, y\rangle\rangle$. The fact that we should expect only two variables to show up should be viewed in analogy with simple singularities in Theorem 2.3; for a full explanation, see [7, Appendix A.13].

5. New applications and results

This section briefly summarises some of the upcoming work [8] which applies noncommutative singularity theory to classify normal forms into noncommutative analogues of the ADE families. We focus here on the properties that we currently know to be numerical, leaving the technical aspects to [8]. In Section 5.2, we then briefly outline the main application to the classification of smooth 3-fold flops, and in Section 5.3, we outline some of the more surprising consequences to curve-counting invariants.

To simplify the discussion, henceforth we consider crepant contractions of 3-folds of only the following two types:



where the contraction map f is projective birational, $\mathrm{R}f_*\mathcal{O}_{\mathcal{X}} = \mathcal{R}$, there is only one curve above the origin, $\mathcal{X}$ is smooth, and the base $\mathcal{R}$ is complete local and Gorenstein.

5.1. Noncommutative forms

In classifying all f with $\mathrm{Jdim}\,\mathcal{J}\mathrm{ac}\,f \leq 1$, the main result [7, Appendix A.13] reduces to one of the following three cases: either $\mathcal{C}\mathrm{rk}(f) \leq 1$, or $\mathcal{C}\mathrm{rk}(f) = 2$ and either $\mathcal{C}\mathrm{rk}_2(f) = 2, 3$. From this, a first coarse numerical subdivision is as follows.

Type	$\mathcal{C}\mathbf{rk}(f)$	$\mathcal{C}\mathbf{rk_2}(f)$	$\mathcal{C}\mathbf{rk_3}(f)$	$\mathcal{C}\mathbf{rk_4}(f)$
A	≤ 1	≤ 1	≤ 1	≤ 1
D	2	2	2	≤ 2
E_6	2	3	4	4

To unpack this in one case, if $\mathcal{C}\mathrm{rk}(f) = 2$ and $\mathcal{C}\mathrm{rk}_2(f) = 3$, then up to isomorphism $f \in \mathbb{C}\langle\langle x, y\rangle\rangle$ has zero linear and quadratic parts, and the derivatives of the cubic part are linearly dependent: that is, $f \cong x^3 +$ terms of order ≥ 4.

However, as stated, the above is not quite the full picture. While these numerical conditions on a potential f do happen to imply $\mathrm{Jdim}\,\mathcal{J}\mathrm{ac}\,f \leq 1$ (but only after classification), there are other families with $\mathrm{Jdim}\,\mathcal{J}\mathrm{ac}\,f = 0$ that are currently defined by more than just numerical coranks, using also analytic data about slices by central elements. These other families are denoted by E_7, $\mathrm{E}_{8,5}$, and $\mathrm{E}_{8,6}$, and [8] explains both the motivation and the technical details.

Given this, we state the following slightly tersely, ignoring the exceptional cases.

Theorem 5.1 ([7, 8]). *Situation ① (flopping). Then* $\mathrm{A}_{\mathrm{con}} \cong \mathcal{J}\mathrm{ac}\,f$*, where* f *is either*

$$\begin{array}{lll} \mathrm{A}_n & x^2 + y^n, & n \geq 2, \\ \mathrm{D}_{n,m} & xy^2 + x^{2m-1} + x^{2n}, & n, m \geq 2, \\ \mathrm{D}_{n,\infty} & xy^2 + x^{2n}, & n \geq 2, \\ \mathrm{E}_{6,n} & x^3 + xy^3 + p(y), & p \in \mathbb{C}[t],\ \mathrm{ord}(p) = n \geq 4, \end{array}$$

or a normal form of type E_7*,* $\mathrm{E}_{8,5}$*, or* $\mathrm{E}_{8,6}$*. (These other type* E *forms also have moduli but not the discrete parameter* n*.)*

As stated, Theorem 5.1 is, strictly speaking, not a list of *normal* forms, since different $p(y)$ in E_6 may give isomorphic Jacobi algebras. As in Arnold's original work [1], the existence of moduli makes normal forms less pressing; the more important point is to cover all cases, which Theorem 5.1 does. Moreover, there are indeed nontrivial moduli in E_6, as we show in [8] that two general members of the family $x^3 + xy^3 + y^4 + ay^5 + by^6$ with $a, b \in \mathbb{C}$ are not isomorphic.

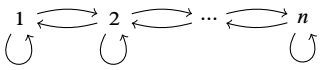
In contrast to Theorem 5.1, the following result does not need to be stated tersely: it turns out, contrary to our own initial expectations, that contraction algebras of crepant divisor-to-curve contractions are very rare, and indeed all can be defined numerically.

Theorem 5.2 ([7,8]). *Situation ② (div→curve). Then* $\mathrm{A}_{\mathrm{con}} \cong \mathcal{J}\mathrm{ac}\, f$*, where* f *is either*

$$\begin{array}{ll} \mathrm{A}_\infty & x^2, \\ \mathrm{D}_{\infty,m} & xy^2 + x^{2m-1}, \quad m \geq 2, \\ \mathrm{D}_{\infty,\infty} & xy^2, \\ \mathrm{E}_{6,\infty} & x^3 + xy^3. \end{array}$$

It is worth pausing to reflect that there are very few examples, namely just one countable family and three exceptional cases. It seems surprising that $\mathrm{Jdim}\, \mathcal{J}\mathrm{ac}\, f = 1$ algebras are significantly rarer than those satisfying $\mathrm{Jdim}\, \mathcal{J}\mathrm{ac}\, f = 0$. The second remarkable aspect of Theorem 5.2 is that every case can be viewed as the $n \to \infty$ limit of a family in Theorem 5.1. This fact is a consequence of the classification, not an input, which we think of as trying to make sense in cold-blooded algebra of the vague intuition that every crepant divisor-to-curve contraction is an infinite limit of flops.

Remark 5.3. In addition to the results presented here, the upcoming paper of Zhang [28] considers the noncommutative singularity theory on the following quiver, where the single loop at each vertex is optional.



Zhang [28] gives various numerical and structural characterisations of 'type A' potentials on this quiver and, amongst other things, proves the realisation conjecture (Conjecture 5.6 below) for this intrinsic class of potentials.

5.2. Structural geometric consequences

Due to Conjecture 4.3, the normal forms in the previous subsection have immediate consequences, once we are able to build the geometry which realises each such potential. The overarching idea to do this elegantly is to use [6], which was itself heavily inspired by the string theory literature.

Theorem 5.4 ([23], [8] ongoing). *Smooth* 3*-fold single-curve flops* $\mathcal{X} \to \mathrm{Spec}\, \mathcal{R}$ *are classified. The classification data can be stated in various forms: by the equation defining* $\mathcal{R}$*, by describing the glue defining* $\mathcal{X}$*, or by the slice of the universal flop.*

The precise statement of the theorem above is a work in progress – noting that Reid's pagoda [23] already handles the type A cases – in part because some pieces of the data are much harder to describe than others. The equation defining $\mathcal{R}$ together with its class group is perhaps the most compelling description, but that seems to be the hardest of all. All the same, from the perspective of users of the data, having multiple descriptions of each flop is the most beneficial outcome; for a commutative analogue, compare [15], either as motivation or provocation.

The various data that can be used to finally classify flops does serve to illustrate, above all, the elegance of the noncommutative forms in Section 5.1. To give a feeling of this, from E_6 onwards the equations defining the base of the flop become vast and unwieldy; as one example, the following is an example of the base of an E_8 flop:

$$\begin{aligned}\operatorname{Spec}\mathcal{R} = (&2041x^{12} - 107109x^{11} + 1390131x^{10} - 6130x^9y - 8432x^9\\ &+ 157127x^8y - 17x^8z + 45x^8 - 654x^7y^2 - 340x^7y + 853x^7z\\ &- x^7 + 21842x^6y^2 - 6x^6y + 5905x^6z + 7157x^6t - 3x^5y^3\\ &- 66x^5y^2 - 93x^5yz - 15x^5z - 2358x^5t + 1090x^4y^3 + x^4y^2\\ &+ 355x^4yz - 9x^4z^2 + 7x^4t - x^3y^3 - 2x^3y^2z - 133x^3yt\\ &+ 15x^3z^2 + y^5 + 41x^2y^4 + 34x^2y^2z + 6x^2z^2 - x^2zt\\ &+ xy^3z - 15xy^2t - 5xzt + z^3 + t^2 = 0).\end{aligned}$$

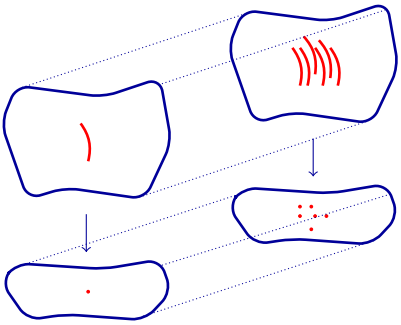
The particular values of at least some of the coefficients in that mighty polynomial are important, but without significant work, it is difficult to say which. The polynomial would, all things considered, be hard to guess without the noncommutative normal form guiding its construction. It is at least reassuring to see that slicing the above by $x = 0$ gives the equation $t^2 + z^3 + y^5 = 0$, which is the E_8 normal form in Theorem 2.3.

5.3. Numerical geometric consequences

There are various numerical consequences of classification in the previous subsection. Fix a flopping contraction $f : \mathcal{X} \to \operatorname{Spec}\mathcal{R}$ with exceptional curve C. Then given any $\beta \in \mathbb{Z}[C]$, so just an integer, Katz [21] defines an associated Gopakumar–Vafa (GV) invariant

$$n_\beta = n_\beta(\mathcal{X}) \in \mathbb{Z}_{\geq 0}$$

by considering a one-parameter deformation



such that

(1) the central fibre $g_0\colon X_0 \to Y_0$ is isomorphic to $f\colon \mathcal{X} \to \operatorname{Spec} \mathcal{R}$,

(2) all other fibres $g_t\colon X_t \to Y_t$ are flopping contractions whose exceptional locus is a disjoint union of $(-1,-1)$-curves.

The idea is then to obtain numerical information by counting the curves in a neighbouring fibre. Regarding the original flopping curve C of f as a curve in the 4-fold $\mathcal{X}$ swept out by the one-parameter deformation, then the GV invariant n_j is defined to be the number of g_t-exceptional $(-1,-1)$-curves C' with curve class $j[C]$. In other words, n_j is the number of $(-1,-1)$-curves C' such that for every line bundle $\mathcal{L}$ on $\mathcal{X}$,

$$\deg(\mathcal{L}|_{C'}) = j \deg(\mathcal{L}|_C).$$

Almost all the n_j are zero. In fact, at most $(n_1, \ldots, n_6)$ are non-zero.

The following is a numerical consequence of the classification of flops via Jacobi algebras in the previous subsections.

Theorem 5.5. *The GV invariants are known for all smooth irreducible* 3*-fold flops.*

Full details will again appear in [8]. To outline one particularly remarkable phenomenon, *all* E_7 flops (equivalently, the so-called length four flops) have the same GV invariants, namely $(6, 5, 2, 1)$. No other values are possible. Similarly, all the Gromov–Witten invariants and all the Donaldson–Thomas invariants are the same for any E_7 flop. Thus, from a numerical perspective, all E_7 flops are identical: it is the extra information of the isomorphism class of the contraction algebra – that is, the class of the noncommutative potential $f \in \mathbb{C}\langle\langle x, y \rangle\rangle$ – that distinguishes them.

5.4. Conjectures

We say that f is *geometric* if it arises from geometry, namely if $\mathcal{J}\mathrm{ac}\, f$ is isomorphic to the contraction algebra of some $\mathcal{X} \to \operatorname{Spec} \mathcal{R}$ in Section 4.1 (either case ① or ②).

Conjecture 5.6 (Geometric realisation conjecture). *Every f whose Jacobi algebra satisfies* $\dim \mathcal{J}\mathrm{ac}\, f \leq 1$ *is geometric.*

Tens of thousands of computer searches (with magma [5] and singular [10]) have failed to uncover counterexamples to Conjecture 5.6, but the methods of [7, 8] do not yet seem strong enough to prove it. At this stage, it seems that a new ingredient in noncommutative singularity theory will be needed, if indeed Conjecture 5.6 is true.

Now let $C \cong \mathbb{P}^1 \subseteq X$, where X is a Calabi–Yau 3-fold. Write Λ_{def} for the noncommutative deformation algebra associated with C, obtained via noncommutative deformation theory. It is conjectured in [6] that in this 3-fold setting, numerical properties of Λ_{def} are sufficient to detect whether or not C contracts.

Conjecture 5.7 (Artin contractibility for 3-folds). *In a suitably local neighbourhood of C in X, the following are equivalent:*

(1) *C contracts to a point and is an isomorphism elsewhere.*

(2) $\dim_{\mathbb{C}} \Lambda_{\mathrm{def}} < \infty$.

The statement of Conjecture 5.7 is given for a single curve, but one of the advantages of the noncommutative deformation theory technology is that it finally permits clean statements for multi-curve configurations; for details, see [6].

It may be the case that the condition $\mathrm{H}^1(\widehat{X}, \mathcal{O}_{\widehat{X}}) = 0$ needs to be added to the assumptions in the conjecture, since this would then more closely match the classical Artin statement for surfaces. However, just as for Conjecture 5.6 above, tens of thousands of computer searches suggest that the condition $\dim_{\mathbb{C}} \Lambda_{\mathrm{def}} < \infty$ is so restrictive that the rational neighbourhood assumption may in fact come for free.

To end, it seems fitting to note the limitations of our view, or, perhaps better, the enormous potential in what has not been covered. The discussion here has considered what we might term noncommutative function germs, that is, elements of $\mathbb{C}\langle\langle x_1, \ldots, x_d \rangle\rangle$ or more generally elements of complete local path algebras on quivers. This is based largely on our admiration and profitable exploitation of a few of the ideas of Arnold's seminal paper [1]. Conventional commutative singularity theory is of course a much larger subject, exemplified by the vast literature, such as [2, 9, 22] to cite a few thousand pages. We have not mentioned notions of stability, k-determinacy, adjacency, versality, more general map germs, vector fields, or a host of other very natural notions. These may or may not have profitable extensions into the noncommutative realm. But for now, we simply do not know, in part due to our

technical limitations, but also since we do not yet know how we would be able to apply them.

Acknowledgements. Truth be told, it was during an otherwise uneventful late afternoon walk in September 2020 between the two authors and a young and crying baby, just past the junction of the footpath that links Palmerston Road to Beechwood Avenue in the suburbs of Coventry, where the idea of a noncommutative singularity theory was born. Life is nothing if not glamorous. However, in that moment, absolutely everything we had been thinking about for almost a decade vastly simplified, and the road forward became crystal clear.

Then, but also much more broadly, maths is fundamentally a collaborative endeavour, and it would have been impossible to be delivering this article without the generous input and insights of many people over the years: from collaborators Jenny August, Will Donovan, Wahei Hara, Yuki Hirano, Osamu Iyama, Navid Nabijou, Ivan Smith; to PhD students Joe Karmazyn, Noah White, Kellan Steele, Okke van Garderen, Sarah Kelleher, Caroline Namanya, Brian Makonzi, Sam Lewis, Zhang Hao, Charlotte Bartram, Marina Godinho, Parth Shimpi; and to postdocs Alice Rizzardo, Agnieszka Bodzenta, Theo Raedschelders, Diletta Martinelli, Ana Garcia Elsener, Franco Rota, Matt Pressland, Emma Lepri, and Timothy De Deyn.

Funding. This work was supported by EPSRC grant EP/R034826/1 and by the ERC Consolidator grant 101001227 (MMiMMa).

References

[1] V. I. Arnold, Local normal forms of functions. *Invent. Math.* **35** (1976), no. 1, 87–109 Zbl 0336.57022 MR 0467795

[2] V. I. Arnold, S. M. Gusein-Zade, and A. N. Varchenko, *Singularities of differentiable maps. Classification of critical points, caustics and wave fronts, Vol. 1*. Mod. Birkhäuser Class., Birkhäuser, Boston, MA, 2012, 382 pp. Zbl 1290.58001 MR 2896292

[3] J. August, On the finiteness of the derived equivalence classes of some stable endomorphism rings. *Math. Z.* **296** (2020), no. 3–4, 1157–1183 Zbl 1473.16004 MR 4159827

[4] A. Bodzenta and A. Bondal, Categorifying non-commutative deformations. *J. Lond. Math. Soc. (2)* **111** (2025), no. 5, article no. e70176, 41 pp. Zbl 08053200 MR 4913180

[5] W. Bosma, J. Cannon, and C. Playoust, The MAGMA algebra system I: the user language. *J. Symbolic Comput.* **24** (1997), no. 3–4, 235–265 Zbl 0898.68039 MR 1484478

[6] G. Brown and M. Wemyss, Derived deformation theory of crepant curves. *J. Topol.* **17** (2024), no. 4, article no. e12359, 42 pp. Zbl 1565.14012 MR 4822926

[7] G. Brown and M. Wemyss, Local normal forms of noncommutative functions. *Forum Math. Pi* **13** (2025), article no. e8, 59 pp. Zbl 1564.14009 MR 4865672

[8] G. Brown and M. Wemyss, The classification of smooth 3-fold fops. In preparation

[9] J. L. Cisneros-Molina, D. L. Tráng, and J. Seade (Eds.), *Handbook of geometry and topology of singularities VII*. Springer, Cham, 2025, 830 pp. Zbl 1560.58001 MR 4890425

[10] W. Decker, G.-M. Greuel, G. Pfister, and H. Schönemann, SINGULAR 4-1-0 — *a computer algebra system for polynomial computations*. 2017, http://www.singular.uni-kl.de visited on 20 April 2026

[11] H. Derksen, J. Weyman, and A. Zelevinsky, Quivers with potentials and their representations I: mutations. *Selecta Math. (N.S.)* **14** (2008), no. 1, 59–119 Zbl 1204.16008 MR 2480710

[12] W. Donovan and M. Wemyss, Noncommutative deformations and flops. *Duke Math. J.* **165** (2016), no. 8, 1397–1474 Zbl 1346.14031 MR 3504176

[13] W. Donovan and M. Wemyss, Contractions and deformations. *Amer. J. Math.* **141** (2019), no. 3, 563–592 Zbl 1420.14022 MR 3956515

[14] W. Donovan and M. Wemyss, Twists and braids for general 3-fold flops. *J. Eur. Math. Soc. (JEMS)* **21** (2019), no. 6, 1641–1701 Zbl 1431.14031 MR 3945738

[15] A. H. Durfee, Fifteen characterizations of rational double points and simple critical points. *Enseign. Math. (2)* **25** (1979), no. 1–2, 131–163 Zbl 0418.14020 MR 0543555

[16] E. Eriksen, An introduction to noncommutative deformations of modules. In *Noncommutative algebra and geometry*, pp. 90–125, Lect. Notes Pure Appl. Math. 243, Chapman & Hall/CRC, Boca Raton, FL, 2006 Zbl 1104.16025 MR 2189988

[17] L. Gerritzen and R. Holtkamp, On Gröbner bases of noncommutative power series. *Indag. Math. (N.S.)* **9** (1998), no. 4, 503–519 Zbl 0930.16017 MR 1691990

[18] Z. Hua and B. Keller, Cluster categories and rational curves. *Geom. Topol.* **28** (2024), no. 6, 2569–2634 Zbl 1564.14006 MR 4817470

[19] Z. Hua and G. Zhou, Quasi-homogeneity of potentials. *J. Noncommut. Geom.* **15** (2021), no. 2, 399–422 Zbl 1481.16032 MR 4325712

[20] G. Jasso, F. Muro, and B. Keller, The derived Auslander–Iyama correspondence. *J. Amer. Math. Soc.* (2026) DOI 10.1090/jams/1079

[21] S. Katz, Genus zero Gopakumar–Vafa invariants of contractible curves. *J. Differential Geom.* **79** (2008), no. 2, 185–195 Zbl 1142.32011 MR 2420017

[22] D. Mond and J. J. Nuño-Ballesteros, *Singularities of mappings. The local behaviour of smooth and complex analytic mappings*. Grundlehren Math. Wiss. 357, Springer, Cham, 2020, 567 pp. Zbl 1448.58032 MR 4321457

[23] M. Reid, Minimal models of canonical 3-folds. In S. Iitaka (Ed.), *Algebraic varieties and analytic varieties (Tokyo, 1981)*, pp. 131–180, Adv. Stud. Pure Math. 1, North-Holland, Amsterdam, 1983 Zbl 0558.14028 MR 0715649

[24] K. Saito, Quasihomogene isolierte Singularitäten von Hyperflächen. *Invent. Math.* **14** (1971), no. 2, 123–142 Zbl 0224.32011 MR 0294699

[25] E. Segal, The A_∞ deformation theory of a point and the derived categories of local Calabi–Yaus. *J. Algebra* **320** (2008), no. 8, 3232–3268 Zbl 1168.18005 MR 2450725

[26] M. Wemyss, Flops and clusters in the homological minimal model programme. *Invent. Math.* **211** (2018), no. 2, 435–521 Zbl 1390.14012 MR 3748312

[27] Y. Yoshino, *Cohen–Macaulay modules over Cohen–Macaulay rings*. London Math. Soc. Lecture Note Ser. 146, Cambridge University Press, Cambridge, 1990, 177 pp. Zbl 0745.13003 MR 1079937

[28] H. Zhang, Local forms for the double A_n quiver. *Math. Z.* **312** (2026), no. 4, article no. 125, 66 pp. MR 5051022

Gavin Brown
Mathematics Institute, University of Warwick, Zeeman Building, Coventry CV4 7AL, UK; g.brown@warwick.ac.uk

Michael Wemyss
School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow G12 8QQ, UK; michael.wemyss@glasgow.ac.uk

EMS prize lectures

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/35

Sketches of nonuniformly elliptic Schauder theory

Cristiana De Filippis

Abstract. Schauder theory is a basic tool in the study of elliptic and parabolic PDEs, asserting that solutions inherit the regularity of the coefficients. It plays a central role in establishing higher regularity for solutions to a broad class of elliptic problems exhibiting ellipticity, including those involving free boundaries. In the linear setting, Schauder theory dates back to the 1920–1930s and is now considered classical. Nonlinear extensions were developed in the 1980s. All these classical results are restricted to uniformly elliptic operators and heavily rely on perturbative techniques – freezing the coefficients and comparing the solution to that of a constant-coefficient problem. However, such methods fail in the nonuniformly elliptic setting, where homogeneous a priori estimates break down and standard iteration arguments no longer apply. Here, we give a brief survey of recent progress including the solution to the long-standing problem of proving the validity of Schauder estimates in the nonlinear, nonuniformly elliptic setting.

Notation. Through this note, $\Omega \subseteq \mathbb{R}^n$, $n \geq 2$, denotes an open, bounded domain with Lipschitz boundary; $\mathrm{B} \equiv B_\varrho \equiv B_\varrho(x_0) \Subset \Omega$ will denote a ball compactly contained into Ω, centred at x_0 with radius $\varrho > 0$. We shall also denote $\mathcal{B}_\varrho = B_\varrho(0_{\mathbb{R}^n})$. Sometimes we shall use symbols "$\gtrsim$", "$\lesssim$" with subscripts, to indicate that a certain inequality holds up to constants whose dependencies are marked in the subfix. With $\Omega_0 \subset \mathbb{R}^n$ being a measurable subset with bounded positive Lebesgue measure $0 < |\Omega_0| < \infty$, $\mathbb{N} \ni k \geq 1$, $\kappa \in \mathbb{R}$ being numbers and $w \colon \Omega_0 \to \mathbb{R}^k$ being an integrable map, we denote

$$(w)_{\Omega_0} \equiv \fint_{\Omega_0} w(x)\,\mathrm{d}x := \frac{1}{|\Omega_0|}\int_{\Omega_0} w(x)\,\mathrm{d}x$$

its integral average and $(w-\kappa)_+ := \max\{w-\kappa, 0\}$ its upper truncation. Finally, whenever introducing various objects instrumental to our presentation, unless clearly stated, we will implicitly assume that all the quantities involved are regular enough to ensure the exposition is well defined.

Mathematics Subject Classification 2020: 35B65 (primary); 49N60, 35J60 (secondary).
Keywords: regularity, nonuniform ellipticity, nonlinear potential theory.

1. Measuring ellipticity: Equations and functionals

The Laplace equation $\Delta u = 0$ is the most prominent among elliptic partial differential equations of the form

$$-\operatorname{div} A(x, Du) = 0 \quad \text{in } \Omega, \tag{1.1}$$

defined by means of a Carathéodory regular[1] vector field $A\colon \Omega \times \mathbb{R}^n \to \mathbb{R}^n$. Ellipticity means that $\partial_z A(x,z)$, whenever it exists, is a (symmetric), positive definite $\mathbb{R}^{n\times n}$-matrix.[2] More precisely, we shall consider (1.1) under the condition

$$g_1(x,|z|)\mathbb{I}_{n\times n} \le \partial_z A(x,z) \le g_2(x,|z|)\mathbb{I}_{n\times n}, \tag{1.2}$$

where the Carathéodory-regular functions $g_1, g_2\colon \Omega \times (0,\infty) \to (0,\infty)$ are the smallest and the largest eigenvalues of $\partial_z A$, respectively. To avoid technical complications, unless otherwise stated, we shall assume that g_1 and g_2 vanish at most when $|z| = 0$, and that $\partial_z A(x,z)$ always exists except when $|z| = 0$. The class of equations (1.1), subject to (1.2), appears in the modelling of a large variety of stationary phenomena such as the theory of electrostatic or electromagnetic potentials, or in the search for vibration modes in elastic structures. We are interested in the regularity of solutions. Measuring the ellipticity of (1.1) provides crucial regularity information. To fix ideas, let us consider the linear equation

$$-\operatorname{div}(\mathcal{A}(x)Du) = 0 \quad \text{in } \Omega, \tag{1.3}$$

where, for simplicity, $\mathcal{A}$ is a measurable, symmetric and positive definite $\mathbb{R}^{n\times n}$-matrix. The standard ellipticity quantifier is the ellipticity ratio $\mathcal{R}$, defined as the ratio of the highest eigenvalue of $\mathcal{A}$ to the lowest eigenvalue of $\mathcal{A}$

$$\mathcal{R}(x) := \frac{\text{highest eigenvalue of } \mathcal{A}(x)}{\text{lowest eigenvalue of } \mathcal{A}(x)}. \tag{1.4}$$

The condition

$$\|\mathcal{R}\|_{L^\infty(\mathrm{B})} \text{ is finite for all balls } \mathrm{B} \Subset \Omega \tag{1.5}$$

[1] This means that $x \mapsto A(x,z)$ is measurable for every choice of $z \in \mathbb{R}^n$, and $z \mapsto A(x,z)$ is continuous for a.e. $x \in \Omega$. It serves to ensure that the composition $x \mapsto A(x, D(x))$ is measurable whenever D is a measurable vector field.

[2] For simplicity, we assume that $\partial_z A$ is symmetric, i.e., $\partial_{z_j} A_i = \partial_{z_i} A_j$ for every $1 \le i, j \le n$. This is the case when A comes from a potential and equation (1.1) is variational, as (1.15) below. As a matter of fact, in this note, we shall consider only the variational case.

classifies (1.3) as a uniformly elliptic PDE [69, Chapter 3]; if otherwise, $\|\mathcal{R}\|_{L^\infty(\mathrm{B})}$ is not finite for at least one ball $\mathrm{B} \Subset \Omega$, equation (1.3) is called nonuniformly elliptic.[3] In order to extend this notion to the nonlinear elliptic setting (1.1)–(1.2), let us consider the simplified equation $-\operatorname{div} A(Du) = 0$ and (formally) differentiate it in the s-direction, $s \in \{1, \dots, n\}$. We find out that every component $v_s := D_s u$ solves

$$-\operatorname{div}(\mathcal{A}(x) D v_s) = 0, \quad \mathcal{A}(x) := \partial_z A(Du(x)), \tag{1.6}$$

i.e., second derivatives of u solve a linear equation with matrix coefficient $\mathcal{A}$ being a nonlinear, measurable function of Du. Back to the general case, the adaptation of (1.4) to the nonlinear framework (1.1) towards gradient regularity is then quite natural [120,125]. Keeping (1.2) and (1.6) in mind, we introduce the pointwise ellipticity ratio for $|z| \neq 0$,

$$\mathcal{R}(x, z) := \frac{g_2(x, |z|)}{g_1(x, |z|)}, \tag{1.7}$$

and, again for $|z| \neq 0$, the nonlocal one [41],

$$\bar{\mathcal{R}}(z, \mathrm{B}) := \frac{\sup_{x \in \mathrm{B}} g_2(x, |z|)}{\inf_{x \in \mathrm{B}} g_1(x, |z|)} \tag{1.8}$$

for any ball $\mathrm{B} \Subset \Omega$. Quantities (1.7) and (1.8) are well defined thanks to the ellipticity condition (1.2) (with the obvious agreement that $\bar{\mathcal{R}}(z, \mathrm{B}) = \infty$ when the denominator in (1.8) is zero[4]). Moreover, by definition,

$$\begin{cases} \mathcal{R}(x, z) \le \bar{\mathcal{R}}(z, \mathrm{B}) \quad \text{for } x \in \mathrm{B}, \\ A(x, z) \equiv A(z) \Longrightarrow \mathcal{R}(x, z) = \bar{\mathcal{R}}(z, \mathrm{B}). \end{cases} \tag{1.9}$$

As suggested in [41], we adopt a slightly more refined taxonomy for uniform ellipticity.

- A is uniformly elliptic if

$$\sup_{|z|>0} \bar{\mathcal{R}}(z, \mathrm{B}) \lesssim_{\mathrm{B}} 1 \tag{1.10}$$

  for all balls $\mathrm{B} \Subset \Omega$.

[3] Since in this note we are mainly interested in local regularity properties, here we are using suitably localized notions of uniform and nonuniform ellipticity, that slightly differ from the classical ones. Specifically, the finiteness of the ellipticity ratio is analysed only on interior balls, rather than on the whole domain as usually done [69], i.e., Ω is taken instead of B in (1.5). This naturally suits interior regularity. Of course, this distinction is not strictly necessary in our setting.

[4] This will never occur in the situations considered here.

- A is softly nonuniformly elliptic if it is pointwise uniformly elliptic, i.e.,

$$\sup_{x\in \mathrm{B},|z|>0} \mathcal{R}(x,z) \lesssim_{\mathrm{B}} 1 \tag{1.11}$$

holds for all balls $\mathrm{B} \Subset \Omega$, but

$$\sup_{|z|>0} \bar{\mathcal{R}}(z,\mathrm{B}) = +\infty \tag{1.12}$$

holds for some ball $\mathrm{B} \Subset \Omega$.

- A is (strongly, pointwise) nonuniformly elliptic if[5]

$$\sup_{x\in \mathrm{B},|z|>0} \mathcal{R}(x,z) = +\infty \tag{1.13}$$

on some ball $\mathrm{B} \Subset \Omega$.

The classical definition of nonuniform ellipticity only prescribes to distinguish between the occurrence of (1.11) and (1.13) (with B replaced by Ω). Instead, here we consider an intermediate notion that on the other hand becomes immaterial in the autonomous case in the second line of (1.9). Specifically, the pointwise ellipticity ratio captures how the growth of the gradient variable affects ellipticity, while the nonlocal one governs the way space-dependent coefficients mix-up with gradients and indicates how much coefficients deviate from merely having a perturbative effect. Let us stress that, when dealing with nonuniformly elliptic problems, it is more natural to consider the variational setting, which provides the correct notion of solution (minimizer) in the natural energy setting. Specifically, we consider variational integrals of the type

$$\mathcal{F}(w,\Omega) := \int_{\Omega} F(x, Dw)\,\mathrm{d}x, \tag{1.14}$$

governed by an elliptic integrand $F\colon \Omega \times \mathbb{R}^n \to [0,+\infty)$. The catch between the functional $\mathcal{F}$ and equation (1.1) is the Euler–Lagrange equation

$$-\mathrm{div}\, \partial_z F(x, Du) = 0 \quad \text{in } \Omega, \tag{1.15}$$

[5] In line with [120], possible alternative definitions of uniform ellipticity can be given by replacing (1.10) and (1.11) with

$$\limsup_{|z|\to\infty} \bar{\mathcal{R}}(z,\mathrm{B}) < \infty \quad \text{and} \quad \limsup_{|z|\to\infty} \sup_{x} \mathcal{R}(x,z) < \infty,$$

respectively. This is according to the fact that one is interested in proving Lipschitz continuity of solutions, after which in several cases the problems in question become uniformly elliptic again. Therefore, one is essentially interested in looking at those situations, where ellipticity weakens and the gradient becomes simultaneously large.

solved by minimizers under reasonable structure conditions. Thanks to (1.15), letting $A = \partial_z F$, the positions in (1.7) and (1.8) can be directly adapted to the variational framework: functions g_1 and g_2 in (1.2) now become lower and upper bounds for the lowest and the highest eigenvalues of the Hessian $\partial_{zz} F$, respectively. It is convenient to clarify the notion of minimality we are adopting.

Definition 1.1 (Minima). A function $u \in W^{1,1}_{\mathrm{loc}}(\Omega)$ is a *(local) minimizer* of the functional $\mathcal{F}$ in (1.14) if, for every ball $\mathrm{B} \Subset \Omega$, $F(\cdot, Du) \in L^1(\mathrm{B})$ and $\mathcal{F}(u, \mathrm{B}) \le \mathcal{F}(w, \mathrm{B})$ holds for every $w \in u + W^{1,1}_0(\mathrm{B})$.

1.1. Master examples

Basic model examples capturing some of the main aspects of the ellipticity notions discussed above are in the following.

1.1.1. Uniform ellipticity. This is the case of equations of the p-Laplace type,

$$\begin{cases} -\mathrm{div}(a(x)|Du|^{p-2}Du) = 0 \quad \text{in } \Omega, \\ 1 < p < \infty, \quad 1 \lesssim a \in L^\infty_{\mathrm{loc}}(\Omega). \end{cases} \tag{1.16}$$

A direct computation yields[6]

$$\mathcal{R}(x,z) \le \frac{\max\{1, p-1\}}{\min\{1, p-1\}}, \quad \bar{\mathcal{R}}(z, \mathrm{B}) \le \frac{\sup_{x \in \mathrm{B}} a(x)}{\inf_{x \in \mathrm{B}} a(x)} \frac{\max\{1, p-1\}}{\min\{1, p-1\}}$$

for all balls $\mathrm{B} \Subset \Omega$. After the foundational contributions of Ural'tseva [130] and Uhlenbeck [129] on the case $a(\cdot) \equiv 1$, maximal regularity for energy solutions[7] to (1.16), i.e., gradient Hölder continuity, was established in the works of DiBenedetto [50] and Manfredi [98]; see also Kuusi and Mingione [90] for related borderline cases.

1.1.2. Soft nonuniform ellipticity. A significant example is the double-phase integral

$$\begin{cases} \mathcal{D}(w, \Omega) := \displaystyle\int_\Omega |Dw|^p + a(x)|Dw|^q \, \mathrm{d}x, \\ 1 < p < q < \infty, \quad 0 \le a \in L^\infty_{\mathrm{loc}}(\Omega). \end{cases} \tag{1.17}$$

[6] In this note, with some abuse of notation, we interchange sup and esssup, the correct meaning being clear from the context.

[7] Energy solutions to (1.1) with $|A(x,z)| \lesssim |z|^{q-1} + 1$, $q > 1$, are those distributional solutions $u \in W^{1,1}_{\mathrm{loc}}(\Omega)$ satisfying the additional requirement $u \in W^{1,q}_{\mathrm{loc}}(\Omega)$, while distributional solutions are only required to satisfy $A(\cdot, Du) \in L^1_{\mathrm{loc}}(\Omega; \mathbb{R}^n)$. These are sometimes called very weak solutions and might be irregular [118]. In the case (1.16), an energy solution u is distributional solution such that $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$.

In this case, we have

$$\begin{cases} \mathcal{R}(x,z) \le \dfrac{q\max\{1,q-1\}}{p\min\{1,p-1\}}, \\ \bar{\mathcal{R}}(z,\mathrm{B}) \le \dfrac{q\max\{1,q-1\}}{p\min\{1,p-1\}}(\|\mathfrak{a}\|_{L^\infty(\mathrm{B})}|z|^{q-p}+1). \end{cases} \tag{1.18}$$

By the first line of (1.18), the integrand governing the functional $\mathcal{D}$ is uniformly elliptic in the classical sense, i.e., (1.11) holds. As a matter of fact, considering $\mathfrak{a}(x)=|x|^\alpha$, we have $\bar{\mathcal{R}}(z,\mathcal{B}_r)\approx r^\alpha|z|^{q-p}+1$ so that (1.12) holds. The possible blow-up of $\bar{\mathcal{R}}$ indicates that the vanishing of the coefficient $\mathfrak{a}$ is solely responsible for the (mild) nonuniformity in (1.17). The functional $\mathcal{D}$ was introduced by Zhikov [82, 132] in the setting of homogenization, and to investigate the occurrence of the Lavrentiev phenomenon. The mix-up of the coefficient $\mathfrak{a}$ and gradient variable in (1.17) might lead to the formation of singularities even in scalar, nondegenerate problems, as shown by Esposito, Leonetti and Mingione [54], Fonseca, Malý and Mingione [57] and Balci, Diening and Surnachev [5,6]. A complete regularity theory was eventually obtained by Baroni, Colombo and Mingione [10, 34]. We will further discuss (1.17) in Section 4.3. Another popular example of softly nonuniformly elliptic functional is the variable exponent one

$$w \mapsto \int_\Omega |Dw|^{\mathfrak{p}(x)}\,\mathrm{d}x, \tag{1.19}$$

where $\mathfrak{p}\colon\Omega\mapsto(1,\infty)$ is a measurable function such that $1<p\le\mathfrak{p}(\cdot)\le q<\infty$, and for which we have

$$\begin{cases} \mathcal{R}(x,z) \le \dfrac{q\max\{1,q-1\}}{p\min\{1,p-1\}}, \\ \bar{\mathcal{R}}(z,\mathrm{B}) \le \dfrac{q\max\{1,q-1\}}{p\min\{1,p-1\}}\max\{|z|,1/|z|\}^{\mathrm{osc}_{\mathrm{B}}\mathfrak{p}}. \end{cases}$$

Maximal regularity for minima of functionals modelled by the one in (1.19) was established by Acerbi and Mingione in [1]. Specifically, for minimizers u of (1.19), a Schauder-type result holds, asserting that the gradient Du is locally Hölder continuous whenever the exponent function $\mathfrak{p}$ itself is Hölder continuous. More examples of softly nonuniformly elliptic functionals are discussed in [4, 74].

1.1.3. Strong nonuniform ellipticity. A mostly celebrated example in this class is given by the area integral,[8]

$$w \mapsto \int_\Omega \sqrt{1+|Dw|^2}\,\mathrm{d}x, \tag{1.20}$$

and the related Euler–Lagrange equation, i.e., the minimal surface equation

$$-\mathrm{div}\left(\frac{Du}{\sqrt{1+|Du|^2}}\right)=0$$

intensively studied, for instance, in the seminal work of Bombieri, De Giorgi and Miranda [19], Ladyzhenskaya and Ural'tseva [94], Simon [121], Giusti [71], Giaquinta, Modica and Souček [68] and Trudinger [127]; see also Bildhauer and Fuchs [17], Beck and Schmidt [12, 13], and Gmeineder and Kristensen [72, 73] for recent advances and more general structures. The pointwise ellipticity ratio of (1.20) (that coincides with the nonlocal one being the functional autonomous) features a quadratic rate of blow-up, i.e., $\mathcal{R}(z)=1+|z|^2$. Analogous considerations hold for more general area-type functionals:

$$w \mapsto \int_\Omega (1+|Dw|^m)^{1/m}\,\mathrm{d}x, \quad m>1,$$

for which $\mathcal{R}(z)\approx 1+|z|^m$ for sufficiently large $|z|$; see [13, 68]. Variational integrals at linear growth are far from being the only strongly nonuniformly elliptic examples. Strong rates of nonuniformity are indeed typical of several functionals arising through various fields, such as fluid dynamics, materials science or nonlinear elasticity. Examples are given by slow-growing functionals [17, 60, 116] (nearly linear growth conditions),

$$w \mapsto \int_\Omega |Dw|\log(1+|Dw|)\,\mathrm{d}x, \tag{1.21}$$

satisfying $\mathcal{R}(z)\approx \log(1+|z|)$ for $|z|$ large, or convex polynomials,

$$\begin{cases} w \mapsto \displaystyle\int_\Omega |Dw|^p + \sum_{i=1}^n \alpha_i(x)|D_i w|^{q_i}\,\mathrm{d}x, \\ 1<p\le q_1\le\cdots\le q_n<\infty, \\ 0\le \alpha_i(\cdot)\in L^\infty_{\mathrm{loc}}(\Omega), \quad i\in\{1,\dots,n\}, \end{cases} \tag{1.22}$$

[8] Beware! In this case, the integrand in question has linear growth, so that the natural ambient function space is BV, and the problems must be formulated accordingly, as for instance in [67, 68]. Therefore, the discussion here is formal and only regards the ellipticity ratio. We refer to [12, 13] for a larger discussion on functionals with linear growth.

in which case we find

$$\mathcal{R}(x,z)+\bar{\mathcal{R}}(z,\mathrm{B})\lesssim_{p,q_i,\|\mathfrak{a}_i\|_{L^\infty(\mathrm{B})}}|z|^{q_n-p}+1$$

for $x\in\mathrm{B}$ (see [21,40,54,59,101]), or integrands at fast exponential growth, i.e.,

$$\begin{cases} w\mapsto\displaystyle\int_\Omega \exp(a(x)|Dw|^p)\,\mathrm{d}x,\\ 1<p<\infty,\quad 1\lesssim a\in L^\infty_{\mathrm{loc}}(\Omega), \end{cases}$$

where

$$\mathcal{R}(x,z)\approx|z|^p+1,\quad \bar{\mathcal{R}}(z,\mathrm{B})\approx(|z|^p+1)e^{(\mathrm{osc}_{\mathrm{B}}a)|z|^p}$$

for $|z|$ large; see [11,41,104]. A novel class of integrals, first considered in [43] and appearing as a combination of classical nearly linear growth ones (1.21) and weighted power terms, is that of log-double-phase problems, i.e.,

$$\begin{cases} w\mapsto\displaystyle\int_\Omega |Dw|\log(1+|Dw|)+\mathfrak{a}(x)|Dw|^q\,\mathrm{d}x,\\ 1<q<\infty,\quad 0\le\mathfrak{a}\in L^\infty_{\mathrm{loc}}(\Omega). \end{cases}\tag{1.23}$$

The related ellipticity ratios behave as hybrids of those of (1.17) and (1.21), i.e.,

$$\begin{cases} \mathcal{R}(x,z)\lesssim_q \log(1+|z|)+1,\\ \bar{\mathcal{R}}(z,\mathrm{B})\lesssim_q \log(1+|z|)+\|\mathfrak{a}\|_{L^\infty(\mathrm{B})}|z|^{q-1}+1. \end{cases}$$

The functional in (1.23), readable as a limiting configuration of (1.17) as $p\to1$ while approaching nearly linear growth conditions as in (1.21), leads to sharp outcomes on the validity of Schauder theory for anisotropic problems whose growth is arbitrarily close to linear [39,43,47] and offers deep insights on subtle singularity phenomena for more general nonautonomous area-type integrals.

2. Polynomial nonuniform ellipticity

A mostly common rate of nonuniformity for equations and functionals is the polynomial one, meaning that the pointwise ellipticity ratio behaves (at infinity) as a positive power of the gradient variable, i.e.,

$$\mathcal{R}(x,z)\approx_{|z|\ge1}|z|^\delta,\quad \delta>0.\tag{2.1}$$

See, for instance, the foundational regularity works of Serrin [119], Ladyzhenskaya and Ural'tseva [94], Ivanov [80], Simon [121], Trudinger [125–127] and Ural'tseva

and Urdaletova [131]. Note that many such results rely on the existence of strong solutions and on the availability of total derivatives, the latter requiring smoothness assumptions on coefficients that automatically rule out the classical Schauder setting. Marcellini's innovative variational approach [100, 101, 103] to nonuniformly elliptic equations (1.1) marked significant progress. The key condition in Marcellini's work is trapping the lowest and highest eigenvalues of $\partial_z A$ between two different powers of the gradient variable in a way to extend the p-Laplace-type behaviour, i.e.,

$$\begin{cases} |z|^{p-2}\mathbb{I}_{n\times n} \lesssim \partial_z A(x,z) \lesssim |z|^{q-2}\mathbb{I}_{n\times n} \ \text{ in the sense of matrices} \\ \text{for some } \ 1 < p \le q \ \text{ and all } \ |z| \ge 1, \ x \in \Omega, \end{cases} \tag{2.2}$$

the growth at infinity of the related ellipticity ratio can be controlled in terms of the difference $q - p$,

$$\mathcal{R}(x,z) \lesssim |z|^{q-p}, \quad |z| \ge 1. \tag{2.3}$$

This naturally relates to (2.1). The growth rate of $\mathcal{R}$ can be tamed by taking $q - p$ sufficiently small, so that the asymptotic in (2.3) suggests that only a moderate blow-up rate of $\mathcal{R}$ gives hope for regular solutions.[9] In fact, building on earlier works of Marcellini [101, 103] and Giaquinta [63], Min-Chun [107] proved that for $n \ge 6$, the function

$$u(x) := \sqrt{\frac{n-4}{24}}\frac{x_n^2}{\sqrt{\sum_{i=1}^{n-1} x_i^2}} - \frac{2}{n-2}\sqrt{\frac{n-4}{24}\sum_{i=1}^{n-1} x_i^2}, \tag{2.4}$$

which is unbounded on the line $(0,\dots,0,x_n)$, minimizes the convex polynomial

$$w \mapsto \int_\Omega |Dw|^2 + \frac{1}{2}|D_n w|^4 \,\mathrm{d}x. \tag{2.5}$$

The functional in (2.5), which is of the type in (1.22), highlights a peculiarity of nonuniformly elliptic problems: already for strongly convex, autonomous functionals

$$\begin{cases} \mathcal{F}_0(w,\Omega) := \displaystyle\int_\Omega F_0(Dw)\,\mathrm{d}x, \\ |z|^p \lesssim F_0(z) \lesssim 1 + |z|^q, \quad 1 < p \le q < \infty, \end{cases} \tag{2.6}$$

satisfying (2.2) with $\partial_z A \equiv \partial_{zz} F_0$, minimizers may be unbounded when a constraint of the type

$$\frac{q}{p} < 1 + \mathrm{o}_n, \quad \lim_n \mathrm{o}_n = 0 \tag{2.7}$$

[9] Conditions (2.2) and (2.3) can be recast in the variational setting by replacing A with $\partial_z F$, and $\partial_z A$ with $\partial_{zz} F$.

is violated. This is an element of sharp contrast with the uniformly elliptic setting. In fact, uniform ellipticity allows the formation of singularities only in vectorial problems [48, 105, 110, 111, 123], while Marcellini, Giaquinta and Min-Chun's counterexamples are scalar. In particular, the regularity of minima of (2.6) does not hold for arbitrary choices of p and q, that actually must satisfy a precise quantitative constraint involving exponents (p, q) and the ambient dimension n. The optimal bound in (2.7) is known only at the 0-order scale: Marcellini [103] exhibited an unbounded minimizer of an integral of type (2.6) provided that

$$q > \frac{p(n-1)}{n-1-p} \quad \text{and} \quad 1 < p < n-1 \tag{2.8}$$

holds, which is consistent with the construction in (2.4) and (2.5): dimensional condition $n \geq 6$ is indeed equivalent to $4 > 2(n-1)/(n-3)$. Local boundedness of minima in violation of (2.8), i.e.,

$$\begin{cases} 1 < q \leq \dfrac{p(n-1)}{n-1-p} & \text{if } 1 < p < n-1, \\ 1 < p \leq q < \infty & \text{if } p \geq n-1, \end{cases}$$

was obtained thirty years later by Hirsch and Schäffner [78] via fine optimization techniques, also relying on methods by Bella and Schäffner [14, 15]. Nonetheless, the analogous issue at the gradient level, in full generality, is still open. Marcellini [103] proved gradient higher differentiability and local Lipschitz continuity for minima of (2.6) with (2.2) in force[10] if

$$\frac{q}{p} < 1 + \frac{2}{n},$$

a constraint later on updated to

$$\frac{q}{p} < 1 + \frac{2}{n-1}$$

by Bella and Schäffner [14, 16, 112]. In case F_0 is an elliptic convex polynomial, the above restrictions were improved to

$$\begin{cases} 2 \leq p \leq q \leq \dfrac{p(n-1)}{n-3} & \text{if } n \geq 4, \\ 2 \leq p \leq q < \infty & \text{if } n \in \{2, 3\} \end{cases} \tag{2.9}$$

by Koch, Kristensen and the author [40]. When $p = 2$, (2.9) matches the threshold determined in [63, 101, 107] and yields sharp gradient regularity; the general case

[10] See [103] for precise assumptions, or see (4.1) below.

$p \neq 2$ is still far from being understood. Aside from the optimal bound under which gradient regularity can be achieved, the theory for autonomous, nonuniformly elliptic functionals is at a very advanced stage. It is therefore natural to wonder what happens when external ingredients come into play, specifically if the Schauder theory still holds for nonuniformly elliptic problems. For a further discussion of nonuniformly elliptic problems and techniques, we refer to [38, 44].

3. Uniformly elliptic Schauder theory

Consider the linear equation with variable coefficients (1.3), where the matrix $\mathcal{A}$ is symmetric (for simplicity), bounded and elliptic, i.e., $\mathbb{I}_{n\times n} \leq \mathcal{A} \in L^{\infty}$. Schauder theory claims

$$\mathcal{A} \in C^{0,\alpha}_{\mathrm{loc}}(\Omega;\mathbb{R}^{n\times n}) \Longrightarrow Du \in C^{0,\alpha}_{\mathrm{loc}}(\Omega;\mathbb{R}^{n}), \quad \alpha \in (0,1). \tag{3.1}$$

That is, the gradient of the solution inherits the same degree of regularity of coefficients; this is optimal (and fails in the borderline case $\alpha = 1$). This kind of results was pioneered by Hopf [79], Giraud [70], Caccioppoli [28] and Schauder [113, 114] in various forms, and today is known as Schauder estimates, with parabolic, fully nonlinear and vectorial analogs [2, 3, 30]. Schauder theory is a basic tool in elliptic and parabolic PDEs and in the calculus of variations, influencing diverse areas of analysis such as nonlinear diffusion, potential theory, field theory or differential geometry. The classical outcome is twofold:

- regularity estimates of type (3.1), local or global, establishing the principle that solutions to nonautonomous equations are as regular as the ingredients allow;
- boundedness or compactness of the inverse of certain elliptic operators;

see [84]. The latter allows proving the existence of solutions via fixed point theorems, that was a standard argument before the introduction of energy methods and Sobolev spaces. Here, we shall be interested in the former. The results in [28, 70, 79, 113, 114] rely on potential theoretic techniques. However, more direct approaches are available today avoiding the use of representation formulas, as we shall see in a few lines directly in the nonlinear case. For instance, Campanato [31] used suitable function spaces, Trudinger [128] used convolution techniques and Leon Simon [122] used blow-up methods. The nonlinear theory dates back to the 1980s, with the work of Giaquinta and Giusti [64–66], DiBenedetto [50] and Manfredi [98]. The underlying principle is essentially to use the fact that Hölder continuity of coefficients allows us to prove the closeness of the original solution u to that of "frozen" problems of the

type

$$\begin{cases} -\mathrm{div}(\mathcal{A}(x_0)Dv) = 0 & \text{in } B_r(x_0), \\ v = u & \text{on } \partial B_r(x_0), \end{cases}$$

on small balls $B_r(x_0) \Subset \Omega$. Let us see an application of this principle directly in the nonlinear case

$$-\mathrm{div}((\mathcal{A}(x)Du \cdot Du)^{(p-2)/2}\mathcal{A}(x)Du) = 0, \tag{3.2}$$

where we consider an energy solution $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$, $p > 1$; note that (3.2) reduces to (1.3) for $p = 2$. For a given ball $B_r(x_0) \Subset \Omega$, the related nonlinear lifting (balayage) can be defined as the solution to the Dirichlet problem

$$\begin{cases} -\mathrm{div}((\mathcal{A}(x_0)Dv \cdot Dv)^{(p-2)/2}\mathcal{A}(x_0)Dv) = 0 & \text{in } B_r(x_0), \\ v = u & \text{on } \partial B_r(x_0). \end{cases}$$

Thanks to Ural'tseva–Uhlenbeck theory [129, 130], Dv is locally Hölder continuous in $B_r(x_0)$, and

$$\begin{cases} \|Dv\|^p_{L^\infty(B_{r/2}(x_0))} \lesssim_{\|\mathcal{A}\|_{L^\infty(B_r(x_0))}} \fint_{B_r(x_0)} |Dv|^p \,\mathrm{d}x, \\ \fint_{B_\sigma(x_0)} |Dv - (Dv)_{B_\sigma(x_0)}|^p \,\mathrm{d}x \\ \qquad \lesssim_{n,p,\|\mathcal{A}\|_{L^\infty(B_r(x_0))}} \left(\frac{\sigma}{\varrho}\right)^{\beta_0 p} \fint_{B_\varrho(x_0)} |Dv - (Dv)_{B_\varrho(x_0)}|^p \,\mathrm{d}x \end{cases} \tag{3.3}$$

hold on all concentric balls $B_\sigma(x_0) \subset B_\varrho(x_0) \Subset B_r(x_0)$ (see, for instance, [98], [91, Theorem 3.2]). Here, $\beta_0 \equiv \beta_0(n, p) \in (0, 1)$ is a universal Hölder continuity exponent, which can be taken to be $\beta_0 = 1$[11] when $p = 2$. Moreover, the α-Hölder continuity of the matrix $\mathcal{A}$ grants the validity of the comparison estimate

$$\fint_{B_r(x_0)} |Du - Dv|^p \,\mathrm{d}x \lesssim_{n,p,[\mathcal{A}]_{0,\alpha;B_r(x_0)}} r^{\alpha \min\{2,p\}} \fint_{B_r(x_0)} |Du|^p \,\mathrm{d}x. \tag{3.4}$$

The combination of (3.3) and (3.4) allows transferring smoothness from v to u

$$\fint_{B_r(x_0)} |Du - (Du)_{B_r}|^p \,\mathrm{d}x \lesssim r^{p\beta}, \quad \beta := \min\{\beta_0, \alpha \min\{2/p, 1\}\} \tag{3.5}$$

[11] In the linear case, this is in fact a classical result of Campanato [31], and the second line of (3.3) holds with $\beta_0 = 1$. When $p \neq 2$, the second line of estimate (3.3) is a rigid form of Ural'tseva–Uhlenbeck theory, and one cannot take $\beta_0 = 1$ in general, by counterexamples.

for all sufficiently small ϱ, that implies that

$$Du \in C^{0,\beta}_{\mathrm{loc}}(\Omega;\mathbb{R}^n) \tag{3.6}$$

via a classical integral characterization due to Campanato and Meyers. Note that in the linear case $p = 2$, we can take $\beta_0 = 1$ and therefore $\beta = \alpha$. This allows us to recover (3.1). The very same circle of ideas applies to equations as in (1.1) satisfying the same growth/ellipticity and oscillation features as the models in (1.16) and (3.2), and to nonautonomous problems governed by general uniformly elliptic operators [97].

Remark 3.1. We identify three aspects of the uniformly elliptic Schauder theory developed between the early 1930s and the end of the 1980s.

- Proof by perturbation. That is, use of the Hölder continuity of coefficients to be quantitatively close to a uniformly elliptic problem with constant (frozen) coefficients (3.4), whose solutions (liftings) enjoy homogeneous excess decay estimates (3.3), and then exploit homogeneity to recover similar integral decay estimates for solutions to the original nonautonomous problem (3.5).
- Smooth data never obstruct the regularity of energy solutions (3.1).
- Lipschitz bounds via gradient Hölder continuity. In the classical theory of uniformly elliptic equations, the standard way to achieve gradient L^∞-bounds in nonautonomous, uniformly elliptic problems with Hölder coefficients as (3.2) is to prove first gradient Hölder continuity for some exponent, and then retain local boundedness.

The last point of Remark 3.1 is of particular interest here. Indeed, when the (nonlinear) equation in question is nondegenerate or nonsingular, the Lipschitz information is employed afterwards to, in a sense, linearize the equation and establish that the gradient of solutions shares the same Hölder exponent as the coefficient. As an example, consider the nondegenerate p-Laplace equation

$$\begin{cases} -\mathrm{div}(a(x)(|Du|^2+1)^{(p-2)/2}Du) = 0 \quad \text{in } \Omega, \\ 1 < p < \infty, \quad 1 \lesssim a \in C^{0,\alpha}_{\mathrm{loc}}(\Omega), \quad \alpha \in (0,1). \end{cases} \tag{3.7}$$

In this case, the technique outlined in this section leads to (3.6). Further regularity techniques, explained in [45, Section 5], eventually lead to

$$Du \in C^{0,\beta}_{\mathrm{loc}}(\Omega;\mathbb{R}^n) \Longrightarrow Du \in L^{\infty}_{\mathrm{loc}}(\Omega;\mathbb{R}^n) \Longrightarrow Du \in C^{0,\alpha}_{\mathrm{loc}}(\Omega;\mathbb{R}^n) \tag{3.8}$$

and therefore to the same sharp result of the linear case notwithstanding[12] $p \neq 2$. Note that the bootstrap in (3.8) only works in the nondegenerate/nonsingular case (3.7). We also stress the fact that the second implication in (3.8) works also when the equations considered are nonuniformly elliptic (but still nondegenerate), as, when the gradient is known to be bounded, equations in a sense become uniformly elliptic again. As remarked above, the first implication in (3.8) is instead essentially the only way to get gradient L^∞-bounds in the presence of Hölder coefficients in the classical theory. We will turn back on these points in Section 4.4, where in fact this way is reversed.

4. Nonuniformly elliptic Schauder theory

Here, the focus will be on nonautonomous variational integrals of the type (1.14) verifying

$$\begin{cases} \overline{z} \mapsto F(x,\overline{z}) \in C^2(\mathbb{R}^n \setminus \{0\}) \cap C^1(\mathbb{R}^n), \\ |z|^p \lesssim F(x,z) \lesssim 1 + |z|^q, \\ \langle \partial_{zz} F(x,z)\xi, \xi \rangle \gtrsim (\mu^2 + |z|^2)^{\frac{p-2}{2}} |\xi|^2, \\ |\partial_{zz} F(x,z)| \lesssim (\mu^2 + |z|^2)^{\frac{p-2}{2}} + (\mu^2 + |z|^2)^{\frac{q-2}{2}}, \\ |\partial_z F(x_1,z) - \partial_z F(x_2,z)| \lesssim |x_1 - x_2|^\alpha (1 + |z|^{q-1}) \end{cases} \tag{4.1}$$

for all $z \in \mathbb{R}^n \setminus \{0\}$, $\xi \in \mathbb{R}^n$, $x, x_1, x_2 \in \Omega$ and some[13] $\mu \in [0,1]$, $\alpha \in (0,1]$; as usual, $1 < p \leq q$. Note that under such assumptions, Definition 1.1 implies that any minimizer automatically belongs to $W^{1,p}_{\mathrm{loc}}(\Omega)$. Moreover, (4.1) implies the following growth bound on the (pointwise) ellipticity ratio:

$$\mathcal{R}(x,z) \lesssim |z|^{q-p} + 1.$$

Note that functionals (1.17), (1.19), (1.22) and (4.18) are all specific instances of (4.1) for suitable choices of p, q, μ and Hölder continuous dependence on x.

[12] In fact, already in the degenerate case (3.2), the standard derivation of (3.6) goes as follows: in a first step, one combines the first line of (3.3) and (3.4) to obtain that Du belongs to a suitable Morrey space. Then one uses this to combine this time the second line of (3.3) and (3.4) to get that Du is locally Hölder continuous with some exponent [98]. In particular, this implies that Du is locally bounded. Finally, one uses this last information to recombine, in a different way, the second line of (3.3) and (3.4) to get (3.6). The derivation of (3.8) in the nondegenerate case (3.7) is more delicate but follows similar steps [45, Section 5].

[13] The parameter μ serves to distinguish the degenerate case $\mu = 0$ from the nondegenerate one $\mu > 0$. We refer to [45] for comments on the assumptions (4.1).

For nonautonomous integrals as (1.14) subject to (4.1) and $p < q$, the regularity of minima unavoidably finds an obstruction in the possible occurrence of Lavrentiev phenomenon, that is, for instance,

$$\inf_{w \in u_0 + W_0^{1,p}(B)} \mathcal{F}(w, B) < \inf_{w \in u_0 + W_0^{1,q}(B)} \mathcal{F}(w, B) \tag{4.2}$$

with $u_0 \in W^{1,\infty}(B)$ for some ball[14] $B \Subset \Omega$. In other words, as shown in [54, Section 3], in connection to (4.2), it is possible that minima u fail even the basic regularity upgrade $u \in W^{1,q}_{\mathrm{loc}}(\Omega)$. The occurrence of phenomena like (4.2) leads to consider a larger class of functionals based on lower semicontinuous envelopes built along sequences of more regular functions. The Lebesgue–Serrin–Marcellini (LSM) relaxation (extension) $\overline{\mathcal{F}}$ of functional $\mathcal{F}$ in (1.14) is defined as

$$\overline{\mathcal{F}}(w, U) := \inf_{\{w_i\} \subset W^{1,q}(U)} \left\{ \liminf_{i \to \infty} \mathcal{F}(w_i, U) : w_i \rightharpoonup w \text{ in } W^{1,p}(U) \right\} \tag{4.3}$$

whenever $U \subset \Omega$ is an open subset. Quantifying the distance between $\mathcal{F}$ and $\overline{\mathcal{F}}$ leads to define the Lavrentiev gap functional $\mathcal{L}_{\mathcal{F}}$ [25] as

$$\mathcal{L}_{\mathcal{F}}(w, B) := \begin{cases} \overline{\mathcal{F}}(w, B) - \mathcal{F}(w, B) & \text{if } \mathcal{F}(w, B) < \infty, \\ 0 & \text{if } \mathcal{F}(w, B) = \infty \end{cases}$$

for every ball $B \Subset \Omega$, which in a sense provides a quantitative measure of the occurrence of phenomena like (4.2). Note that

- $\mathcal{F}(\cdot, U) \le \overline{\mathcal{F}}(\cdot, U)$. Indeed, the convexity of $z \mapsto F(\cdot, z)$ implied by the third line of (4.1) grants the $W^{1,p}$-weak lower semicontinuity of $\mathcal{F}$. It follows that $\mathcal{L}_{\mathcal{F}}$ is non-negative.
- $\mathcal{F}(\cdot, B) = \overline{\mathcal{F}}(\cdot, B)$ for every ball $B \Subset \Omega$, when $\mathcal{F}$ is autonomous, i.e., the integrand F does not depend on x. This follows from a simple convolution argument [54, Lemma 12].
- $\mathcal{F}(\cdot, U) = \overline{\mathcal{F}}(\cdot, U)$ on $W^{1,q}(U)$. Therefore, $\overline{\mathcal{F}}$ can be thought of as an extension of $\mathcal{F}$, when this last functional is initially defined on the smaller space $W^{1,q}(U)$, to the whole $W^{1,p}(U)$. This explains the use of the word extension[15] when referring to $\mathcal{F}$.

[14] By standard density arguments, (4.2) does not occur when $p = q$.

[15] The LSM relaxation was first introduced by Marcellini in [100, 102] to describe cavitation phenomena in nonlinear elasticity, and it is connected to similar, earlier constructions of Lebesgue [96] and Serrin [117]. It was eventually intensively developed in the literature [20, 37, 39, 42, 46, 54, 56, 58, 73, 101, 115]. Variants of the definition in (4.3) are possible

The LSM relaxation naturally comes along with a notion of minimality.

Definition 4.1 (Relaxed minima). A function $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ is a *(local) minimizer* of $\overline{\mathcal{F}}$ in (4.3) if, for every ball $\mathrm{B} \Subset \Omega$, $\overline{\mathcal{F}}(u,\mathrm{B}) < \infty$ and $\overline{\mathcal{F}}(u,\mathrm{B}) \le \overline{\mathcal{F}}(w,\mathrm{B})$ hold for every $w \in u + W^{1,p}_0(\mathrm{B})$.

The approach used in [54], eventually widely adopted afterwards in the literature, prescribes to prove regularity of (local) minimizers of $\overline{\mathcal{F}}$[16] rather than those of $\mathcal{F}$. But then

- When
$$\frac{q}{p} < 1 + \frac{\alpha}{n}, \tag{4.4}$$
minima of $\overline{\mathcal{F}}$ also (locally) minimize $\mathcal{F}$. Therefore, $\overline{\mathcal{F}}$ can be used to detect regular minima of $\mathcal{F}$. In a sense, drawing a parallel with the classical theory of elliptic equations, minima of $\mathcal{F}$ detected by $\overline{\mathcal{F}}$ correspond to distributional solutions to (1.1) that are also energy solutions. See Theorem 4.3.
- When the Lavrentiev gap functional vanishes on a minimizer u of the original functional $\mathcal{F}$, that is, when $\mathcal{L}_{\mathcal{F}}(u,\mathrm{B}) = 0$ holds for all balls $\mathrm{B} \Subset \Omega$, then
$$\overline{\mathcal{F}}(u,\mathrm{B}) = \mathcal{F}(u,\mathrm{B}) + \mathcal{L}_{\mathcal{F}}(u,\mathrm{B}) = \mathcal{F}(u,\mathrm{B}) \le \mathcal{F}(w,\mathrm{B}) \le \overline{\mathcal{F}}(w,\mathrm{B}) \tag{4.5}$$
holds whenever $w \in u + W^{1,p}_0(\mathrm{B})$. Therefore, u is also a minimizer of the $\overline{\mathcal{F}}$, and its regularity properties can be derived from the regularity results available for minima of $\overline{\mathcal{F}}$. See Corollary 4.4. The Lavrentiev gap actually vanishes in a variety of situations, as for instance detailed in [54, Section 5]. See Corollary 4.6.

Within this framework, specifically thought for nonautonomous functionals, we shall describe key aspects in a rich regularity theory initiated by the seminal work of Esposito, Leonetti and Mingione [54]. This theory explores the interplay between coefficient regularity, the growth behavior of the ellipticity ratio and the presence of the Lavrentiev phenomenon, ultimately culminating in the recent establishment of Schauder estimates for nonuniformly elliptic problems by Mingione and the author in [42,45].

---

and, often, useful. For instance, the following local version is largely studied in the literature:

$$\overline{\mathcal{F}}_{\mathrm{loc}}(w,U) := \inf_{\{w_i\} \subset W^{1,q}_{\mathrm{loc}}(U)} \left\{ \liminf_{i \to \infty} \mathcal{F}(w_i,U) : w_i \rightharpoonup w \text{ in } W^{1,p}(U) \right\}.$$

For both $\overline{\mathcal{F}}$ and $\overline{\mathcal{F}}_{\mathrm{loc}}$ very interesting measure representation results are available [56]. We refer to [115] for further variants and related results. Most of the results for $\overline{\mathcal{F}}$ presented here extend to $\overline{\mathcal{F}}_{\mathrm{loc}}$.

[16] This, in a sense, automatically excludes the occurrence of phenomena like (4.2) and allows us to concentrate on a priori estimates for more regular solutions.

4.1. Lavrentiev phenomenon

A classic in the calculus of variations, the Lavrentiev gap phenomenon occurs whenever variational integrals admit different infima if considered on the full class of admissible functions or on a smaller, yet dense subclass of more regular functions (cf. (4.2)). This apparent anomaly was first observed by Lavrentiev [95]. Later on, Manià [99] considered the functional

$$\mathcal{I}_1(w) := \int_0^1 (x - w^3)^2 |w'|^6 \, \mathrm{d}x,$$

when defined on the two different function spaces

$$\mathcal{W}^\infty := \{w \in W^{1,\infty}(0,1) : w(0) = 0,\ w(1) = 1\},$$
$$\mathcal{W}^1 := \{w \in W^{1,1}(0,1) : w(0) = 0,\ w(1) = 1\}.$$

Obviously, $u(x) = x^{1/3}$ minimizes $\mathcal{I}_1$ on $\mathcal{W}^1$, while explicit computations yield

$$\inf_{w \in \mathcal{W}^\infty} \mathcal{I}_1(w) > 0 = \inf_{w \in \mathcal{W}^1} \mathcal{I}_1(w) = \mathcal{I}_1(u)$$

(cf. [36, Lemma 4.43]). Manià's example cannot be detected via standard finite element methods (i.e., taking piecewise affine functions that are Lipschitz continuous). One could argue that $\mathcal{I}_1$ lacks coercivity. This issue was eventually fixed by Ball and Mizel [9] by looking at functional

$$w \mapsto \int_0^1 \varepsilon |w'|^2 + (x^4 - w^6)^2 |w'|^m \, \mathrm{d}x,$$

with $m \geq 27$, and $\varepsilon > 0$, that nonetheless presents the same pathological behavior as $\mathcal{I}_1$ in (4.2). See also [77] for the gap problem in stochastic control theory and [8] for numerical approximation schemes that spot lower energy singular minimizers despite the fact that the cost of any sequence in the admissible class of Lipschitz continuous functions is bounded away from the true minimum value.[17]

4.2. Approximation in energy

The (absence of) Lavrentiev phenomenon admits other, more application-friendly and general formulations beyond the one in (4.2):

- coincidence between the functional $\mathcal{F}$ and the LSM relaxation $\overline{\mathcal{F}}$;
- density of smooth maps in the related Lagrangian space[18] [82, Chapter 14].

[17] Note that in this case, when a minimizer over the smoother admissible class exists, the usual approximation schemes converge to this suboptimal solution.

[18] That is, the class of all functions $w \in W^{1,1}(B)$ with finite energy $\mathcal{F}(w, B) < \infty$.

Indeed, notice that whenever $w \in W^{1,p}(B)$ is such that $\overline{\mathcal{F}}(w, B) < \infty$, then there exists a sequence $w_i \rightharpoonup w$ in $W^{1,p}(B)$ such that $\mathcal{F}(w_i, B) \to \overline{\mathcal{F}}(w, B)$. It follows that $\mathcal{F}(w, B) = \overline{\mathcal{F}}(w, B)$ if and only if there exists a sequence

$$w_i \rightharpoonup w \quad \text{weakly in } W^{1,p}(B) \qquad \text{and} \qquad \mathcal{F}(w_i, B) \to \mathcal{F}(w, B). \tag{4.6}$$

This is called approximation in energy. If the integrand F is nonautonomous, the occurrence of the Lavrentiev phenomenon connects to the way space-depending coefficients and gradient variable interact; in particular, (4.6) may be violated even in simple cases. A paradigm of such irregularity phenomena is the celebrated Zhikov's checkerboard example [82,132]; consider the functional in (1.19) with $\Omega = \mathcal{B}_1 \subset \mathbb{R}^2$ and, with $x \equiv (x_1, x_2) \in \mathbb{R}^2$, set

$$\mathfrak{p}(x) := \begin{cases} p & \text{if } x_1 x_2 < 0, \\ q & \text{if } x_1 x_2 \geq 0, \end{cases} \qquad 1 < p < 2 < q. \tag{4.7}$$

In polar coordinates $(x_1, x_2) = (\varrho \cos\theta, \varrho \sin\theta)$, define

$$w_*(x) := \begin{cases} 1 & \text{if } x_1, x_2 > 0, \\ \sin\theta & \text{if } x_2 > 0 > x_1, \\ 0 & \text{if } x_1, x_2 < 0, \\ \cos\theta & \text{if } x_1 > 0 > x_2 \end{cases} \tag{4.8}$$

(see Figure 1). The Lagrangian space generated by the functional in (1.19) is given by

$$W^{1,\mathfrak{p}(\cdot)}(\mathcal{B}_1) := \{w \in W^{1,1}(\mathcal{B}_1) : |Dw|^{\mathfrak{p}(\cdot)} \in L^1(\mathcal{B}_1)\}.$$

The integral in (1.19) is clearly well defined and finite on $W^{1,\mathfrak{p}(\cdot)}(\mathcal{B}_1)$. Zhikov proved that despite $w_* \in W^{1,\mathfrak{p}(\cdot)}(\mathcal{B}_1)$, energy approximation by smooth maps in the sense

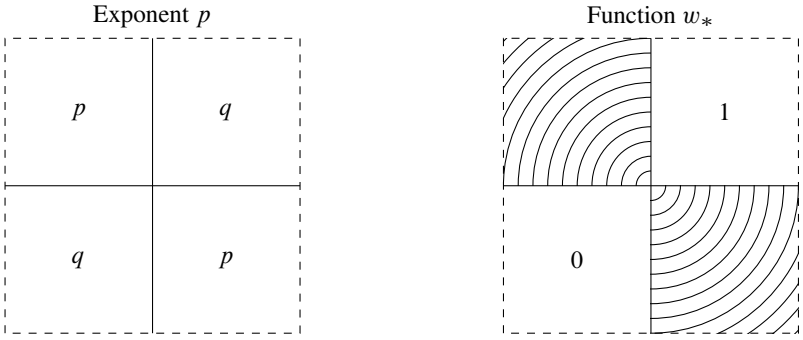


Figure 1. Zhikov's checkerboard. Source [6].

of (4.6) fails. This discussion highlights an interplay between gradient variable and space-depending coefficient, that is stronger than the one observed in the case of the p-Laplace energy: indeed, now coefficients no longer have a perturbative role and might dramatically affect the growth/ellipticity features of the integral.[19] We will deepen these facts in Section 4.3; here, we would simply note the huge amount of information obtained from the analysis of the Lavrentiev phenomenon: functional analytic properties of Lagrangian spaces, relaxation of functionals and, what matters most here, the possibility of constructing SOLAs,[20] i.e., building up energy approximations of general nonautonomous functionals, a crucial aspect of the analysis described in Section 4.4 (see comments after Theorem 4.5).

4.3. Soft nonuniform ellipticity

Before focusing on classical, strong nonuniform ellipticity in the sense of (1.13), let us consider the softly nonuniformly elliptic case of the double-phase functional $\mathcal{D}$ in (1.17), where from now on we assume that

$$0 \leq a \in C^{0,\alpha}_{\mathrm{loc}}(\Omega), \quad \alpha \in (0,1]. \tag{4.9}$$

Behind a rather simple structure, energy $\mathcal{D}$ hides several interesting (ir)regularity aspects. In fact, building on previous two-dimensional constructions of Zhikov [133, 134] concerning the Lavrentiev phenomenon, it was discovered by Esposito, Leonetti and Mingione [54] that, whenever exponents (p,q) and α verify

$$p < n < n+\alpha < q \quad \left(\Longrightarrow \frac{q}{p} > 1 + \frac{\alpha}{n} \right), \tag{4.10}$$

it is possible to construct an integrand as in (1.17) and (4.9), and a Lipschitz-regular boundary datum $u_0 \in W^{1,\infty}(\mathcal{B}_1)$, such that the solution to the corresponding Dirichlet problem exhibits a one-point singularity that forbids its membership to $W^{1,q}_{\mathrm{loc}}(\mathcal{B}_1)$, i.e.,

$$u \mapsto \min_{w \in u_0 + W^{1,p}_0(\mathcal{B}_1)} \mathcal{D}(w, \mathcal{B}_1), \quad u \notin W^{1,q}_{\mathrm{loc}}(\mathcal{B}_1). \tag{4.11}$$

An even more striking example was shown by Fonseca, Malý and Mingione [57], who, under the same conditions of [54] (but obviously with a different, much more

[19] In fact, this already happens in the examples (1.17), (1.19) and (1.23).

[20] Solution obtained by limiting approximations. These are solutions obtained as limits, in suitable sense, of solutions to more regular problems, like for instance minima of approximating uniformly elliptic functionals.

complex coefficient $\mathfrak{a}$), exhibited a solution to (4.11) whose set of its essential discontinuities forms a Cantor-type fractal of almost maximal Hausdorff dimension $n-p$. This result was eventually sharpened[21] by Balci, Diening and Surnachev [5,6], who constructed bounded minima of $\mathcal{D}$ whose set of essential discontinuities has Hausdorff dimension equal to $n-p$, and therefore do not belong to any better Sobolev space than $W^{1,p}(B_1)$. All in all, there are minima of convex, scalar, uniformly elliptic (in the classical sense) variational integrals that are as bad as any other competitor! This oddity confirms how delicate regularity issues are when dealing with nonuniform ellipticity, even if only a soft form of it is present. The key idea behind these examples is that the coefficient $\mathfrak{a}$ vanishes where certain low-integrable competitor maps w_* (no better than $W^{1,p}$) are nonconstant, i.e., $\mathfrak{a}(x)|Dw_*(x)|^q = 0$. These malicious competitors in (4.11) decrease the total energy below the energy of any possible $W^{1,q}$-competitor once highly varying traces ($m > 0$ and $-m$) on the upper and lower parts of $\partial\mathcal{B}_1$ are chosen; see Figure 2.

This finally creates a nonvanishing Lavrentiev gap, and the situation is similar to the one described in Section 4.2; see (4.7) and (4.8). Choosing a coefficient $\mathfrak{a}$ with a fractal-type geometry allows us to distribute the corresponding singularities on a Cantor fractal with maximal Hausdorff dimension. In particular, the results in [5,57] highlight that, in the presence of very mild types of nonuniformity as the one in the first line of (1.18), smooth coefficients do no longer guarantee regular solutions

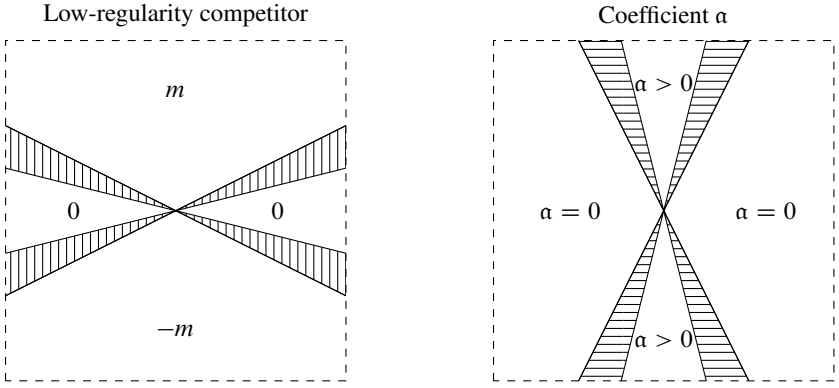


Figure 2. Low-regular competitor vs. coefficient $\mathfrak{a}$. Figure 2 is a modification of the one in [6].

[21] More precisely, in [57] for every $\varepsilon > 0$, the authors construct an example where the Hausdorff dimension of the set of essential discontinuity points of the minimizer is larger than $n-p-\varepsilon$, still keeping (4.10) and with $q-p < \alpha+\varepsilon$. Note that any minimizer of $\mathcal{D}$ belongs to $W^{1,p}_{\mathrm{loc}}(\Omega)$, and that the Hausdorff dimension of the set of essential discontinuities of any $W^{1,p}_{\mathrm{loc}}$-function cannot exceed $n-p$.

(cf. (3.1) and Remark 3.1). All such examples rely on the fact that under the considered condition the Lavrentiev gap of $\mathcal{D}$ does not vanish $\mathcal{L}_{\mathcal{D}} \neq 0$. Condition (4.10) violates (4.4), under which maximal gradient regularity is consequently expected. The first general regularity result is again in [54], where it is proved that if u minimizes $\mathcal{D}$, then[22]

$$\begin{cases} \dfrac{q}{p} \leq 1 + \dfrac{\alpha}{n} & \Longrightarrow \mathcal{L}_{\mathcal{D}}(\cdot, \mathrm{B}) = 0 \quad \forall\, \mathrm{B} \Subset \Omega, \\ & \Longrightarrow u \in W^{1,q+\delta}_{\mathrm{loc}}(\Omega), \quad 0 < \delta \equiv \delta(n, p, q, \alpha). \end{cases} \tag{4.12}$$

Then in [10, 34] it is proved that

$$\frac{q}{p} \leq 1 + \frac{\alpha}{n} \Longrightarrow u \in C^{1,\beta}_{\mathrm{loc}}(\Omega), \quad 0 < \beta \equiv \beta(n, p, q, \alpha) < 1. \tag{4.13}$$

This is the best amount of regularity achievable in degenerate problems [130], already when $\mathfrak{a} \equiv 0$. Let us highlight that the strategy in [10, 34] is still perturbative. The soft nonuniformity of $\mathcal{D}$ in (1.18) still allows us to use certain perturbation methods of the type sketched in Section 3, or, at least, "half" of them. More precisely, with reference to the "Freeze & Compare" scheme in Section 3, the main difficulty here lies in deriving homogeneous comparison estimates (cf. (3.4)), the validity of homogeneous reference estimates for suitable liftings being guaranteed by the pointwise uniform ellipticity in the first line of (1.18). Indeed, the proof of (4.13) conceptually goes as follows.

- Using pointwise uniform ellipticity in the first line of (1.18). The "frozen" integrand with constant coefficient $\mathfrak{a}_0$

$$P_0(z) := |z|^p + \mathfrak{a}_0 |z|^q, \quad \mathfrak{a}_0 \in [0, \infty)$$

  is uniformly elliptic, and its ellipticity ratio is bounded independently of $\mathfrak{a}_0$ by the first line of (1.18). With $B_r(x_0) \Subset \Omega$ being an arbitrary ball, this implies that the solution $v \in u + W^{1,p}_0(B_r(x_0))$ of the Dirichlet problem,

$$v \mapsto \min_{u + W^{1,p}_0(B_r(x_0))} \int_{B_r(x_0)} P_0(Dw)\, \mathrm{d}x, \tag{4.14}$$

[22] The additional equality case $q/p = 1 + \alpha/n$ is a typical effect of the pointwise uniform ellipticity of the double-phase integrand, allowing us to get some extra small initial gradient integrability.

features intrinsically homogeneous excess decay estimates of type (3.3), i.e.,

$$\begin{cases} \|P_0(Dv)\|_{L^\infty(B_{r/2}(x_0))} \lesssim_{n,p,q} \displaystyle\fint_{B_r} P_0(Dv)\,\mathrm{d}x, \\ \displaystyle\fint_{B_\sigma(x_0)} P_0(Dv-(Dv)_{B_\sigma(x_0)})\,\mathrm{d}x \\ \qquad\qquad \lesssim_{n,p,q} \left(\frac{\sigma}{\varrho}\right)^{\beta_1} \displaystyle\fint_{B_\varrho(x_0)} P_0(Dv-(Dv)_{B_\varrho(x_0)})\,\mathrm{d}x \end{cases} \tag{4.15}$$

for all concentric balls $B_\sigma(x_0) \subset B_\varrho(x_0) \Subset B_r(x_0)$ and with $\beta_1 \equiv \beta_1(n,p,q) \in (0,1)$ being independent of $\mathtt{a}_0$; compare with (3.3), to which (4.15) reduces when $\mathtt{a}_0 = 0$. See [52,97] for a proof. Estimate (4.15) holds for any choice of $1 < p \le q$; no upper bound on q/p is needed.

- Rebalancing the soft nonuniform ellipticity in the second line of (1.18). This is the point where the optimal bound $q/p \le 1 + \alpha/n$ crucially enters the game. With M large enough, to be quantitatively chosen, the p-phase is said to occur when[23]

$$\inf_{B_r(x_0)} \mathtt{a} \le M[\mathtt{a}]_{0,\alpha;B_r(x_0)} r^\alpha. \tag{4.16}$$

  The q-phase instead occurs otherwise. In the p-phase, the original minimizer is compared to the solution of (4.14) with $\mathtt{a}_0 = 0$. In the (p,q)-phase, instead, one chooses $\mathtt{a}_0 = \inf_{B_r(x_0)} \mathtt{a}$. Via a delicate comparison scheme based either on reverse Hölder inequalities [34] or on quantitative harmonic-type approximation lemmas [10], one arrives at

$$\fint_{B_{r/2}(x_0)} P_0(Du-Dv)\,\mathrm{d}x \lesssim r^{\beta_2} \fint_{B_r(x_0)} P_0(Du)\,\mathrm{d}x \tag{4.17}$$

  for some $\beta_2 \in (0,1)$.

- Estimates (4.15) and (4.17) are homogeneous estimates when considered with respect to the intrinsic quantity $P_0(Du)$ and, as such, can now be matched and iterated in the setting of a delicate exit time argument based on the occurrence of (4.16) at various scales. This eventually yields gradient Hölder continuity for u.

The viewpoint in [10, 34] has been further developed; see [4, 7, 24, 26, 27, 55, 83] and references therein for a nonexhaustive list of relevant contributions. In particular, Hästö and Ok [74–76] extended the outcome of [10] to a very large family of

[23] As usual

$$[\mathtt{a}]_{0,\alpha;B_r(x_0)} = \sup_{x,y\in B_r(x_0);x\neq y} \frac{|\mathtt{a}(x)-\mathtt{a}(y)|}{|x-y|^\alpha}.$$

softly nonuniformly elliptic problems, again, by means of a perturbative approach that strongly relies on the pointwise uniform ellipticity of the class of integrands considered. The whole body of techniques described above heavily relies on pointwise uniform ellipticity (soft nonuniform ellipticity) and fails to deliver results in the presence of strong nonuniform ellipticity (1.13), even for the most basic model examples.

Remark 4.2. It is worth noting that in the counterexample provided in [57], the parameter α can be taken as any positive number $\alpha > 0$, and not necessarily restricted to the interval $\alpha \in (0, 1]$, provided that p and q are chosen accordingly, as in (4.10).[24] This shows that pathological minimizers can arise even when the coefficients have arbitrary degrees of smoothness, as long as p and q are sufficiently far apart.

4.4. Nonuniformly elliptic Schauder estimates

The discussion in Section 3 highlights that, while for PDEs driven by uniformly elliptic operators as in (1.10), the classical Schauder implication

$$\text{Hölder continuous coefficients} \implies \text{Hölder continuous first derivatives}$$

holds for energy solutions, already for the mild nonuniformity in (1.11) and (1.12) regularity of solutions is no longer guaranteed in general, not even by smooth ingredients. However, as seen in Section 4.3, perturbation techniques are still available for soft nonuniformly elliptic equations or functionals, the main difficulty being getting close, in a homogeneous fashion, to a suitably "frozen" Dirichlet problem, whose solution features homogeneous excess decay estimates. Such a circle of ideas immediately breaks down already for the very simple model

$$\mathcal{F}_a(w, \Omega) := \int_\Omega a(x) F_0(Dw)\,\mathrm{d}x, \quad 1 \lesssim a \in C^{0,\alpha}(\Omega),\ \alpha \in (0, 1], \tag{4.18}$$

where F_0 is as in (4.1). In fact, recalling (1.7), in view of (4.1), the best bound on the ellipticity ratio we have is still

$$\mathcal{R}(x, z) \approx \frac{\text{highest eigenvalue of } a(x)\partial_{zz} F_0(z)}{\text{lowest eigenvalue of } a(x)\partial_{zz} F_0(z)} \lesssim |z|^{q-p} + 1.$$

Therefore, any hope of recovering uniformly elliptic proof schemes as in Section 4.3 vanishes. The validity of the Schauder theory in the nonuniformly elliptic setting is a classical problem raised at several stages in the literature. For instance, in his MR

[24] In this case, one takes $a \in C^{[\alpha],\{\alpha\}}$.

review[25] of Giaquinta and Giusti's paper [66], Lieberman remarks how the authors' approach does not apply when uniform ellipticity conditions are violated, unless a priori Lipschitz bounds are assumed. However, these gradient estimates are available only if equations or functionals are uniformly elliptic, or if an unnatural amount of smoothness is imposed on coefficients and heavy regularity is a priori assumed on solutions. In other terms, regularity for nonuniformly elliptic problems with Hölder continuous coefficients could be delivered only for a priori Lipschitz continuous solutions.[26] The central role of gradient bounds is also remarked by Ivanov [81, pp. 7 and 15]: in view of the results of Ladyzhenskaya and Ural'tseva [93], the issue of solvability of boundary value problems for a nonuniformly elliptic or parabolic equation reduces to the possibility of constructing a priori estimates of the L^{∞}-norm of the gradients of solutions for a suitable one-parameter family of similar equations. This means finding uniform a priori gradient estimates for regularized problems. The importance of deriving Lipschitz bounds in nonuniformly elliptic problems as a starting point for higher regularity was also stressed by Giaquinta and Giusti [66, p. 56]. In addition to these considerations, Schauder theory does not always hold even in the case of soft nonuniform ellipticity (1.11)–(1.12) (cf. Section 4.3). This emphasizes how delicate the issue of establishing Schauder theory is for strongly nonuniformly elliptic problems (1.13). Such a long-standing problem has been settled in [42, 45], through a novel approach based on a direct proof of gradient boundedness, that does not make any use of Hölder gradient estimates.[27] This alters the classical paradigm described in the third point of Remark 3.1. The following results appear in [45].

Theorem 4.3. *Under assumptions* (4.1) *and* (4.4)*, let* $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ *be a local minimizer of the LSM-extension* $\overline{\mathcal{F}}$ *of the functional* $\mathcal{F}$ *in* (1.14)*. Then*

- Du *is locally Hölder continuous in* Ω.
- u *is also a minimizer of* $\mathcal{F}$.

Recalling (4.5), we have the following corollary.

Corollary 4.4. *Under assumptions* (4.1) *and* (4.4)*, let* $u \in W^{1,p}_{\mathrm{loc}}(\Omega)$ *be a local minimizer of the functional* $\mathcal{F}$ *in* (1.14)*. If* $\mathcal{L}_{\mathcal{F}}(u, \mathrm{B}) = 0$ *holds for all balls* $\mathrm{B} \Subset \Omega$*, then* u *is a minimizer of* $\overline{\mathcal{F}}$ *and hence* Du *is locally Hölder continuous in* Ω.

[25] *Math. Rev.* MR0749677.

[26] On the other hand, in such cases the problems considered are essentially again uniformly elliptic as in the present setting the ellipticity ratio blows-up only when the gradient blows-up.

[27] The same approach also leads to settling the long-standing issue of proving gradient Hölder continuity of minima of nondifferentiable, nonuniformly elliptic variational integrals. The additional difficulty in this case is that in general the Euler–Lagrange equation is not available. See [42].

If integrand F is nondegenerate or nonsingular, in the sense that $\mu > 0$ in (4.1), we recover the classical Schauder-type implication (3.1).

Theorem 4.5. *In Theorem 4.3 and Corollary 4.4, assume that $\mu > 0$ and $\alpha \in (0,1)$ in (4.1), and that $\partial_{zz} F$ is continuous. Then $Du \in C^{0,\alpha}_{\mathrm{loc}}(\Omega;\mathbb{R}^n)$.*

In particular, Theorems 4.3 and 4.5 highlight that the LSM-extension $\overline{\overline{\mathcal{F}}}$ provides a selection principle capable of identifying/producing regular minimizers of the original functional $\mathcal{F}$, while excluding those that are irregular due to the appearance of Lavrentiev-type phenomena. In Corollary 4.4, the role of the assumption $\mathcal{L}_{\mathcal{F}}(u,\mathrm{B}) = 0$ is rather natural: via the occurrence of the approximation in energy, it is possible to approximate the original minimizer with a sequence of minimizers of more regular functionals for which a priori regularity estimates, in fact described in the next section, apply. The Lavrentiev gap functional vanishes in a large variety of situations, for instance as detailed in [54, Section 5]. As an example, a condition of the type $F(x,z) \approx G(z)$, for $|z|$ large, where $G\colon \mathbb{R}^n \to [0,\infty)$ is a convex function, guarantees that $\mathcal{L}_{\mathcal{F}} \equiv 0$. As a consequence, we have the following corollary.

Corollary 4.6. *Let $u \in W^{1,1}(\Omega)$ be a minimizer of the functional in (4.18), where F_0 satisfies the first to fourth lines of (4.1) and (4.4) holds. Then, Du is locally Hölder continuous in Ω. If $\mu > 0$ and $\partial_{zz} F$ is continuous, then $u \in C^{1,\alpha}_{\mathrm{loc}}(\Omega)$ when $\alpha < 1$.*

As a matter of fact, an interesting twist is that, on several occasions, assumption (4.4) itself guarantees that the Lavrentiev gap vanishes. For instance, this happens when $F(x,z) \approx |z|^p + a(x)|z|^q$ holds for large $|z|$ and (4.4) is satisfied. This last fact does not even require that F satisfies (4.1) and holds for general Carathéodory integrands; compare with (4.12).

4.5. Key steps in the proof of Theorem 4.3

The proof offers a general body of methods and perspectives that might be useful in a number of problems with lack of ellipticity and/or regularity of external ingredients. The main arguments leading to Theorem 4.3 can be outlined as follows.[28]

- *Almost Lipschitz continuity via fractional Moser's iteration.* The gradient of minima belongs to L^t_{loc} for all $1 \le t < \infty$ via a combined use of interpolation

[28] For simplicity, we describe how a priori estimates are obtained; the results then follow by approximation procedures for which we refer to [45]. Keep in mind the comments made after Theorem 4.5.

inequalities and a fractional Moser-type iteration. The first step is showing that a relaxed bound on q/p, i.e.,

$$q < p + \begin{cases} p\alpha/2 & \text{if } p < 2, \\ \alpha & \text{if } 2 \le p \le n, \\ p\alpha/n & \text{if } n < p \end{cases} \tag{4.19}$$

implies that any (locally) bounded minimizer has a locally integrable gradient with an arbitrary large, finite power (local boundedness of minima under the assumptions used here follows by [78]). In this respect, (4.19) is optimal by counterexamples working for bounded minimizers [54], where minima fail to be $W^{1,q}$-regular; compare (4.19) with (4.10). In the autonomous case,[29] results of this type were known, and classical, starting by the work of Ladyzhenskaya and Ural'tseva [94], see also [10, 32, 33, 39, 131], where dimension-free gap bounds are proved to be effective when in the presence of bounded minimizers. The techniques used for them require differentiating the Euler–Lagrange equation (1.15), which is impossible in the presence of Hölder coefficients. In [45], higher integrability estimates are instead achieved by means of fine Besov spaces techniques, sometimes employed in uniformly elliptic problems with a certain lack of ellipticity, such as degenerate PDEs in the Heisenberg group [53], or in nonlocal problems [18, 22, 23, 51, 61, 62]. The underlying principle is that dealing with strong nonuniform ellipticity requires similar efforts as those needed to compensate for the lack of strong forms of monotonicity. This is delicate to handle and heavily relies on a sharp numerology to keep the optimal conditions (4.19). Specifically, we prove that

$$\left\| \frac{\tau_h^2 u}{|h|^{1+\varepsilon_i}} \right\|_{L^{t_i}} \le c_i \tag{4.20}$$

holds for $h \in \mathbb{R}^n$, $|h| > 0$ sufficiently small, and with integro-differentiability exponents $t_i \nearrow \infty$, $0 < \varepsilon_i \searrow 0$ and $c_i \to \infty$. In (4.20), τ_h^2 denotes the double finite differences operator in the direction[30] h. Via basic embedding properties of Besov spaces, (4.20) implies that $Du \in L^{t_i}_{\mathrm{loc}}$ for every i, and therefore in all Lebesgue spaces with finite exponent, locally. We stress that the validity of (4.19) guarantees that sequence $\{t_i\}_{i\in\mathbb{N}}$ in (4.20) diverges. Estimates leading

[29] with the exception of the specific double-phase case, see [10].

[30] With $\mathtt{t} > 0$, $h \in \mathbb{R}^n$, set $\Omega_{\mathtt{t}} := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > \mathtt{t}\}$; the finite difference operators $\tau_h \colon L^1(\Omega;\mathbb{R}^n) \mapsto L^1(\Omega_{|h|};\mathbb{R}^n)$, $\tau_h^2 \colon L^1(\Omega;\mathbb{R}^n) \mapsto L^1(\Omega_{2|h|};\mathbb{R}^n)$ are defined as $\tau_h w(x) := w(x+h) - w(x)$ and $\tau_h^2 w := \tau_h(\tau_h w)$, respectively.

to (4.20) implement a fractional Moser-type iteration that does not directly lead to $Du \in L^{\infty}_{\mathrm{loc}}$, as the integrability gain from t_i to t_{i+1} is linear rather than geometric. A fractional Moser's iteration is a natural idea to use in this setting as it relies on the basic observation that α-Hölder continuity means almost α-fractional differentiability, in the sense that

$$x \mapsto \partial_z F(x,z) \overset{(4.1)}{\overset{(4\text{th line})}{\in}} C^{0,\alpha} \implies x \mapsto \partial_z F(x,z) \in W^{s,t}$$

for all $s \in (0,\alpha)$, $t \in [1,\infty)$, thus aligning with classical Moser's iteration in the differentiable setting, where in fact one uses that $x \mapsto \partial_z F(x,z) \in W^{1,\infty}$.

- *Fractional Caccioppoli inequalities on level sets.* A standard formulation of gradient Caccioppoli inequalities on level sets for autonomous, quasilinear uniformly elliptic PDEs, such as $\operatorname{div}(|Du|^{p-2}Du) = 0$, reads as

$$\varrho^2 \fint_{B_{\varrho/2}(x_0)} |D(|Du|^p - \kappa)_+|^2 \lesssim \fint_{B_{\varrho}(x_0)} (|Du|^p - \kappa)_+^2 \,\mathrm{d}x \tag{4.21}$$

for any $\kappa \geq 0$ and all balls $B_{\varrho}(x_0) \Subset \Omega$. Estimate (4.21) builds on the classical Bernstein method, that prescribes differentiating the equation to prove that suitable convex functions of $|Du|$, such as for instance $|Du|^p$, are subsolutions to linear, uniformly elliptic equations with measurable coefficients; hence, they satisfy (4.21). In this respect, see the seminal work of Uhlenbeck [129]. This yields local boundedness for $|Du|$ since (4.21) grants its membership to upper De Giorgi classes, i.e., classical De Giorgi's iteration applies. This argument breaks down in the presence of merely Hölder continuous coefficients, as the Euler–Lagrange equation is no longer differentiable. Nonetheless, the fractional, renormalized Caccioppoli inequality on level sets

$$\begin{aligned}\varrho^{2\beta-n}[(|Du|^p - \kappa)_+]^2_{\beta,2;B_{\varrho/2}(x_0)} &\\ \lesssim M^{2\mathtt{b}} \fint_{B_{\varrho}(x_0)} (|Du|^p - \kappa)_+^2 \,\mathrm{d}x &+ M^{2\mathtt{b}}\varrho^{2\alpha} \fint_{B_{\varrho}(x_0)} 1 + |Du|^m \,\mathrm{d}x\end{aligned} \tag{4.22}$$

holds whenever $M \geq \|Du\|_{L^{\infty}(B_{\varrho}(x_0))}$ for some $\beta \in (0,1)$, $m \in (1,\infty)$, on any ball $B_{\varrho}(x_0) \Subset \Omega$. In (4.22), the number $\mathtt{b} \equiv \mathtt{b}(n,p,q,\alpha)$ is an explicit function such that $\mathtt{b} \to 0$ when $q \to p$. There are three major differences between the classical (4.21) and the formulation (4.22), as introduced in [42, 45].[31] First, the

[31] Fractional Caccioppoli-type inequalities in the setting of nonlinear potential theory were pioneered by Mingione [108, 109], where also fractional De Giorgi's classes are used. See also [29, 35, 49] for other instances of fractional De Giorgi classes in the nonlocal setting.

fractional derivatives of $(|Du|^p - \kappa)_+$ appearing in (4.22) replace the full derivatives on the left-hand side of (4.21), which are no longer available due to the nondifferentiability of the coefficients. Second, the third term in (4.22) is a "size" term that captures the contribution of the coefficients and requires careful handling throughout the analysis. Third, the renormalization constant M is introduced to make the first two terms in (4.22) appear homogeneous in $|Du|^p$, mirroring the structure of (4.21), and thereby enabling iteration. This strategy also draws on an approach developed by Beck and Mingione [11], which prescribes – when working with inequalities like (4.21) in the nonuniformly elliptic setting – to exploit the fact that constants deteriorate only polynomially during De Giorgi iterations used to prove L^∞-estimates.[32] Accordingly, careful tracking of the quantity $M^{2\mathtt{b}}$ throughout the iteration process raises the possibility of ultimately reabsorbing it. In this respect, exponent $\mathtt{b}$ accounts for the power-type growth of the ellipticity ratio, while $\beta \equiv \beta(\alpha) \in (0,1)$ records the rate of Hölder continuity of coefficients. The sharp threshold (4.4) plays a pivotal role in linking these analytic features, ultimately yielding local Lipschitz regularity for minimizers. Crucially, achieving this under the optimal threshold (4.4) requires that (4.22) is employed in its sharpest possible form, ensuring that no loss of information occurs at any stage of the derivation. The construction of the renormalized, fractional De Giorgi classes in (4.22) is based on a Littlewood–Paley-type argument traceable to the dyadic decomposition of Besov functions [124], but implemented at a nonlinear level. It was first introduced by Kristensen and Mingione [86, 87] in the study of singular sets of minima of multiple integrals. In the present setting, the usual atoms appearing in dyadic decompositions are replaced by solutions to nonlinear, autonomous problems, whose regularity is quantified via suitable improved forms of related a priori estimates due to Bella and Schäffner [14, 16, 112]. This grants the sharp value of exponent $\mathtt{b}$ in (4.22), that allows us to preserve the optimal threshold in (4.4).

- *Nonlinear potentials.* A key aspect of the De Giorgi-type iteration described before lies in its potential theoretic nature. In fact, the size information encoded in the third term of (4.22) is carried along iterations by a general class of nonlinear Havin–Maz'ya–Wolff-type potentials [106] of the form

$$\mathbf{P}_\sigma^\vartheta(f; x_0, r) := \int_0^r \varrho^\sigma \left(\fint_{B_\varrho(x_0)} |f| \, \mathrm{d}x \right)^\vartheta \frac{\mathrm{d}\varrho}{\varrho},$$

[32] This behavior is less widely known, as constants deteriorate exponentially instead when proving Hölder estimates, which are the ultimate aim of the De Giorgi–Nash–Moser theory.

with $\sigma > 0$, $\vartheta \geq 0$ and $f \in L^1(B_r(x_0))$. For specific choices of σ and ϑ, potential $\mathbf{P}_\sigma^\vartheta$ becomes classical Riesz and Wolff potentials, that are usually employed when investigating fine properties of solutions to linear or nonlinear, nonhomogeneous PDEs. More precisely, potentials act as "ghosts" of the representation formula,[33] granting sharp pointwise bounds for solutions and their gradients under minimal conditions on external data; see [85, 88, 89, 91]. This means that potentials have been mainly used to deal with certain borderline regularity conditions on external ingredients. In [39, 42, 43, 45], instead, potentials are employed as carriers of size information along iterations, and eventually apply their mapping properties [42, Section 4] to control the resulting quantities. Also at this stage, in the absence of gradient higher integrability, condition (4.4) is crucial to ensure that potentials embed in the right Lebesgue spaces, [42, 43].

The approach sketched above yields local Lipschitz continuity for minima. In turn, gradient boundedness makes the nonuniformity of functional (1.14) (and of its LSM-relaxation (4.3)) immaterial, so gradient Hölder continuity is achieved by a delicate adaptation of more standard perturbation methods [42, Section 10], [45, Section 5]. See also the third point in Remark 3.1.

Acknowledgements. The author thanks Professor Anna Balci for sharing the drawings in Figures 1 and 2.

Funding. This work was supported by the European Research Council, through the ERC StG project NEW, No. 101220121, and by the University of Parma, through the action "Bando di Ateneo 2024 per la ricerca".

References

[1] E. Acerbi and G. Mingione, Regularity results for a class of functionals with non-standard growth. *Arch. Ration. Mech. Anal.* **156** (2001), no. 2, 121–140 Zbl 0984.49020 MR 1814973

[2] S. Agmon, A. Douglis, and L. Nirenberg, Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Comm. Pure Appl. Math.* **12** (1959), no. 4, 623–727 Zbl 0093.10401 MR 0125307

[3] S. Agmon, A. Douglis, and L. Nirenberg, Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. II. *Comm. Pure Appl. Math.* **17** (1964), no. 1, 35–92 Zbl 0123.28706 MR 0162050

[33] Representation formulas via fundamental solutions are obviously not available in nonlinear problems.

[4] S. Baasandorj and S.-S. Byun, Regularity for Orlicz phase problems. *Mem. Amer. Math. Soc.* **308** (2025), no. 1556, 131 pp. Zbl 1573.49002 MR 4896165

[5] A. Balci, L. Diening, and M. Surnachev, Scalar minimizers with maximal singular sets and lack of Meyers property. In D. Apushkinskaya, A. Laptev, A. I. Nazarov, and H. Shahgholian (eds.), *Friends in partial differential equation. The Nina N. Uraltseva 90th anniversary volume*, pp. 1–43, EMS Press, Berlin, 2025 Zbl 08132144 MR 4967630

[6] A. K. Balci, L. Diening, and M. Surnachev, New examples on Lavrentiev gap using fractals. *Calc. Var. Partial Differential Equations* **59** (2020), no. 5, article no. 180, 34 pp. Zbl 1453.35082 MR 4153906

[7] A. K. Balci, C. Ortner, and J. Storn, Crouzeix–Raviart finite element method for non-autonomous variational problems with Lavrentiev gap. *Numer. Math.* **151** (2022), no. 4, 779–805 Zbl 1501.65105 MR 4453291

[8] J. M. Ball and G. Knowles, A numerical method for detecting singular minimizers. *Numer. Math.* **51** (1987), no. 2, 181–197 Zbl 0636.65064 MR 0890031

[9] J. M. Ball and V. J. Mizel, One-dimensional variational problems whose minimizers do not satisfy the Euler–Lagrange equation. *Arch. Ration. Mech. Anal.* **90** (1985), no. 4, 325–388 Zbl 0585.49002 MR 0801585

[10] P. Baroni, M. Colombo, and G. Mingione, Regularity for general functionals with double phase. *Calc. Var. Partial Differential Equations* **57** (2018), no. 2, article no. 62, 48 pp. Zbl 1394.49034 MR 3775180

[11] L. Beck and G. Mingione, Lipschitz bounds and nonuniform ellipticity. *Comm. Pure Appl. Math.* **73** (2020), no. 5, 944–1034 Zbl 1445.35140 MR 4078712

[12] L. Beck and T. Schmidt, On the Dirichlet problem for variational integrals in BV. *J. Reine Angew. Math.* **2013** (2013), no. 674, 113–194 Zbl 1260.49079 MR 3010549

[13] L. Beck and T. Schmidt, Interior gradient regularity for BV minimizers of singular variational problems. *Nonlinear Anal.* **120** (2015), 86–106 Zbl 1319.49054 MR 3348048

[14] P. Bella and M. Schäffner, On the regularity of minimizers for scalar integral functionals with (p, q)-growth. *Anal. PDE* **13** (2020), no. 7, 2241–2257 Zbl 1460.49027 MR 4175825

[15] P. Bella and M. Schäffner, Local boundedness and Harnack inequality for solutions of linear nonuniformly elliptic equations. *Comm. Pure Appl. Math.* **74** (2021), no. 3, 453–477 Zbl 1469.35073 MR 4201290

[16] P. Bella and M. Schäffner, Lipschitz bounds for integral functionals with (p, q)-growth conditions. *Adv. Calc. Var.* **17** (2024), no. 2, 373–390 Zbl 1541.35173 MR 4726518

[17] M. Bildhauer and M. Fuchs, Convex variational problems with linear growth. In S. Hildebrandt and H. Karcher (eds.), *Geometric analysis and nonlinear partial differential equations*, pp. 327–344, Springer, Berlin, 2003 Zbl 1046.35020 MR 2008345

[18] V. Bögelein, F. Duzaar, N. Liao, G. Molica Bisci, and R. Servadei, Regularity for the fractional p-Laplace equation. *J. Funct. Anal.* **289** (2025), no. 9, article no. 111078, 69 pp. Zbl 08065953 MR 4919132

[19] E. Bombieri, E. De Giorgi, and M. Miranda, Una maggiorazione a priori relativa alle ipersuperfici minimali non parametriche. *Arch. Ration. Mech. Anal.* **32** (1969), no. 4, 255–267 Zbl 0184.32803 MR 0248647

[20] G. Bouchitté, I. Fonseca, and J. Malý, The effective bulk energy of the relaxed energy of multiple integrals below the growth exponent. *Proc. Roy. Soc. Edinburgh Sect. A* **128** (1998), no. 3, 463–479 Zbl 0907.49008 MR 1632814

[21] P. Bousquet and L. Brasco, Lipschitz regularity for orthotropic functionals with nonstandard growth conditions. *Rev. Mat. Iberoam.* **36** (2020), no. 7, 1989–2032 Zbl 1460.35162 MR 4163990

[22] L. Brasco and E. Lindgren, Higher Sobolev regularity for the fractional p-Laplace equation in the superquadratic case. *Adv. Math.* **304** (2017), 300–354 Zbl 1364.35055 MR 3558212

[23] L. Brasco, E. Lindgren, and A. Schikorra, Higher Hölder regularity for the fractional p-Laplacian in the superquadratic case. *Adv. Math.* **338** (2018), 782–846 Zbl 1400.35049 MR 3861716

[24] M. Bulíček, P. Gwiazda, and J. Skrzeczkowski, On a range of exponents for absence of Lavrentiev phenomenon for double phase functionals. *Arch. Ration. Mech. Anal.* **246** (2022), no. 1, 209–240 Zbl 1497.49049 MR 4487513

[25] G. Buttazzo and V. J. Mizel, Interpretation of the Lavrentiev phenomenon by relaxation. *J. Funct. Anal.* **110** (1992), no. 2, 434–460 Zbl 0784.49006 MR 1194993

[26] S.-S. Byun and J. Oh, Regularity results for generalized double phase functionals. *Anal. PDE* **13** (2020), no. 5, 1269–1300 Zbl 1477.49057 MR 4149062

[27] S.-S. Byun, J. Ok, and K. Song, Hölder regularity for weak solutions to nonlocal double phase problems. *J. Math. Pures Appl. (9)* **168** (2022), 110–142 Zbl 1504.35104 MR 4515255

[28] R. Caccioppoli, Sulle equazioni ellittiche a derivate parziali con n variabili indipendenti. *Atti Accad. Naz. Lincei Rend. Lincei Mat. Appl.* **19** (1934), 83–89 Zbl 0009.06802

[29] L. Caffarelli, C. H. Chan, and A. Vasseur, Regularity theory for parabolic nonlinear integral operators. *J. Amer. Math. Soc.* **24** (2011), no. 3, 849–869 Zbl 1223.35098 MR 2784330

[30] L. A. Caffarelli and X. Cabré, *Fully nonlinear elliptic equations*. Amer. Math. Soc. Colloq. Publ. 43, American Mathematical Society, Providence, RI, 1995, 104 pp. Zbl 0834.35002 MR 1351007

[31] S. Campanato, Equazioni ellittiche del II° ordine e spazi $\mathcal{L}^{(2,\lambda)}$. *Ann. Mat. Pura Appl. (4)* **69** (1965), no. 1, 321–381 Zbl 0145.36603 MR 0192168

[32] M. Carozza, J. Kristensen, and A. Passarelli di Napoli, Higher differentiability of minimizers of convex variational integrals. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **28** (2011), no. 3, 395–411 Zbl 1245.49052 MR 2795713

[33] H. J. Choe, Interior behaviour of minimizers for certain functionals with nonstandard growth. *Nonlinear Anal.* **19** (1992), no. 10, 933–945 Zbl 0786.35040 MR 1192273

[34] M. Colombo and G. Mingione, Regularity for double phase variational problems. *Arch. Ration. Mech. Anal.* **215** (2015), no. 2, 443–496 Zbl 1322.49065 MR 3294408

[35] M. Cozzi, Regularity results and Harnack inequalities for minimizers and solutions of nonlocal problems: a unified approach via fractional De Giorgi classes. *J. Funct. Anal.* **272** (2017), no. 11, 4762–4837 Zbl 1366.49040 MR 3630640

[36] B. Dacorogna, *Direct methods in the calculus of variations*. 2nd edn., Appl. Math. Sci. 78, Springer, New York, 2008, 622 pp. Zbl 1140.49001 MR 2361288

[37] C. De Filippis, Quasiconvexity and partial regularity via nonlinear potentials. *J. Math. Pures Appl. (9)* **163** (2022), 11–82 Zbl 1537.35123 MR 4438896

[38] C. De Filippis, μ-ellipticity and nonautonomous integrals. *Eur. Math. Soc. Mag.* **137** (2025), 4–15 Zbl 08130155 MR 4990156

[39] C. De Filippis, F. De Filippis, and M. Piccinini, Bounded minimizers of double phase problems at nearly linear growth. 2026, arXiv:2411.14325v2, to appear in *J. Eur. Math. Soc.*

[40] C. De Filippis, L. Koch, and J. Kristensen, Quantified Legendreness and the regularity of minima. *Arch. Ration. Mech. Anal.* **248** (2024), no. 4, article no. 69, 51 pp. Zbl 1546.49069 MR 4779533

[41] C. De Filippis and G. Mingione, Lipschitz bounds and nonautonomous integrals. *Arch. Ration. Mech. Anal.* **242** (2021), no. 2, 973–1057 Zbl 1483.49050 MR 4331020

[42] C. De Filippis and G. Mingione, Nonuniformly elliptic Schauder theory. *Invent. Math.* **234** (2023), no. 3, 1109–1196 Zbl 1542.35140 MR 4665778

[43] C. De Filippis and G. Mingione, Regularity for double phase problems at nearly linear growth. *Arch. Ration. Mech. Anal.* **247** (2023), no. 5, article no. 85, 50 pp. Zbl 1525.35089 MR 4630451

[44] C. De Filippis and G. Mingione, Nonuniform ellipticity in variational problems and regularity. *Notices Amer. Math. Soc.* **72** (2025), no. 9, 989–999 Zbl 08137377 MR 4965436

[45] C. De Filippis and G. Mingione, The sharp growth rate in nonuniformly elliptic Schauder theory. *Duke Math. J.* **174** (2025), no. 9, 1775–1848 Zbl 1571.49030 MR 4917742

[46] C. De Filippis and B. Stroffolini, Singular multiple integrals and nonlinear potentials. *J. Funct. Anal.* **285** (2023), no. 2, article no. 109952, 76 pp. Zbl 1516.31026 MR 4575686

[47] F. De Filippis and M. Piccinini, Regularity for multi-phase problems at nearly linear growth. *J. Differential Equations* **410** (2024), 832–868 Zbl 1548.49044 MR 4790887

[48] E. De Giorgi, Un esempio di estremali discontinue per un problema variazionale di tipo ellittico. *Boll. Unione Mat. Ital. (4)* **1** (1968), 135–137 Zbl 0155.17603 MR 0227827

[49] A. Di Castro, T. Kuusi, and G. Palatucci, Local behavior of fractional p-minimizers. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **33** (2016), no. 5, 1279–1299 Zbl 1355.35192 MR 3542614

[50] E. DiBenedetto, $C^{1+\alpha}$ local regularity of weak solutions of degenerate elliptic equations. *Nonlinear Anal.* **7** (1983), no. 8, 827–850 Zbl 0539.35027 MR 0709038

[51] L. Diening, K. Kim, H.-S. Lee, and S. Nowak, Higher differentiability for the fractional p-Laplacian. *Math. Ann.* **391** (2025), no. 4, 5631–5693 Zbl 1564.35068 MR 4884557

[52] L. Diening, B. Stroffolini, and A. Verde, Everywhere regularity of functionals with ϕ-growth. *Manuscripta Math.* **129** (2009), no. 4, 449–481 Zbl 1168.49035 MR 2520895

[53] A. Domokos, Differentiability of solutions for the non-degenerate p-Laplacian in the Heisenberg group. *J. Differential Equations* **204** (2004), no. 2, 439–470 Zbl 1065.35103 MR 2085543

[54] L. Esposito, F. Leonetti, and G. Mingione, Sharp regularity for functionals with (p, q) growth. *J. Differential Equations* **204** (2004), no. 1, 5–55 Zbl 1072.49024 MR 2076158

[55] L. Ferreri, L. Spolaor, and B. Velichkov, Unique continuation for nonlinear variational problems. 2024, arXiv:2408.00405v1

[56] I. Fonseca and J. Malý, Relaxation of multiple integrals below the growth exponent. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **14** (1997), no. 3, 309–338 Zbl 0868.49011 MR 1450951

[57] I. Fonseca, J. Malý, and G. Mingione, Scalar minimizers with fractal singular sets. *Arch. Ration. Mech. Anal.* **172** (2004), no. 2, 295–307 Zbl 1049.49015 MR 2058167

[58] I. Fonseca and P. Marcellini, Relaxation of multiple integrals in subcritical Sobolev spaces. *J. Geom. Anal.* **7** (1997), no. 1, 57–81 Zbl 0915.49011 MR 1630777

[59] I. Fragalà, F. Gazzola, and B. Kawohl, Existence and nonexistence results for anisotropic quasilinear elliptic equations. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **21** (2004), no. 5, 715–734 Zbl 1144.35378 MR 2086756

[60] M. Fuchs and G. Mingione, Full $C^{1,\alpha}$-regularity for free and constrained local minimizers of elliptic variational integrals with nearly linear growth. *Manuscripta Math.* **102** (2000), no. 2, 227–250 Zbl 0995.49023 MR 1771942

[61] P. Garain and E. Lindgren, Higher Hölder regularity for mixed local and nonlocal degenerate elliptic equations. *Calc. Var. Partial Differential Equations* **62** (2023), no. 2, article no. 67, 36 pp. Zbl 1505.35072 MR 4530314

[62] P. Garain and E. Lindgren, Higher Hölder regularity for the fractional p-Laplace equation in the subquadratic case. *Math. Ann.* **390** (2024), no. 4, 5753–5792 Zbl 1555.35079 MR 4816120

[63] M. Giaquinta, Growth conditions and regularity, a counterexample. *Manuscripta Math.* **59** (1987), no. 2, 245–248 Zbl 0638.49005 MR 0905200

[64] M. Giaquinta and E. Giusti, On the regularity of the minima of variational integrals. *Acta Math.* **148** (1982), 31–46 Zbl 0494.49031 MR 0666107

[65] M. Giaquinta and E. Giusti, Differentiability of minima of non-differentiable functionals. *Invent. Math.* **72** (1983), no. 2, 285–298 Zbl 0513.49003 MR 0700772

[66] M. Giaquinta and E. Giusti, Global $C^{1,\alpha}$-regularity for second order quasilinear elliptic equations in divergence form. *J. Reine Angew. Math.* **1984** (1984), no. 351, 55–65 Zbl 0528.35014 MR 0749677

[67] M. Giaquinta, G. Modica, and J. Souček, Functionals with linear growth in the calculus of variations. I. *Comment. Math. Univ. Carolin.* **20** (1979), no. 1, 143–156 Zbl 0409.49006 MR 0526154

[68] M. Giaquinta, G. Modica, and J. Souček, Functionals with linear growth in the calculus of variations. II. *Comment. Math. Univ. Carolin.* **20** (1979), no. 1, 157–172 Zbl 0409.49007 MR 0526154

[69] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*. Grundlehren Math. Wiss. 224, Springer, Berlin, 1983, 513 pp. Zbl 0562.35001 MR 0737190

[70] G. Giraud, Sur le problème de Dirichlet généralisé (deuxième mémoire). *Ann. Sci. Éc. Norm. Supér. (3)* **46** (1929), 131–245 JFM 55.0285.02 MR 1509295

[71] E. Giusti, On the equation of surfaces of prescribed mean curvature. Existence and uniqueness without boundary conditions. *Invent. Math.* **46** (1978), no. 2, 111–137 Zbl 0381.35035 MR 0487722

[72] F. Gmeineder and J. Kristensen, Partial regularity for BV minimizers. *Arch. Ration. Mech. Anal.* **232** (2019), no. 3, 1429–1473 Zbl 1411.26016 MR 3928753

[73] F. Gmeineder and J. Kristensen, Quasiconvex functionals of (p, q)-growth and the partial regularity of relaxed minimizers. *Arch. Ration. Mech. Anal.* **248** (2024), no. 5, article no. 80, 125 pp. Zbl 1548.49045 MR 4795764

[74] P. Hästö and J. Ok, Maximal regularity for local minimizers of non-autonomous functionals. *J. Eur. Math. Soc. (JEMS)* **24** (2022), no. 4, 1285–1334 Zbl 1485.49044 MR 4397041

[75] P. Hästö and J. Ok, Regularity theory for non-autonomous partial differential equations without Uhlenbeck structure. *Arch. Ration. Mech. Anal.* **245** (2022), no. 3, 1401–1436 Zbl 1511.35172 MR 4467321

[76] P. Hästö and J. Ok, Regularity theory for non-autonomous problems with a priori assumptions. *Calc. Var. Partial Differential Equations* **62** (2023), no. 9, article no. 251, 28 pp. Zbl 1529.35109 MR 4658589

[77] A. C. Heinricher and V. J. Mizel, A stochastic control problem with different value functions for singular and absolutely continuous control. In *Proceedings of the 1986 25th IEEE Conference on Decision and Control, Athens*, pp. 134–139, IEEE, 1986

[78] J. Hirsch and M. Schäffner, Growth conditions and regularity, an optimal local boundedness result. *Commun. Contemp. Math.* **23** (2021), no. 3, article no. 2050029, 17 pp. Zbl 1458.49009 MR 4216424

[79] E. Hopf, Bemerkungen zu einem Satze von S. Bernstein aus der Theorie der elliptischen Differentialgleichungen. *Math. Z.* **29** (1929), no. 1, 744–745 JFM 55.0289.01 MR 1545036

[80] A. V. Ivanov, Local estimates of the maximum modulus of the first derivatives of the solutions of quasilinear nonuniformly elliptic and nonuniformly parabolic equations and of systems of general form (in Russian). *Tr. Mat. Inst. Steklova* **110** (1970), 45–64. English translation. *Proc. Steklov Inst. Math.* **110** (1970), 48–71 Zbl 0247.35049 MR 0293241

[81] A. V. Ivanov, Second-order quasilinear degenerate and nonuniformly elliptic and parabolic equations (in Russian). *Tr. Mat. Inst. Steklova* **160** (1982), 286 pp. English translation. *Proc. Steklov Inst. Math.* **160** (1984), no. 1, 287 pp. Zbl 0517.35001 Zbl 0544.35003 MR 0753230

[82] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of differential operators and integral functionals*. Springer, Berlin, 1994, 570 pp. Zbl 0838.35001 MR 1329546

[83] A. Karppinen and M. Lee, Hölder continuity of the minimizer of an obstacle problem with generalized Orlicz growth. *Int. Math. Res. Not. IMRN* **2022** (2022), no. 19, 15313–15354 Zbl 1527.49027 MR 4490955

[84] S. Kichenassamy, Schauder-type estimates and applications. In M. Chipot and P. Quittner (eds.), *Handbook of differential equations. Stationary partial differential equations, Vol. 3*, pp. 401–464, North-Holland, Amsterdam, 2006 Zbl 1192.35003

[85] T. Kilpeläinen and J. Malý, The Wiener test and potential estimates for quasilinear elliptic equations. *Acta Math.* **172** (1994), no. 1, 137–161 Zbl 0820.35063 MR 1264000

[86] J. Kristensen and G. Mingione, The singular set of ω-minima. *Arch. Ration. Mech. Anal.* **177** (2005), no. 1, 93–114 Zbl 1082.49036 MR 2187315

[87] J. Kristensen and G. Mingione, The singular set of minima of integral functionals. *Arch. Ration. Mech. Anal.* **180** (2006), no. 3, 331–398 Zbl 1116.49010 MR 2214961

[88] T. Kuusi and G. Mingione, Universal potential estimates. *J. Funct. Anal.* **262** (2012), no. 10, 4205–4269 Zbl 1252.35097 MR 2900466

[89] T. Kuusi and G. Mingione, Linear potentials in nonlinear potential theory. *Arch. Ration. Mech. Anal.* **207** (2013), no. 1, 215–246 Zbl 1266.31011 MR 3004772

[90] T. Kuusi and G. Mingione, A nonlinear Stein theorem. *Calc. Var. Partial Differential Equations* **51** (2014), no. 1–2, 45–86 Zbl 1316.35132 MR 3247381

[91] T. Kuusi and G. Mingione, Vectorial nonlinear potential theory. *J. Eur. Math. Soc. (JEMS)* **20** (2018), no. 4, 929–1004 Zbl 1394.35206 MR 3779689

[92] O. A. Ladyzhenskaya and N. N. Ural'tseva, Quasi-linear elliptic equations and variational problems with many independent variables (in Russian). *Uspehi Mat. Nauk* **16** (1961), no. 1(97), 19–90. English translation. *Russian Math. Surveys* **16** (1961), no. 1, 17–91 Zbl 0142.37602 MR 0149075

[93] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and quasilinear elliptic equations*. Izdat. Nauka, Moscow, 1964; Academic Press, New York, 1968, 495 pp. Zbl 0164.13002 MR 0244627

[94] O. A. Ladyzhenskaya and N. N. Ural'tseva, Local estimates for gradients of solutions of non-uniformly elliptic and parabolic equations. *Comm. Pure Appl. Math.* **23** (1970), no. 4, 677–703 Zbl 0193.07202 MR 0265745

[95] M. Lavrentieff, Sur quelques problèmes du calcul des variations. *Ann. Mat. Pura Appl.* **4** (1927), no. 1, 7–28 JFM 53.0481.02 MR 1553097

[96] H. Lebesgue, Intégrale, longueur, aire. *Ann. Mat. Pura Appl.* **7** (1902), no. 1, 231–359 JFM 33.0307.02

[97] G. M. Lieberman, The natural generalization of the natural conditions of Ladyzhenskaya and Ural'tseva for elliptic equations. *Comm. Partial Differential Equations* **16** (1991), no. 2–3, 311–361 Zbl 0742.35028 MR 1104103

[98] J. J. Manfredi, Regularity for minima of functionals with p-growth. *J. Differential Equations* **76** (1988), no. 2, 203–212 Zbl 0674.35008 MR 0969420

[99] B. Manià, Sopra un esempio di Lavrentieff. *Boll. Unione Mat. Ital.* **13** (1934), no. 3, 147–153 Zbl 0009.15803

[100] P. Marcellini, On the definition and the lower semicontinuity of certain quasiconvex integrals. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **3** (1986), no. 5, 391–409 Zbl 0609.49009 MR 0868523

[101] P. Marcellini, Regularity of minimizers of integrals of the calculus of variations with non standard growth conditions. *Arch. Ration. Mech. Anal.* **105** (1989), no. 3, 267–284 Zbl 0667.49032 MR 0969900

[102] P. Marcellini, The stored-energy for some discontinuous deformations in nonlinear elasticity. In F. Colombini, A. Marino, L. Modica, and S. Spagnolo (eds.), *Partial differential equations and the calculus of variations. Essays in Honor of Ennio De Giorgi Volume 2*, pp. 767–786, Progr. Nonlinear Differential Equations Appl. 1, Birkhäuser, Boston, 1989 Zbl 0679.73006 MR 1034028

[103] P. Marcellini, Regularity and existence of solutions of elliptic equations with p,q-growth conditions. *J. Differential Equations* **90** (1991), no. 1, 1–30 Zbl 0724.35043 MR 1094446

[104] P. Marcellini, Everywhere regularity for a class of elliptic systems without growth conditions. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (4)* **23** (1996), no. 1, 1–25 Zbl 0922.35031 MR 1401415

[105] V. G. Maz'ya, Examples of nonregular solutions of quasilinear elliptic equations with analytic coefficients (in Russian). *Funktsional. Anal. i Prilozhen.* **2** (1968), no. 3, 53–57. English translation. *Funct. Anal. Appl.* **2** (1968), no. 3, 230–234 Zbl 0179.43601 MR 0237946

[106] V. G. Maz'ya and V. P. Havin, Non-linear potential theory (in Russian). *Uspehi Mat. Nauk* **27** (1972), no. 6(168), 67–138. English translation. *Russian Math. Surveys* **27** (1972), no. 6, 71–148 Zbl 0247.31010 Zbl 0269.31004 MR 0409858

[107] H. Min-Chun, Some remarks on the minimizers of variational integrals with non standard growth conditions. *Boll. Unione Mat. Ital. A* **6** (1992), 91–101 Zbl 0768.49022 MR 1164739

[108] G. Mingione, The Calderón–Zygmund theory for elliptic problems with measure data. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **6** (2007), no. 2, 195–261 Zbl 1178.35168 MR 2352517

[109] G. Mingione, Gradient potential estimates. *J. Eur. Math. Soc. (JEMS)* **13** (2011), no. 2, 459–486 Zbl 1217.35077 MR 2746772

[110] C. Mooney and O. Savin, Some singular minimizers in low dimensions in the calculus of variations. *Arch. Ration. Mech. Anal.* **221** (2016), no. 1, 1–22 Zbl 1338.49040 MR 3483890

[111] J. Nečas, Example of an irregular solution to a nonlinear elliptic system with analytic coefficients and conditions for regularity. In *Theory of nonlinear operators (Proc. Fourth Internat. Summer School, Acad. Sci., Berlin, 1975)*, pp. 197–206, Akademie, Berlin, 1977 Zbl 0372.35031 MR 0509483

[112] M. Schäffner, Lipschitz bounds for nonuniformly elliptic integral functionals in the plane. *Proc. Amer. Math. Soc.* **152** (2024), no. 11, 4717–4727 Zbl 1551.35214 MR 4802625

[113] J. Schauder, Numerische Abschätzungen in elliptischen linearen Differentialgleichungen. *Stud. Math.* **5** (1934), 34–42 Zbl 0010.20701

[114] J. Schauder, Über lineare elliptische Differentialgleichungen zweiter Ordnung. *Math. Z.* **38** (1934), no. 1, 257–282 Zbl 0008.25502 MR 1545448

[115] T. Schmidt, Regularity of relaxed minimizers of quasiconvex variational integrals with (p,q)-growth. *Arch. Ration. Mech. Anal.* **193** (2009), no. 2, 311–337 Zbl 1173.49032 MR 2525120

[116] G. A. Seregin and J. Frehse, Regularity of solutions to variational problems of the deformation theory of plasticity with logarithmic hardening. In N. N. Uraltseva (ed.), *Proceedings of the St. Petersburg mathematical society, Volume V*, pp. 127–152, Amer. Math. Soc. Transl. Ser. 2 193, American Mathematical Society, Providence, RI, 1999 Zbl 0920.00027 MR 1736908

[117] J. Serrin, On the definition and properties of certain variational integrals. *Trans. Amer. Math. Soc.* **101** (1961), 139–167 Zbl 0102.04601 MR 0138018

[118] J. Serrin, Pathological solutions of elliptic differential equations. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (3)* **18** (1964), no. 3, 385–387 Zbl 0142.37601 MR 0170094

[119] J. Serrin, The problem of Dirichlet for quasilinear elliptic differential equations with many independent variables. *Philos. Trans. Roy. Soc. Lond. Ser. A* **264** (1969), no. 1153, 413–496 Zbl 0181.38003 MR 0282058

[120] L. Simon, *Interior gradient bounds for non-uniformly elliptic partial equations of divergence form*. Ph.D. thesis, University of Adelaide, 1971

[121] L. Simon, Interior gradient bounds for non-uniformly elliptic equations. *Indiana Univ. Math. J.* **25** (1976), no. 9, 821–855 Zbl 0346.35016 MR 0412605

[122] L. Simon, Schauder estimates by scaling. *Calc. Var. Partial Differential Equations* **5** (1997), no. 5, 391–407 Zbl 0946.35017 MR 1459795

[123] V. Sverák and X. Yan, Non-Lipschitz minimizers of smooth uniformly convex functionals. *Proc. Natl. Acad. Sci. USA* **99** (2002), no. 24, 15269–15276 Zbl 1106.49046 MR 1946762

[124] H. Triebel, *The structure of functions*. Monogr. Math. 97, Birkhäuser, Basel, 2001, 425 pp. Zbl 0984.46021 MR 1851996

[125] N. S. Trudinger, The Dirichlet problem for nonuniformly elliptic equations. *Bull. Amer. Math. Soc.* **73** (1967), 410–413 Zbl 0149.32403 MR 0214916

[126] N. S. Trudinger, On the regularity of generalized solutions of linear, non-uniformly elliptic equations. *Arch. Ration. Mech. Anal.* **42** (1971), no. 1, 50–62 Zbl 0218.35035 MR 0344656

[127] N. S. Trudinger, Harnack inequalities for nonuniformly elliptic divergence structure equations. *Invent. Math.* **64** (1981), no. 3, 517–531 Zbl 0501.35012 MR 0632988

[128] N. S. Trudinger, A new approach to the Schauder estimates for linear elliptic equations. In M. Cowling, C. Meaney, and W. Moran (eds.), *Miniconference on operator theory and partial differential equations (North Ryde, 1986)*, pp. 52–59, Proc. Centre Math. Appl. Austral. Nat. Univ. 14, Australian National University, Centre for Mathematical Analysis, Canberra, 1986 Zbl 0641.35023 MR 0912926

[129] K. Uhlenbeck, Regularity for a class of non-linear elliptic systems. *Acta Math.* **138** (1977), no. 3–4, 219–240 Zbl 0372.35030 MR 0474389

[130] N. N. Ural'tseva, Degenerate quasilinear elliptic systems (in Russian). *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* **7** (1968), 184–222. English translation. *Semin. Math. V. A. Steklov Math. Inst. Leningrad* **7** (1968), 83–99 Zbl 0196.12501 Zbl 0199.42502 MR 0244628

[131] N. N. Ural'tseva and A. B. Urdaletova, The boundedness of the gradients of generalized solutions of degenerate quasilinear nonuniformly elliptic equations (in Russian). *Vestn. Leningr. Univ. Mat. Mekh. Astronom.* **19** (1983), 50–56. English translation. *Vestn. Leningr. Univ. Math* **16** (1984), 263–270 Zbl 0569.35029 MR 0725829

[132] V. V. Zhikov, Averaging of functionals of the calculus of variations and elasticity theory (in Russian). *Izv. Akad. Nauk SSSR Ser. Mat.* **50** (1986), no. 4, 675–710. English translation. *Math. USSR-Izv.* **29** (1987), no. 1, 33–66 Zbl 0599.49031 MR 0864171

[133] V. V. Zhikov, On Lavrentiev's phenomenon. *Russ. J. Math. Phys.* **3** (1995), no. 2, 249–269 Zbl 0910.49020 MR 1350506

[134] V. V. Zhikov, On some variational problems. *Russ. J. Math. Phys.* **5** (1997), no. 1, 105–116 Zbl 0917.49006 MR 1486765

Cristiana De Filippis
Dipartimento SMFI, Università di Parma, Viale delle Scienze 53/a, Campus, 43124 Parma, Italy; cristiana.defilippis@unipr.it

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/36

An introduction to representations of p-adic groups

Jessica Fintzen

Abstract. An explicit understanding of the (category of all smooth, complex) representations of p-adic groups provides an important tool not just within representation theory. It also has applications to number theory and other areas, and, in particular, it enables progress on various different forms of the Langlands program. In this write-up of the author's ECM 2024 colloquium-style talk, we will introduce p-adic groups and explain how the category of representations of p-adic groups decomposes into subcategories, called Bernstein blocks. We also provide an overview of what we know about the structure of these Bernstein blocks including a sketch of recent results of the author with Adler, Mishra and Ohara that allow us to reduce a lot of problems about the (category of) representations of p-adic groups to problems about representations of finite groups of Lie type, where answers are often already known or are at least easier to achieve. Moreover, we provide an overview of what is known about the construction of supercuspidal representations, which are the building blocks of all smooth representations and whose construction is also the key to obtaining the above results about the structure of the whole category of smooth representations. We will, in particular, focus on recent advances which include the work of the author mentioned in the EMS prize citation as well as a hint toward her recent joint work with David Schwein.

This article is a written version of the talk that the author gave during the 9th ECM. For more extended and precise survey articles, see [14] (for a general math audience), [15] (for a more specialized audience, including graduate students working in number theory and representation theory) and [16].

1. p-adic numbers and p-adic groups

In order to talk about the representation theory of p-adic groups, we first explain what p-adic groups are, for which we should first explain what "p-adic" even means.

Mathematics Subject Classification 2020: 22E50 (primary); 11F27, 22E35, 20C08, 20C20 (secondary).
Keywords: smooth representations of p-adic groups, supercuspidal representations, Bernstein blocks, Hecke algebras.

Throughout this article, p will denote a prime number. Once we have chosen our favorite prime number p, we can define the p-adic absolute value of the integers $\mathbb{Z}$ that measures how often an integer is divisible by p.

Definition 1.1. Let r be a non-zero integer coprime to p, and let s be a non-negative integer. The *p-adic absolute value* $|p^s \cdot r|_p$ of the integer $p^s \cdot r$ is defined to be

$$|p^s \cdot r|_p = \left(\frac{1}{p}\right)^s.$$

We observe that the more often an integer is divisible by p, the smaller is its p-adic absolute value.

Definition 1.2. The *p-adic integers* $\mathbb{Z}_p$ are defined to be the completion of the integers $\mathbb{Z}$ by the p-adic absolute value.

This means that a p-adic integer is of the form

$$a_0 + a_1 \cdot p + a_2 \cdot p^2 + a_3 \cdot p^3 + \cdots \quad \text{for some integers } a_i \ (0 \leq a_i < p),$$

because $|p^n|_p = \left(\frac{1}{p}\right)^n$ goes to zero as n goes to infinity.

Definition 1.3. The *p-adic numbers* $\mathbb{Q}_p$ are the fraction field of the p-adic integers $\mathbb{Z}_p$ or, equivalently, the completion of the rational numbers $\mathbb{Q}$ with respect to p-adic absolute value (that can be extended to the rational numbers).

Similarly to the p-adic integers, a p-adic number can be represented as

$$a_{-n} \cdot p^{-n} + \cdots + a_0 + a_1 \cdot p + a_2 \cdot p^2 + \cdots \quad \text{with } a_i \in \{0, \ldots, p-1\}.$$

While the p-adic numbers are, like the real numbers $\mathbb{R}$, a completion of the rational numbers, the topological space of the p-adic numbers arising from the p-adic absolute value is very different from the topology on the real numbers. The p-adic numbers turn out to be totally disconnected. To illustrate the topology on the 3-adic integers as an example, note that the integers 0, 1 and 2 all have 3-adic distance 1 from each other. The integers 0, 3, and 6 all have 3-adic distance $\frac{1}{3}$ from each other, and similarly the integers 1, $1 + 3 = 4$, $1 + 2 \cdot 3 = 7$ all have 3-adic distance $\frac{1}{3}$ from each other, while any of the integers 0, 3 or 6 (pictured in the lower left triangle of the middle figure in Figure 1) has 3-adic distance 1 from any of the integers 1, $1 + 3 = 4$, $1 + 2 \cdot 3 = 7$ (pictured in the lower right triangle of the middle figure in Figure 1). Hence it is impossible to embed the integers with their 3-adic distance in the Euclidean plane, but Figure 1 tries to give a visualization of the 3-adic distances of the integers of the form $a_0 + a_1 \cdot 3 + a_2 \cdot 3^2$ with $a_0, a_1, a_2 \in \{0, 1\}$

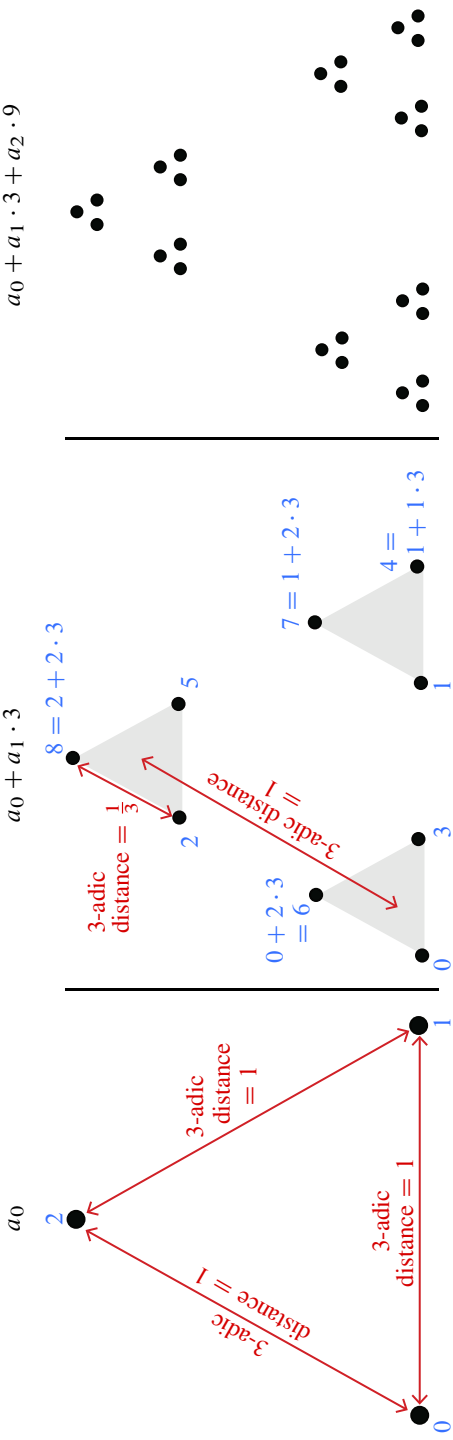


Figure 1. A visualization of the 3-adic integers based on [9].

following [9]. We hope that these figures allow the reader to imagine the fractal-like structure of the 3-adic integers by adding more and more 3-adic "digits" in the presentation $a_0 + a_1 \cdot 3 + a_2 \cdot 3^2 + a_3 3^3 + \cdots$. In particular, the p-adic integers are a compact group whose underlying topological space is totally disconnected.

As mentioned above, the real numbers $\mathbb{R}$ and the p-adic numbers $\mathbb{Q}_p$ arise both as completions of the rational numbers $\mathbb{Q}$, but for a different absolute value. Since the real numbers are connected, more precisely, they are a line, while the p-adic numbers are totally disconnected, we have seen that the resulting topological spaces are very different. Another key difference between these fields consists of the following observation: The real numbers contain only one compact subgroup under addition, which is the trivial subgroup consisting only of the element 0. The p-adic numbers, on the other hand, contain infinitely many compact (and open) subgroups under addition, e.g., the p-adic integers $\mathbb{Z}_p$ form a compact, open subgroup of $\mathbb{Q}_p$, and similarly $p \cdot \mathbb{Z}_p$ is a compact, open subgroup, and more generally $p^n \mathbb{Z}_p$ for any integer n is a compact, open subgroup of $\mathbb{Q}_p$. This means in particular that the element 0 in $\mathbb{Q}_p$ has a basis of open neighborhoods consisting of compact, open subgroups. This property translates into similar properties for p-adic groups that shape the flavor of the study of their representation theory.

Before introducing p-adic groups, we would like to mention that the p-adic numbers have a sister field that behaves very similarly to the p-adic numbers: the Laurent series $\mathbb{F}_p((t))$ over a finite field $\mathbb{F}_p$ with p elements, elements of which can be written as

$$a_{-n} \cdot t^{-n} + \cdots + a_0 + a_1 \cdot t + a_2 \cdot t^2 + \cdots \quad \text{with } a_i \in \mathbb{F}_p,$$

where t is an abstract variable. There is an obvious similarity between $\mathbb{Q}_p$ and $\mathbb{F}_p((t))$ based on the way we presented their elements, and the reader is welcome to choose whichever field they feel more comfortable with when reading this article. The key difference between the two fields is that the field of p-adic numbers $\mathbb{Q}_p$ has characteristic zero, i.e., it contains the rational numbers $\mathbb{Q}$ as a subfield, while the field of Laurent series $\mathbb{F}_p((t))$ has characteristic p, i.e., it contains the finite field $\mathbb{F}_p$ with p elements as a subfield in which $p = 0$. From now on, we write F for either of the two fields $\mathbb{Q}_p$ or $\mathbb{F}_p((t))$. While we focus on these two fields in this article, everything covered here also works for finite extensions of these two fields, which are called *non-archimedean local fields*.

In order to introduce p-adic groups, we recall that examples of real Lie groups include:

- the general linear group $\mathrm{GL}_n(\mathbb{R})$, i.e., $n \times n$-matrices whose determinant is non-zero with the group action being multiplication,
- the special linear group $\mathrm{SL}_n(\mathbb{R})$, which is the subgroup of $\mathrm{GL}_n(\mathbb{R})$ consisting of those matrices whose determinant is 1,

- the special orthogonal group $\mathrm{SO}_n(\mathbb{R})$, which is the subgroup of $\mathrm{SL}_n(\mathbb{R})$ consisting of those matrices A satisfying $AA^T = \mathrm{Id}$, where Id denotes the identity matrix and A^T the transpose of A; in other words, these are the matrices that preserve the standard inner product,
- the symplectic group $\mathrm{Sp}_{2n}(\mathbb{R})$, which is the subgroup of $\mathrm{GL}_{2n}(\mathbb{R})$ that preserves a symplectic form.

Replacing the real numbers $\mathbb{R}$ in the above examples by the field F, we obtain examples of what we call *p-adic groups*: $\mathrm{GL}_n(F)$, i.e., $\mathrm{GL}_n(\mathbb{Q}_p)$ and $\mathrm{GL}_n(\mathbb{F}_p((t)))$, $\mathrm{SL}_n(F)$, $\mathrm{SO}_n(F)$, $\mathrm{Sp}_{2n}(F)$. This means p-adic groups are nice subgroups of matrix groups with entries in either the field of p-adic numbers $\mathbb{Q}_p$ or its sister field $\mathbb{F}_p((t))$. Similar to real Lie groups, there are also p-adic groups of type G_2, F_4, E_6, E_7, E_8, and these groups are in general classified using combinatorial data (root data) and additional structure related to the absolute Galois group of F. Note that we use the terminology "p-adic groups" for these nice subgroups of $\mathrm{GL}_n(F)$ even though F might be either the p-adic numbers $\mathbb{Q}_p$ or the Laurent series $\mathbb{F}_p((t))$ over the finite field $\mathbb{F}_p$. For the experts, the results we discuss below hold for groups $\underline{G}(F)$, where $\underline{G}$ is a connected reductive group over F that splits over a tamely ramified field extension of F, but the reader new to this topic is encouraged to just think of the explicit examples mentioned above. For a more detailed introduction to p-adic groups including precise definitions, we refer the interested reader to the survey article [14, Section 2].

From now on, we denote by G a p-adic group.

2. Representations of p-adic groups

Definition 2.1. A *smooth representation* of G is a group homomorphism

$$\pi : G \to \mathrm{Aut}_{\mathbb{C}}(V)$$

for some complex vector space V such that for every $v \in V$, there exists an open subgroup $K_v \subset G$ with $\pi(k)(v) = v \; \forall \, k \in K_v$.

The vector space V underlying a smooth representation of G is usually infinite dimensional. If $G = \mathrm{SL}_2(\mathbb{Q}_p)$, then we can take as the K_v appearing in the above definition subgroups of the form $\left(\begin{smallmatrix} 1+p^m\mathbb{Z}_p & p^m\mathbb{Z}_p \\ p^m\mathbb{Z}_p & 1+p^m\mathbb{Z}_p \end{smallmatrix}\right)_{\det=1}$ for some positive integers m.

Smooth representations of p-adic groups have been studied for more than 50 years, with a long-term goal of understanding the whole category of all smooth representations of p-adic groups as explicitly as possible.

Long-term goal. Understand the category of all smooth representations of G.

Before explaining the structure of the category of representations of p-adic groups, we would like to provide the reader with some examples and a procedure to construct many representations that is used to approach the above long-term goal.

2.1. Example of a representation

Let $G = \mathrm{SL}_2(\mathbb{Q}_p)$ with subgroup $B = \left(\begin{smallmatrix} \mathbb{Q}_p & \mathbb{Q}_p \\ 0 & \mathbb{Q}_p \end{smallmatrix}\right)_{\det=1}$.

We set

$$\begin{aligned} V &:= \{f : B\backslash G \to \mathbb{C} \mid f \text{ locally constant}\} \\ &\simeq \{f : \mathbb{P}^1(\mathbb{Q}_p) \to \mathbb{C} \mid f \text{ locally constant}\}, \end{aligned}$$

which is an infinite-dimensional complex vector space, where $\mathbb{P}^1(\mathbb{Q}_p)$ denotes the projective line over $\mathbb{Q}_p$. The map $\pi : G \to \mathrm{Aut}_{\mathbb{C}}(V)$ defining our smooth representation is given as follows: For $g \in G$, we have

$$\begin{aligned} \pi(g) : \quad & V & \longrightarrow \quad & V, \\ \pi(g) : \quad & x \mapsto f(x) & \longmapsto \quad & x \mapsto f(xg). \end{aligned}$$

The requirement that the functions in V are locally constant implies that this is a smooth representation.

2.2. Parabolic induction

We would like to generalize the example of the previous subsection. To do so, we introduce the following class of subgroups of G which will replace the subgroup B above by.

Definition 2.2. A *parabolic subgroup* of $G = \mathrm{GL}_n(F)$ (or $\mathrm{SL}_n(F)$, $\mathrm{SO}_n(F)$ or $\mathrm{Sp}_{2n}(F)$) is a subgroup of the form

$$g \cdot \begin{pmatrix} \star & \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star & \star \\ & & \star & \star & \star & \star \\ & & & \star & \star & \star \\ & 0 & & \star & \star & \star \\ & & & \star & \star & \star \end{pmatrix} \cdot g^{-1}$$

for some choice of number and sizes of blocks (e.g., in the centered expression in this definition, we chose three blocks of sizes $2 \times 2, 1 \times 1, 3 \times 3$ for $n = 6$ to illustrate the shape of the upper-block-triangular matrices that we consider) and some choice of $g \in G$.

For a more abstract general definition, we refer the interested reader to [14, Section 2].

Any parabolic subgroup admits a decomposition as a semi-direct product of another p-adic group, called a *Levi subgroup* and denoted by M below, corresponding to (the g-conjugate of) block-diagonal matrices, and a normal subgroup, called the *unipotent radical* of the parabolic subgroup and denoted by U below, corresponding to (the g-conjugate of) the strictly-upper-block-triangular matrices with ones on the diagonal:

$$\underbrace{g\begin{pmatrix} \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ & & \star & \star & \star \\ & 0 & & \star & \star & \star \\ & & & \star & \star & \star \end{pmatrix} g^{-1}}_{P} = \underbrace{g\begin{pmatrix} \star & \star & & & \\ \star & \star & & 0 & \\ & & \star & & \\ & 0 & & \star & \star & \star \\ & & & \star & \star & \star \end{pmatrix} g^{-1}}_{M} \cdot \underbrace{g\begin{pmatrix} 1 & 0 & \star & \star & \star \\ 0 & 1 & \star & \star & \star \\ & & 1 & \star & \star & \star \\ & 0 & & 1 & 0 & 0 \\ & & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{pmatrix} g^{-1}}_{U}.$$

If P is a proper parabolic subgroup, i.e., $P \subsetneq G$, then the Levi subgroup M is a smaller p-adic group. We now introduce a construction to build smooth representations of G from smooth representations of M.

Definition 2.3 (Parabolic induction). Let $P = M \ltimes U$ be a parabolic subgroup of G. Let $\sigma : M \to \mathrm{Aut}_{\mathbb{C}}(V_\sigma)$ be a smooth representation of M. The *parabolic induction* $\mathrm{Ind}_P^G V_\sigma$ is the representation of G with vector space

$$\mathrm{Ind}_P^G V_\sigma := \{f : G \to V_\sigma \mid f(mug) = \sigma(m)(f(g))\ \forall\, m \in M,\, u \in U,\, g \in G;\\ f \text{ locally constant}\},$$

and for all $g \in G$,

$$\begin{aligned} \mathrm{Ind}_P^G \sigma(g) : \quad \mathrm{Ind}_P^G V_\sigma \quad &\longrightarrow \quad \mathrm{Ind}_P^G V_\sigma, \\ \mathrm{Ind}_P^G \sigma(g) : \quad x \mapsto f(x) \quad &\longmapsto \quad x \mapsto f(xg). \end{aligned}$$

Note that we recover the example discussed in Section 2.1 if we set $G = \mathrm{SL}_2(\mathbb{Q}_p)$, $P = B$ and $V_\sigma = \mathbb{C}$ with $\sigma : G \mapsto 1 \in \mathbb{C}^\times = \mathrm{Aut}_{\mathbb{C}}(\mathbb{C})$.

2.3. Supercuspidal representations: The building blocks

Definition 2.4. A smooth representation $\pi : G \to \mathrm{Aut}_{\mathbb{C}}(V)$ is called *irreducible* if it has exactly two subrepresentations, i.e., if there exist exactly two subspaces $W \subseteq V$ (which are $\{0\}$ and V) that are preserved under $\pi(g)$ for all $g \in G$.

Definition 2.5. An irreducible smooth representation $\pi : G \to \mathrm{Aut}_{\mathbb{C}}(V)$ is called *supercuspidal* if it is not a subrepresentation of $\mathrm{Ind}_P^G V_\sigma$ for every proper parabolic subgroup $P = M \ltimes U \subsetneq G$ and every irreducible smooth representation $\sigma : M \to \mathrm{Aut}_{\mathbb{C}}(V_\sigma)$ of M.

Fact 2.6. Let $\pi : G \to \mathrm{Aut}_{\mathbb{C}}(V)$ be an irreducible smooth representation of G. Then there exists a parabolic subgroup $P = M \ltimes U \subseteq G$ and a supercuspidal representation (σ, V_σ) of M such that V is a subrepresentation of $\mathrm{Ind}_P^G V_\sigma$, which we write as

$$V \hookrightarrow \mathrm{Ind}_P^G V_\sigma.$$

Thus, the supercuspidal representations form the building blocks of all smooth representations of p-adic groups. Therefore, our long-term goal to understand all representations of p-adic groups can be divided into two steps: constructing all supercuspidal representations and understanding how the whole category of smooth representations is built up from the supercuspidal representations.

Problem 1. Construct all supercuspidal representations.

Solving this first problem of constructing all supercuspidal representations explicitly has plenty of applications within the representation theory of p-adic groups, e.g., when studying how representations of G behave when we restrict them to subgroups, or how they behave under various operations done to representations of p-adic groups. It is also necessary to describe supercuspidal representations explicitly if one wants to construct an explicit local Langlands correspondence, which means attaching to each representation of a p-adic group a number-theoretic parameter (roughly a representation of the absolute Galois group of F) that bridges representation theory and number theory. There are also applications to automorphic forms, which are a vast generalization of modular forms, see [20] for an example, and many more applications beyond the representation theory of p-adic groups itself.

The vague answer to Problem 1 is that we can do this under minor assumptions. We will make this more precise in Section 2.6. For now, we first assume that we can construct supercuspidal representations, and we try to understand how the whole category of smooth representations of G looks.

2.4. The category of smooth representations

By Bernstein [4], the category $\mathrm{Rep}(G)$ of all smooth representations of G decomposes into a product of smaller subcategories as follows:

$$\underbrace{\mathrm{Rep}(G)}_{\substack{\text{smooth}\\ \text{representations}}} = \prod_{(M,\sigma)/\sim} \underbrace{\mathrm{Rep}(G)_{[M,\sigma]}}_{\substack{\text{Bernstein}\\ \text{block}}}. \tag{2.1}$$

The decomposition is indexed by pairs consisting of a Levi subgroup M of (a parabolic subgroup of) G and a supercuspidal representation σ of M, where the pairs (M, σ) are considered up to an equivalence relation that ensures that each Bernstein block appears only once in the product. More precisely, two pairs (M, σ) and (M', σ') are equivalent if there exist $g \in G$ and a one-dimensional representation $\phi : M \to \mathbb{C}^\times$ that is trivial on all compact subgroups of M such that

$$M = gM'g^{-1} \quad \text{and} \quad \sigma(m) = \sigma'(g^{-1}mg) \cdot \phi(m) \quad \text{for } m \in M.$$

Given such a pair (M, σ), the attached *Bernstein block* $\mathrm{Rep}(G)_{[M,\sigma]}$ consists of all those representations all of whose irreducible subquotients are contained in some $\mathrm{Ind}_{P'}^G \sigma'$ for some parabolic P' with Levi M' and with (M', σ') equivalent to (M, σ). In particular, the Bernstein block $\mathrm{Rep}(G)_{[M,\sigma]}$ contains the parabolic induction $\mathrm{Ind}_P^G \sigma$ for any parabolic subgroup P with Levi subgroup M, and it can also be characterized as the block that contains $\mathrm{Ind}_P^G \sigma$. Moreover, the Bernstein blocks are indecomposable, i.e., they are not themselves products of smaller non-trivial subcategories, which is why they merit being called *blocks*.

For example, if $G = \mathrm{SL}_2(\mathbb{Q}_p)$, then up to conjugation, we have two choices for M:

$$M = G \quad \text{or} \quad M = T := \left\{ \begin{pmatrix} \boxed{t} & 0 \\ 0 & \boxed{t^{-1}} \end{pmatrix} \mid t \in \mathbb{Q}_p^\times \right\}.$$

In the case $M = G$, let $\sigma : G \to \mathrm{Aut}_{\mathbb{C}}(V_\sigma)$ be a supercuspidal representation of G. Then the Bernstein block $\mathrm{Rep}(\mathrm{SL}_2(\mathbb{Q}_p))_{[\mathrm{SL}_2(\mathbb{Q}_p),\sigma]}$ contains as objects arbitrary finite and infinite direct sums of σ, i.e., σ, $\sigma \oplus \sigma$, $\sigma \oplus \sigma \oplus \sigma, \ldots$ are objects of this Bernstein block. Since σ is irreducible, by Schur's lemma, the only morphisms between V_σ and V_σ that commute with $\sigma(g)$ for all $g \in G$ are multiplication by a scalar; in other words, $\mathrm{Hom}_G(V_\sigma, V_\sigma) = \mathbb{C}$. From this, one can deduce the structure of the morphisms between all other objects in this Bernstein block.

The Bernstein blocks with more complicated structures arise when one considers $M = T$. As an example, we write triv for the trivial one-dimensional representation of T, i.e., $V_{\mathrm{triv}} = \mathbb{C}$ and $\mathrm{triv} : T \mapsto 1 \in \mathbb{C}^\times = \mathrm{Aut}_{\mathbb{C}}(V_{\mathrm{triv}})$. The Bernstein block $\mathrm{Rep}(\mathrm{SL}_2(\mathbb{Q}_p))_{[T,\mathrm{triv}]}$ is called the *principal block* and contains by definition $\mathrm{Ind}_B^G \mathrm{triv}$, where $B = \left(\begin{smallmatrix} \mathbb{Q}_p & \mathbb{Q}_p \\ 0 & \mathbb{Q}_p \end{smallmatrix}\right)_{\det=1}$ as above. The representation $\mathrm{Ind}_B^G \mathrm{triv}$ contains the trivial one-dimensional representation triv of G as a subrepresentation, and the quotient $\mathrm{Ind}_B^G \mathrm{triv} / \mathrm{triv}$ is another irreducible smooth representation of G that is called the Steinberg representation and is denoted by St. Hence the trivial representation and the Steinberg representation are both also contained in the principal block, and so is their sum triv $\oplus$St. However, the sum triv $\oplus$St is not isomorphic to the parabolic induction $\mathrm{Ind}_B^G \mathrm{triv}$, even though both contain the same subquotients, so the principal block exhibits a more complicated structure.

Nevertheless, it turns out that we can describe this structure rather well as follows: The principal block $\mathrm{Rep}(\mathrm{SL}_2(\mathbb{Q}_p))_{[T,\mathrm{triv}]}$ is equivalent to the category of (unital right) modules over the affine Hecke algebra $\mathcal{H}_{\mathrm{aff}}$ that is defined as follows: A basis for the complex vector space underlying the affine Hecke algebra $\mathcal{H}_{\mathrm{aff}}$ is given by

$$\{T_w \mid w \in W_{\mathrm{aff}} := \langle s_0, s_1 \mid s_i^2 = 1 \rangle\}.$$

Note that the abstractly defined group W_{aff} is the group of transformations of a plane that is generated by two reflections across two parallel hyperplanes. The relations that describe the multiplication in the affine Hecke algebra are generated by

$$T_{s_i} T_w = \begin{cases} T_{s_i w}, & \ell(s_i w) > \ell(w), \\ p T_{s_i w} + (p-1) T_w, & \ell(s_i w) < \ell(w), \end{cases}$$

where ℓ denotes the length function $W_{\mathrm{aff}} \to \mathbb{Z}_{\geq 0}$ that assigns to an element w in the affine Weyl group W_{aff} the length of the shortest expression $s_0 s_1 s_0 \cdots$ or $s_1 s_0 s_1 \cdots$ that equals w. Finally, the modules of the affine Hecke algebra $\mathcal{H}_{\mathrm{aff}}$ are well understood.

By the Bernstein decomposition (2.1), describing the category of smooth representations reduces to the following problem, in addition to Problem 1.

Problem 2. Describe the Bernstein block $\mathrm{Rep}(G)_{[M,\sigma]}$ explicitly.

2.5. Bernstein blocks: Hecke algebras and reduction to depth zero

We sketch an answer to Problem 2 based on our recent preprints [2, 3] that are joint work with Adler, Mishra and Ohara and that were still works in progress during the ECM. For a more detailed description, see the survey [16].

The answer to Problem 2, assuming the prime p is not very small (analogous to the assumptions made for the construction of supercuspidal representations discussed in more detail in Section 2.6 below), which we will assume from now on, consists of two parts. The first result is of independent interest.

Vague Theorem 2.7 (Adler–Fintzen–Mishra–Ohara [2, 3]). *Given a Bernstein block* $\mathrm{Rep}(G)_{[M,\sigma]}$*, we have an equivalence of Bernstein blocks*

$$\mathrm{Rep}(G)_{[M,\sigma]} \simeq \mathrm{Rep}(G^0)_{[M^0,\sigma_0]}$$

for some subgroup $G^0 \subseteq G$ *(which becomes a Levi subgroup over a field extension), some Levi subgroup* M^0 *of a parabolic subgroup of* G^0 *and some depth zero supercuspidal representation* σ_0 *of* M^0*.*

We will say a few more words about *depth zero* supercuspidal representations below, but the main feature of these representations is that they essentially correspond to representations of finite groups of Lie type. Hence the above result allows us to

reduce a lot of questions in the representation theory of p-adic groups and the explicit and categorical local Langlands program to representations of depth zero, where answers are often either already known or much easier to tackle. Such a result has previously only been available for some special cases of G and in the two "extreme" cases of $M = T$, where T denotes a split maximal torus, which means in terms of Definition 2.2 that all the blocks have size 1×1, by Roche [32], and $M = G$, by Ohara [30].

The second part to answer Problem 2 is that we already understood the structure of the Bernstein block $\mathrm{Rep}(G^0)_{[M^0,\sigma_0]}$ very explicitly in many cases since the early 1990s, thanks to Morris [25], and the general case is now treated in [3]: We can describe each depth zero Bernstein block for all p-adic groups explicitly as modules over an explicit algebra $\mathcal{H}$ with explicit generators and relations. More precisely, we proved that the Hecke algebra $\mathcal{H}$ is isomorphic to a semi-direct product of an affine Hecke algebra $\mathcal{H}_{\mathrm{aff}}$ (generalizing the affine Hecke algebra that we have seen in the example above, allowing more general affine Weyl groups) with a twisted group algebra $\mathbb{C}[\Omega, \mu]$; see [3, Notation 3.10.8] for more precise definitions.

Vague Theorem 2.8 (Morris [25] and Adler–Fintzen–Mishra–Ohara [3]).

$$\mathrm{Rep}(G^0)_{[M^0,\sigma_0]} \simeq \mathbb{C}[\Omega, \mu] \ltimes \mathcal{H}_{\mathrm{aff}} - \mathrm{mod}.$$

Thus, combining Vague Theorems 2.7 and 2.8, we obtain the following result:

$$\mathrm{Rep}(G)_{[M,\sigma]} \simeq \mathrm{Rep}(G^0)_{[M^0,\sigma_0]} \simeq \mathbb{C}[\Omega, \mu] \ltimes \mathcal{H}_{\mathrm{aff}} - \mathrm{mod}.$$

In fact, the first equivalence (Vague Theorem 2.7) is proven via an explicit isomorphism of Hecke algebras. Since modules over affine Hecke algebras and more general modules over the semi-direct products $\mathbb{C}[\Omega, \mu] \ltimes \mathcal{H}_{\mathrm{aff}}$ are well studied and understood, this allows us to also understand the structure of the previously mysterious Bernstein blocks.

2.6. Construction of supercuspidal representations for general p-adic groups

We return to Problem 1, the construction of (all) supercuspidal representations of p-adic groups.

For about 50 years, mathematicians have tried to construct these supercuspidal representations. In 1979, Carayol [8] gave a construction of supercuspidal representations that provides all supercuspidal representations of $\mathrm{GL}_n(F)$ for n a prime number different from p. In 1986, Moy [27] proved that Howe's construction [22] from the 1970s exhausts all supercuspidal representations of $\mathrm{GL}_n(F)$ if n is an integer coprime to p. In the early 1990s, Bushnell and Kutzko extended these constructions to obtain all supercuspidal representations of $\mathrm{GL}_n(F)$ for arbitrary n [7]. Similar methods have later been exploited by Stevens [36] to construct all supercuspidal representations of

classical groups, i.e., symplectic, orthogonal and unitary groups, under the assumption that $p \neq 2$, and by Sécherre and Stevens [33] to construct all supercuspidal representations of inner forms of $\mathrm{GL}_n(F)$, i.e., $\mathrm{GL}_m(D)$ for some division algebra D. There has also been a lot of prior work of mathematicians solving special cases; for a few more details, see, e.g., [15].

The picture has been less complete for arbitrary reductive groups. The introduction of the Moy–Prasad filtration in the 1990s spurred remarkable progress. The work of Moy and Prasad built upon the innovative Bruhat–Tits theory introduced in the 1970s/1980s: In [5,6], Bruhat and Tits defined a building $\mathcal{B}(G,F)$ associated with G on which G acts. For each point y in $\mathcal{B}(G,F)$, they constructed a compact subgroup $G_{y,0}$ of G, called a *parahoric subgroup*, which is (up to finite index) the stabilizer G_y in G of the point y. In [28,29], Moy and Prasad defined a filtration of these parahoric subgroups by smaller, normal subgroups

$$G_{y,0} \rhd G_{y,s_1} \rhd G_{y,s_2} \rhd G_{y,s_3} \rhd \cdots,$$

where $0 < s_1 < s_2 < \cdots$ are real numbers depending on y. These subgroups play a crucial role in the study and construction of supercuspidal representations. In some sense, one can heuristically imagine the subgroups G_{y,s_i} as the stabilizer of all points in the Bruhat–Tits building that are within a distance s_i of y. We set $G_{y,r} := G_{y,s_i}$ for $s_{i-1} < r \leq s_i$. The quotient $G_{y,0}/G_{y,s_1}$ is a finite group of Lie type.

For example, if we take $G = \mathrm{SL}_2(\mathbb{Q}_p)$, then the Bruhat–Tits building is an infinite tree with valency $p+1$; see Figure 2 (for $p=2$). Let y be the barycenter of a maximal facet, i.e., the center of an edge of the infinite tree, and let x be a vertex. Then (by choosing an appropriate basis) the associated Moy–Prasad filtrations at the points x and y look like the following:

$$G_{x,0} = \begin{pmatrix} \mathbb{Z}_p & \mathbb{Z}_p \\ \mathbb{Z}_p & \mathbb{Z}_p \end{pmatrix}_{\det=1}, \qquad G_{y,0} = \begin{pmatrix} \mathbb{Z}_p & p\mathbb{Z}_p \\ \mathbb{Z}_p & \mathbb{Z}_p \end{pmatrix}_{\det=1},$$

$$G_{y,0.5} = \begin{pmatrix} 1+p\mathbb{Z}_p & p\mathbb{Z}_p \\ \mathbb{Z}_p & 1+p\mathbb{Z}_p \end{pmatrix}_{\det=1},$$

$$G_{x,1} = \begin{pmatrix} 1+p\mathbb{Z}_p & p\mathbb{Z}_p \\ p\mathbb{Z}_p & 1+p\mathbb{Z}_p \end{pmatrix}_{\det=1}, \qquad G_{y,1} = \begin{pmatrix} 1+p\mathbb{Z}_p & p^2\mathbb{Z}_p \\ p\mathbb{Z}_p & 1+p\mathbb{Z}_p \end{pmatrix}_{\det=1},$$

$$G_{y,1.5} = \begin{pmatrix} 1+p^2\mathbb{Z}_p & p^2\mathbb{Z}_p \\ p\mathbb{Z}_p & 1+p^2\mathbb{Z}_p \end{pmatrix}_{\det=1},$$

$$G_{x,2} = \begin{pmatrix} 1+p^2\mathbb{Z}_p & p^2\mathbb{Z}_p \\ p^2\mathbb{Z}_p & 1+p^2\mathbb{Z}_p \end{pmatrix}_{\det=1}, \qquad G_{y,2} = \begin{pmatrix} 1+p^2\mathbb{Z}_p & p^3\mathbb{Z}_p \\ p^2\mathbb{Z}_p & 1+p^2\mathbb{Z}_p \end{pmatrix}_{\det=1},$$

$$\vdots, \qquad \vdots.$$

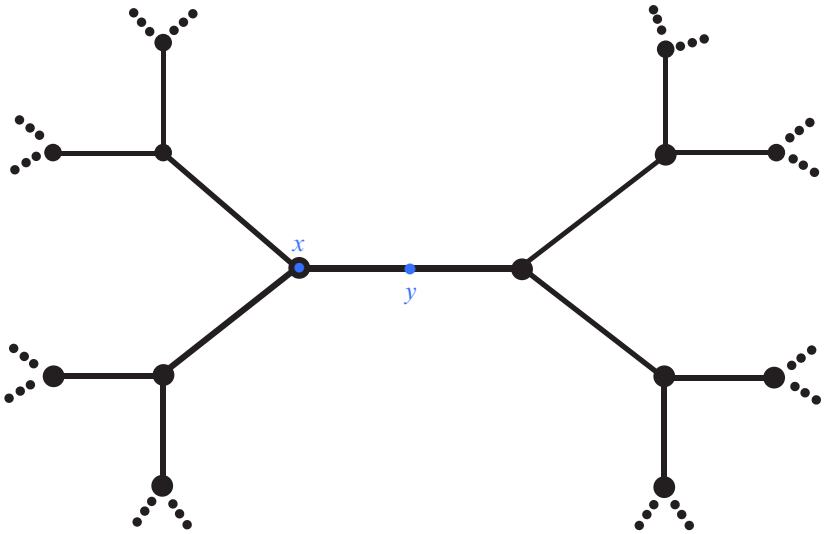


Figure 2. Bruhat–Tits building for $G = \mathrm{SL}_2(\mathbb{Q}_2)$.

Based on this filtration, Moy and Prasad introduced in [28,29] the notion of *depth* of a representation, which measures the first occurrence of a fixed vector in a given representation. More precisely, it is the smallest non-negative rational number r such that the representation contains a vector fixed under $G_{y,r+} := \bigcup_{s>r} G_{y,s}$ for some $y \in \mathcal{B}(G, F)$. In [29], Moy and Prasad gave a classification of depth zero representations, showing, roughly speaking, that they correspond to representations of finite groups of Lie type, the group $G_{y,0}/G_{y,0+}$. A similar result was obtained shortly afterward by Morris [26] using Hecke algebra techniques.

The first construction of positive-depth supercuspidal representations for general p-adic groups[1] was given by Adler [1] in 1998 and generalized by Yu [37] in 2001. Since then, Yu's construction has been widely used. However, about 20 years later, it was noticed by Spice that Yu's proof relies on a misprinted (and therefore false) statement in [21], and therefore, it became uncertain for a few years whether the representations constructed by Yu are irreducible and supercuspidal. In [10], we illustrate the impact of this false statement on Yu's proof by providing a counterexample to Proposition 14.1 and Theorem 14.2 of [37], and at the same time, we offer a different argument for the second half of Yu's proof that avoids the false statements and show that, nevertheless, Yu's construction yields supercuspidal representations.

However, Proposition 14.1 and Theorem 14.2 of [37] are the main intertwining results in [37] that not only formed the heart of the proof of supercuspidality but

[1]For experts, as mentioned above, the groups considered are reductive groups over non-archimedean local fields that split over a tamely ramified field extension.

are also crucial for other applications. In [17], in joint work with Kaletha and Spice, we therefore construct a quadratic character that allows us to twist Yu's construction so that Yu's initial proof and also the intertwining results hold true for Yu's construction twisted by our quadratic character. This technical result has been the key to calculating Harish-Chandra character formulas [34, 35], constructing a local Langlands correspondence for a large class of representations [23] and obtaining Vague Theorems 2.7 and 2.8 above.

Since many results in our area are only valid for the representations constructed by Yu, one might hope that all supercuspidal representations arise in this way. In 2007, Kim [24] showed that Yu's representations indeed include all supercuspidal representations if F has characteristic zero, i.e., if we work with $\mathbb{Q}_p$, but not with $\mathbb{F}_p((t))$, and if in addition the prime p is very large. In 2021, we used completely different and much more explicit techniques in [13] to show that *all* supercuspidal representations of G are obtained from Yu's construction under much more mild assumptions on p, more precisely that p does not divide the order of the absolute Weyl group of G, which excludes for exceptional groups at most the primes 2, 3, 5 and 7. In particular, our result holds for any F; we do not distinguish between $\mathbb{Q}_p$ and its sister $\mathbb{F}_p((t))$. Moreover, in [12], we illustrate that these assumptions are optimal in general (at least if one considers a slightly more general setup than we discuss in these notes). This settles the long search for all supercuspidal representations under minor assumptions.

On the other hand, this also means that there are more representations to be found for small p. In 2014, Reeder and Yu [31] gave a new construction of some supercuspidal epipelagic representations. Epipelagic representations are representations of smallest positive depth. In [11, 18], we showed that the input for Reeder–Yu's construction also exists for small primes p, which provides examples of positive-depth supercuspidal representations that do not arise from Yu's construction. These representations have already found various applications in the study of automorphic representations and the inverse Galois problem. However, because these representations are very special and far from exhaustive, many more representations remain to be discovered in the future, a large class of which are the subject of current work in progress. A first big step in this direction is the preprint [19] from this year in which we extend in joint work with David Schwein the construction by Yu to also work for $p = 2$, a question that had remained open for about 25 years due to the use of the theory of Heisenberg–Weil representations in Yu's construction, which relied on $p \neq 2$. We also remove a superfluous genericity condition by overcoming the challenges of working with disconnected groups, which aligns our construction more closely with potential geometric approaches. This yields new supercuspidal representations of arbitrary depth (for small primes p).

All the supercuspidal representations constructed above are of the following form:

$$\text{c-ind}_K^G V_\rho := \{f : G \to V_\rho \mid f(kg) = \rho(k)(f(g)) \ \forall g \in G,\ k \in K \\ f \text{ compactly supported mod center}\}$$

for some compact-mod-center, open subgroup $K \subset G$, e.g., $\mathrm{SL}_2(\mathbb{Z}_p) \subset \mathrm{SL}_2(\mathbb{Q}_p)$, and some finite-dimensional representation (ρ, V_ρ) of K. The difficulty lies in finding appropriate K and ρ that yield (all) supercuspidal representations.

Acknowledgments. The author thanks the referee for spotting some typos.

Funding. The author was partially supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 950326).

References

[1] J. D. Adler, Refined anisotropic K-types and supercuspidal representations. *Pacific J. Math.* **185** (1998), no. 1, 1–32 Zbl 0924.22015 MR 1653184

[2] J. D. Adler, J. Fintzen, M. Mishra, and K. Ohara, Reduction to depth zero for tame p-adic groups via Hecke algebra isomorphisms. 2024, arXiv:2408.07805v1

[3] J. D. Adler, J. Fintzen, M. Mishra, and K. Ohara, Structure of Hecke algebras arising from types. 2024, arXiv:2408.07801v1

[4] J. N. Bernstein, Le "centre" de Bernstein. In *Representations of reductive groups over a local field*, pp. 1–32, Travaux en Cours, Hermann, Paris, 1984 Zbl 0599.22016 MR 0771671

[5] F. Bruhat and J. Tits, Groupes réductifs sur un corps local. *Publ. Math. Inst. Hautes Études Sci.* **41** (1972), no. 1, 5–251 Zbl 0254.14017 MR 0327923

[6] F. Bruhat and J. Tits, Groupes réductifs sur un corps local: II. Schémas en groupes. Existence d'une donnée radicielle valuée. *Publ. Math. Inst. Hautes Études Sci.* **60** (1984), 197–376 Zbl 0597.14041 MR 0756316

[7] C. J. Bushnell and P. C. Kutzko, *The admissible dual of GL(N) via compact open subgroups*. Ann. of Math. Stud. 129, Princeton University Press, Princeton, NJ, 1993, 313 pp. Zbl 0787.22016 MR 1204652

[8] H. Carayol, Représentations supercuspidales de GL_n. *C. R. Acad. Sci. Paris Sér. A* **288** (1979), no. 1, A17–A20 Zbl 0398.22024 MR 0522009

[9] A. A. Cuoco, Visualizing the p-adic integers. *Amer. Math. Monthly* **98** (1991), no. 4, 355–364 Zbl 0748.11059 MR 1103194

[10] J. Fintzen, On the construction of tame supercuspidal representations. *Compos. Math.* **157** (2021), no. 12, 2733–2746 Zbl 1495.22009 MR 4357723

[11] J. Fintzen, On the Moy–Prasad filtration. *J. Eur. Math. Soc. (JEMS)* **23** (2021), no. 12, 4009–4063 Zbl 1492.22014 MR 4321207

[12] J. Fintzen, Tame tori in p-adic groups and good semisimple elements. *Int. Math. Res. Not. IMRN* **2021** (2021), no. 19, 14882–14904 Zbl 1507.22034 MR 4324731

[13] J. Fintzen, Types for tame p-adic groups. *Ann. of Math. (2)* **193** (2021), no. 1, 303–346 Zbl 1492.22013 MR 4199732

[14] J. Fintzen, Representations of p-adic groups. In *Current developments in mathematics 2021*, pp. 1–42, Int. Press, Somerville, MA, 2023 Zbl 1531.22007 MR 4649682

[15] J. Fintzen, Supercuspidal representations: construction, classification, and characters. In *The Langlands program: I. Representations, Shimura varieties, and Shtukas*, pp. 51–85, Proc. Sympos. Pure Math. 112, American Mathematical Society, Providence, RI, 2025 MR 5007738

[16] J. Fintzen, Reduction from representations of p-adic groups to representations of finite groups. In *Proceedings of the International Congress of Mathematicians*, 2026, https://www.math.uni-bonn.de/people/fintzen/FintzenICM.pdf visited on 23 April 2026

[17] J. Fintzen, T. Kaletha, and L. Spice, A twisted Yu construction, Harish–Chandra characters, and endoscopy. *Duke Math. J.* **172** (2023), no. 12, 2241–2301 Zbl 1536.22033 MR 4654051

[18] J. Fintzen and B. Romano, Stable vectors in Moy–Prasad filtrations. *Compos. Math.* **153** (2017), no. 2, 358–372 Zbl 1373.22024 MR 3705228

[19] J. Fintzen and D. Schwein, Construction of tame supercuspidal representations in arbitrary residue characteristic. 2025, arXiv:2501.18553v2

[20] J. Fintzen and S. W. Shin, Congruences of algebraic automorphic forms and supercuspidal representations. *Camb. J. Math.* **9** (2021), no. 2, 351–429 Zbl 1509.11040 MR 4325284

[21] P. Gérardin, Weil representations associated to finite fields. *J. Algebra* **46** (1977), no. 1, 54–101 Zbl 0359.20008 MR 0460477

[22] R. E. Howe, Tamely ramified supercuspidal representations of Gl_n. *Pacific J. Math.* **73** (1977), no. 2, 437–460 Zbl 0404.22019 MR 0492087

[23] T. Kaletha, Supercuspidal L-packets. [v1] 2019, [v2] 2021, arXiv:1912.03274v2, to appear in *Ann. Math.*

[24] J.-L. Kim, Supercuspidal representations: an exhaustion theorem. *J. Amer. Math. Soc.* **20** (2007), no. 2, 273–320 Zbl 1111.22015 MR 2276772

[25] L. Morris, Tamely ramified intertwining algebras. *Invent. Math.* **114** (1993), no. 1, 1–54 Zbl 0854.22022 MR 1235019

[26] L. Morris, Level zero **G**-types. *Compos. Math.* **118** (1999), no. 2, 135–157 Zbl 0937.22011 MR 1713308

[27] A. Moy, Local constants and the tame Langlands correspondence. *Amer. J. Math.* **108** (1986), no. 4, 863–930 Zbl 0597.12019 MR 0853218

[28] A. Moy and G. Prasad, Unrefined minimal K-types for p-adic groups. *Invent. Math.* **116** (1994), no. 1, 393–408 Zbl 0804.22008 MR 1253198

[29] A. Moy and G. Prasad, Jacquet functors and unrefined minimal K-types. *Comment. Math. Helv.* **71** (1996), no. 1, 98–121 Zbl 0860.22006 MR 1371680

[30] K. Ohara, Hecke algebras for tame supercuspidal types. *Amer. J. Math.* **146** (2024), no. 1, 277–293 Zbl 1535.22057 MR 4691489

[31] M. Reeder and J.-K. Yu, Epipelagic representations and invariant theory. *J. Amer. Math. Soc.* **27** (2014), no. 2, 437–477 Zbl 1284.22011 MR 3164986

[32] A. Roche, Types and Hecke algebras for principal series representations of split reductive p-adic groups. *Ann. Sci. Éc. Norm. Sup. (4)* **31** (1998), no. 3, 361–413 Zbl 0903.22009 MR 1621409

[33] V. Sécherre and S. Stevens, Représentations lisses de $\mathrm{GL}_m(D)$. IV. Représentations supercuspidales. *J. Inst. Math. Jussieu* **7** (2008), no. 3, 527–574 Zbl 1140.22014 MR 2427423

[34] L. Spice, Explicit asymptotic expansions for tame supercuspidal characters. *Compos. Math.* **154** (2018), no. 11, 2305–2378 Zbl 1409.22015 MR 3867302

[35] L. Spice, Explicit asymptotic expansions in p-adic harmonic analysis II. *Invent. Math.* **244** (2026), no. 3, 1147–1249 Zbl 08209953 MR 5048794

[36] S. Stevens, The supercuspidal representations of p-adic classical groups. *Invent. Math.* **172** (2008), no. 2, 289–352 Zbl 1140.22016 MR 2390287

[37] J.-K. Yu, Construction of tame supercuspidal representations. *J. Amer. Math. Soc.* **14** (2001), no. 3, 579–622 Zbl 0971.22012 MR 1824988

Jessica Fintzen
Mathematisches Institut, University of Bonn, Endenicher Allee 60, 53115 Bonn, Germany;
fintzen@math.uni-bonn.de

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/37

Recent progress on the problem of soliton resolution

Jacek Jendrej

Abstract. Dispersive partial differential equations are evolution equations (that is, involving the time variable) whose solutions preserve the energy but can still decay in large time due to the fact that various frequencies propagate with distinct velocities. In some cases, there exist special solutions called *solitons*, which do not change their shape as time passes. The soliton resolution conjecture predicts that, apart from exceptional cases, solitons are the only obstruction to the decay of solutions. More precisely, every solution eventually decomposes into a superposition of solitons and a decaying term called *radiation*.

We discuss the conjecture in the context of the critical wave maps equation, which is the analog of the wave equation for maps from $\mathbb{R}^2$ to $\mathbb{S}^2$. The solitons correspond to *harmonic maps*, which were classified by Eells and Wood in 1976. We consider *equivariant* solutions, which are solutions having a specific symmetry preserved by the flow. In a joint work with Andrew Lawrie, we prove that soliton resolution holds for these solutions. Our proof hinges on an analysis of *collisions* of solitons and an appropriate localized Lyapunov functional, which together allow us to prove a *no-return lemma* for multisoliton configurations.

Building on some of these ideas, we solve, in a joint work with Andrew Lawrie and Wilhelm Schlag, an analogous problem for the harmonic map heat flow of Eells and Sampson (1964) *without assuming any symmetry* of the initial data.

1. The phenomenon of dispersion

This section is devoted to standard introductory material. For a comprehensive introduction to the topic, the reader can consult, for instance, [37, 63].

1.1. The wave equation

Consider the *wave equation* in dimension $1 + 2$:

$$c^{-2}\partial_t^2\psi(t,x_1,x_2) = \partial_{x_1}^2\psi(t,x_1,x_2) + \partial_{x_2}^2\psi(t,x_1,x_2), \quad (t,x_1,x_2) \in \mathbb{R}^{1+2}. \tag{W}$$

The positive number $c > 0$ is the wave speed. Let us assume for simplicity that ψ is real valued, but it could just as well be vector valued. We will always write $\mathbb{R}^{1+2}$

Mathematics Subject Classification 2020: 35B40 (primary); 37K40 (secondary).
Keywords: dispersive PDE, soliton, wave map.

instead of $\mathbb{R}^3$ in order to stress that there is one time dimension and two space dimensions. Equation (W) is equivalent to requiring that ψ is a *critical point* of the Lagrangian

$$\mathcal{L}(\psi) := \iiint \left(\frac{1}{2c^2}(\partial_t \psi)^2 - \frac{1}{2}(\partial_{x_1}\psi)^2 - \frac{1}{2}(\partial_{x_2}\psi)^2 \right) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}t. \tag{1.1}$$

The precise meaning of this assertion is the following. Let $\zeta : \mathbb{R}^{1+2} \to \mathbb{R}$ be a smooth compactly supported function and $\psi_\varepsilon := \psi + \varepsilon\zeta$, where ε is a small real number. We then have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\varepsilon}|_{\varepsilon=0}\mathcal{L}(\psi_\varepsilon) &= \iiint \left(c^{-2}\partial_t\psi\partial_t\zeta - \partial_{x_1}\psi\partial_{x_1}\zeta - \partial_{x_2}\psi\partial_{x_2}\zeta\right) \mathrm{d}x_1\mathrm{d}x_2\mathrm{d}t \\ &= -\iiint \zeta\left(c^{-2}\partial_t^2\psi - \partial_{x_1}^2\psi - \partial_{x_2}^2\psi\right) \mathrm{d}x_1\mathrm{d}x_2\mathrm{d}t, \end{aligned}$$

where the last step is integration by parts. The left-hand side can be interpreted as the directional derivative of $\mathcal{L}$ at ψ in the direction ζ. Hence, we see that all the directional derivatives vanish if and only if ψ satisfies (W).

Equation (W) appears in many physical contexts, the most familiar being the evolution in time of a small disturbance of the membrane of a drum. It should be understood that the membrane extends to infinity and occupies the whole horizontal plane, and $\psi(t, x_1, x_2)$ is the vertical displacement at time t of the element of the membrane whose horizontal coordinates are (x_1, x_2). This is most easily seen using the Lagrangian formulation of classical mechanics. Indeed, let B be an infinitesimal ball in the horizontal plane, centered at (x_1, x_2). Then the area of the part of the membrane located above B is given by

$$|B|\sqrt{1 + (\partial_{x_1}\psi)^2 + (\partial_{x_2}\psi)^2} \simeq |B|\left(1 + \frac{1}{2}(\partial_{x_1}\psi)^2 + \frac{1}{2}(\partial_{x_2}\psi)^2\right),$$

where the last step uses the assumption that the disturbance is "small" in the sense that $|\nabla_{x_1,x_2}\psi| \ll 1$. Letting h be the height of the membrane and ρ its mass density, the leading term of the kinetic energy of this part of the membrane is thus

$$\rho h |B| (\partial_t \psi)^2.$$

The potential (elasticity) energy is proportional to the extension of the area and thus equals at the leading order

$$N_{\mathrm{rr}}^* |B| \left(\frac{1}{2}(\partial_{x_1}\psi)^2 + \frac{1}{2}(\partial_{x_2}\psi)^2\right),$$

where N_{rr}^* is the proportionality coefficient called the radial membrane resultant. In classical mechanics, the Lagrangian density is the difference of the kinetic and the

potential energy densities. Hence, up to a multiplicative constant, we obtain (1.1) with the wave speed $c := \sqrt{\frac{N^*_{\mathrm{tt}}}{\rho h}}$. Note that the total energy (again up to a multiplicative constant) is given by

$$E(\psi) := E_{\text{kinetic}} + E_{\text{potential}} = \iint \left(\frac{1}{2c^2}(\partial_t \psi)^2 + \frac{1}{2}(\partial_{x_1} \psi)^2 + \frac{1}{2}(\partial_{x_2} \psi)^2 \right) \mathrm{d}x_1 \mathrm{d}x_2$$

and is a conservation law (a quantity independent of time).

Remark 1.1. By an appropriate choice of units, one can assume that

$$c = 1,$$

which we will always do in the sequel. We will also write

$$x := (x_1, x_2), \quad \nabla := (\partial_{x_1}, \partial_{x_2}), \quad \Delta := \partial^2_{x_1} + \partial^2_{x_2}, \quad \mathrm{d}x := \mathrm{d}x_1 \mathrm{d}x_2.$$

Mechanical intuition suggests that in order to determine the evolution in time of the disturbance of the membrane, we need to specify the *initial conditions* consisting of the initial positions $\psi(0, x)$ and the initial velocities $\partial_t \psi(0, x)$ of all the elements of the drum. One can indeed prove that for any such initial conditions, equation (W) has a unique solution for all time, which moreover depends continuously on the initial conditions (in an appropriate sense which we will not make precise here), which is referred to as *global well-posedness*.

Having once more recourse to the intuition from mechanics, we can expect that if the membrane is initially perturbed only in a bounded region and flat elsewhere, then this disturbance will propagate in various directions, resulting in a decay of its amplitude, namely

$$\lim_{t \to \infty} \sup_{x \in \mathbb{R}^2} |\psi(t, x)| = 0, \tag{1.2}$$

which is referred to as *radiative behavior* and is related to the notion of the *group velocity*, which we now briefly present.

Consider *plane wave solutions* $\psi(t, x) = A \cos(k \cdot x - \omega t)$ with $k \in \mathbb{R}^2$ (the wave number), $\omega \in \mathbb{R}$ (the frequency) and $A \in \mathbb{R}$ (the amplitude). Plugging this formula into (W), we find the *dispersion relation*

$$\omega^2 = |k|^2 \Leftrightarrow \omega = \pm |k|.$$

We choose the sign "plus" and consider the so-called *slowly varying envelope ansatz*. We introduce slow time and space variables $(T, X) = (\varepsilon t, \varepsilon x)$, where $\varepsilon \ll 1$, and set

$$\psi(t, x) = A(T, X) \cos(k \cdot x - |k| t).$$

Then a quick computation using the chain rule yields

$$0 = \partial_t^2 \psi - \Delta \psi = 2\varepsilon |k| \left(\partial_T A + \frac{k}{|k|} \cdot \nabla_X A \right) \sin(k \cdot x - |k| t) + O(\varepsilon^2).$$

Hence, at main order, $A(T, X) = A_0(T - v \cdot X)$; in other words, the amplitude function is transported with the *group velocity* $v = v(k) = \frac{k}{|k|}$.

The above computation is the precise meaning of the claim that "the wave number k propagates with velocity $v(k)$". We see that all the wave numbers $k \neq 0$ propagate with speed 1, but the direction in which they propagate depends on k. Hence, the energy of the wave is dispersed in various directions, leading to (1.2). We say that (W) is an example of a *dispersive* linear PDE.

1.2. A few generalities on linear dispersive PDEs

The above computation leading to the computation of the group velocity can be carried out in much greater generality, which we prefer to leave to the reader. With this preparation, we can state the following heuristic definition.

Definition 1.2. A linear PDE is called *dispersive* if

(i) it is an evolution equation: it involves the time variable t and the space variable x,

(ii) various frequencies propagate with distinct velocities, in the sense that the group velocity defined analogously as above is not constant on any open ball of the momentum space.

No mathematically rigorous definition exists which would cover all the interesting cases, so we will limit ourselves to the non-rigorous definition given above. Examples of linear dispersive PDEs include

- the wave equation (W) and its analogs in higher space dimensions,
- the Schrödinger equation,
- the Klein–Gordon equation,

but the list could be made longer. All these examples are *time reversible* and have a *conserved energy*, and yet smooth localized initial conditions lead to *radiative behavior* as $t \to \infty$, for reasons similar to the ones exposed above in the case of (W).

Remark 1.3. Radiative behavior crucially depends on the fact that the spatial domain is the whole Euclidean space (or in any case that it is unbounded). The notion of group velocity makes sense also for bounded domains, but we will not discuss this case here.

Remark 1.4. In some sense, for a linear dispersive PDE, the trivial solution $\{\psi \equiv 0\}$ is the *global attractor* of the flow.

2. The nonlinear setting: Dispersion, solitons, soliton resolution

For a comprehensive introduction to the topic of this section, the reader can consult the monographs [51,58].

2.1. Wave maps

Wave maps are nonlinear, geometric analogs of linear waves in the case of maps taking values in a Riemannian manifold and not a Euclidean space. We consider here wave maps $\Psi : \mathbb{R}^{1+2} \to \mathbb{S}^2 \subset \mathbb{R}^3$. In classical mechanics, a constrained mechanical system is obtained from the same Lagrangian as for the ambient system; see [1, Chapter 4]. Following this principle and recalling (1.1), we say that a map $\Psi : \mathbb{R}^{1+2} \to \mathbb{S}^2 \subset \mathbb{R}^3$ is a *wave map* if it is a critical point of the Lagrangian

$$\mathcal{L}(\Psi) = \frac{1}{2} \iiint (|\partial_t \Psi|^2 - |\partial_{x_1} \Psi|^2 - |\partial_{x_2} \Psi|^2) \, \mathrm{d}x\mathrm{d}t, \tag{2.1}$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^3$. Similarly, as in Section 1.1, we can consider $\Psi_\varepsilon = \Psi + \varepsilon Z$, where $Z : \mathbb{R}^{1+2} \to \mathbb{R}^3$ is smooth and compactly supported. In order not to violate at main order, the condition that Ψ_ε has values in $\mathbb{S}^2$, it is necessary and sufficient to require that

$$Z(t,x) \perp \Psi(t,x) \quad \text{for all } (t,x). \tag{2.2}$$

As in Section 1.1, we have

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}|_{\varepsilon=0} \mathcal{L}(\Psi_\varepsilon) = - \iiint Z \cdot (\partial_t^2 \Psi - \Delta \Psi) \, \mathrm{d}x\mathrm{d}t.$$

This quantity vanishes for all Z satisfying (2.2) if and only if

$$\partial_t^2 \Psi(t,x) - \Delta \Psi(t,x) = \mu(t,x) \Psi(t,x) \quad \text{for some } \mu(t,x) \in \mathbb{R}.$$

Differentiating twice the identity $\Psi \cdot \Psi = 1$, we obtain

$$\Psi \cdot (\partial_t^2 \Psi - \Delta \Psi) = -|\partial_t \Psi|^2 + |\partial_{x_1} \Psi|^2 + |\partial_{x_2} \Psi|^2,$$

so we can write the wave map equation $\mathbb{R}^{1+2} \to S^2$ as follows:

$$\partial_t^2 \Psi - \Delta \Psi = -(|\partial_t \Psi|^2 - |\partial_{x_1} \Psi|^2 - |\partial_{x_2} \Psi|^2) \Psi. \tag{WM}$$

Similarly, as in the linear case, the following *total energy* is a conservation law for wave maps:

$$E(\Psi) := \iint_{\mathbb{R}^2} \left(\frac{1}{2} |\partial_t \Psi|^2 + \frac{1}{2} |\partial_{x_1} \Psi|^2 + \frac{1}{2} |\partial_{x_2} \Psi|^2 \right) \mathrm{d}x.$$

Equation (WM) has trivial constant in space-time solutions $\Psi(t,x) = \omega_0 \in \mathbb{S}^2$. If we linearize around such a solution, that is, if we write $\Psi = \omega_0 + \varepsilon\Phi$ with $\varepsilon \ll 1$ and plug into (WM), at main order, we obtain $\partial_t^2\Phi - \Delta\Phi = 0$ so that each component of Φ satisfies equation (W), which indicates that small perturbations of a constant solution should exhibit a radiative behavior so that the whole wave map should converge to a constant. It was proved in the works of Tataru [60] and Tao [56,57] that, in an appropriate sense, this is indeed the case.

For the study of long-time behavior of solutions of (WM), *criticality* is an important (and helpful) property. Let $\lambda > 0$ and consider

$$\Psi_\lambda(t,x) := \Psi(t/\lambda, x/\lambda).$$

It is clear from (WM) (or from the Lagrangian) that Ψ is a wave map if and only if Ψ_λ is a wave map. Moreover,

$$E(\Psi_\lambda) = E(\Psi).$$

For this reason, equation (WM) is called *energy critical* and its solutions *critical wave maps*.

Remark 2.1. In general, a problem is *subcritical* if it becomes a "small data problem" when rescaling (zooming) to a small region. It is called *supercritical* if such a zoom makes it large. It is called *critical* if the size of the data remains unchanged.

2.2. Harmonic maps

One might wonder if every, not necessarily small, solution of (WM) has radiative behavior. The answer is "no" for a simple reason: there exist non-trivial static (time-independent) solutions. Namely, inserting $\Psi(t,x) = \omega(x)$ into (WM), we obtain the *critical harmonic maps equation*:

$$-\Delta\omega = |\nabla\omega|^2\omega, \quad \omega : \mathbb{R}^2 \to \mathbb{S}^2 \subset \mathbb{R}^3.$$

Its solutions are referred to as *harmonic maps* $\mathbb{R}^2 \to \mathbb{S}^2$. It was established by Eells and Wood [17] and Hélein [23] that harmonic maps of finite energy correspond to *rational functions* $\mathbb{S}^2 \to \mathbb{S}^2$ and their complex conjugates (we identify $\mathbb{R}^2$ with $\mathbb{S}^2$ using the stereographic projection).

Remark 2.2. It was proved by Krieger, Schlag and Tataru [34], Rodnianski and Sterbenz [47] and Raphaël and Rodnianski [46] that solutions of large energy can even cease to exist in finite time. Equation (WM) is thus locally well posed but not globally well posed.

2.3. A few generalities on nonlinear dispersive PDEs

A nonlinear PDE is called dispersive if it is related to a linear dispersive PDE. Most frequently, "related" means "obtained through linearization around trivial solutions", like in the case of wave maps discussed above.

Nonlinear dispersive PDEs frequently appear in physics, for example, in the study of water waves and nonlinear optics; see [63, Chapters 12, 13, 16, 17]. Typical examples are Hamiltonian systems; in particular, they are *time reversible* and have a *conserved energy*.

One is often interested in the *dynamical behavior* of solutions of a given nonlinear PDE, by which we mean their asymptotic description as time becomes large (for solutions defined for all time; if they are not, one studies the limit as the time converges to the maximal time of existence of the solution). Among the most common questions of this type is the *problem of stability*, which can be formulated as follows.

Problem. Do small solutions of a nonlinear dispersive PDE exhibit radiative behavior? In other words, does the flow restricted to small solutions have a trivial attractor, like in the linear case (see Remark 1.4)?

The intuitive reasoning is that small solutions should behave in the same way as the solutions of the linearized problem, which have radiative behavior as we saw in Section 1.2.

2.4. Solitons and soliton resolution

The notion of a soliton is somewhat controversial; see [40, Section 1.5]. We adopt the following definition.

Definition 2.3. A soliton is a solution of an evolution PDE which does not change its shape in the course of time (it can, however, change its position).

Harmonic maps from Section 2.2 are examples of solitons for (WM). Solitons moving at constant velocity can be obtained using the Lorentz invariance of (WM).

Solitons do not exhibit radiative behavior. The problem of *soliton resolution* is to prove that they are the only obstruction to radiative behavior. However, it would be too naive to expect that any solution is either radiative or a soliton. Rather, one expects that any solution eventually decomposes into a superposition of solitons which interact sufficiently weakly (for example, they could travel with distinct velocities). Such superpositions are called *multisoliton configurations*.

Problem (Soliton resolution conjecture). For a given nonlinear dispersive PDE, does any solution converge in large time to a sum of a multisoliton and a radiative term?

In other words, does the flow have a simple global attractor related to multisoliton configurations?

The soliton resolution is inspired by

- numerical simulations (see Fermi, Pasta, Ulam and Tsingou [18] and Zabusky and Kruskal [64]),
- theory of *completely integrable systems* (see Segur and Ablowitz [49] and Eckhaus and Schuur [15]),
- analogous *elliptic* and *parabolic bubbling* phenomena (see the works of Sacks and Uhlenbeck [48] and Struwe [54]).

Remark 2.4. Even with such a vague formulation, the soliton resolution is not expected to hold for all nonlinear dispersive PDEs. For example, the sine-Gordon equation has so-called breather solutions, which do not fall into the regime of soliton resolution.

Remark 2.5. Strictly speaking, soliton resolution only provides an "upper bound" on the global attractor, in the sense that it does not say anything about the types of multisoliton configurations which can be realized by the evolution.

Our main goal is to provide an example of "natural" nonlinear dispersive PDEs for which we can prove soliton resolution. Even though we cannot provide a full description of the global attractor, we will see that it contains configurations consisting of more than one soliton.

3. Soliton resolution for equivariant energy-critical wave maps

3.1. Equivariant maps

This PDE will be obtained from (WM) by restricting the flow to a certain subclass of all the maps $\mathbb{R}^2 \to \mathbb{S}^2$ which is preserved by the flow. They are called *equivariant maps* and are defined in the following way. We fix $k \in \{1, 2, \ldots\}$ and consider maps of the form

$$\Psi(t, r\cos\theta, r\sin\theta) = (\sin(\psi(t,r))\cos(k\theta), \sin(\psi(t,r))\sin(k\theta), \cos(\psi(t,r))), \tag{3.1}$$

where $r > 0$ and $\psi(t,r) \in \mathbb{R}$. Plugging this into (WM), we find the following *scalar equation*:

$$\partial_t^2\psi(t,r) - \partial_r^2\psi(t,r) - \frac{1}{r}\partial_r\psi(t,r) + \frac{k^2}{2r^2}\sin(2\psi(t,r)) = 0, \tag{WM$_k$}$$

where $r > 0$ is the radial coordinate. Note that ψ and $\psi + 2\pi\ell$ represent the same map Ψ for any $\ell \in \mathbb{Z}$.

Remark 3.1. In the non-geometric context, it is common to consider spherically symmetric solutions. Equivariant solutions are analogous objects in the geometric setting of (WM). More generally, whenever symmetries of a given equation lead to invariance of a certain class of states, it is a well-known technique to study the restriction of the system to this subclass.

One can check that under the substitution (3.1), the Lagrangian (2.1) becomes

$$\mathcal{L} := \pi \iint \left((\partial_t \psi)^2 - (\partial_r \psi)^2 - \frac{k^2 \sin(\psi)^2}{r^2} \right) r \mathrm{d}r \, \mathrm{d}t.$$

Its critical points are thus k-equivariant wave maps, a fact which can easily be checked directly. The kinetic energy and the potential energy are

$$E_{\text{kinetic}} := \pi \int_0^\infty (\partial_t \psi)^2 \, r \mathrm{d}r,$$
$$E_{\text{potential}} := \pi \int_0^\infty \left((\partial_r \psi)^2 + \frac{k^2 \sin(\psi)^2}{r^2} \right) r \mathrm{d}r.$$

Their sum is the total energy. It is a conserved quantity.

We always consider strong solutions of finite energy (that is, strong limits of sequences of smooth solutions in the topology induced by the energy, locally uniformly in time). Their existence and uniqueness for any finite energy initial conditions were obtained in [20, 50]. It can be deduced from Strichartz estimates for the wave equation; see, for example, [7, Section 2] in the case $k \in \{1, 2\}$. If $k \geq 3$ is large, Strichartz estimates from [43] can be applied; see [26, Section 2]. Finite energy solutions of (WM_k) are not guaranteed to exist for all time. I denote by $(T_-, T_+) \subset \mathbb{R}$ the maximal time interval on which the solution exists.

3.2. Multibubble (multisoliton) configurations

Recalling the discussion from Section 2.2, the only k-equivariant harmonic maps correspond to rational functions $az^k, az^{-k}, a\overline{z}^k$ and $a\overline{z}^{-k}$ with $a > 0$. In order to represent these maps in the context of the scalar equation (WM_k), it is convenient to denote

$$Q_\lambda(r) := 2 \arctan\left(\frac{r^k}{\lambda^k} \right), \quad \lambda > 0.$$

Then the stationary solutions of (WM_k) are

$$Q_\lambda + 2\pi\ell, \quad -Q_\lambda + 2\pi\ell, \quad \pi + Q_\lambda + 2\pi\ell, \quad \pi - Q_\lambda + 2\pi\ell, \qquad a > 0, \ell \in \mathbb{Z}.$$

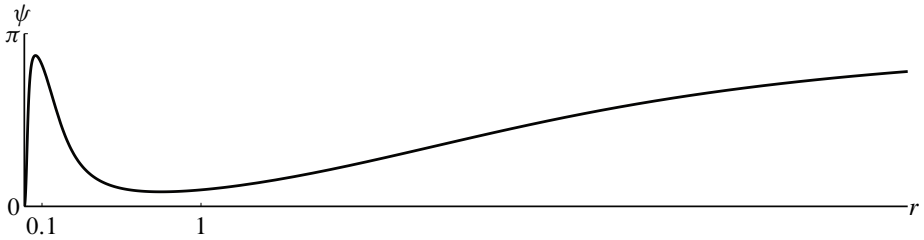


Figure 1. A multibubble configuration with three bubbles.

In this context of equation (WM_k), solitons are also called *bubbles*. Note that, with our notational conventions, λ is the spatial scale of the bubble Q_λ. For a given number of bubbles M, an integer m and scales $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M$, we define a *multibubble configuration* by

$$Q(\lambda_1, \ldots, \lambda_M) := m\pi + \sum_{j=1}^{M} \pm Q_{\lambda_j} \tag{3.2}$$

(in the notation, we skip the dependence on m, which is not going to be essential here). One should think of the scales as satisfying $\lambda_1 \ll \cdots \ll \lambda_M$ so that each bubble is separated in scale from all the others. Figure 1 shows a multibubble with $(\lambda_1, \lambda_2, \lambda_3) = \left(\frac{1}{10}, \frac{1}{2}, 5\right)$.

3.3. Soliton resolution

Our main result can be formulated as follows.

Theorem 3.2 (Jendrej–Lawrie [28]). *Let ψ be a solution of (WM_k) defined for all $t \in (0, \infty)$. As $t \to \infty$, $(\psi(t,\cdot), \partial_t \psi(t,\cdot))$ decomposes into a superposition of*

- *multibubble configuration,*
- *radiation, corresponding to a solution of the linear wave equation (W),*
- *remainder whose energy converges to 0.*

Remark 3.3. We also prove a similar result in the case of a finite maximal time of existence of the solution.

Remark 3.4. The case $k = 1$ was settled by Duyckaerts, Kenig, Martel and Merle [11] using a different approach (so-called channels of energy; see below).

Remark 3.5. It was proved in [25] that for all $k \geq 2$, there exists a solution containing two bubbles. For these solutions, the radiation component vanishes. It was

proved in [28] that for $k = 1$, multibubble solutions with vanishing radiation component do not exist. Existence of multibubble solutions with a non-vanishing radiation component is an open problem for all k.

The history of progress in understanding the dynamical behavior of large solutions of (WM_k) is quite long. Fundamental results were obtained in [4, 5, 52, 53], see also [51, Chapter 8], the main conclusion being the *decay of energy at the self-similar scale*, which in particular excludes self-similar blow-up, but also, as proved by Struwe [55], leads to *bubbling*: if ψ is a solution of (WM_k) which blows up in finite time T_+, then there exist sequences $t_n \to T_+$ and $0 < \lambda_n \ll T_+ - t_n$ such that

$$(\psi(t_n, \lambda_n \cdot), \lambda_n \partial_t \psi(t_n, \lambda_n \cdot)) \to n\pi \pm Q,$$

the convergence being understood in the topology induced by the energy *locally* (on bounded sets).

The bubbling also implies that a solution whose energy is smaller than the energy of Q cannot blow up. It was proved by Côte, Kenig, Lawrie and Schlag [7] that such a solution actually has radiative behavior. Above this threshold energy, finite time blow-up can occur, as was proved in the works [34, 46, 47] already mentioned above. Important progress toward the soliton resolution conjecture was obtained in [8]. *Sequential* soliton resolution, that is, convergence to a superposition of solitons for a sequence of times, was proved by Côte [6] for $k \in \{1, 2\}$ and Jia and Kenig [30] for $k \geq 3$. Similar results without imposing equivariant symmetry assumptions, but with a less precise description of the radiation, were obtained by Grinis [21].

In [26], continuous in time resolution was proved at the minimal possible energy level allowing for existence of a two-bubble. As a relatively simple consequence, continuous in time resolution was proved in [27] under the assumption that the solution contains at most two bubbles.

For the closely related energy-critical wave equation, scattering below the ground state energy threshold was proved by Kenig and Merle [32], establishing together with [31] the so-called *Kenig–Merle route map*. In the radially symmetric case, the soliton resolution conjecture was proved by Duyckaerts, Kenig and Merle in space dimension 3 in [12], in any odd space dimension in [13] and in dimension 4 in [11] (in collaboration with Martel). All these works used the *channels of energy* introduced in [12].

In the non-radial case, sequential soliton resolution was proved by Duyckaerts, Jia, Kenig and Merle [10].

3.4. Main ideas of the proof

Let ψ be a solution of (WM_k) defined for all $t \in (0, \infty)$. Thanks to the sequential soliton resolution results [6, 30], we know that there exists a sequence $t_n \to \infty$ such that

$(\psi(t_n,\cdot),\partial_t\psi(t_n,\cdot))$ decomposes into a superposition of a multibubble configuration, radiation and a small remainder. It thus suffices to prove a *no-return lemma*: if a multibubble configuration is destroyed (we say that a *collision* happens), it cannot recover its shape (note the analogy with non-existence of homoclinic/heteroclinic orbits). A similar idea is present in the works of Duyckaerts and Merle [14], Nakanishi and Schlag [42], Krieger, Nakanishi and Schlag [33] for a *single soliton* which is linearly unstable. In our case, interactions between solitons play a role similar to the linear instability in those works. This idea has already been used in [26] in the special case where there are only two bubbles and the radiative component vanishes.

The no-return lemma is obtained by the classical method of Lyapunov functionals. In our case, a Lyapunov functional can be obtained from the *virial identity*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^\infty \chi_R\,\partial_t\psi\, r\partial_r\psi\, r\mathrm{d}r = -\int_0^\infty \chi_R(\partial_t\psi)^2\, r\mathrm{d}r + O(\mathcal{E}_R). \tag{3.3}$$

Here, χ_R is a standard cut-off function at scale R, and $\mathcal{E}_R$ can be thought of as the amount of energy in the annulus with radii R and $2R$. In order to identify the relevant spatial and temporal scales at which identity (3.3) should be applied, we introduce the notion of *collision intervals*, *inner bubbles* and *outer bubbles*. This is done as follows.

Let $\mathbf{d}(t)$ be the (appropriately defined) distance to (multibubble + radiation) at time t. We say that $[a_n,b_n]$ is a sequence of collision intervals if $\mathbf{d}(a_n)\to 0$, $\mathbf{d}(b_n)\to 0$ but $\mathbf{d}(c_n)\not\to 0$ for some $c_n\in[a_n,b_n]$. The inner bubbles are the ones which enter into a collision, and the outer bubbles are the ones which do not collide. We let K be the smallest natural number K such that there exists an infinite sequence of collision intervals with K inner bubbles.

We now claim that the duration of the collision has to be at least of order λ_K. Indeed, by assumption, at time a_n the solution is close to a multibubble configuration. At a timescale much smaller than λ_K, by finite speed of propagation, the bubble of scale λ_K cannot be destroyed. This, however, would contradict the minimality of K, and the claim follows.

We thus have an infinite sequence of collision intervals $[a_n,b_n]$ with K inner bubbles. Let us neglect for a moment the term $O(\mathcal{E}_R)$ in (3.3) and integrate the virial identity in time over a collision interval $[a_n,b_n]$ with R of order λ_K (we assume for simplicity that λ_K does not change much on any given collision interval; the possibility of important variations in time of λ_K introduces additional complications in the proof). The absolute value of the left-hand side of (3.3) can be seen to be $\ll\lambda_K$ at the endpoints of the interval (indeed, the solution being close to a multibubble configuration implies in particular that the kinetic part satisfies $\|\partial_t\psi\|_{L^2}\ll 1$, and it suffices to apply the Cauchy–Schwarz inequality). We thus conclude that the mean kinetic energy at space-timescale λ_K converges to 0 as $n\to\infty$.

Finally, we have what we call the *compactness lemma*: if the mean kinetic energy at timescale ρ vanishes, then the solution is close to a multibubble up to scale $\gg \rho$. Intuitively, vanishing of the kinetic energy implies that the solution is "almost stationary" at a given scale, and the bubbles are the only stationary solutions. Our proof of this lemma is very closely related to the already mentioned sequential resolution results. It crucially uses the *profile decomposition* method of Bahouri and Gérard [2] (see also [3, 19, 38, 39, 41] for other notable works leading to the development of the profile decomposition method).

There remains the problem of absorbing the error term $O(\mathcal{E}_R)$ in the virial identity (3.3). In order to accomplish this step, we derive an ODE system on the scales of the bubbles, resembling an n-body problem. It is easy to understand how the relevant n-body problem should look using the Lagrangian formulation again. We consider a multibubble $Q(\lambda_1, \dots, \lambda_M)$ defined by (3.2) and define the reduced Lagrangian

$$\widetilde{\mathcal{L}}(\lambda_1, \dots, \lambda_M) := \mathcal{L}(Q(\lambda_1, \dots, \lambda_M)).$$

One then expects that critical points of $\widetilde{\mathcal{L}}$ should describe, at least to the leading order, the evolution of the scales of the bubbles. Such a reduction goes by the name of the *modulation method*. In our case, in order to justify that the bubbles indeed evolve as predicted by the ODE, we use the refined modulation parameters introduced in [24]. The ODE system allows us to obtain upper bounds on the length of collision intervals and absorb the error term in (3.3). We note that a similar reduction to a 2-body problem was already used in [26].

4. Bubble decomposition for the harmonic map heat flow

4.1. Harmonic map heat flow

We consider the gradient flow of maps $\Psi : \mathbb{R}^2 \to \mathbb{S}^2 \subset \mathbb{R}^3$ for the energy functional

$$E(\Psi) := \int_{\mathbb{R}^2} \frac{1}{2} |\nabla \Psi|^2 \, \mathrm{d}x.$$

We obtain the following nonlinear heat equation called the *harmonic map heat flow*:

$$\partial_t \Psi = \mathcal{T}(\Psi) := \Delta \Psi + |\nabla \Psi|^2 \Psi, \quad \Psi : [0, T_+) \times \mathbb{R}^2 \to \mathbb{S}^2. \tag{HMHF}$$

It was introduced by Eells and Sampson [16] and served as an inspiration for the Ricci flow of Hamilton [22].

Like for the wave maps equation (WM), the stationary solutions of (HMHF) are given by the harmonic maps $\mathbb{R}^2 \to \mathbb{S}^2$. Convergence to multibubble configurations *for a time sequence* was proved by Qing and Tian [45]; see also Struwe [54],

Qing [44], Ding and Tian [9], Wang [62], Qing and Tian [45], Lin and Wang [35] and Topping [61], as well as the book by Lin and Wang [36].

4.2. Continuous time bubble decomposition

We prove that convergence toward the set of multibubble configurations holds in continuous time.

Theorem 4.1 (Jendrej–Lawrie–Schlag [29]). *Let Ψ be a solution of* (HMHF) *defined for all $t \in (0, \infty)$. As $t \to \infty$, $\Psi(t)$ decomposes into a superposition of*

- *a multibubble configuration,*
- *a remainder converging to* 0 *in the energy space and uniformly.*

Remark 4.2. We also prove a corresponding result in the case $T_+ < \infty$.

Remark 4.3. Unlike for wave maps, here we impose no symmetry assumptions on the initial data.

4.3. Some proof ideas

Let $\mathbf{d}(t)$ be the (appropriately defined) distance to a multibubble at time t. We say that $[a_n, b_n]$ is a sequence of *collision intervals* if $\mathbf{d}(a_n) \to 0$, but $\mathbf{d}(b_n) \nrightarrow 0$. In order to identify the relevant spatial and temporal scales, we use the *energy quantization* for harmonic maps: $E(\omega) = 4\pi|\deg(\omega)|$. We can thus let K be the smallest natural number for which there are collision intervals with colliding bubbles of total energy $4K\pi$ (and the other bubbles remaining intact throughout the collision). We then claim that the duration of the collision has to be at least of order λ^2, where λ is the largest scale of the colliding bubbles. The reason is again the propagation speed: the heat equation cannot move a substantial portion of energy by distance λ during a timescale much smaller than λ^2 (for the linear heat equation, this follows from the well-known expression for the heat kernel; the argument is more involved in the nonlinear case, but the result remains true).

We thus have an infinite sequence of collision intervals $[a_n, b_n]$, with colliding bubbles of total energy $4K\pi$. Consider now the energy identity $\frac{\mathrm{d}}{\mathrm{d}t} E(t) = -\|\mathcal{T}(\Psi(t))\|_{L^2}^2$. Upon integration in time over the collision intervals $[a_n, b_n]$ and taking into account that $b_n - a_n \gtrsim \lambda_n$, where λ_n is the largest among the scales of the colliding bubbles, we obtain $\|\mathcal{T}(\Psi(t))\|_{L^2} \ll \lambda_n^{-1}$ as $n \to \infty$. However, we have the following *compactness lemma*, which is essentially a reformulation of the results of Qing and Tian [45] on almost harmonic maps: if $\Phi_n : \mathbb{R}^2 \to \mathbb{S}^2$ and $\|\mathcal{T}(\Phi_n)\|_{L^2} \ll \rho_n^{-1}$, then Φ_n converges to multibubbles in disks of radius ρ_n. Letting ρ_n be of order λ_n, we obtain a contradiction with the fact that the bubbles collide

(and thus are not close to a multibubble configuration on some part of the collision interval).

Proving uniform convergence requires additional arguments of a different kind, in particular the use of the Strichartz estimates for the heat flow proved by Tao [59].

Acknowledgments. I am grateful to Andrew Lawrie and Wilhelm Schlag for helpful comments.

Funding. This work was partially supported by the ERC Starting Grant "INSOLIT" 101117126.

References

[1] V. I. Arnold, *Mathematical methods of classical mechanics*. 2nd edn., Grad. Texts in Math. 60, Springer, New York, 1989, 520 pp. Zbl 1571.70001 MR 0997295

[2] H. Bahouri and P. Gérard, High frequency approximation of solutions to critical nonlinear wave equations. *Amer. J. Math.* **121** (1999), no. 1, 131–175 Zbl 0919.35089 MR 1705001

[3] H. Brezis and J. M. Coron, Convergence of solutions of H-systems or how to blow bubbles. *Arch. Ration. Mech. Anal.* **89** (1985), no. 1, 21–56 Zbl 0584.49024 MR 0784102

[4] D. Christodoulou and A. S. Tahvildar-Zadeh, On the asymptotic behavior of spherically symmetric wave maps. *Duke Math. J.* **71** (1993), no. 1, 31–69 Zbl 0791.58105 MR 1230285

[5] D. Christodoulou and A. S. Tahvildar-Zadeh, On the regularity of spherically symmetric wave maps. *Comm. Pure Appl. Math.* **46** (1993), no. 7, 1041–1091 Zbl 0744.58071 MR 1223662

[6] R. Côte, On the soliton resolution for equivariant wave maps to the sphere. *Comm. Pure Appl. Math.* **68** (2015), no. 11, 1946–2004 Zbl 1343.35155 MR 3403756

[7] R. Côte, C. E. Kenig, A. Lawrie, and W. Schlag, Characterization of large energy solutions of the equivariant wave map problem. I. *Amer. J. Math.* **137** (2015), no. 1, 139–207 Zbl 1315.35130 MR 3318089

[8] R. Côte, C. E. Kenig, A. Lawrie, and W. Schlag, Characterization of large energy solutions of the equivariant wave map problem. II. *Amer. J. Math.* **137** (2015), no. 1, 209–250 Zbl 1315.35131 MR 3318090

[9] W. Ding and G. Tian, Energy identity for a class of approximate harmonic maps from surfaces. *Comm. Anal. Geom.* **3** (1995), no. 4, 543–554 Zbl 0855.58016 MR 1371209

[10] T. Duyckaerts, H. Jia, C. Kenig, and F. Merle, Soliton resolution along a sequence of times for the focusing energy critical wave equation. *Geom. Funct. Anal.* **27** (2017), no. 4, 798–862 Zbl 1391.35276 MR 3678502

[11] T. Duyckaerts, C. Kenig, Y. Martel, and F. Merle, Soliton resolution for critical co-rotational wave maps and radial cubic wave equation. *Comm. Math. Phys.* **391** (2022), no. 2, 779–871 Zbl 1491.35292 MR 4397184

[12] T. Duyckaerts, C. Kenig, and F. Merle, Classification of the radial solutions of the focusing, energy-critical wave equation. *Camb. J. Math.* **1** (2013), no. 1, 75–144 Zbl 1308.35143 MR 3272053

[13] T. Duyckaerts, C. Kenig, and F. Merle, Soliton resolution for the radial critical wave equation in all odd space dimensions. *Acta Math.* **230** (2023), no. 1, 1–92 Zbl 1520.35103 MR 4567713

[14] T. Duyckaerts and F. Merle, Dynamics of threshold solutions for energy-critical wave equation. *Int. Math. Res. Pap. IMRP* **2008** (2008), article no. rpn002, 67 pp. Zbl 1159.35043 MR 2416069

[15] W. Eckhaus and P. Schuur, The emergence of solitons of the Korteweg–de Vries equation from arbitrary initial conditions. *Math. Methods Appl. Sci.* **5** (1983), no. 1, 97–116 Zbl 0518.35074 MR 0690898

[16] J. Eells, Jr. and J. H. Sampson, Harmonic mappings of Riemannian manifolds. *Amer. J. Math.* **86** (1964), no. 1, 109–160 Zbl 0122.40102 MR 0164306

[17] J. Eells and J. C. Wood, Restrictions on harmonic maps of surfaces. *Topology* **15** (1976), no. 3, 263–266 Zbl 0328.58008 MR 0420708

[18] E. Fermi, J. Pasta, S. Ulam, and M. Tsingou, *Studies of nonlinear problems. I*. Los Alamos Scientific Lab., Report no. LA-1940, 1955

[19] P. Gérard, Description du défaut de compacité de l'injection de Sobolev. *ESAIM Control Optim. Calc. Var.* **3** (1998), 213–233 Zbl 0907.46027 MR 1632171

[20] J. Ginibre, A. Soffer, and G. Velo, The global Cauchy problem for the critical non-linear wave equation. *J. Funct. Anal.* **110** (1992), no. 1, 96–130 Zbl 0813.35054 MR 1190421

[21] R. Grinis, Quantization of time-like energy for wave maps into spheres. *Comm. Math. Phys.* **352** (2017), no. 2, 641–702 Zbl 1382.35164 MR 3627409

[22] R. S. Hamilton, Three-manifolds with positive Ricci curvature. *J. Differential Geom.* **17** (1982), no. 2, 255–306 Zbl 0504.53034 MR 0664497

[23] F. Hélein, *Harmonic maps, conservation laws and moving frames*. 2nd edn. (Translated from the 1996 French original. With a Foreword by James Eells), Cambridge Tracts in Math. 150, Cambridge University Press, Cambridge, 2002, 264 pp. Zbl 1010.58010 MR 1913803

[24] J. Jendrej, Construction of two-bubble solutions for the energy-critical NLS. *Anal. PDE* **10** (2017), no. 8, 1923–1959 Zbl 1375.35492 MR 3694010

[25] J. Jendrej, Construction of two-bubble solutions for energy-critical wave equations. *Amer. J. Math.* **141** (2019), no. 1, 55–118 Zbl 1433.35198 MR 3904767

[26] J. Jendrej and A. Lawrie, Two-bubble dynamics for threshold solutions to the wave maps equation. *Invent. Math.* **213** (2018), no. 3, 1249–1325 Zbl 1411.35205 MR 3842064

[27] J. Jendrej and A. Lawrie, Continuous time soliton resolution for two-bubble equivariant wave maps. *Math. Res. Lett.* **29** (2022), no. 6, 1745–1766 Zbl 1523.37077 MR 4589375

[28] J. Jendrej and A. Lawrie, Soliton resolution for energy-critical wave maps in the equivariant case. *J. Amer. Math. Soc.* **38** (2025), no. 3, 783–875 Zbl 1562.35394 MR 4890658

[29] J. Jendrej, A. Lawrie, and W. Schlag, Continuous in time bubble decomposition for the harmonic map heat flow. *Forum Math. Pi* **13** (2025), article no. e4, 37 pp. Zbl 1566.35122 MR 4862091

[30] H. Jia and C. Kenig, Asymptotic decomposition for semilinear wave and equivariant wave map equations. *Amer. J. Math.* **139** (2017), no. 6, 1521–1603 Zbl 1391.35278 MR 3730929

[31] C. E. Kenig and F. Merle, Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrödinger equation in the radial case. *Invent. Math.* **166** (2006), no. 3, 645–675 Zbl 1115.35125 MR 2257393

[32] C. E. Kenig and F. Merle, Global well-posedness, scattering and blow-up for the energy-critical focusing non-linear wave equation. *Acta Math.* **201** (2008), no. 2, 147–212 Zbl 1183.35202 MR 2461508

[33] J. Krieger, K. Nakanishi, and W. Schlag, Global dynamics away from the ground state for the energy-critical nonlinear wave equation. *Amer. J. Math.* **135** (2013), no. 4, 935–965 Zbl 1307.35170 MR 3086065

[34] J. Krieger, W. Schlag, and D. Tataru, Renormalization and blow up for charge one equivariant critical wave maps. *Invent. Math.* **171** (2008), no. 3, 543–615 Zbl 1139.35021 MR 2372807

[35] F. Lin and C. Wang, Energy identity of harmonic map flows from surfaces at finite singular time. *Calc. Var. Partial Differential Equations* **6** (1998), no. 4, 369–380 Zbl 0908.58008 MR 1624304

[36] F. Lin and C. Wang, *The analysis of harmonic maps and their heat flows*. World Scientific, Hackensack, NJ, 2008, 267 pp. Zbl 1203.58004 MR 2431658

[37] F. Linares and G. Ponce, *Introduction to nonlinear dispersive equations*. Universitext, Springer, New York, 2014, 301 pp. MR 3308874

[38] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The locally compact case. Part 1. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **1** (1984), no. 2, 109–145 Zbl 0541.49009 MR 0778970

[39] P.-L. Lions, The concentration-compactness principle in the calculus of variations. The limit case. Part 2. *Rev. Mat. Iberoam.* **1** (1985), no. 2, 45–121 Zbl 0704.49006 MR 0850686

[40] N. Manton and P. Sutcliffe, *Topological solitons*. Cambridge Monogr. Math. Phys., Cambridge University Press, Cambridge, 2004, 493 pp. Zbl 1100.37044 MR 2068924

[41] F. Merle and L. Vega, Compactness at blow-up time for L^2 solutions of the critical nonlinear Schrödinger equation in 2D. *Int. Math. Res. Not. IMRN* **1998** (1998), no. 8, 399–425 Zbl 0913.35126 MR 1628235

[42] K. Nakanishi and W. Schlag, Global dynamics above the ground state energy for the focusing nonlinear Klein–Gordon equation. *J. Differential Equations* **250** (2011), no. 5, 2299–2333 Zbl 1213.35307 MR 2756065

[43] F. Planchon, J. G. Stalker, and A. S. Tahvildar-Zadeh, L^p estimates for the wave equation with the inverse-square potential. *Discrete Contin. Dyn. Syst.* **9** (2003), no. 2, 427–442 Zbl 1031.35092 MR 1952384

[44] J. Qing, On singularities of the heat flow for harmonic maps from surfaces into spheres. *Comm. Anal. Geom.* **3** (1995), no. 2, 297–315 Zbl 0868.58021 MR 1362654

[45] J. Qing and G. Tian, Bubbling of the heat flows for harmonic maps from surfaces. *Comm. Pure Appl. Math.* **50** (1997), no. 4, 295–310 Zbl 0879.58017 MR 1438148

[46] P. Raphaël and I. Rodnianski, Stable blow up dynamics for the critical co-rotational wave maps and equivariant Yang–Mills problems. *Publ. Math. Inst. Hautes Études Sci.* **115** (2012), 1–122 Zbl 1284.35358 MR 2929728

[47] I. Rodnianski and J. Sterbenz, On the formation of singularities in the critical $O(3)$ σ-model. *Ann. of Math. (2)* **172** (2010), no. 1, 187–242 Zbl 1213.35392 MR 2680419

[48] J. Sacks and K. Uhlenbeck, The existence of minimal immersions of 2-spheres. *Ann. of Math. (2)* **113** (1981), no. 1, 1–24 Zbl 0462.58014 MR 0604040

[49] H. Segur and M. J. Ablowitz, Asymptotic solutions and conservation laws for the nonlinear Schrödinger equation. I. *J. Math. Phys.* **17** (1976), no. 5, 710–713 MR 0450822

[50] J. Shatah and M. Struwe, Well-posedness in the energy space for semilinear wave equations with critical growth. *Int. Math. Res. Not. IMRN* **1994** (1994), no. 7, 303–309 Zbl 0830.35086 MR 1283026

[51] J. Shatah and M. Struwe, *Geometric wave equations*. Courant Lect. Notes Math. 2, American Mathematical Society, Providence, RI, 2000, 153 pp. Zbl 1051.35500 MR 1674843

[52] J. Shatah and A. Tahvildar-Zadeh, Regularity of harmonic maps from the Minkowski space into rotationally symmetric manifolds. *Comm. Pure Appl. Math.* **45** (1992), no. 8, 947–971 Zbl 0769.58015 MR 1168115

[53] J. Shatah and A. S. Tahvildar-Zadeh, On the Cauchy problem for equivariant wave maps. *Comm. Pure Appl. Math.* **47** (1994), no. 5, 719–754 Zbl 0811.58059 MR 1278351

[54] M. Struwe, On the evolution of harmonic mappings of Riemannian surfaces. *Comment. Math. Helv.* **60** (1985), no. 1, 558–581 Zbl 0595.58013 MR 0826871

[55] M. Struwe, Equivariant wave maps in two space dimensions. *Comm. Pure Appl. Math.* **56** (2003), no. 7, 815–823 Zbl 1033.53019 MR 1990477

[56] T. Tao, Global regularity of wave maps. I. Small critical Sobolev norm in high dimension. *Int. Math. Res. Not. IMRN* **2001** (2001), no. 6, 299–328 Zbl 0983.35080 MR 1820329

[57] T. Tao, Global regularity of wave maps. II. Small energy in two dimensions. *Comm. Math. Phys.* **224** (2001), no. 2, 443–544 Zbl 1020.35046 MR 1869874

[58] T. Tao, *Nonlinear dispersive equations. Local and global analysis*. CBMS Reg. Conf. Ser. Math. 106, American Mathematical Society, Providence, RI, 2006, 373 pp. Zbl 1106.35001 MR 2233925

[59] T. Tao, Global regularity of wave maps IV. Absence of stationary or self-similar solutions in the energy class. [v1] 2008, [v2] 2009, arXiv:0806.3592v2

[60] D. Tataru, On global existence and scattering for the wave maps equation. *Amer. J. Math.* **123** (2001), no. 1, 37–77 Zbl 0979.35100 MR 1827277

[61] P. Topping, Winding behaviour of finite-time singularities of the harmonic map heat flow. *Math. Z.* **247** (2004), no. 2, 279–302 Zbl 1067.53055 MR 2064053

[62] C. Wang, Bubble phenomena of certain Palais–Smale sequences from surfaces to general targets. *Houston J. Math.* **22** (1996), no. 3, 559–590 Zbl 0879.58019 MR 1417632

[63] G. B. Whitham, *Linear and nonlinear waves*. Pure Appl. Math., John Wiley & Sons, Inc., New York, 1999, 636 pp. Zbl 0940.76002 MR 1699025

[64] N. J. Zabusky and M. D. Kruskal, Interaction of "solitons" in a collisionless plasma and the recurrence of initial states. *Phys. Rev. Lett.* **15** (1965), no. 6, 240–243 Zbl 1201.35174

Jacek Jendrej
Institut de Mathématiques de Jussieu, Sorbonne Université, 4 place Jussieu, 75005 Paris, France; jendrej@imj-prg.fr

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/38

Recent progress on sparse equidistribution

Adam M. Kanigowski

Abstract. Given a dynamical system (X, T) and a sequence of natural numbers (a_n), one can study the distribution of the sequence $\{T^{a_n}x\}$ for a given $x \in X$. In this note, we survey recent progress on this topic. We also discuss some open problems.

1. Introduction

Classical ergodic theory studies long-term evolution of a measure-preserving system (X, T, μ). That is, given a point $x \in X$ (or a subset $A \in X$), one is interested in the behavior of the orbit $\{T^n x\}_{n\in\mathbb{N}}$ (or the sets $\{T^n A\}_{n\in\mathbb{N}}$). Two of the most fundamental theorems in the theory are the ergodic theorems by von Neumann [50] (mean ergodic theorem) and Birkhoff [3] (pointwise ergodic theorem). They give a qualitative asymptotics for the behavior of orbits of typical points in the system. Recall that Birkhoff ergodic theorem asserts in particular that for all points $x \in X$ (outside of a measure 0 set) and for every $f \in L^1(X, \mu)$, the ergodic averages $1/N \sum_{n\le N} f(T^n x)$ converge to a limit $f^*(x)$. Moreover, if the measure μ is ergodic, then $f^*(x) = \int_X f d\mu$. Often, one is also interested in studying a fixed initial condition, i.e., given a topological dynamical system (X, T), $x \in X$ and $f \in C(X)$, one is interested in asymptotic behavior of $1/N \sum_{n\le N} f(T^n x)$. In general, one cannot expect asymptotics for all points in the system; however, there are some additional conditions under which one can guarantee convergence for all points (unique ergodicity is the best-known such condition).

Birkhoff's ergodic theorem has an analogue for averaging along some sparse (structured) subsets of the integers. In his fundamental work, Bourgain [4] has shown that if $P(\cdot)$ is a non-constant polynomial with integer coefficients, then for every measure-theoretic dynamical system (X, T, μ) and every $f \in L^2(X, \mu)$, the averages given by $1/N \sum_{n\le N} f(T^{P(n)}x)$ converge almost surely to a limit. This has been generalized by Wierdl [52] to averages along primes. In this note, we will be

Mathematics Subject Classification 2020: 37A25 (primary); 37C40 (secondary).
Keywords: dynamical systems, ergodic theory, equidistribution.

interested in the everywhere counterpart of classical averages. More precisely, given a sequence of natural numbers (a_n), a topological dynamical system (X, T), a point $x \in X$ and $f \in C(X)$, one is interested in (topological) conditions under which averages $1/N \sum_{n\le N} f(T^{a_n}x)$ have a limit. Let us mention a result from [26] where the authors show that if (a_n) is an almost sparse[1] sequence, then there exists a minimal, uniquely ergodic, entropy zero system (X, T), a point $x \in X$ and $g \in C(X)$ such that the averages $1/N \sum_{n\le N} g(T^{a_n}x)$ do not have a limit. Therefore, it seems some stronger (quantitative) conditions on the system (X, T) are needed to guarantee convergence. Recently, there has been significant progress in understanding sparse equidistribution in dynamical systems. In this note, we will highlight some classical results, discuss recent developments and formulate some open problems. We will focus on the cases where a_n is (1) a polynomial sequence (this will be discussed in Section 2), (2) a sequence of primes and k-almost primes (this will be discussed in Section 3) and (3) an arithmetic function, like the prime counting function $\Omega(\cdot)$ (this will be discussed in Section 4). In Section 5, we will formulate some open problems.

2. Polynomial sequences

In this section, we will consider the case $a_n = P(n)$, where $P \in \mathbb{Z}[x]$ is a polynomial with integer coefficients. We will also discuss some special cases of generalized polynomials, i.e., $a_n = n^r$, where $r \in \mathbb{R}_+$. It turns out that there are not too many methods to study polynomial sequences, especially for degree ≥ 2. One classical way of studying polynomial orbits is the so-called van der Corput inequality, which we now recall.

Lemma 2.1 (van der Corput). *Let (b_n) be a sequence of complex numbers bounded in absolute value by 1. Then for every $1 \le H \le N$, we have*

$$\left|\frac{1}{N}\sum_{n\le N} b_n\right| \ll \left(\frac{1}{H}\sum_{h\le H}\left|\frac{1}{N}\sum_{n\le N} b_{n+h}\overline{b_n}\right|\right)^{1/2} + \frac{1}{H^{1/2}} + \frac{H^{1/2}}{N^{1/2}}.$$

Given a dynamical system (X, T), a point $x \in X$ and $f \in C(X)$, one can apply the above inequality to the sequence $b_n = f(T^n x)$. Then using the above inequality, one replaces the original problem of studying $1/N \sum_{n\le N} f(T^{P(n)}x)$ by

$$\frac{1}{N}\sum_{n\le N} f(T^{P(n)}x)\overline{f(T^{P(n+h)}x)}. \tag{2.1}$$

[1]This means that the average gap in the sequence goes to ∞. Primes and polynomials of degree > 1 form an almost sparse sequence.

Usually, the last expression is more complicated than the original one (as now one has a bilinear polynomial average); however, in some very special cases, it actually simplifies the problem. Below, we provide a classical example that shows how this simplification works.

Example 2.2. Let $X = \mathbb{T} = \mathbb{R}/\mathbb{Z}$, $T = R_\alpha(x) = x + \alpha \bmod 1$, $\alpha \notin \mathbb{Q}$, $f_k(x) = e^{2\pi i k x}$, and let $P(z) = z^2$. Note that the original expression in this case reduces to studying $\sum_{n\le N} e^{2\pi i k n^2 \alpha}$, and the one in (2.1) becomes $\sum_{n\le N} e^{2\pi i k (n^2-(n+h)^2)\alpha}$ and $n^2 - (n+h)^2 = -h^2 - 2nh$ becomes a linear expression in n which is much easier to study.

In fact, the above example is the main idea behind Weyl equidistribution theorem [51].

Theorem 2.3. *For $(R_\alpha, \mathbb{T})$ with $\alpha \notin \mathbb{Q}$, polynomial orbits equidistribute for every initial condition $x \in \mathbb{T}$.*

Recall that a sequence $(a_n) \in X$ is equidistributed with respect to a measure κ if for every $f \in C(X)$,

$$\frac{1}{N}\sum_{n\le N} f(a_n) \to \int f d\kappa \quad \text{as } N \to \infty.$$

The main idea is to use repeatedly the van der Corput inequality to reduce the degree of the polynomial. In fact, from the dynamical point of view, the reason why (2.1) is simpler in the case of rotations is that the characters f_k are eigenfunctions of the system, i.e., $f_k(R_\alpha x) = e^{2\pi i k\alpha} f_k(x)$. One can further explore this idea (of generalized eigenfunctions) to study polynomial orbits for nil-rotations on nilgroups. This was done by Leibman [31] and by Green and Tao [22] in a quantitative way. More precisely, we have the following.

Theorem 2.4. *Let G be a nilpotent Lie group, and let Γ be a lattice in G. Let $X = G/\Gamma$, and let $g \in G$. Define $T = T_g(x\Gamma) = gx\Gamma$. Then for every $P \in \mathbb{Z}[x]$ and for any $x \in X$, the polynomial orbit $\{T^{P(n)}x\}$ is equidistributed in $\overline{\{T^n x\}}_{n\in\mathbb{Z}}$.*

As mentioned above, one way to prove the above theorem is to use the existence of so-called generalized eigenfunctions for nil-translations. However, when there are no (generalized) eigenfunctions in the system, then essentially always the expression in (2.1) becomes more difficult than the initial problem (as it is still a genuine bilinear expression). In fact, the two classes above were, until recently, the only classes of systems for which equidistribution along the sequence of squares (for example) was known. In particular, there were no examples of weakly mixing systems (X, T)

(roughly those with no eigenfunctions) which would be equidistributed along, for example, the sequence $\{n^2\}$. Moreover, since usually if one can apply the van der Corput inequality once, one can apply it repeatedly, there were no systems known which would be equidistributed at $\{n^2\}$ but would fail to equidistribute at $\{n^3\}$. This has changed recently with the works of Kasprzak [28,29]. In [28], the author showed existence of weakly mixing dynamical systems which are equidistributed at squares (and other polynomial sequences). The main new mechanism in [28] was to study systems which are rigid in a strong sense, i.e., for some sequence (n_k), $T^{n_k} \to \mathrm{Id}$ with a polynomial rate. This allowed us to circumvent the lack of eigenfunctions and instead rely on rigidity and arithmetic properties of polynomial sequences in arithmetic progressions and short intervals. In particular, the author worked directly with the polynomial expression and not the bilinear version (which in weakly mixing case is much harder to study). Also in [29], the author constructed symbolic dynamical systems (X, T) which are equidistributed at, for example, squares but fail to equidistribute at, say, cubes. The systems constructed in [29] belong to the class of so-called Toeplitz systems.

For polynomials of degree ≥ 2, the van der Corput inequality and the method from [28] are essentially the only known methods to study polynomial sequences. In particular, no single mixing system (X, T) is known which would equidistribute along the sequence of squares. Also there are algebraic examples for which equidistribution is not known. This is known as the Margulis–Shah conjecture. This conjecture in the simplest case asks about distribution of polynomial orbits of the horocycle flow on finite volume quotients of $\mathrm{SL}(2,\mathbb{R})$. Some progress toward this conjecture was made (see the following section), but it is limited to sub-quadratic polynomials. We mention more open questions related to that in Section 5.

2.1. Sub-quadratic polynomials

As mentioned above, starting with polynomials of degree 2, there are no general methods to approach, for example, mixing systems. However, there are some recently developed methods for polynomials of the form $\{n^{1+\delta}\}$, $\delta \in (0,1)$. It will be convenient in this section to restrict our attention to flows, i.e., an $\mathbb{R}$-parameter family (T_t) acting on a space X (since one can then talk about $\{T_{n^{1+\delta}}x\}$).

The first result of this type comes with the seminal work of Venkatesh [48], in which the author studied equidistribution properties of horocycle flows acting on quotients of $\mathrm{SL}(2,\mathbb{R})$ at times $n^{1+\delta}$ with a small positive δ. We will now recall the needed definitions. Let $G = \mathrm{SL}(2,\mathbb{R}) = \left\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) : ad - bc = 1\right\}$, and let Γ be a lattice in G (a discrete subgroup such that the quotient has finite volume). Let $h_t = \left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$ be the horocycle flow acting on $X = G/\Gamma$ by $h_t(x\Gamma) = (h_t x)\Gamma$. Horocycle flows are one of the fundamental objects in dynamical systems, ergodic theory and number theory.

Their dynamical and ergodic properties have been extensively studied. As mentioned above, due to their algebraic nature and tame (polynomial) behavior, it is generally believed that horocycle flows should equidistribute at polynomial times. This is in fact the content of the already mentioned Margulis–Shah conjecture (see Section 5). In [48], the following result was proved.

Theorem 2.5. *Let (h_t) act on X, where X is compact. There exists $\delta_\Gamma > 0$ such that for every $0 < \delta < \delta_\Gamma$, every $x \in X$, the orbit $\{h_{n^{1+\delta}} x\}$ is equidistributed with respect to the Haar measure on X.*

In fact, the method above gives (in many instances) $\delta_\Gamma = 1/48$; however, the author did not optimize it. This theorem was generalized by Flaminio–Forni–Tanis [16] to get rid of the dependence on Γ. More precisely, the authors showed that in the compact setting, the sequence $\{h_{n^{1+\delta}} x\}$ equidistributes for any $\delta < 1/12$. In the work of Streck [45], the author generalized the above results to the case of non-compact X (where the formulation is more delicate due to existence of periodic orbits). The approach in [48] uses so-called twisted ergodic integrals in which one twists the classical ergodic integral by an exponential phase. In fact, the two main properties of the horocycle flow that are used in the proof are polynomial mixing and polynomial equidistribution. Let us make a general definition to explain the general result.

Definition 2.6. Let (T_t) be a C^∞ flow on a smooth compact manifold M that preserves a measure μ. We say that (T_t) is *polynomially mixing on C^k* if there exists $\eta > 0$ and $C_k > 0$ such that for any $\phi, \psi \in C^k(M)$,

$$\left| \int_X \phi \circ T_t \cdot \psi d\mu - \int_X \phi d\mu \int_X \phi d\mu \right| \leq C_k \|\phi\|_k \|\psi\|_k t^{-\eta}.$$

We say that (T_t) has *polynomially decaying ergodic averages on C^k* if there exist $\kappa > 0$ and $C'_k > 0$ such that for any $\phi \in C^k(M)$,

$$\left| \int_0^R \phi(T_t x) dt - R \int_X \phi d\mu \right| \leq C'_k \|\phi\|_k R^{1-\kappa}.$$

It is known that horocycle flows are both polynomially mixing on C^k [23, 39] and have polynomially decaying ergodic averages on C^k [15, 46]. The main idea in the method of Venkatesh is to use the above two properties. In fact, as was noted in a paper by Auer [1], this approach works in the general case. We have the following proposition.

Proposition 2.7. *Let (T_t) be a uniquely ergodic flow on M which is polynomially mixing and has polynomially decaying averages (for some $k \in \mathbb{N}$). Assume that (T_t)*

has polynomially growing derivatives.[2] *Then there exists* $0 < \delta_0 = \delta_0(\eta, \kappa) < 1$ *such that* $\{h_{n^{1+\delta}}x\}$ *is equidistributed (with respect to the unique ergodic measure) for any* $x \in X$ *and any* $0 < \delta < \delta_0$.

As noted above, the δ_0 depends on the mixing rate and the equidistribution rate. The above methods cannot go beyond the case $\delta_0 = 1/2$, and so other methods are needed to cover the cases $\{n^{1+\delta}\}$, $\delta \geq 1/2$. This problem remains widely open, and there are currently no new methods to prove equidistribution in this case. However, if one is interested in density of horocyclic orbits at polynomial times instead of equidistribution, one can employ a different approach to cover the cases $\delta \in [0, 1)$. In a recent preprint [27], the authors have shown that if $X = \mathrm{SL}(2, \mathbb{R})/\Gamma$ is not compact, then for every non-periodic point $x \in X$, the orbit $\{h_{n^{1+\delta}}x\}$ is dense.

Theorem 2.8. *Let* Γ *be a non-compact lattice in* $\mathrm{SL}(2, \mathbb{R})$. *Then for every* $\delta \in (0, 1)$, *every point* $x \in X = \mathrm{SL}(2, \mathbb{R})/\Gamma$ *which is not periodic for* (h_t), *the orbit* $\{h_{n^{1+\delta}}x\}$ *is dense in* X.

Of course, the case of small $\delta > 0$ is already covered by [45] (with equidistribution not only density), so the above theorem becomes most interesting for δ close to 1. The methods of showing density are different from those that show equidistribution for small $\delta > 0$. The main idea in the proof of the above theorem is an approximation scheme, along a sequence of times, of a long enough piece of an orbit of x by a periodic orbit. For this, one uses quantitative equidistribution of horocycle flows together with non-divergence results and harmonic analysis methods on the circle to study the twist in the approximation. In particular, this method does not apply to the case of compact X (as there are no periodic points in this case). This approximation however stops working at the endpoint $\delta = 1$: one can show that the optimal piece of an orbit of $x \in X$ which is approximated by a periodic orbit at time $T > 0$ cannot in general be longer than $c \cdot T^{1/2}$ (for some small $c > 0$). In particular, this interval contains at most one square, and so one cannot use it to study square orbits (or any polynomial orbits with polynomials of degree ≥ 2). A similar approach can be used to study orbits of the horocycle flow sampled at $\Omega(n)$ (the prime factors counting function). We discuss this case in Section 4.

[2]This means that $\|\phi \circ T_t\| \ll_k t^{kd}\|\phi\|_k$, where $\|\cdot\|_k$ is the C^k norm. This holds for the horocycle flow.

3. Primes and almost primes

In this section, we will discuss results on so-called prime number theorems in dynamics. Given a dynamical system (X, T) and a point $x \in X$, one is interested in the behavior of the orbit $\{T^p x\}_{p \text{ is prime}}$. More generally, one also considers orbits along k-almost primes, i.e., numbers with at most k-prime factors. At first look, studying primes seems related to Sarnak's conjecture (see Section 4 for details). However, the methods used to study Sarnak's conjecture require softer input, both on the dynamical and number theory side. The number-theoretic tools (for example, [7, 8, 30, 34, 47]) that are used in the context of Sarnak's conjecture rely on the fact that most integers are composite and thus factor. These methods cannot be used for studying primes or k-almost primes. A dynamical prime number theorem has so far been established in some algebraic and symbolic examples: rotations on $\mathbb{T}$ [49], nilsystems [22], Rudin–Shapiro sequences [37], enumeration systems [5, 21], certain finite rank symbolic systems [6], automata [40]. The main (classical) tool used to establish a prime number theorem is the method of type I and type II sums. We recall one of the formulations due to Duke, Friedlander and Iwaniec [10].

Theorem 3.1 (A criterion for PNT). *Let (x_n) be a bounded sequence. If for every $\varepsilon > 0$, both of the following conditions hold:*

- *type I sums (of level $1/2$): for $D \le N^{1/2-\varepsilon}$,*

$$\frac{1}{D} \sum_{D \le d \le 2D} \left| \frac{d}{N} \sum_{n \le N/d} x_{dn} \right| \ll_\varepsilon \frac{1}{\log^3 N},$$

- *type II sums (of level $1/3$): for $D \le N^{1/3-\varepsilon}$,*

$$\frac{1}{D^2} \sum_{D < d_1, d_2 < 2D} \frac{1}{\min(N/d_1, N/d_2)} \left| \sum_{n \le \min(N/d_1, N/d_2)} x_{d_1 n} \overline{x_{d_2 n}}, \right| \ll_\varepsilon \frac{1}{\log^{22} N},$$

then

$$\sum_{p \le N} x_p = o(\pi(N)).$$

We should mention that there are variations in the above criterion, where one can adjust a bit between the level of distribution of type I and type II sums. However, what is important, and also makes the problem often very difficult, is that one needs quantitative information on both of these sums and the step, i.e., the numbers d, d_1, d_2 depend on N. Let us discuss the case of type II sums (type I sums are usually easier).

Notice that if we apply the criterion to $x_n = f(T^n x)$ for some dynamical system (X, T), $x \in X$, $f \in C(X)$, then type II sums are expressions of the form

$$\sum_{n \leq \min(N/d_1, N/d_2)} f(T^{d_1 n} x)\overline{f(T^{d_2 n} x)},$$

where the step d_1, d_2 can be of size $N^{1/3-\varepsilon}$. This makes it impossible to apply any classical tools from ergodic theory when one fixes the transformation and then takes longer and longer averages. This also highlights the difference between Sarnak's conjecture and primes, as in the former case, the DKBSZ orthogonality criterion, allows one to fix d_1, d_2 and then let N go to ∞. Therefore, for all the cases where a prime number theorem was established, the authors used some quantitative information about the system that would usually allow one to reduce type II sums to type I sums (as is the case for rotations) and then study type I sums (arithmetic progressions) using either the algebraic or symbolic structure of the system.

Relatively recently, the authors in [26] developed a new method to study primes that avoids studying type I and type II sums for the system directly but rather approximates the system with a sequence of periodic systems (with a polynomial twist in the phase) and then relies on results on type I and type II for polynomial exponential sums. In fact, the authors also need to use type III sums as the level for type II sums is barely sufficient. The main theorem in [26] is the following.

Theorem 3.2. *Let $X = \mathbb{T}^2$ and $g : \mathbb{T} \to \mathbb{R}$ be an analytic function. Let $T : \mathbb{T}^2 \to \mathbb{T}^2$ be given by $T(x, y) = (x + \alpha, y + g(x))$. If T is uniquely ergodic, then it satisfies a prime number theorem.*

A weaker question one might ask is density of prime orbits, i.e., is it true that for a given $x \in X$, the orbit $\{T^p x\}_{p \text{ is prime}}$ becomes dense in X? A still weaker question is on non-concentration of primes, i.e., can $\{T^p x\}$ be trapped in a ball of small radius? In fact, to study these questions (especially non-concentration), one can use sieve methods, i.e., the upper bound sieve [24]. This was used by Sarnak and Ubis [43] to show non-concentration of prime orbits for the horocycle flow for $\Gamma = \mathrm{SL}(2, \mathbb{Z})$. The main benefit is that the upper bound sieve only uses type I sums (arithmetic progressions). These in turn can be studied using quantitative equidistribution results for horocycle flows. The main result in [43] is the following.

Theorem 3.3. *Let $\Gamma = \mathrm{SL}(2, \mathbb{Z})$. Let $x \in X = \mathrm{SL}(2, \mathbb{R})/\Gamma$ be a non-periodic point. Then $\mu_{\mathrm{Haar}}(\overline{P_x}) \geq 1/10$, where $P_x = \{h_p x\}_{p \text{ prime}}$.*

This result was generalized by Streck [44] to general non-compact lattices in $\mathrm{SL}(2, \mathbb{R})$. As mentioned in [43], the upper bound sieve and quantitative equidistribution could be used to show density of orbits sampled at k-primes for k large enough.

This was done by McAdam, who proved the following (we state the result for the horocycle flow only; the result in [38] is true for horospherical actions in $\mathrm{SL}(n,\mathbb{R})$).

Theorem 3.4. *Let Γ be co-compact. Then there exists a $k = k(\Gamma) \in \mathbb{N}$ such that the orbit of every $x \in X = \mathrm{SL}(2,\mathbb{R})/\Gamma$ is dense when sampled at k-primes (recall that these are numbers with at most k-prime factors).*

It should be mentioned that the above results use type I sums only, and so the k is large: as already mentioned in [43], the methods work for $k \geq 10$. In what follows, we will discuss the case of 2-primes for which we recently obtained equidistribution results for horocycle flows.

3.1. Equidistribution along semi-primes

In this section, we discuss equidistribution along 2-primes. Recall that these are integers of the form $p_1 \cdot p_2$ with p_i prime. It seems that to study 2-primes, type I information is not sufficient. In a recent paper [17], the authors established a criterion that allows one to study 2-primes for some classes of dynamical systems. The precise criterion is the following.

Proposition 3.5 ([17]). *Let $\varepsilon > 0$ be given. For all $N \geq 2$, let $\mathcal{P}_{\varepsilon,N}$ be the set of primes in the interval $[\exp(\log^{\varepsilon} N), \exp(\log^{1-\varepsilon} N)]$ with $\exp(\log^{\varepsilon} N) > (\log N)^{1000}$, and let $\mathcal{S}_{\varepsilon,N}$ be a subset of $\mathcal{P}_{\varepsilon,N}$ with the property that in every dyadic interval $[P, 2P]$, we discard at most $\ll P/(\log N)^{100}$ primes. Let (a_n) be a sequence with $|a_n| \leq 1$. Suppose that, for $N \geq 2$,*

$$\sum_{n \leq N} a_{nq_1}\overline{a_{nq_2}} \ll \frac{N}{(\log N)^{100}} \tag{3.1}$$

for primes $q_1, q_2 \in \mathcal{S}_{\varepsilon,N}$ with $1/5 \leq q_1/q_2 \leq 5$ and $q_1 \neq q_2$. Then,

$$\sum_{pq \leq N} a_{pq} \ll \varepsilon \cdot \sum_{pq \leq N} 1 + \frac{1}{\varepsilon^{53}} \cdot \frac{N}{(\log N)^{50}}\,. \tag{3.2}$$

Moreover, assume that for every $M \in [N^{9/10}, N]$, for N large enough,

$$\sum_{n \leq M} a_{nq_1}\overline{a_{nq_2}} \ll \frac{M}{(\log\log M)^{10}}$$

holds for all $q_1 \neq q_2$, with $q_1, q_2 \in [e^{(\log\log\log N)^3}, e^{(\log\log N)^{10}}]$ and $1/5 \leq q_1/q_2 \leq 5$. Then for every multiplicative function ν with $|\nu| \leq 1$, for all $N \in \mathbb{N}$ large enough,

$$\left|\sum_{n \leq N} \nu(n) a_n\right| \ll N(\log\log N)^{-4}\,.$$

The main parts of the criterion are (3.1) and (3.2). The part on multiplicative functions is a quantitative version of the orthogonality criterion already studied in [7]. Notice that (3.1) is still a bound on type II sums. However, the main difference with respect to the criterion for primes is that the level of distribution is much lower. In fact, it follows from the statement that if there exists $\gamma > 0$ such that for any two primes $q_1, q_2 \leq N^\gamma$, we have that (3.1) holds, then we can deduce equidistribution along 2-primes (recall that classically for primes, one needs $\gamma = 1/3$). Therefore, this general criterion seems to be applicable to a wider class of dynamical systems. In [17], the authors use this criterion to show equidistribution of horocycle flows along 2-primes. The main result in [17] is the following.

Theorem 3.6. *Let Γ be a co-compact arithmetic lattice in $\mathrm{SL}(2,\mathbb{R})$, or let $\Gamma = \mathrm{SL}(2,\mathbb{Z})$. Then for every non-periodic point $x \in X = \mathrm{SL}(2,\mathbb{R})/\Gamma$, the orbit $\{h_{p_1\cdot p_2}x\}_{p_1,p_2 \text{ primes}}$ is equidistributed in X.*

The main input for the theorem is a quantitative type II sums estimate (at any polynomial level). This reduces (in the case of horocycle flows) to studying quantitatively deviation of ergodic averages for the product flow $h_t \times h_t$ on $X \times X$. Until recently, this was a major open problem in homogeneous dynamics. Qualitative results were established in the landmark work of M. Ratner in the 1980s [41]; however, no rates could be obtained from Ratner's work. In a recent breakthrough of Lindenstrauss–Mohammadi–Wang [32] (see also [53] for related work), the authors obtained a quantitative result for the square of the horocycle flow (with a polynomial error). To state the main result of [32], let $a_t = \left(\begin{smallmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{smallmatrix}\right)$ be the geodesic flow on X. Notice the following renormalization relation between the geodesic and horocycle flows:

$$a_t h_s = h_{e^t s} a_t.$$

Let H denote the diagonal embedding of $\mathrm{SL}(2,\mathbb{R})$. The main result of [32] is as follows.

Theorem 3.7. *Assume Γ is an arithmetic lattice. For every $(x,y) \in X \times X$ and large enough R (depending explicitly on X), for any $T \geq R^{A_1}$, at least one of the following holds:*

(1) *For every $\varphi \in C_c^\infty(X \times X)$, we have*

$$\left| \frac{1}{T} \int_0^T \varphi(h_t \times h_t(x,y)) - \int \varphi d\mu \times \mu \right| \leq S(\varphi) R^{-\kappa_2},$$

where $S(\varphi)$ is a certain Sobolev norm.

(2) *There exists a periodic* $z \in X \times X$ *with* $\mathrm{vol}(Hz) \leq R^{A_1}$, *and for every* $r \in [0, T]$, *there exists* $g \in \mathrm{SL}(2,\mathbb{R})^2$ *with* $\|g\| \leq R^{A_1}$ *so that*

$$d_{X\times X}((h_s \times h_s)(x,y), gHz) \leq R^{A_1}\left(\frac{|s-r|}{T}\right)^{1/A_2} \quad \text{for all } s \in [0,T].$$

(3) *For every* $r \in [0, T]$ *and* $t \in [\log R, \log T]$ *the injectivity radius at* $(a_{-t} \times a_{-t})((h_r \times h_r)(x, y))$ *is at most* $R^{A_1} e^{-t}$.

The constants A_1, A_2 *and* κ *are positive and depend on* X *but not on* (x, y).

We should point out that the meaning of periodic orbit in the above theorem is that it is a closed H-invariant subspace (not necessarily a periodic point). We will not define some of the terms in the above theorem (like volume of a periodic orbit). The above theorem is optimal in the following sense: it is known and not so hard to see that if a point satisfies (2) or (3), then it cannot be polynomially equidistributed (i.e., it cannot satisfy (1)). The significance of the above result lies in the fact that the straightforward obstructions to (1) are the only obstructions. Let us now briefly comment on how to approach Theorem 3.6. We use the criterion for 2-primes which requires studying ergodic averages of length N for the map $h_{q_1} \times h_{q_2}$, where q_i can be as large as N^{ε}. Notice that using renormalization (at times $\log q_i$), one can reduce this problem to studying ergodic averages for the map $h_1 \times h_1$ at a point $(a_{-\log q_1}x, a_{-\log q_2}x)$ and the function $F(y_1, y_2) = f \circ a_{\log q_1}(y_1) \cdot f \circ a_{\log q_2}(y_2)$ (where we need a polynomial gain over the trivial bound for the averages). This puts us in the position to apply Theorem 3.7 with $R = N^{\kappa}$ for some small κ. Using this, we get that either we have a polynomial saving in the ergodic average for $h_1 \times h_1$ (and we can apply the criterion) or there are q_1, q_2 such that the point $(a_{-\log q_1}x, a_{-\log q_2}x)$ satisfies (2) or (3). Here, the approach is different in the case of co-compact lattices and $\mathrm{SL}(2,\mathbb{Z})$. First note that if the lattice is co-compact, then (3) can never hold since the injectivity radius is uniformly bounded below on X. Moreover, using some algebraic considerations, we show that the point $(a_{-\log q_1}x, a_{-\log q_2}x)$ never satisfies (2) for co-compact arithmetic lattices. The argument for this can be thought of as a quantitative version of the corresponding argument in [7], where the authors showed that $(a_{-\log q_1}x, a_{-\log q_2}x)$ never lies on an invariant subvariety. In our case, we need to relate the largeness of the q_i's with the volume of the subvariety.

The most difficult part of the paper is studying (2) and (3) for points given by $(a_{-\log q_1}x, a_{-\log q_2}x)$ and the lattice $\mathrm{SL}(2,\mathbb{Z})$. In this case, it can happen that both (2) and (3) are satisfied, but this gives us enough restrictions on the starting point x to complete the argument using a different approach. To skip the technical details, we will only focus on highlighting the main difficulties in both cases. The model behavior for which (3) holds is if x itself is very close to a periodic point (for (h_t)) of

length $N^{\kappa'}$. In this case (skipping lots of details), the analysis boils down to the analysis on the circle and the Hardy–Littlewood method. However, since the period might be polynomially large, we cannot use directly classical results from number theory, for example, the Siegel–Walfisz theorem (which only works for periods of the form $\log^A N$). We use stronger results on prime in arithmetic progressions of polynomial size, where the result is twisted by a quadratic character. More precisely, we prove (let $\pi_2(N)$ denote the number of semi-primes up to N) the following.

Proposition 3.8. *Let $J \subset [0,T]$ be an interval of length $|J| \geq T^{1-\kappa}$, and let $r < T^{\kappa}$. Then*

$$|\{pq \in J : pq \equiv a \mod r\}| = 1_{(a,r)=1}\frac{\pi_2(J)}{\varphi(r)} + \chi(a)\cdot\frac{F_{J,r,T}}{\varphi(r)} + \mathrm{O}\left(\frac{1}{\varphi(r)}\kappa\pi_2(J)\right),$$

where χ is a quadratic real character and $F_{J,r,T}$ is such that $|F_{J,r,T}| \leq \pi_2(J)$ for all large enough T.

If the period is irrational (or not too close to rational), we apply minor arc-type estimates, but there is an extra problem to overcome: we need to get polynomial saving. If the point $(a_{-\log q_1}x, a_{-\log q_2}x)$ satisfies (2), then we use Ratner's theory to relate the periodic orbit Hz with a lattice Γ_{q_1,q_2} and the corresponding horocycle flow on $X_{q_1,q_2} = \mathrm{SL}(2,\mathbb{R})/\Gamma_{q_1,q_2}$. In this case, we need a quantitative equidistribution result for the horocycle flow with polynomial saving which would be uniform over all q_1, q_2 up to size N^{ε}. This requires a new input, as earlier results [15, 46] obtain bounds for a fixed lattice. Using the methods of [15], we prove the following.

Theorem 3.9. *Let $X_\Gamma = \mathrm{SL}(2,\mathbb{R})/\Gamma$. For every $s \geq 4$, there exists a constant $C_s > 0$ such that for every function $f \in W^s(X_\Gamma)$ and for all $(x,T) \in X_\Gamma \times [1,+\infty)$,*

$$\left|\frac{1}{T}\int_0^T f\circ h_t(x)dt - \int_{X_\Gamma} f\,d\,\mathrm{vol}_\Gamma\right| \leq C_s\|f\|_{W^s(X_\Gamma)}\max\left\{\mathrm{inj}_\Gamma^{-1}, e^{(1-\nu_\Gamma)/2 d_\Gamma(a_{\log T}(x))}\right\}T^{-(1-\nu_\Gamma)/2},$$

where $\nu_\Gamma = \sqrt{1-4\mu_\Gamma}$ and $\mu_\Gamma > 0$ is the bottom of the spectrum of the Laplace operator and $d_\Gamma(\cdot)$ denotes the distance to e in $X = \mathrm{SL}(2,\mathbb{R})/\Gamma$.

In Theorem 3.9, the volume $d\,\mathrm{vol}_\Gamma$ is normalized, while the Sobolev spaces are defined with respect to the constant curvature metric (whose volume is not normalized). As one can see, these estimates depend explicitly on the lattice and involve the injectivity radius of the compact part M_Γ and the number ν_Γ. We then show that the lattices Γ_{q_1,q_2} are congruence lattices, and so we have a uniform bound on ν_Γ and also on inj_Γ. We also need to use a result of [11] to get estimates on closeness of vol_Γ to the product Haar measure.

4. Arithmetic functions

In this section, we briefly discuss recent developments around correlation between dynamically defined sequences and multiplicative functions. This area has seen a big development in the last 15 years. We will focus only on main techniques that led to the progress and not so much on the particular results. For a detailed discussion, we refer the reader to [14]. In the last subsection, we will discuss the case of the $\Omega(\cdot)$ function (prime counting function). The interest in correlation of multiplicative functions with dynamical sequences was initiated by the following conjecture of Sarnak [42].

Conjecture 4.1 (Sarnak). *Let (X, T) be a dynamical system of topological entropy zero. Let $\mu(\cdot)$ be the Möbius function, i.e., $\mu(n) = 0$ if n is divisible by a square and otherwise $\mu(n) = (-1)^k$, where k is the number of different prime divisors of n. Then for any $f \in C(X)$ and any $x \in X$,*

$$\sum_{n\le N} \mu(n) f(T^n x) = o(N). \tag{4.1}$$

Even though this conjecture is open in full generality, there has been substantial progress in verifying the conjecture for various (natural) classes of dynamical systems (see [14]). The first main tool is the so-called orthogonality criterion [7, 8, 30] which allows one to study correlations with (general) multiplicative functions by studying type II sums qualitatively, i.e., to show that (4.1) holds, it is enough to show that for two different prime numbers q_1 and q_2,

$$\sum_{n\le N} f(T^{q_1 n} x)\overline{f(T^{q_2 n} x)} = o(N).$$

The fact that we only need qualitative information simplifies the situation dramatically as now one is interested in classical ergodic averages for the map $T^{q_1} \times T^{q_2}$. This is strongly connected to the theory of joinings in ergodic theory introduced by H. Furstenberg [20]. The other source of progress around Sarnak's conjecture is related to results on behavior of multiplicative functions in short intervals. Here, the breakthrough came with the work of Matomäki and Radziwiłł [37], who proved the following result.

Theorem 4.2. *Let $f : \mathbb{N} \to [-1, 1]$ be a multiplicative function. Then for any $2 \le H \le X$,*

$$\left| \frac{1}{H} \sum_{n\in[x,x+H]} f(n) - \frac{1}{X} \sum_{n\in[X,2X]} f(n) \right| = o_{H\to\infty}(1)$$

for all but $o(X)$ of the intervals $[x, x + H]$.

The strength of the above theorem is that the H can go to ∞ arbitrarily slowly with X. These types of results are beyond rich (or sometimes not true) for k-primes or polynomials. These results, together with related works, [35, 36] allow for many classes of dynamical systems (X, T) to reduce the problem to qualitative behavior of T on typical short intervals where one can again use tools from ergodic theory. We also mention other methods to study Sarnak's conjecture such as Furstenberg systems [18, 19], and we refer the reader to these papers for more details.

4.1. The function Ω

Here, we discuss some results on distribution of orbits in dynamical systems along the sequence $\Omega(n)$. Recall that if $n = \prod_{i=1}^{k} p_i^{\alpha_i}$ is the prime factorization of n, then $\Omega(n) = \sum_{i=1}^{k} \alpha_i$. In [2], Bergelson and Richter proved the following general result.

Theorem 4.3. *Let (X, T) be uniquely ergodic (with the unique measure μ). Then for every $f \in C(X), x \in X$,*

$$\left| \frac{1}{N} \sum_{n \leq N} f(T^{\Omega(n)} x) - \int_X f d\mu \right| = o(1).$$

This theorem is general as the only assumption is unique ergodicity. It is also necessary: in [33], Loyd provided an example of a non-uniquely ergodic topological system for which the above averages do not converge for some $x_0 \in X$. More generally, for a dynamical system (X, T) and $x \in X$, one defines the empirical measures $\nu_N(x) = 1/N \sum_{n \leq N} \delta_{T^n x}$. One can then define $\mathrm{Acc}(x)$ to be all the accumulation points of the sequence $\{\nu_N(x)\}$ (note that all measures in $\mathrm{Acc}(x)$ are T-invariant). For a uniquely ergodic system, it follows that $\mathrm{Acc}(x)$ is a singleton for any $x \in X$. Analogously, one can define $\nu_N^{\Omega}(x) = 1/N \sum_{n \leq N} \delta_{T^{\Omega(n)} x}$ and $\mathrm{Acc}^{\Omega}(x)$ to be the accumulation points of the measures $\{\nu_N^{\Omega}(x)\}$. It follows from [2] that all the elements of $\mathrm{Acc}^{\Omega}(x)$ are T-invariant. Then one can rephrase the above theorem by saying that for uniquely ergodic systems, $\mathrm{Acc}^{\Omega}(x) = \mathrm{Acc}(x)$. The relation between the sets $\mathrm{Acc}^{\Omega}(x)$ and $\mathrm{Acc}(x)$ can be studied for general dynamical systems (X, T). In [25], the authors show that one of the inclusions persists for ergodic measures.

Theorem 4.4. *Let (X, T) be any topological dynamical system. Then for any $x \in X$,*

$$M_T^e(X) \cap \mathrm{Acc}(x) \subset \mathrm{Acc}^{\Omega}(x).$$

In other words, the theorem says that if a point $x \in X$ is quasi-generic for an ergodic measure μ, then the measure μ can be seen as a limit of a subsequence of Ω-empirical measures for x. It would be interesting to know if the ergodicity assumption can be dropped from the above theorem (see Section 5), but the current method

of proof uses ergodicity in an essential way. The main number-theoretic input behind this theorem is results on statistical behavior of Ω. It follows from Erdös and Kac [13] that the values of Ω become normally equidistributed.

Theorem 4.5 (Erdös–Kac theorem). *For $A < B \in \mathbb{R}$,*

$$\lim_{N\to\infty} \frac{1}{N} \#\left\{ n \le N : A \le \frac{\Omega(n) - \log\log N}{\sqrt{\log\log N}} \le B \right\} = \frac{1}{\sqrt{2\pi}} \int_A^B e^{-t^2/2} dt.$$

In fact, we need a stronger result (a local limit theorem). This was proven by Erdös [12, Theorem II].

Given $k \in \mathbb{N}$, let $\mathbb{P}_k$ denote the set of k-almost primes, those integers satisfying $\Omega(n) = k$. Let $\pi_N(k)$ denote the number of k-almost primes less than or equal to N. Then

$$\pi_N(k) = \#\{n \le N : \Omega(n) = k\} = \sum_{n \in \mathbb{P}_k \cap [N]} 1.$$

The result of Erdös provides a uniform estimate for $\pi_N(k)$ for certain values of k. For $N \in \mathbb{N}$, define the interval I_N by

$$I_N := \left[\log\log N - C\sqrt{\log\log N}, \log\log N + C\sqrt{\log\log N}\right],$$

where $C \ge 1$ is some fixed constant. Define

$$R_N := \log\log N + C\sqrt{\log\log N}$$

to be the right-hand endpoint of this interval.

Lemma 4.6 ([12]). *The function $\pi_N(k)$ satisfies*

$$\pi_N(k) = \frac{N}{\log N} \frac{(\log\log N)^{k-1}}{(k-1)!} (1 + o_{N\to\infty}(1))$$

uniformly for $k \in I_N$.

Using the above number-theoretic results along with ergodic-theoretic methods, we can reduce the statement to usual ergodicity on short intervals. This can be done for a specific sequence of times (depending on x).

4.1.1. Horocycle flows at $\Omega(n)$. Recall that if $X = \mathrm{SL}(2,\mathbb{R})/\Gamma$ is compact, then the horocycle flow (and also the time 1 map) is uniquely ergodic, and therefore, the theorem of Bergelson and Richter applies. However, in the non-compact case, the flow is no longer uniquely ergodic, and therefore, it makes sense to ask what happens with the horocyclic orbits along the sequence $\Omega(n)$. To state the result, we need some

definitions. Let Γ be such that X is not compact. It is known that for some $k \geq 1$, the space X has k (inequivalent) cusps, and for every cusp C_i, there is a one-parameter family of periodic measures corresponding to C_i. Denote by $\{\nu_s^i\}_{s\in\mathbb{R},i\leq k}$ the set of periodic measures for the horocycle flow. By the work of Dani and Smillie [9], all the ergodic invariant measures for the horocycle flow are either μ or one of the $\{\nu_s^i\}$, i.e., for every $x \in X$, $\mathrm{Acc}(x)$ is a singleton, and it is either equal to μ or $\nu_s^{(i)}$. The main result of [25] describes the set of accumulation points for averages along $\Omega(\cdot)$.

Theorem 4.7. *For any non-periodic point $x \in X$, we have the following:*

$$\mathrm{Acc}^{\Omega}(x) \subset \{\mu\} \cup \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{r^2}{2}} \nu^{(i)}_{s_0+2\log(|1+cr|)} dr : 0 < c,\ i \leq k,\ s_0 \in \mathbb{R} \right\}. \tag{4.2}$$

Moreover, there exists $D' > 0$ such that for every $i \leq k$, $s_0 \in \mathbb{R}$, there exists[3] *$C(s_0) > 0$ and $s' \in [s_0, s_0 + 1]$ such that for some $u' \in [C(s_0), D'C(s_0)]$, we have*

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{r^2}{2}} \nu^{(i)}_{s'+2\log(|1+u'r|)} dr \in \mathrm{Acc}^{\Omega}(x). \tag{4.3}$$

Finally, $\mu \in \mathrm{Acc}^{\Omega}(x)$.

The difference between (4.2) and (4.3) is the following: (4.2) states that any accumulation point belongs to the set on the left-hand side with *any* parameters $c > 0, i \leq k, s_0 > 0$. On the other hand, (4.3) states that given i, s_0, one can find an accumulation point *bounded distance* from them, i.e., $s' \in [s_0, s_0 + 1]$ and $u' \in [C(s_0), 2C(s_0)]$. It is interesting to point out that for a periodic point with rational period, the corresponding measure for the time 1 map of the horocycle flow is an atomic measure on finitely many points. The above result shows that these measures never appear in the decomposition for $\mathrm{Acc}^{\Omega}(x)$ for non-periodic $x \in X$, i.e., all $\{\nu_s^i\}$ appearing above are atomless. Note that the case of a periodic $x \in X$ follows from theorems on equidistribution of orbits of rotations for averages along $\Omega(\cdot)$, in particular $\mathrm{Acc}^{\Omega}(x)$ is a singleton in this case.

In the proof, we use the fact that if for a time T, the Ω orbit of length T is not equidistributed in space, then it needs to spend a positive proportion of time near a periodic orbit. But then it follows that it has to spend all its time sliding along periodic orbits of varying lengths. This is the heuristic for (4.2). To show (4.3), we need to argue that along a sequence of times, the regular orbit of length T of the horocycle flow stays $\sim T^{1/2}$ close to a periodic point. This dynamical part of the argument for

[3] In fact, $C(\cdot)$ depends continuously on s_0.

the above theorem is similar to results on density of orbits $\{h_{n^{2-\delta}}(x)\}$. We need to show that along a sequence of times, the orbit of a generic point comes sufficiently close to a given periodic orbit such that we can then use normal distribution of $\Omega(\cdot)$ to deduce (4.3).

5. Open problems

In this section, we will formulate some open problems and questions related to sparse equidistribution.

Problem 5.1. Give an example of a smooth, zero entropy, diffeomorphism (f, M) which is uniquely ergodic and for which a prime number theorem fails, i.e., there is $x \in M$ and $\phi \in C^\infty(M)$ such that the averages $1/\pi(N) \sum_{p \le N} \phi(f^p x)$ do not have a limit. The same can be formulated for polynomials and k-almost primes.

The counterexamples in [26] are only continuous, and in the setting of [26], it seems impossible to make them smooth.

Problem 5.2. Find a uniquely ergodic system (X, T) such that it is mixing (with respect to the only invariant measure), and for every $x \in X$ and every $\phi \in C(X)$, the averages $1/N \sum_{n \le N} \phi(T^{n^2} x)$ converge. The same can be asked for more general polynomials of degree ≥ 2.

The only known examples of systems which equidistribute for $\{n^2\}$ are systems which have a non-trivial Kronecker factor (and so are not even weakly mixing) or systems which are rigid (and hence are not mixing).

Question 5.3. Let $\{X_n\}$ be a sequence of random variables that satisfies the local limit theorem. For a system (X, T), $x \in X$, study the averages $1/N \sum_{n \le N} \phi(T^{X_n(\omega)} x)$. Is the analogue of the result in [2] for $\Omega(\cdot)$ true for $X_n(\omega)$ (almost surely)?

Question 5.4. In the same setting as above, what are the (almost sure) accumulation points of $\{h_{X_n(\omega)}(x)\}_{n \in \mathbb{N}}$, where (h_t) is the horocycle flow on $\mathrm{SL}(2, \mathbb{R})/\mathrm{SL}(2, \mathbb{Z})$ (or more generally non-compact lattices)? Is the analogue of the result in [25] true in this more general setting?

It seems that for both results [2, 25], the main number-theoretic input on $\Omega(\cdot)$ is that it satisfies the local limit theorem, and so it should be possible to study the above questions in the more general setting.

Question 5.5. Let $\mathbf{u}_t$ be a horospherical flow on G/Γ (non-compact). In which generality is the analogue of [27] true in this setting? That is, if $\mathbf{u}_t$ is d-dimensional, is it true that the orbit of every $x \in X$ generic for the Haar measure is dense along $(n_1^{2-\delta}, \dots, n_d^{2-\delta})$? What is the optimal $\delta < 1$?

Question 5.6. In the same setting as above, when does $\mathbf{u}_t$ satisfy a 2-almost prime number theorem, i.e., distribution along $(p_1^1 \cdot p_2^1, \dots, p_1^d \cdot p_2^d)$, where p_i^j are primes.

Question 5.7. Does the analogue of Theorem 4.4 hold for non-ergodic measures?

Acknowledgments. I would like to thank the anonymous referee for useful comments on the text.

Funding. A. M. K. was partially supported by NSF grant DMS-2247572 and by a grant from the Priority Research Area SciMat under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

[1] M. Auer, Weak mixing and sparse equidistribution. [v1] 2024, [v3] 2025, arXiv: 2408.15238v3

[2] V. Bergelson and F. K. Richter, Dynamical generalizations of the prime number theorem and disjointness of additive and multiplicative semigroup actions. *Duke Math. J.* **171** (2022), no. 15, 3133–3200 Zbl 1514.37018 MR 4497225

[3] G. D. Birkhoff, Proof of the ergodic theorem. *Proc. Natl. Acad. Sci. USA* **17** (1931), no. 12, 655–660 Zbl 0003.25602

[4] J. Bourgain, An approach to pointwise ergodic theorems. In J. Lindenstrauss and V. D. Milman (eds.), *Geometric Aspects of Functional Analysis. Israel Seminar (GAFA) 1986–1987*, pp. 204–223, Lecture Notes in Math. 1317, Springer, Berlin, 1988 Zbl 0662.47006 MR 0950982

[5] J. Bourgain, Möbius–Walsh correlation bounds and an estimate of Mauduit and Rivat. *J. Anal. Math.* **119** (2013), no. 1, 147–163 Zbl 1357.11092 MR 3043150

[6] J. Bourgain, On the correlation of the Moebius function with rank-one systems. *J. Anal. Math.* **120** (2013), no. 1, 105–130 Zbl 1358.37017 MR 3095150

[7] J. Bourgain, P. Sarnak, and T. Ziegler, Disjointness of Moebius from horocycle flows. In H. Farkas, R. Gunning, M. Knopp, and B. Taylor (eds.), *From Fourier analysis and number theory to Radon transforms and geometry. In Memory of Leon Ehrenpreis*, pp. 67–83, Dev. Math. 28, Springer, New York, 2013 Zbl 1336.37030 MR 2986954

[8] H. Daboussi, Remarques sur les fonctions multiplicatives. In *Séminaire Delange-Pisot-Poitou (Théorie des nombres, Fasc. 1), 18e année: 1976/77*, Exp. No. 4, Secrétariat Mathématique, Paris, 1977, 3 pp. Zbl 0378.10028 MR 0551333

[9] S. G. Dani and J. Smillie, Uniform distribution of horocycle orbits for Fuchsian groups. *Duke Math. J.* **51** (1984), no. 1, 185–194 Zbl 0547.20042 MR 0744294

[10] W. Duke, J. B. Friedlander, and H. Iwaniec, Equidistribution of roots of a quadratic congruence to prime moduli. *Ann. of Math. (2)* **141** (1995), no. 2, 423–441 Zbl 0840.11003 MR 1324141

[11] M. Einsiedler, G. Margulis, A. Mohammadi, and A. Venkatesh, Effective equidistribution and property (τ). *J. Amer. Math. Soc.* **33** (2020), no. 1, 223–289 Zbl 1466.11049 MR 4066475

[12] P. Erdös, On the integers having exactly k prime factors. *Ann. of Math. (2)* **49** (1948), no. 1, 53–66 Zbl 0030.29604 MR 0023279

[13] P. Erdös and M. Kac, The Gaussian law of errors in the theory of additive number theoretic functions. *Amer. J. Math.* **62** (1940), no. 1, 738–742 Zbl 0024.10203 MR 0002374

[14] S. Ferenczi, J. Kułaga-Przymus, and M. Lemańczyk, Sarnak's conjecture: what's new. In S. Ferenczi, J. Kułaga-Przymus, and M. Lemańczyk (eds.), *Ergodic theory and dynamical systems in their interactions with arithmetics and combinatorics. CIRM Jean–Morlet Chair, Fall 2016*, pp. 163–235, Lecture Notes in Math. 2213, Springer, Cham, 2018 Zbl 1421.37003 MR 3821717

[15] L. Flaminio and G. Forni, Invariant distributions and time averages for horocycle flows. *Duke Math. J.* **119** (2003), no. 3, 465–526 Zbl 1044.37017 MR 2003124

[16] L. Flaminio, G. Forni, and J. Tanis, Effective equidistribution of twisted horocycle flows and horocycle maps. *Geom. Funct. Anal.* **26** (2016), no. 5, 1359–1448 Zbl 1369.37004 MR 3568034

[17] G. Forni, A. Kanigowski, and M. Radziwiłł, Horocycle flow at product of two primes. 2024, arXiv:2409.16687v1

[18] N. Frantzikinakis and B. Host, The logarithmic Sarnak conjecture for ergodic weights. *Ann. of Math. (2)* **187** (2018), no. 3, 869–931 Zbl 1400.11129 MR 3779960

[19] N. Frantzikinakis and B. Host, Furstenberg systems of bounded multiplicative functions and applications. *Int. Math. Res. Not. IMRN* **2021** (2021), no. 8, 6077–6107 Zbl 1483.37006 MR 4251272

[20] H. Furstenberg, *Recurrence in ergodic theory and combinatorial number theory*. M. B. Porter Lect., Princeton University Press, Princeton, NJ, 1981 Zbl 0459.28023 MR 0603625

[21] B. Green, On (not) computing the Möbius function using bounded depth circuits. *Combin. Probab. Comput.* **21** (2012), no. 6, 942–951 Zbl 1279.11095 MR 2981162

[22] B. Green and T. Tao, The quantitative behaviour of polynomial orbits on nilmanifolds. *Ann. of Math. (2)* **175** (2012), no. 2, 465–540 Zbl 1251.37012 MR 2877065

[23] R. Howe, On a notion of rank for unitary representations of the classical groups. In A. F. Talamanca (ed.), *Harmonic analysis and group representation*, pp. 223–331, Liguori, Naples, 1982 MR 0777342

[24] H. Iwaniec and E. Kowalski, *Analytic number theory*. Amer. Math. Soc. Colloq. Publ. 53, American Mathematical Society, Providence, RI, 2004, 615 pp. Zbl 1059.11001 MR 2061214

[25] A. Kanigowski and K. Loyd, Averages along Ω. Preprint

[26] A. Kanigowski, M. Lemańczyk, and M. Radziwiłł, Prime number theorem for analytic skew products. *Ann. of Math. (2)* **199** (2024), no. 2, 591–705 Zbl 1546.11139 MR 4713020

[27] A. Kanigowski and M. Radziwiłł, Density of orbits of horocycle flows at sub-quadratic polynomial times. 2025, arXiv:2510.21964v1

[28] K. Kasprzak, Equidistribution of orbits at polynomial times in rigid dynamical systems. 2025, arXiv:2501.05538v1

[29] K. Kasprzak, Distribution of polynomial orbits in Toeplitz systems. *Ergodic Theory Dynam. Systems* **46** (2026), no. 7, 1800–1818 Zbl 08210511 MR 5036714

[30] I. Kátai, A remark on a theorem of H. Daboussi. *Acta Math. Hungar.* **47** (1986), no. 1–2, 223–225 Zbl 0607.10034 MR 0836415

[31] A. Leibman, Pointwise convergence of ergodic averages for polynomial actions of $\mathbb{Z}^d$ by translations on a nilmanifold. *Ergodic Theory Dynam. Systems* **25** (2005), no. 1, 215–225 Zbl 1080.37004 MR 2122920

[32] E. Lindenstrauss, A. Mohammadi, and Z. Wang, Effective equidistribution for some one parameter unipotent flows. [v1] 2022, [v2] 2025, arXiv:2211.11099v2, to appear in *Ann. of Math. (2)*

[33] K. Loyd, A dynamical approach to the asymptotic behavior of the sequence $\Omega(n)$. *Ergodic Theory Dynam. Systems* **43** (2023), no. 11, 3685–3706 Zbl 1533.37019 MR 4651580

[34] 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

[35] K. Matomäki, M. Radziwiłł, and T. Tao, An averaged form of Chowla's conjecture. *Algebra Number Theory* **9** (2015), no. 9, 2167–2196 Zbl 1377.11109 MR 3435814

[36] K. Matomäki, M. Radziwiłł, and T. Tao, Fourier uniformity of bounded multiplicative functions in short intervals on average. *Invent. Math.* **220** (2020), no. 1, 1–58 Zbl 1459.11186 MR 4071407

[37] C. Mauduit and J. Rivat, Prime numbers along Rudin–Shapiro sequences. *J. Eur. Math. Soc. (JEMS)* **17** (2015), no. 10, 2595–2642 Zbl 1398.11121 MR 3420517

[38] T. McAdam, Almost-prime times in horospherical flows on the space of lattices. *J. Mod. Dyn.* **15** (2019), 277–327 Zbl 1470.37052 MR 4042163

[39] C. C. Moore, Exponential decay of correlation coefficients for geodesic flows. In C. C. Moore (ed.), *Group representations, ergodic theory, operator algebras, and mathematical physics (Proc. Conf. Hon. G. W. Mackey, Berkeley, Calif., 1984)*, pp. 163–181, Math. Sci. Res. Inst. Publ. 6, Springer, New York, 1987 Zbl 0625.58023 MR 0880376

[40] C. Müllner, Automatic sequences fulfill the Sarnak conjecture. *Duke Math. J.* **166** (2017), no. 17, 3219–3290 Zbl 1439.11089 MR 3724218

[41] M. Ratner, Horocycle flows, joinings and rigidity of products. *Ann. of Math. (2)* **118** (1983), no. 2, 277–313 Zbl 0556.28020 MR 0717825

[42] P. Sarnak, *Three lectures on the Möbius function, randomness and dynamics*. 2011, http://publications.ias.edu/sarnak/ visited on 12 May 2026

[43] P. Sarnak and A. Ubis, The horocycle flow at prime times. *J. Math. Pures Appl. (9)* **103** (2015), no. 2, 575–618 Zbl 1307.11048 MR 3298371

[44] L. Streck, Non-concentration of primes in $\Gamma\backslash \mathrm{PSL}_2(\mathbb{R})$. *Israel J. Math.* **269** (2025), no. 1, 1–29 Zbl 08124162 MR 4982566

[45] L. Streck, On equidistribution of polynomial sequences in quotients of $PSL_2(\mathbb{R})$. *Ergodic Theory Dynam. Systems* **46** (2026), no. 3, 845–856 Zbl 08159100 MR 5027337

[46] A. Strömbergsson, On the deviation of ergodic averages for horocycle flows. *J. Mod. Dyn.* **7** (2013), no. 2, 291–328 Zbl 1360.37088 MR 3106715

[47] T. Tao, The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi* **4** (2016), article no. e8, 36 pp. Zbl 1383.11116 MR 3569059

[48] A. Venkatesh, Sparse equidistribution problems, period bounds and subconvexity. *Ann. of Math. (2)* **172** (2010), no. 2, 989–1094 Zbl 1214.11051 MR 2680486

[49] I. M. Vinogradov, *The method of trigonometrical sums in the theory of numbers* (in Russian). Tr. Mat. Inst. Steklova 23, 1947, 109 pp. Zbl 0041.37002 MR 0029417

[50] J. von Neumann, Zur Operatorenmethode in der klassischen Mechanik. *Ann. of Math. (2)* **33** (1932), no. 3, 587–642 Zbl 0005.12203 MR 1503078

[51] H. Weyl, Über die Gleichverteilung von Zahlen mod. Eins. *Math. Ann.* **77** (1916), no. 3, 313–352 JFM 46.0278.06 MR 1511862

[52] M. Wierdl, Pointwise ergodic theorem along the prime numbers. *Israel J. Math.* **64** (1988), no. 3, 315–336 Zbl 0695.28007 MR 0995574

[53] L. Yang, Effective version of Ratner's equidistribution theorem for $\mathrm{SL}(3,\mathbb{R})$. *Ann. of Math. (2)* **202** (2025), no. 1, 189–264 Zbl 1575.37004 MR 4927736

Adam M. Kanigowski
Department of Mathematics, University of Maryland, College Park, MD 20742, USA; Faculty of Mathematics and Computer Science, Jagiellonian University, Lojasiewicza 6, 30-348 Krakow, Poland; akanigow@umd.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/39

Recent progress in the study of Latin squares

Richard Montgomery

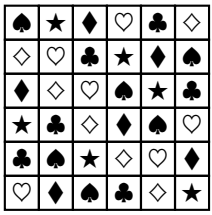
Abstract. A Latin square of order n is an n by n grid filled with n symbols so that each symbol appears exactly once in each row and each column. From their early origins in recreational mathematics, Latin squares have been formally studied for their combinatorial properties since the late 18th century. More recently, tools from extremal and probabilistic combinatorics have been applied to Latin squares to make progress on a range of old questions. This brief survey will concentrate on the topics covered by the author's talk at the 9th European Congress of Mathematics, while highlighting recent relevant surveys in the area.

1. Introduction and brief history

A Latin square of order n is an n by n grid filled with n symbols so that each symbol appears exactly once in each row and each column (see Figure 1). A key example is the Cayley table of any group G of order n, which forms a Latin square of order n: indexing both the rows and columns by G, the cell of the n by n grid indexed by (g, h) is filled with $g \circ h$. Latin squares have strong links to design theory and combinatorial designs, as well as connections to error-correcting codes, 2-dimensional permutations and finite projective planes. They have a long history in early mathematics, where often their appearance was connected to magic squares; for an overview of this history, we recommend the historical survey of Andersen [1]. More recently, Latin squares have quietly gained wide familiarity due to Sudoku puzzles, the solutions of which form a Latin square of order 9 (with an additional condition on 9 disjoint 3 by 3 subsquares).

The formal study of Latin squares for their mathematical properties began with a long paper by Euler [12] in 1782. This study was motivated by a recreational question at the time, which is now known as Euler's 36 officers problem. Given 36 officers, representing 6 different ranks held in 6 different regiments, can the officers stand in a 6 by 6 grid so that there is an officer of each rank and each regiment in each row

Mathematics Subject Classification 2020: 05B15 (primary); 05-02 (secondary).
Keywords: transversals, Latin squares, decompositions, graph theory.



0	1	2	3	4	5	6
1	2	3	4	5	6	0
2	3	4	5	6	0	1
3	4	5	6	0	1	2
4	5	6	0	1	2	3
5	6	0	1	2	3	4
6	0	1	2	3	4	5

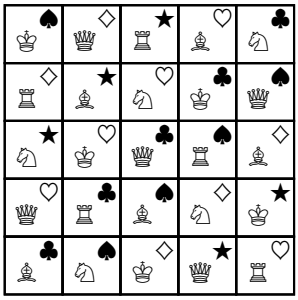
8	2	4	6	5	9	7	3	1
6	5	1	7	2	3	9	4	8
9	7	3	4	1	8	2	6	5
2	1	5	8	9	4	3	7	6
7	9	8	5	3	6	4	1	2
4	3	6	2	7	1	5	8	9
1	6	9	3	4	5	8	2	7
5	4	2	1	8	7	6	9	3
3	8	7	9	6	2	1	5	4

Figure 1. Latin squares of order 6, 7 and 9, respectively, where the middle square is the Cayley table for $\mathbb{Z}_7$, and the square on the right is an underlying grid for a Sudoku puzzle.

and column? For example, a solution to the corresponding problem with 5 ranks and 5 regiments is presented in Figure 2. In this solution, the ranks of the officers alone form a Latin square, while the regiments alone form another. This gives two Latin squares of order $n = 5$, where the n^2 possibilities for pairs of symbols from the two sets each occur exactly once in a pair of corresponding cells. Such pairs of Latin squares are known as *mutually orthogonal*.

Euler found examples of pairs of mutually orthogonal Latin squares of order n for each $n \neq 2 \bmod 4$, using Latin and Greek alphabets, respectively, as symbols in the pair of Latin squares, which gave rise to their name. He could, however, find no solution to the 36 officers problem and conjectured that, moreover, there are no pairs of mutually orthogonal Latin squares of order n for each $n = 2 \bmod 4$. This can be easily seen to be true for $n = 2$. Euler could not confirm his conjecture for $n = 6$, but eventually, in 1900, Tarry [47] proved that, indeed, no such pairs had eluded Euler. Perhaps surprisingly, however, Euler's conjecture is false! It was disproved first for $n = 22$ and $n = 50$ by Bose and Shrikhande [3] in 1959, before they extended this with Parker [4] to find counterexamples for every $n = 2 \bmod 4$ with $n \geq 10$. In particular, Parker's counterexample for $n = 10$ was one of the earliest counterexamples found with computer assistance.

Since these results, Latin squares have been intensively studied as a combinatorial object on their own terms and as a key part of design theory. In the last



$d\ \eta$	$e\ \alpha$	$f\ \beta$	$g\ \gamma$	$a\ \delta$	$b\ \varepsilon$	$c\ \zeta$
$f\ \gamma$	$g\ \delta$	$a\ \varepsilon$	$b\ \zeta$	$c\ \eta$	$d\ \alpha$	$e\ \beta$
$a\ \zeta$	$b\ \eta$	$c\ \alpha$	$d\ \beta$	$e\ \gamma$	$f\ \delta$	$g\ \varepsilon$
$c\ \beta$	$d\ \gamma$	$e\ \delta$	$f\ \varepsilon$	$g\ \zeta$	$a\ \eta$	$b\ \alpha$
$e\ \varepsilon$	$f\ \zeta$	$g\ \eta$	$a\ \alpha$	$b\ \beta$	$c\ \gamma$	$d\ \delta$
$g\ \alpha$	$a\ \beta$	$b\ \gamma$	$c\ \delta$	$d\ \varepsilon$	$e\ \zeta$	$f\ \eta$
$b\ \delta$	$c\ \varepsilon$	$d\ \zeta$	$e\ \eta$	$f\ \alpha$	$g\ \beta$	$a\ \gamma$

Figure 2. Two pairs of mutually orthogonal Latin squares. On the left is a solution for the $n = 5$ version of the (unsolvable) 36 officers problem, with chess pieces denoting 5 different ranks and symbols denoting 5 different regiments. On the right, following Euler's initial convention, a Latin square using Latin letters is entered in the grid along with a Latin square using Greek letters, where each possible combination of pairs like (d, η) appears in the grid exactly once.

decade, new progress has been made by studying them in an equivalent form using edge-coloured graphs, allowing the application of techniques and perspectives from extremal and probabilistic combinatorics. This equivalence is seen as follows. Given a Latin square L of order n, form the edge-coloured graph $H(L)$ by taking a vertex representing each row and each column and, for each entry in L, putting an edge between the row and column of the entry with the symbol of the entry as its 'colour' (see Figure 3). In $H(L)$, the conditions of the Latin square imply that an edge of every colour appears exactly once at every vertex. That no two edges of one colour share a vertex means that the graph is *properly coloured*, and so we have a proper colouring of $H(L)$ with n colours. The graph formed is a *complete bipartite graph*, i.e., its vertices split into two classes so that its edges are exactly every possible edge between these two classes. The Latin squares of order n can then be easily seen to be equivalent to the complete bipartite graphs with n vertices in each class which are properly coloured using n colours.

This survey will cover some of the recent progress made in the study of Latin squares using perspectives from graph theory under this equivalence, with a particular focus on the topics of the author's talk at the 9th European Congress of Mathematics (Seville, 2024) in conjunction with the award of an EMS prize. A recent survey by the author for a more specialist readership can be found at [34], as well as an article for a wide audience at [35]. The study of edge-coloured graphs more broadly has recently seen a great deal of activity and many additional applications. For an overview, we

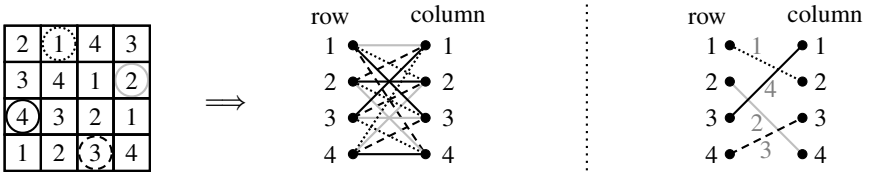


Figure 3. On the left, a Latin square of order 4 is converted into a complete bipartite properly coloured graph with 4 vertices in each class and edges with colours 1, 2, 3 and 4 drawn with dotted, grey, dashed and black lines, respectively. On the right, the perfect rainbow matching corresponding to the transversal highlighted in the Latin square on the left is shown.

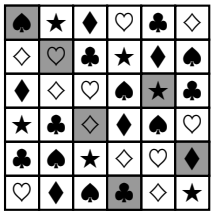
recommend the survey by Sudakov [46] (Chapter 3 of the current volume), as well as the recent survey by Pokrovskiy [40], aimed at a more specialist audience.

2. Transversals in Latin squares

A *transversal* (sometimes known as a *full transversal*) is a collection of n cells in a Latin square so that each row, column, and symbol is used exactly once (see Figure 4). As was known by Euler, not every Latin square of order n has a transversal when $n = 0 \bmod 2$. The canonical example of this is the Cayley table for $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. To see this, let G be $\mathbb{Z}_n$, or any abelian group of order n, and let T be a set of n pairs (g, h) such that $(g, h) \mapsto g$, $(g, h) \mapsto h$ and $(g, h) \mapsto g + h$ each forms a bijection from T to G. Thus, T maps out a transversal in the Latin square $L(G)$ corresponding to the Cayley table for G. Then,

$$\sum_{v\in G} v = \sum_{(g,h)\in T} (g + h) = \sum_{(g,h)\in T} g + \sum_{(g,h)\in T} h = 2\sum_{v\in G} v,$$

and therefore, $\sum_{v\in G} v = 0$. Thus, $L(G)$ can only have a transversal if $\sum_{v\in G} v = 0$, and in particular, if $G = \mathbb{Z}_{2m}$ for any $m \in \mathbb{N}$, then $L(G)$ has no transversal.



0	1	2	3	4	5
1	2	3	4	5	0
2	3	4	5	0	1
3	4	5	0	1	2
4	5	0	1	2	3
5	0	1	2	3	4

Figure 4. Latin squares of order 6, with a transversal highlighted on the left and a partial transversal with 5 cells highlighted on the right.

In 1947, Paige [39] proved that this condition is sufficient for $L(G)$ to have a transversal for any finite abelian group G. Considering the problem for non-abelian groups G, Paige [38] proved that if $L(G)$ has a transversal, then its elements can be ordered so that their product is the identity element and conjectured that this is a sufficient condition for the existence of a transversal. In 1955, Hall and Paige [18] showed moreover that this condition is equivalent to $\sum_{x\in G} x = 0$ in the abelianisation $G^{\mathrm{ab}} = G/G'$ of G, where G' is the commutator subgroup of G, which became known as the *Hall–Paige condition*. Hall and Paige showed that this is equivalent to the condition that all the 2-Sylow subgroups of G are trivial or non-cyclic and conjectured that it implies the existence of a transversal, as follows.

Conjecture 2.1 (The Hall–Paige conjecture). *For any finite group G, $L(G)$ has a transversal if and only if G satisfies the Hall–Paige condition.*

Conjecture 2.1 was proved in 2009 using the classification of finite simple groups, through a combination of work by Wilcox [50], Evans [13] and Bray (see [50], and [6] for the later publication). In 2022, Eberhard, Manners and Mrazović [9] gave an alternative proof of the Hall–Paige conjecture for sufficiently large groups, using tools from analytic number theory. Their proof not only yields an impressive asymptotic formula for the number of transversals in the Cayley table of a group but can also reduce some of the complexity of the original proof of the Hall–Paige conjecture by taking care of the largest examples.

The Hall–Paige conjecture was also proved for large groups using combinatorial methods by Müyesser and Pokrovskiy [37], using the equivalent form in edge-coloured graphs outlined in Section 1. Let L be a Latin square of order n, and let $H(L)$ be the corresponding edge-coloured complete bipartite n by n graph. Then the cells in a transversal T of L correspond directly to a set E of edges in $H(L)$ such that the edges in E share no vertex and each colour appears exactly once (see Figure 3). Then E is a *perfect rainbow matching* in $H(L)$.

For Latin squares more generally, the natural extremal question is: given a Latin square of order n, how close can we get to a transversal? That is, how large a *partial transversal* can we find, where a partial transversal in a Latin square is a collection of cells that share no row, column or symbol (see Figure 4). In 1967, Ryser [44] conjectured that every Latin square of odd order has a transversal, and subsequently Brualdi suggested that every Latin square of even order n should have a partial transversal with $n-1$ cells. In 1975, Stein made some related conjectures, and in recent years, the following has become known as the Ryser–Brualdi–Stein conjecture.

Conjecture 2.2 (The Ryser–Brualdi–Stein conjecture). *Every Latin square of order n has a partial transversal with $n-1$ cells and a full transversal if n is odd.*

Progress towards the Ryser–Brualdi–Stein conjecture has concentrated on showing that increasingly large partial transversals exist in any Latin square of order n. After initial bounds by Koksma [24] and Drake [8], in 1978 Brouwer, de Vries and Wieringa [7] and Woolbright [51] showed that Conjecture 2.2 holds approximately, both showing that every Latin square of order n has a partial transversal with at least $n - \sqrt{n}$ cells. In 1982, Shor [45] gave a proof that every Latin square of order n has a partial transversal with $n - O(\log^2 n)$ cells, though this had an error later corrected by Hatami and Shor [19]. After this bound had essentially stood for almost 40 years, Keevash, Pokrovskiy, Sudakov and Yepremyan [23] made a breakthrough, working from the perspective of edge-coloured graphs. They showed that every Latin square of order n has a partial transversal with $n - O(\log n / \log\log n)$ cells. The current state of the art is a result of the current author [33] that every Latin square of order n contains a partial transversal with $n - 1$ cells when n is sufficiently large, as follows.

Theorem 2.3. *There is some n_0 such that every Latin square of order $n \geq n_0$ has a partial transversal with $n - 1$ cells.*

That Theorem 2.3 is tight when n is even is shown not only by the Cayley tables mentioned above but also by further extremal examples. In 1894, Maillet [30] showed that a wide range of extremal examples can be constructed from the Cayley tables for $\mathbb{Z}_{2m}$, where $m \in \mathbb{N}$. Rather than the direct algebraic structure of $\mathbb{Z}_{2m}$, these examples come with a partition of the symbols into $2m$ sets, where the sets have some algebraic structure between them. All the known extremal examples have some algebraic structure of this sort, but it need not hold perfectly. Indeed, in 2017, Cavenagh and Wanless made the lovely observation that you can make slight adjustments to the Cayley tables for $\mathbb{Z}_{2m}$ under some simple rule without creating a transversal. This imperfection in the algebraic structure of some extremal examples, along with the range of extremal examples known, suggests that determining the truth or otherwise of Conjecture 2.2 will be rather difficult.

The key to the progress made on Theorem 2.3 in [33] is taking a general Latin square and partitioning the symbols carefully in order to find some approximate algebraic structure, if possible, before finding a large partial transversal using a range of old and new combinatorial tools. Most of the large partial transversal is found using the semi-random method, as introduced by Rödl [41] in the 1980s. To modify this to find a partial transversal with $n - 1$ cells, the proof in [33] uses absorption, a general method for finding large structures in combinatorial objects that was codified by Rödl, Ruciński and Szemerédi [43] in 2006. In particular, it uses a form of absorption known as distributive absorption, as introduced by the current author [32], which has allowed absorption to be used in many different settings to find large rigid substructures. The proof of Theorem 2.3 uses this in conjunction with a new ‘addition structure’, which

determines which row, column and symbol is omitted from the partial transversal to achieve the right conditions for the new implementation of absorption. A further example of the graph theory tools used in [33] is that involving sublinear expansion. Sublinear expansion was introduced by Komlós and Szemerédi [25, 26] in the 1990s and has since become a critical tool that has been used in a range of results in extremal graph theory, as seen in the recent survey by Letzter [29]. In [33], sublinear expansion is used to partition the symbols into sets with some approximate algebraic structure.

3. Transversals in random Latin squares

The natural counterpart to Conjecture 2.2 from the perspective of probabilistic combinatorics is to determine whether we can expect a random Latin square to have a transversal. To make this concrete, for each $n \in \mathbb{N}$, let $\mathcal{L}(n)$ be the set of Latin squares of order n which use the symbols $1, \dots, n$, and pick $L_n \in \mathcal{L}(n)$ uniformly at random. Does the probability that L_n has a transversal tend to 1 as n tends to infinity? In other words, does a random Latin square have a transversal *with high probability* (whp)?

This question is perhaps at first glance quite innocuous, as the known Latin squares without a transversal (as discussed in Section 2) have, or are close to having, a precise algebraic structure, while a random Latin square is likely to have no such structure. However, it turns out that a random Latin square is a difficult object to study. Effectively, this is because it is hard to make small alterations to the symbols in a Latin square and reach another Latin square. It took therefore until 2020 before this question was answered by Kwan [27], who confirmed that a random Latin square L_n has a transversal whp. To do so, Kwan approximated L_n using a modification of the random triangle removal process, using breakthrough methods of Keevash [21, 22] on the existence of designs and the analysis of the random triangle removal process due to Bohman, Frieze and Lubetzky [2].

Random Latin squares have also been approached using switching methods, as in the pioneering work of Godsil and McKay [15] and McKay and Wanless [31] in the 1990s. Using these methods, and building on their work with Kühn and Osthus [17], Gould and Kelly [16] gave in 2023 an alternative proof that a random Latin square has a transversal whp. Another, entirely different, proof of this was given by Eberhard, Manners and Mrazović [10], who moreover gave an asymptotically tight bound on the number of transversals which occur in a random Latin square whp. This used their methods from analytic number theory and gave asymptotically tight bounds on the number of transversals in Latin squares satisfying certain quasirandom conditions. These conditions are satisfied by random Latin square whp, as shown by Eberhard,

Manners and Mrazović using work of Kwan, Sah and Sawhney [28], who in particular applied the deletion method of Rödl and Ruciński [20, 42] very effectively in the setting of random Latin squares.

With these results, we now have three distinct methods to show that a random Latin square has a transversal whp, but we may be able to find much more. More specifically, are we likely to be able to find many disjoint transversals (i.e., which share no common cells) in a random Latin square whp? Most ambitiously, does the random Latin square L_n have a collection of n disjoint transversals whp? Such a collection of disjoint transversals would then use each cell of the Latin square exactly once, giving a decomposition of the Latin square into transversals. In 1990, van Rees [48] conjectured that a random Latin square does not have such a decomposition whp, but more recently, Wanless and Webb [49] observed that numerical estimates suggest it does whp.

Kwan's methods to find one transversal in a random Latin square whp can be used to find many disjoint transversals whp. Further progress in this direction, however, was not made directly on Latin squares but in the closely related setting of Steiner triple systems. Relevant definitions for this can be found, for example, in [27], but pertinent here is that, while not proved formally for Latin squares, we will state the analogous results from [28, 36] in the setting of Latin squares (particularly because these translated results now follow from Theorem 3.1 later), as, while not formally checked, the methods can be expected to translate across. In 2017, Morris [36] used Kwan's methods to show, correspondingly, that there is some $\varepsilon > 0$ such that a random Latin square contains at least εn disjoint transversals whp. Developing these methods much further and using additional tools, in 2020, Ferber and Kwan [14] then showed that, correspondingly, a random Latin square contains $(1 - o(1))n$ disjoint transversals whp.

Combining and developing many of the techniques mentioned so far, including recent implementations of the semi-random method by Ehard, Glock and Joos [11], and using an intricate absorption argument, Bowtell and the current author [5] were very recently able to confirm that a random Latin square of order n will indeed have n disjoint transversals whp, as follows.

Theorem 3.1. *The probability that the random Latin square L_n has a decomposition into disjoint transversals tends to 1 as n tends to infinity.*

Decompositions of Latin squares into disjoint transversals are very strongly linked to where this survey began. Indeed, a pair of mutually orthogonal Latin squares of order n, as studied by Euler, is exactly equivalent to a Latin square of order n which can be decomposed into disjoint transversals. In a pair of mutually orthogonal Latin squares of order n, the entries of each different symbol in the first Latin square

mark out a transversal in the second Latin square (see Figure 2), and thus, the n different symbols in the first Latin square together mark out a decomposition of the second Latin square into transversals (and vice versa). In the context of Euler's disproved conjecture that there are no Latin squares of order $n = 2 \bmod 4$ which have a decomposition into disjoint transversals, Theorem 3.1 shows that, as n gets large, counterexamples to Euler's conjecture are so common that a Latin square of order $n = 2 \bmod 4$ is much more likely to be a counterexample than not! That we now know this is thanks to a raft of techniques from extremal and probabilistic graph theory, techniques that can be expected to form the foundation of further results in the study of Latin squares.

Funding. This work was supported by the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement no. 947978).

References

[1] L. D. Andersen, The history of Latin squares. In R. Wilson and J. J. Watkins (eds.), *Combinatorics: ancient and modern*. Oxford University Press, Oxford, 2013, 381 pp. MR 3204727

[2] T. Bohman, A. Frieze, and E. Lubetzky, Random triangle removal. *Adv. Math.* **280** (2015), 379–438 Zbl 1312.05123 MR 3350225

[3] R. C. Bose and S. S. Shrikhande, On the falsity of Euler's conjecture about the non-existence of two orthogonal Latin squares of order $4t + 2$. *Proc. Nat. Acad. Sci. USA* **45** (1959), no. 5, 734–737 Zbl 0085.00902 MR 0104590

[4] R. C. Bose, S. S. Shrikhande, and E. T. Parker, Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture. *Canad. J. Math.* **12** (1960), 189–203 Zbl 0093.31905 MR 0122729

[5] C. Bowtell and R. Montgomery, Almost every Latin square has a decomposition into transversals. 2025, arXiv:2501.05438v1

[6] J. N. Bray, Q. Cai, P. J. Cameron, P. Spiga, and H. Zhang, The Hall–Paige conjecture, and synchronization for affine and diagonal groups. *J. Algebra* **545** (2020), 27–42 Zbl 1485.20002 MR 4044687

[7] A. E. Brouwer, A. J. de Vries, and R. M. A. Wieringa, A lower bound for the length of partial transversals in a Latin square. *Nieuw Arch. Wisk. (3)* **26** (1978), no. 2, 330–332 Zbl 0395.05012 MR 0480083

[8] D. A. Drake, Maximal sets of Latin squares and partial transversals. *J. Statist. Plann. Inference* **1** (1977), no. 2, 143–149 Zbl 0392.05014 MR 0485434

[9] S. Eberhard, F. Manners, and R. Mrazović, An asymptotic for the Hall–Paige conjecture. *Adv. Math.* **404, part A** (2022), article no. 108423, 73 pp. Zbl 1494.05018 MR 4416136

[10] S. Eberhard, F. Manners, and R. Mrazović, Transversals in quasirandom Latin squares. *Proc. Lond. Math. Soc. (3)* **127** (2023), no. 1, 84–115 Zbl 1525.05013 MR 4611404

[11] S. Ehard, S. Glock, and F. Joos, Pseudorandom hypergraph matchings. *Combin. Probab. Comput.* **29** (2020), no. 6, 868–885 Zbl 1466.05147 MR 4173135

[12] L. Euler, Recherches sur un nouvelle espéce de quarrés magiques. In *Verhandelingen uitgegeven door het zeeuwsch Genootschap der Wetenschappen te Vlissingen, Volume 9*, pp. 85–239, 1782

[13] A. B. Evans, The admissibility of sporadic simple groups. *J. Algebra* **321** (2009), no. 1, 105–116 Zbl 1166.20012 MR 2469351

[14] A. Ferber and M. Kwan, Almost all Steiner triple systems are almost resolvable. *Forum Math. Sigma* **8** (2020), article no. e39, 24 pp. Zbl 1453.05018 MR 4170629

[15] C. D. Godsil and B. D. McKay, Asymptotic enumeration of Latin rectangles. *J. Combin. Theory Ser. B* **48** (1990), no. 1, 19–44 Zbl 0687.05010 MR 1047551

[16] S. Gould and T. Kelly, Hamilton transversals in random Latin squares. *Random Structures Algorithms* **62** (2023), no. 2, 450–478 Zbl 1522.05019 MR 4545866

[17] S. Gould, T. Kelly, D. Kühn, and D. Osthus, Almost all optimally coloured complete graphs contain a rainbow Hamilton path. *J. Combin. Theory Ser. B* **156** (2022), 57–100 Zbl 1490.05147 MR 4419500

[18] M. Hall and L. J. Paige, Complete mappings of finite groups. *Pacific J. Math.* **5** (1955), no. 4, 541–549 Zbl 0066.27703 MR 0079589

[19] P. Hatami and P. W. Shor, A lower bound for the length of a partial transversal in a Latin square. *J. Combin. Theory Ser. A* **115** (2008), no. 7, 1103–1113 Zbl 1159.05303 MR 2450332

[20] S. Janson and A. Ruciński, The infamous upper tail. *Random Structures Algorithms* **20** (2002), no. 3, 317–342 Zbl 0996.60023 MR 1900611

[21] P. Keevash, Counting designs. *J. Eur. Math. Soc. (JEMS)* **20** (2018), no. 4, 903–927 Zbl 1386.05015 MR 3779688

[22] P. Keevash, The existence of designs. [v1] 2014, [v4] 2024, arXiv:1401.3665v4

[23] 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), no. 8, 288–321 Zbl 1490.05269 MR 4408405

[24] K. K. Koksma, A lower bound for the order of a partial transversal in a Latin square. *J. Combin. Theory* **7** (1969), no. 1, 94–95 Zbl 0172.01504 MR 0239988

[25] J. Komlós and E. Szemerédi, Topological cliques in graphs. *Combin. Probab. Comput.* **3** (1994), no. 2, 247–256 Zbl 0809.05080 MR 1288443

[26] J. Komlós and E. Szemerédi, Topological cliques in graphs. II. *Combin. Probab. Comput.* **5** (1996), no. 1, 79–90 Zbl 0846.05023 MR 1395694

[27] M. Kwan, Almost all Steiner triple systems have perfect matchings. *Proc. Lond. Math. Soc. (3)* **121** (2020), no. 6, 1468–1495 Zbl 1462.05045 MR 4144368

[28] M. Kwan, A. Sah, and M. Sawhney, Large deviations in random Latin squares. *Bull. Lond. Math. Soc.* **54** (2022), no. 4, 1420–1438 Zbl 1522.05020 MR 4488316

[29] S. Letzter, Sublinear expanders and their applications. In F. Fischer and R. Johnson (eds.), *Surveys in Combinatorics 2024*, pp. 89–130, London Math. Soc. Lecture Note Ser. 493, Cambridge University Press, Cambridge, 2024 Zbl 1573.05203 MR 4749039

[30] E. Maillet, Sur les carrés latins d'Euler. *Assoc. Franc. Caen* **23** (1894), 244–252 JFM 26.0239.01

[31] B. D. McKay and I. M. Wanless, Most Latin squares have many subsquares. *J. Combin. Theory Ser. A* **86** (1999), no. 2, 322–347 Zbl 0948.05014 MR 1685535

[32] R. Montgomery, Spanning trees in random graphs. *Adv. Math.* **356** (2019), article no. 106793, 92 pp. Zbl 1421.05080 MR 3998769

[33] R. Montgomery, A proof of the Ryser–Brualdi–Stein conjecture for large even n. 2023, arXiv:2310.19779v1

[34] R. Montgomery, Transversals in Latin squares. In F. Fischer and R. Johnson (eds.), *Surveys in Combinatorics 2024*, pp. 131–158, London Math. Soc. Lecture Note Ser. 493, Cambridge University Press, Cambridge, 2024 Zbl 1573.05060 MR 4749040

[35] R. Montgomery, Latin squares via graph theory. *Lond. Math. Soc. Newsl.* (2025), no. 515, 19–23 Zbl 1562.05023 MR 4907803

[36] P. Morris, *Random Steiner triple systems*. Master's thesis, Freie Universität, Berlin, 2017

[37] A. Müyesser and A. Pokrovskiy, A random Hall–Paige conjecture. *Invent. Math.* **240** (2025), no. 3, 779–867 Zbl 1565.20065 MR 4902157

[38] L. Paige, Complete mappings of finite groups. *Pacific J. Math.* **1** (1951), no. 1, 111–116 Zbl 0043.02404 MR 0043080

[39] L. J. Paige, A note on finite Abelian groups. *Bull. Amer. Math. Soc.* **53** (1947), no. 6, 590–593 Zbl 0033.15101 MR 0020990

[40] A. Pokrovskiy, Rainbow subgraphs and their applications. In A. Nixon and S. Prendiville (eds.), *Surveys in Combinatorics 2022*, pp. 191–214, London Math. Soc. Lecture Note Ser. 481, Cambridge University Press, Cambridge, 2022 Zbl 1529.05072 MR 4421403

[41] V. Rödl, On a packing and covering problem. *European J. Combin.* **6** (1985), no. 1, 69–78 Zbl 0565.05016 MR 0793489

[42] V. Rödl and A. Ruciński, Threshold functions for Ramsey properties. *J. Amer. Math. Soc.* **8** (1995), no. 4, 917–942 Zbl 0846.05079 MR 1276825

[43] V. Rödl, A. Ruciński, and E. Szemerédi, A Dirac-type theorem for 3-uniform hypergraphs. *Combin. Probab. Comput.* **15** (2006), no. 1–2, 229–251 Zbl 1082.05057 MR 2195584

[44] H. Ryser, Neuere Probleme der Kombinatorik. *Vorträge über Kombinatorik, Oberwolfach* (1967), 69–91

[45] P. W. Shor, A lower bound for the length of a partial transversal in a Latin square. *J. Combin. Theory Ser. A* **33** (1982), no. 1, 1–8 Zbl 0489.05012 MR 0665651

[46] B. Sudakov, Restricted subgraphs of edge-colored graphs and applications. In *European Congress of Mathematics, Sevilla, 15–19 July, 2024*, pp. 57–86, EMS Press, 2026

[47] G. Tarry, *Le problème des 36 officiers*. Secrétariat de l'Association française pour l'avancement des sciences, 1900 JFM 32.0219.04

[48] G. H. J. van Rees, Subsquares and transversals in Latin squares. *Ars Combin.* **29** (1990), 193–204 Zbl 0718.05014 MR 1412875

[49] I. M. Wanless and B. S. Webb, The existence of Latin squares without orthogonal mates. *Des. Codes Cryptogr.* **40** (2006), no. 1, 131–135 Zbl 1180.05023 MR 2226288

[50] S. Wilcox, Reduction of the Hall–Paige conjecture to sporadic simple groups. *J. Algebra* **321** (2009), no. 5, 1407–1428 Zbl 1172.20024 MR 2494397

[51] D. E. Woolbright, An $n \times n$ Latin square has a transversal with at least $n - \sqrt{n}$ distinct symbols. *J. Combin. Theory Ser. A* **24** (1978), no. 2, 235–237 Zbl 0368.05012 MR 0472562

Richard Montgomery
Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK;
richard.montgomery@warwick.ac.uk

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/40

Fourier uniqueness pairs and interpolation formulas

Danylo Radchenko

Abstract. This is a short, informal survey of recent analytic and number-theoretic constructions of Fourier uniqueness pairs and interpolation formulas, as well as their applications in sphere packing and energy minimization problems.

1. Sphere packing and energy minimization problems

One of the main motivations for the recent works on Fourier uniqueness pairs and interpolation formulas comes from two problems in geometric optimization: sphere packing and energy minimization. We begin by briefly describing them.

Problem 1.1 (Sphere packing). How much of $\mathbb{R}^d$ can be covered by non-overlapping equal spheres?

To formulate this more precisely, we will say that a discrete closed subset $\mathcal{C} \subset \mathbb{R}^d$ is a sphere packing if $|x - y| \geq 2$ for all $x, y \in \mathcal{C}$ with $x \neq y$. If we think of $\mathcal{C}$ as the set of centers of spheres, then the condition simply says that unit spheres with centers at $\mathcal{C}$ do not overlap. For any configuration $\mathcal{C}$, we define its *center density* by

$$\rho(\mathcal{C}) = \liminf_{R\to\infty} \frac{|\mathcal{C} \cap B_R|}{\mathrm{Vol}(B_R)}.$$

This quantity counts the average number of points of $\mathcal{C}$ per unit volume. The sphere packing problem then asks to maximize $\rho(\mathcal{C})$ over all sphere packings $\mathcal{C}$.

For simplicity, one may restrict oneself to periodic configurations $\mathcal{C}$, i.e., sets of the form $\mathcal{C} = \bigsqcup_{s\in S} s + \Lambda$ for some lattice $\Lambda \subset \mathbb{R}^d$ and some finite set of shifts S. A simple approximation argument shows that the supremum of densities remains the same if we restrict to such sphere packings, so not much is lost in making this assumption. For periodic configurations, we have $\rho(\mathcal{C}) = |S|/|\Lambda|$, where $|\Lambda| = \mathrm{Vol}(\mathbb{R}^d/\Lambda)$ is the covolume of the lattice Λ.

Mathematics Subject Classification 2020: 42A38 (primary); 42B10, 52C17 (secondary).
Keywords: Fourier analysis, uniqueness pairs, modular forms, sphere packing, energy minimization.

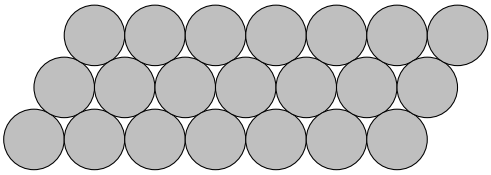


Figure 1. Part of an optimal sphere packing in $\mathbb{R}^2$.

When restricted even further to lattice packings, with $\mathcal{C} = \Lambda$ for some lattice Λ, the problem becomes equivalent to the question of determining the Hermite constant in dimension d. The Hermite constant γ_d is the maximum of the squared norm of the shortest non-zero vector of Λ taken over all lattices Λ of covolume 1. The values of γ_d for $d \leq 8$ are known from the works of Lagrange ($d = 2$; see Figure 1), Gauss ($d = 3$), Korkin–Zolotarev ($d = 4, 5$; see [18,19]), and Blichfeldt ($d = 6, 7, 8$; see [3]). Despite the enormous increase in computing power since the 1930s, no values of γ_d for $d \geq 9$ have been computed to this day, with the exception of γ_{24} (see [8]), which requires a completely different approach.

Coming back to the general question, in [11], Conway and Sloane gave a conjectural description of what all the optimal sphere packings should be in dimensions up to 9, conjecturing among other things that for $d \leq 8$ the optimal density is attained by a root lattice of type A_2, A_3, D_4, D_5, E_6, E_7, or E_8. Here, the only proof (excluding the trivial case $d = 1$) going back to the pre-computer age is the case $d = 2$ due to Thue; the proof for $d = 3$ is a very difficult computer-assisted proof due to Hales [16], and for $d = 4, 5, 6$, and 7, the problem is open.

The real surprise came from the breakthrough work of Viazovska [30], who solved the problem for $d = 8$. Shortly afterward, her work was extended in [10], solving the sphere packing problem for $d = 24$. While the proofs for $d = 2$ and $d = 3$ are based on local analysis of configurations, the proofs for $d = 8$ and $d = 24$ are based on linear programming, which connects the sphere packing problem with an optimization problem in Fourier analysis.

Before we explain the connection to Fourier analysis, let us describe a more general problem, energy minimization, where this connection figures more prominently. A potential (or potential function) is any function $p\colon \mathbb{R}_{>0} \to \mathbb{R}$ from positive reals to reals. The p-energy of a configuration $\mathcal{C}$, roughly speaking, is the average of pairwise interactions between particles, where the interaction between $x, y \in \mathcal{C}$ is given by $p(|x - y|)$. For a periodic configuration $\mathcal{C} = S + \Lambda$, one can define the energy by

$$E_p(\mathcal{C}) = \frac{1}{|S|} \sum_{u,v \in S} \sum_{x \in \Lambda \smallsetminus \{v-u\}} p(|x + u - v|),$$

provided that the sum converges. We will be interested only in two classes of potentials: the Riesz potentials $p(r) = r^{-s}$, $s > d$, and the Gaussian potentials $p(r) = e^{-\alpha r^2}$, $\alpha > 0$, for which absolute convergence is guaranteed.

Problem 1.2 (Energy minimization). Given a potential function $p\colon (0, +\infty) \to \mathbb{R}$, what is the minimal possible p-energy of a configuration $\mathcal{C} \subset \mathbb{R}^d$ with center density ρ?

This problem is closely related to the sphere packing problem. For example, if p is the indicator function of the interval $[0, 2)$, then the p-energy of $\mathcal{C}$ vanishes (for periodic $\mathcal{C}$) if and only if $\mathcal{C}$ is a sphere packing. Moreover, one can easily show that if $\mathcal{C}$ is optimal for $p(r) = r^{-s}$ for all $s > s_0$ for some s_0, then $\mathcal{C}$ is an optimal sphere packing as well, and an analogous statement holds for Gaussian potentials $p(r) = e^{-\alpha r^2}$ for $\alpha > \alpha_0$.

In view of the scarcity of results on general sphere packings, one expects even less to be known about energy minimization. For example, in dimension 2, the expectation is that A_2 should be optimal for both Gaussians and Riesz potentials. In the class of lattices, by the works of Rankin [27], Cassels [5], Diananda [12], and Ennola [14], the optimality is known for Riesz potentials and by the work of Montgomery [21] for the Gaussians. However, as of this writing, the optimality of the A_2-lattice in the class of all periodic configurations is not known for any Riesz or Gaussian potentials.

In some dimensions, the optimal configuration for Gaussian energy $p(r) = e^{-\alpha r^2}$ appears to depend on the parameter α; for instance, this is the expected behavior for $d = 3$. By contrast, the conjectural situation in dimension 2 is that A_2 is the optimal configuration for a wide class of potentials, including all Gaussians. Following Cohn and Kumar [7], we call a configuration $\mathcal{C}$ *universally optimal* if it minimizes p-energy for all $p\colon (0, \infty) \to \mathbb{R}$ that are completely monotone functions of squared distance. By a theorem of Bernstein on completely monotone functions, it is equivalent to requiring optimality of p-energy for all Gaussian potentials. In [7], the following conjecture was made.

Conjecture 1.3 (Cohn–Kumar). *In dimensions* 2, 8, *and* 24, *the configurations given by the A_2-lattice, the E_8-lattice, and the Leech lattice, respectively, are universally optimal.*

More precisely, Cohn and Kumar conjectured that the universal optimality in these cases can be certified via a linear programming bound, which we will explain shortly. In [7], they also prove that for $d = 1$, the set of integers $\mathbb{Z}$ is universally optimal, although this result was later discovered to be known from an earlier work of Ventevogel and Nijboer [29].

The key tool for proving lower bounds for energy and the connection to Fourier analysis is given in the following theorem.

Theorem 1.4 (Cohn–Kumar). *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a Schwartz function such that*

$$\begin{aligned} f(x) &\le p(|x|), \quad && x \neq 0, \\ \widehat{f}(\xi) &\ge 0, && \xi \in \mathbb{R}^d. \end{aligned}$$

Then any subset $\mathcal{C} \subset \mathbb{R}^d$ of density ρ satisfies

$$E_p(\mathcal{C}) \ge \rho \widehat{f}(0) - f(0)\,.$$

Here, we normalize Fourier transform in $\mathbb{R}^d$ by $\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i \xi \cdot x} dx$. The proof of the above result, for the case when $\mathcal{C}$ is a lattice Λ, is a one-line calculation:

$$\sum_{x \in \Lambda \smallsetminus \{0\}} p(|x|) \ge \sum_{x \in \Lambda \smallsetminus \{0\}} f(x) = -f(0) + \frac{1}{|\Lambda|} \sum_{\xi \in \Lambda^*} \widehat{f}(\xi) \ge \frac{\widehat{f}(0)}{|\Lambda|} - f(0)\,,$$

where we used the conditions for f and $\widehat{f}$ together with the Poisson summation formula (Λ^* denotes the lattice dual to Λ). A proof for general periodic configurations is completely analogous and is only a few lines longer.

The above result allows one to get lower bounds for energy by constructing certain auxiliary functions, but since it does not prescribe any method for constructing suitable f, it is not clear how good a bound it is. Jumping ahead, it turns out that, at least in some cases, the bound that Theorem 1.4 gives is optimal. Indeed, analyzing the above proof gives the following stronger claim: if there exists a function f such that

$$\begin{cases} f(x) \le p(|x|), & x \neq 0, \\ f(x) = p(|x|), & x \in \Lambda \smallsetminus \{0\}, \end{cases} \quad \text{and} \quad \begin{cases} \widehat{f}(\xi) \ge 0, & \xi \neq 0, \\ \widehat{f}(\xi) = 0, & \xi \in \Lambda^* \smallsetminus \{0\}, \end{cases}$$

then Λ is optimal for p-energy. The precise version of the conjecture of Cohn and Kumar is that such optimal auxiliary functions f witnessing a tight bound for Gaussian energy exist in dimensions 2, 8, and 24.

Let $d = 8$ and $\Lambda = \Lambda_{E_8}$ be the E_8-lattice. It is even and self-dual:

$$\Lambda^* = \Lambda, \quad |x|^2 \in 2\mathbb{Z} \quad \text{for all } x \in \Lambda\,.$$

Moreover, Λ has vectors of length $\sqrt{2n}$ for all $n \ge 0$. Assuming that the function f above is radial (which does not reduce generality, since all conditions are linear and radially symmetric, so any solution can be made radial by averaging over all rotations), we get the following rough picture; see Figure 2.

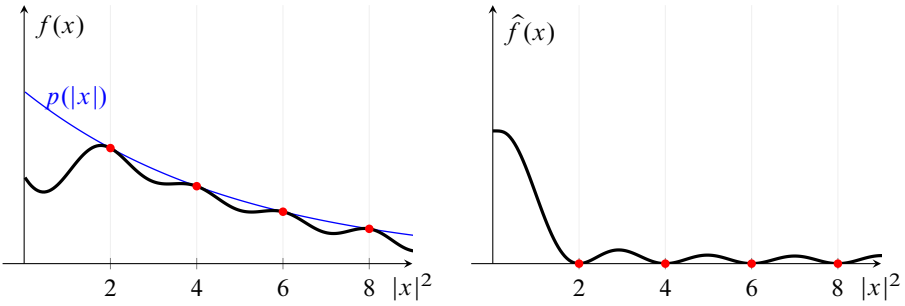


Figure 2. Behavior of any optimal auxiliary function f for $d = 8$.

In particular, we see that if such a function f exists, then $f\big(\sqrt{2n}\big) = p\big(\sqrt{2n}\big)$, $f'\big(\sqrt{2n}\big) = p'\big(\sqrt{2n}\big)$, $\widehat{f}\big(\sqrt{2n}\big) = \widehat{f}'\big(\sqrt{2n}\big) = 0$, $n \geq 1$. This leads to the question of how much freedom there is in prescribing the values of f, f', $\widehat{f}$, and $\widehat{f}'$ at $\sqrt{2n}$, $n \geq 1$. The following theorem from [9] gives a fairly complete answer to this question.

Theorem 1.5. *For $d \in \{8, 24\}$, there exist two sequences of radial Schwartz functions $a_n, b_n \in \mathcal{S}(\mathbb{R}^d)$, $n \geq 0$, such that for any radial Schwartz function f, we have*

$$f(x) = \sum_{n \geq n_0} \big(a_n(x) f\big(\sqrt{2n}\big) + b_n(x) f'\big(\sqrt{2n}\big) + \widehat{a_n}(x) \widehat{f}\big(\sqrt{2n}\big) + \widehat{b_n}(x) \widehat{f}'\big(\sqrt{2n}\big)\big).$$

Here, $n_0 = 1$ for $d = 8$ and $n_0 = 2$ for $d = 24$. Moreover, for any four rapidly decaying sequences of numbers $f\big(\sqrt{2n}\big)$, $f'\big(\sqrt{2n}\big)$, $\widehat{f}\big(\sqrt{2n}\big)$, and $\widehat{f}'\big(\sqrt{2n}\big)$, the right-hand side of the above formula defines a radial Schwartz function extending these values.

Based on an explicit form of the above interpolation formula, the cases $d = 8$ and $d = 24$ of the Cohn–Kumar conjecture were resolved in [9]. The case $d = 2$ remains widely open, and it is now known that there can be no direct analogue of the interpolation formula with $\sqrt{2n}$ replaced by the lengths of vectors in the A_2-lattice, and it is not clear how to remedy the above approach.

2. Fourier uniqueness pairs

The interpolation formula from Theorem 1.5 has some very interesting features but is rather complicated, both because it is only stated for radial functions in 8 and 24 dimensions, and it involves not only the values of f and its Fourier transform but also the values of their derivatives. The following simpler and more basic result was

inspired by Theorem 1.5 and proved a few years earlier in [23] (while Theorem 1.5 was still a conjecture), and it will serve as a better entry point for discussing Fourier uniqueness pairs.

Theorem 2.1. *There exists a sequence of even Schwartz functions $a_n\colon \mathbb{R} \to \mathbb{R}$ with the property that for every even Schwartz function $f : \mathbb{R} \to \mathbb{R}$, we have*

$$f(x) = \sum_{n=0}^{\infty} a_n(x) f\big(\sqrt{n}\big) + \sum_{n=0}^{\infty} \widehat{a}_n(x) \widehat{f}\big(\sqrt{n}\big). \tag{2.1}$$

Moreover, for any pair of rapidly decaying sequences $(x_n)_{n\geq 0}$ and $(y_n)_{n\geq 0}$ satisfying $\sum_{n\in\mathbb{Z}} x_{n^2} = \sum_{n\in\mathbb{Z}} y_{n^2}$, there exists a unique even Schwartz function f with $f\big(\sqrt{n}\big) = x_n$ and $\widehat{f}\big(\sqrt{n}\big) = y_n$, and it is given by $f = \sum_{n\geq 0} x_n a_n + \sum_{n\geq 0} y_n \widehat{a_n}$.

There is also an analogous statement for odd Schwartz functions that also involves the values $f'(0)$ and $\widehat{f}'(0)$, but we will concentrate on the even case. The proof of the above result given in [23] is based on an explicit construction of the functions a_n, whose defining property is that they satisfy $a_n\big(\sqrt{m}\big) = \delta_{m,n}$, $\widehat{a_n}\big(\sqrt{m}\big) = 0$ for $n \geq 0$, $m \geq 1$. The special functions a_n end up being closely related to weakly holomorphic modular forms for the Hecke theta group. Without going into details here, the function a_0 is given by the contour integral

$$a_0(x) = \frac{1}{4} \int_{-1}^{1} \theta^3(z) e^{\pi i z x^2}\, dz,$$

where $\theta(z) = \sum_{n\in\mathbb{Z}} e^{\pi i n^2 z}$, and a similar formula exists for all a_n.

As an immediate corollary of the theorem, we have the following result.

Corollary 2.2. *If an even Schwartz function f satisfies $f\big(\sqrt{n}\big) = \widehat{f}\big(\sqrt{n}\big) = 0$, $n \geq 0$, then $f = 0$.*

Note that we may omit the condition $\widehat{f}(0) = 0$ since by Poisson summation, we have $\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \widehat{f}(n)$. With this adjustment, the above statement is tight in the sense that if we omit one of the $f\big(\sqrt{m}\big) = 0$, $m \geq 0$, or $\widehat{f}\big(\sqrt{m}\big) = 0$, $m \geq 1$, then the claim fails. Indeed, for $f = a_m$, we have $f\big(\sqrt{n}\big) = 0$ for all $n \neq m$ and $\widehat{f}\big(\sqrt{n}\big) = 0$ for all $n \geq 1$.

Both known proofs (the one from [23] and a more recent proof in the work of Gerbelli-Gauthier and Venkatesh [15]) of this corollary involve modular forms, whose appearance may seem quite surprising. This corollary naturally leads us to the following concept.

Definition 2.3. Let $\mathcal{H}$ be a subspace of the space of continuous absolutely integrable functions on $\mathbb{R}$. We say that a pair of closed subsets $A, B \subseteq \mathbb{R}$ is a *Fourier uniqueness pair* (for $\mathcal{H}$) if for any $f \in \mathcal{H}$ if $f|_A = 0$ and $\widehat{f}\,|_B = 0$, then $f = 0$. If both A and B cannot be replaced by a proper closed subset while keeping the above property, we say that the uniqueness pair is *tight*.

If (A, B) is a uniqueness pair, there exists an inverse map $(f|_A, \widehat{f}\,|_B) \mapsto f$, which we will call a Fourier interpolation formula (there may be subtleties in determining the domain of this map, which we will largely ignore). The above definition can certainly be extended, for example, by considering general locally compact abelian or even non-abelian groups or relaxing the restrictions on $\mathcal{H}$, but it does capture most of the interesting examples of uniqueness pairs that follow.

Example 1. Our first example is related to classical results in time-frequency analysis. Namely, taking $A = \{x : |x| \geq 1/2\}$ and $B = \mathbb{Z}$ gives a (tight) uniqueness pair. The proof of this fact, together with an interpolation formula, is immediate upon noting that (for a suitable class of functions)

$$f(x) = \sum_{m\in\mathbb{Z}} \widehat{f}(m) e^{2\pi i m x} - \sum_{n\neq 0} f(x+n),$$

which is a form of the Poisson summation formula. If we are in the situation when $f|_{|x|\geq 1/2} = 0$, taking the Fourier transform of the above equation would give us the famous Whittaker–Shannon formula (for $\widehat{f}$ with finitely supported f):

$$\widehat{f}(x) = \sum_{m\in\mathbb{Z}} \widehat{f}(m) \frac{\sin(\pi(x-m))}{\pi(x-m)}.$$

Example 2. Our second example concerns discrete Fourier transform. Let p be a prime, and for $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$, we define its Fourier transform by

$$\widehat{f}(\xi) = \frac{1}{\sqrt{p}} \sum_{x=0}^{p-1} f(x) e^{-2\pi i x\xi/p}.$$

In [28], Tao proved the following uncertainty principle: if $f : \mathbb{Z}/p\mathbb{Z} \to \mathbb{C}$ is non-zero, then $|\mathrm{supp}(f)| + |\mathrm{supp}(\widehat{f})| \geq p + 1$. This can be reformulated as saying that for any $A, B \subseteq \mathbb{Z}/p\mathbb{Z}$, (A, B) is a uniqueness pair if and only if $|A| + |B| \geq p$, and (A, B) is a tight uniqueness pair if and only if $|A| + |B| = p$. Therefore, in this case, we get a complete classification of uniqueness pairs.

Example 3. Our third example is the pair $A = \{\sqrt{n}\}_{n\geq 0}$, $B = A \setminus \{0\}$, with which we started this discussion. It is a tight uniqueness pair for the space of even Schwartz

functions by Theorem 2.1, together with an interpolation formula (2.1). This is the first example of a uniqueness pair where the sets A and B are both discrete.

Example 4. Our last example is related to zeros of L-functions. Fix $\varepsilon > 0$, and let $\mathcal{H}_1$ be the space of functions f analytic in the strip $\{z : |\operatorname{Im} z| < 1/2 + \varepsilon\}$ such that

$$\sup_{|y|<1/2+\varepsilon} \int_{-\infty}^{\infty} |f(x+iy)|(1+|x|)dx < \infty.$$

In [4], the following interpolation formula was proved.

Theorem 2.4. *There exist two sequences of rapidly decaying even entire functions $U_n(z)$, $n = 1, 2, \dots$, and $V_{\rho,j}(z)$, $0 \le j < m_\rho$, with ρ ranging over the nontrivial zeros of $\zeta(s)$ such that for every even function f in $\mathcal{H}_1$ and z with $|\operatorname{Im} z| < 1/2$,*

$$f(z) = \sum_{n=1}^{\infty} \hat{f}\left(\frac{\log n}{4\pi}\right) U_n(z) + \lim_{m\to\infty} \sum_{0<\gamma\le T_m} \sum_{j=0}^{m_\rho - 1} f^{(j)}\left(\frac{\rho - \frac{1}{2}}{i}\right) V_{\rho,j}(z)$$

for some universal sequence of positive numbers $T_m \to \infty$.

This gives rise to a Fourier uniqueness pair.

Corollary 2.5. *If $\mathfrak{Z}$ denotes the set of nontrivial zeros of $\zeta(s)$, then*

$$A = \left\{\frac{\rho - \frac{1}{2}}{i} : \rho \in \mathfrak{Z}\right\}, \quad B = \left\{\pm\frac{\log n}{4\pi} : n \in \mathbb{N}\right\}$$

is a tight uniqueness pair for $\mathcal{H}_1$.

For this corollary, we need to refine the definition of a uniqueness pair to accommodate the fact that A may not lie on the real line, if Riemann hypothesis fails, or have points with multiplicities (in which case $f|_A = 0$ includes vanishing of higher derivatives, according to a multiplicity of a point in A), if the simple zeros conjecture fails, so unconditionally we only know that A is a multiset contained in a strip $|\operatorname{Im} z| < 1/2$.

Perhaps surprisingly, the proof again relies on modular forms. It also works for several other classes of L-functions: $L(\chi, s)$ for Dirichlet characters χ (with $\log n$ replaced by $\log(qn)$), $\zeta_K(s)$ for any imaginary quadratic field K (with $\log n$ replaced by $\log(4|D|n^2)$), and $L(f, s)$ for a cusp form f of weight k and level N (with $\log n$ replaced by $\log(4Nn^2)$). One could expect that there should be analogous statements for all L-functions in the Selberg class, but the strategy used in [4] cannot work for any L-functions of degree higher than 2.

The uniqueness pairs from the last two examples account for almost all of the tight uniqueness pairs (with discrete A and B) that are currently known. The proof of

Theorem 2.1 relies heavily on the arithmetic nature of the sequence $\sqrt{n}$ and does not generalize to other sequences in any direct way (although a slight modification of the argument also works for the sequence $\left(\sqrt{n+a}\right)_n$ for some range of a). A perturbation argument due to Ramos and Sousa [25] shows that for sufficiently small ε_n, the sequence $\left(\sqrt{n+\varepsilon_n}\right)_n$ has similar properties, but it does not allow us to get very far from the original sequence. As we will see below, the situation becomes better if we relax the tightness condition, allowing the use of more analytic techniques.

3. Connection to modular forms

Before describing the recent analytic results on Fourier uniqueness pairs, it is worthwhile to give at least a very rough idea of how the proof of Theorem 2.1 involves modular forms. For a more detailed exposition, we refer to [23], [4, §3].

The first step is to note that it suffices to prove (2.1) for f from some subset of the Schwartz space with a dense linear span. A good choice turns out to be to consider all $f(x) = e^{\pi i \tau x^2}$ with $\operatorname{Im}(\tau) > 0$, since the Fourier transform of a Gaussian is again a Gaussian. Denoting

$$F(\tau) = \sum_{n \geq 0} a_n(x) e^{\pi i n \tau}, \quad G(\tau) = \sum_{n \geq 0} \widehat{a_n}(x) e^{\pi i n \tau},$$

equation (2.1) can then be rewritten as

$$F(\tau) + \left(\frac{\tau}{i}\right)^{-1/2} G\left(\frac{-1}{\tau}\right) = e^{\pi i \tau x^2}.$$

(Here, we used the fact that $\widehat{f}(\xi) = (\tau/i)^{-1/2} e^{\pi i(-1/\tau)\xi^2}$.) Now let us fix $x \in \mathbb{R}$ and consider the above identity as a functional equation in τ. Since both F and G are 2-periodic, we get the following system of equations:

$$\begin{cases} F(\tau) + \left(\frac{\tau}{i}\right)^{-k} G\left(\frac{-1}{\tau}\right) = \varphi(\tau), \\ F(\tau) - F(\tau+2) = 0, \\ G(\tau) - G(\tau+2) = 0, \end{cases} \tag{3.1}$$

where $\varphi(\tau) = e^{\pi i \tau x^2}$ and $k = 1/2$. It is not hard to see that (2.1) is in fact equivalent to (3.1) if we additionally require F and G to be of moderate growth in the upper half-plane $\mathbb{H}$, meaning that they should be bounded by some polynomial in $(\operatorname{Im} \tau)^{-1}$ and $|\tau|$ for $\tau \in \mathbb{H}$.

It turns out that (3.1) can be solved for any $k \geq 0$ and any φ that is holomorphic and of moderate growth in $\mathbb{H}$, with F and G in the same class, and moreover the solution is unique up to a finite-dimensional space of modular forms (which correspond to

solutions of the homogeneous system with $\varphi = 0$). For $k > 2$, this can be derived from the results of Knopp [17] on Eichler cohomology, but it is hard to turn his proof into an explicit formula. We instead use a contour integral construction inspired by [13]. Let us explain, ignoring all technical details, how this can be achieved. Note that symmetrizing $F_\pm = F \pm G$, we get equations in a single unknown function:

$$\begin{cases} F_\pm(\tau) \pm \left(\frac{\tau}{i}\right)^{-k} F_\pm\left(\frac{-1}{\tau}\right) = \varphi(\tau) \pm \left(\frac{\tau}{i}\right)^{-k} \varphi\left(\frac{-1}{\tau}\right), \\ F_\pm(\tau) - F_\pm(\tau+2) = 0. \end{cases}$$

Let us first look at the case of F_- and $k = 0$:

$$\begin{cases} F_-(\tau) - F_-\left(\frac{-1}{\tau}\right) = \varphi(\tau) - \varphi\left(\frac{-1}{\tau}\right), \\ F_-(\tau) = F_-(\tau+2). \end{cases}$$

The two fractional linear transformations $\tau \mapsto \tau + 2$ and $\tau \mapsto -1/\tau$ generate the so-called theta group Γ_θ, a fundamental group for which is depicted in Figure 3.

There is a conformal map J mapping the fundamental domain onto the slit plane $\mathbb{C} \setminus (-\infty, 1]$ as shown in Figure 3 (this function J is the so-called Hauptmodul for Γ_θ). The composition $A(t) = F_-(J^{-1}(t))$ is then holomorphic in $\overline{\mathbb{C}} \setminus [0,1]$, with a jump discontinuity across $[0,1]$ prescribed by $\varphi(J^{-1}(t))$. The solution is then given by the Sokhotski–Plemelj theorem, which says that A can be computed as an integral against the Cauchy kernel: $A(s) = \int_0^1 g(s)/(t-s)ds$, where $g(s)$ is the jump at $t = s$. Translating back into τ variable, this gives a formula for F_- as a contour integral over the semicircle bounding the fundamental domain.

It turns out that the general case can be derived from the special case $k = 0$ and sign "$-$": first, taking instead of $\varphi(\tau)$ the function $\varphi(\tau)/\theta^{2k}(\tau)$, where $\theta(\tau) = \sum_{n\in\mathbb{Z}} e^{\pi i n^2 \tau}$ is the Jacobi theta function, we get to the case of general k, for which one uses the modular transformation $\theta(\tau) = (\tau/i)^{-1/2}\theta(-1/\tau)$. A similar trick with division by $1 - 2\lambda(\tau)$ (where $\lambda(\tau)$ is the modular lambda invariant) allows one to get

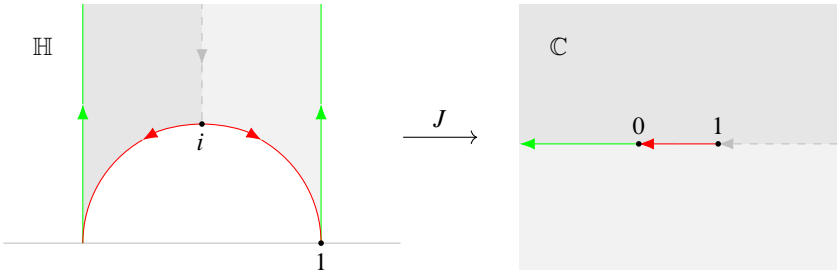


Figure 3. Fundamental domain for Γ_θ and conformal map J.

solutions for F_+ from the ones for F_-. After all of these manipulations, one arrives at an explicit formula

$$F_\pm(\tau) = \int_{-1}^{1} \mathcal{K}_k^\pm(\tau, z)\varphi(z)dz,$$

where $\mathcal{K}_k^\pm(\tau, z)$ is a certain modular kernel. This formula can then be extended by analytic continuation to give a solution in all of $\mathbb{H}$, and since everything is explicit, going back to the original question, one can derive integral representations for all of the functions $a_n(x)$ appearing in (2.1).

4. Further results and questions

We conclude by discussing the recent analytic results on uniqueness pairs and some open questions.

4.1. Uniqueness pairs from powers of integers

In [24], Ramos and Sousa have constructed the first discrete uniqueness pairs by purely analytic techniques. Their results concern uniqueness pairs of the form $A = \{\pm cn^\alpha\}_{n\geq 0}$, $B = \{\pm dn^\beta\}_{n\geq 0}$, where $\alpha, \beta, c, d > 0$. They have proved that (A, B) is a uniqueness pair, provided that $(\alpha, \beta) \in \Omega$, where Ω is a certain open subset in $\{\alpha, \beta > 0, \alpha + \beta < 1\}$.

More recently, Kulikov, Nazarov, and Sodin [20] obtained a much more precise description:

$$\begin{aligned} \alpha + \beta < 1 \quad &\Rightarrow \quad (A, B) \text{ is a uniqueness pair,} \\ \alpha + \beta > 1 \quad &\Rightarrow \quad (A, B) \text{ is a non-uniqueness pair.} \end{aligned}$$

Moreover, in the case $\alpha + \beta = 1$, they obtained a precise threshold value for cd:

$$\alpha + \beta = 1: \quad \begin{aligned} 2cd < \alpha^{-\alpha}\beta^{-\beta} \quad &\Rightarrow \quad (A, B) \text{ is a uniqueness pair,} \\ 2cd > \alpha^{-\alpha}\beta^{-\beta} \quad &\Rightarrow \quad (A, B) \text{ is a non-uniqueness pair.} \end{aligned}$$

The results of [20] in fact extend to other sequences that behave roughly like n^α. However, they do not seem to reach the critical case, leaving the following problem.

Question. For $\alpha + \beta = 1$ and $2cd = \alpha^{-\alpha}\beta^{-\beta}$, is $A = \{\pm cn^\alpha\}_{n\geq 0}$, $B = \{\pm dn^\beta\}_{n\geq 0}$ a uniqueness pair? If so, is it a tight uniqueness pair, and is there a well-behaved interpolation formula?

Note that for $\alpha = \beta = 1/2$, the answer is provided by Theorem 2.1.

4.2. Discrete uniqueness pairs in $\mathbb{R}^d$

The situation in $\mathbb{R}^d$ for $d > 1$ is more complicated, and for discrete A, B, there seem to be no known explicit interpolation formulas akin to Theorem 2.1. Nevertheless, in [26], Ramos and Stoller proved that discrete uniqueness pairs exist by deforming interpolation formulas for the case when $A = B$ is a union of concentric spheres. More recently, Adve [1] established that for any $\alpha + \beta = 1$, the pair

$$A = \{c|n|^{-\alpha} n : n \in \mathbb{Z}^d \setminus \{0\}\}, \quad B = \{c|n|^{-\beta} n : n \in \mathbb{Z}^d \setminus \{0\}\} \tag{4.1}$$

is a uniqueness pair if $c < c_0$ and a non-uniqueness pair if $c > c_1$ for some constants $0 < c_0 < c_1$. See Figure 4 below for a picture for $\alpha = \beta = 1/2$ and $A = B$. Note that in Adve's example (for $d = 2$ and $\alpha = \beta = 1/2$), the number of points in A in a ball of radius R is approximately CR^4 for some fixed $C > 0$. In [22], a number-theoretic (based on the ring of integers in a real quadratic field) construction of a family of non-uniqueness pairs was given that also have $C'R^4$ points in a ball of radius R, and C' can be arbitrarily large. This shows that it is not just the "size" of the sets A and B that is important for (A, B) to be a uniqueness pair, but also their "regularity".

Note that in $\mathbb{R}$, the transition from uniqueness pairs to non-uniqueness pairs in (4.1) as c goes from 0 to ∞ has a sharp threshold at certain value $c = c_{\text{crit}}$. In $\mathbb{R}^d$ for $d > 1$, there is a range of values $c \in (c_0, c_1)$ for which it is not clear whether (4.1) is a uniqueness pair or not.

Question. Is there a sharp transition value of c for all $d > 1$?

Question. Are there tight uniqueness pairs in $\mathbb{R}^d$ with discrete A, B for any $d > 1$?

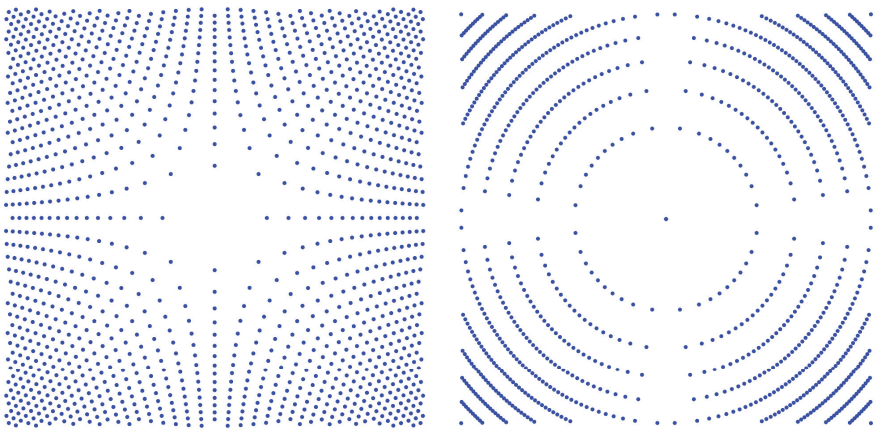


Figure 4. Left: Adve's uniqueness pair. Right: a non-uniqueness pair constructed from the ring of integers of $\mathbb{Q}(\sqrt{89})$. For the general case of $\mathbb{Q}(\sqrt{D})$ as $D \to \infty$, these non-uniqueness sets get higher and higher density in the union of concentric circles $\{x : |x|^2 \in \mathbb{Z}_{\geq 0}\}$.

4.3. Linear programming for energy and conformal bootstrap

To conclude the story, we return to sphere packing and energy minimization problems with which we started and present an important conjecture relating the linear programming bound of Cohn–Kumar to Fourier interpolation formulas in all dimensions ≥ 4.

To find an optimal function for the use in Theorem 1.4, we have the following optimization problem (CK_α): maximize the value $\widehat{f}(0) - f(0)$ over all $f : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$\begin{aligned} f(x) &\leq e^{-\alpha\pi|x|^2}, \quad && x \in \mathbb{R}^d, \\ \widehat{f}(\xi) &\geq 0, && \xi \in \mathbb{R}^d. \end{aligned}$$

In [9], the following conjecture was formulated.

Conjecture 4.1. *For $d \geq 4$, the solution to (CK_α) is unique and is given by a radial Schwartz function. The points where $f(r) = e^{-\alpha\pi r^2}$ is attained form a discrete set $r_1 < r_2 < \cdots$ with $r_n^2 \sim 2n$, $n \to \infty$, that does not depend on α. The condition $\widehat{f}(r) = 0$ holds for exactly the same values $r = r_n$. Finally, there exists a linear interpolation formula recovering any radial Schwartz function $f : \mathbb{R}^d \to \mathbb{R}$ from the values $f(r_n)$, $f'(r_n)$, $\widehat{f}(r_n)$, $\widehat{f}'(r_n)$, $n \geq 1$.*

This conjecture suggests a natural generalization of Theorem 1.5, directly connected to the linear programming bound in all dimensions $d \geq 4$, not just $d = 8$ and $d = 24$. It can be further refined by describing what happens to the solutions of the dual linear program. In the dual setting, one considers the space of generalized theta functions

$$\vartheta(\tau) = 1 + \sum_{n \geq 1} c_n e^{2\pi i \lambda_n \tau}$$

with $\rho_n > cn$ for some $c > 0$ and $c_n \geq 0$ such that ϑ satisfies

$$\vartheta\left(\frac{-1}{\tau}\right) = \left(\frac{\tau}{i}\right)^{d/2} \vartheta(\tau).$$

The dual optimization problem (CK_α^*) is then to minimize the value of $\vartheta(i\alpha)$ in the space of all such series for some fixed $\alpha > 0$. The above conjecture can then be extended to say that for every $d \geq 4$, there exists a *universally optimal* theta function Θ_d that minimizes the value at $i\alpha$ for all $\alpha > 0$. This function has an expansion

$$\Theta_d(\tau) = 1 + \sum_{n \geq 1} a_n e^{2\pi i \rho_n \tau},$$

with $\rho_n = \rho_n^{(d)} = r_n^2/2$ (where r_n are specified in the above conjecture) and some $a_n \geq 0$. By the results of [9], for $d = 8$ and $d = 24$, these functions are known to be

$$\Theta_8(\tau) = 1 + 240e^{2\pi i\tau} + 2160e^{4\pi i\tau} + \cdots,$$
$$\Theta_{24}(\tau) = 1 + 196560e^{4\pi i\tau} + 16773120e^{6\pi i\tau} + \cdots,$$

the theta functions of the E_8 and the Leech lattice.

In [2], the above optimization problem is shown to be related to conformal bootstrap in conformal field theory, and extensive numerical data is presented on the values of ρ_n and a_n for a large range of dimensions. Of particular interest both in conformal bootstrap and in the sphere packing problem is the behavior of the smallest value ρ_1. For the sphere packing, it should correspond to the optimal value of the Cohn–Elkies LP bound for sphere packing [6]. This gives further motivation for the study of the two optimization problems above, since a good understanding of their structure could give further insights into the sphere packing problem in large dimensions.

Funding. The author acknowledges funding from the European Union (European Research Council, FourIntExP, 101078782).

References

[1] A. Adve, Density criteria for Fourier uniqueness phenomena in $\mathbb{R}^d$. 2023, arXiv: 2306.07475v1

[2] N. Afkhami-Jeddi, H. Cohn, T. Hartman, D. de Laat, and A. Tajdini, High-dimensional sphere packing and the modular bootstrap. *J. High Energy Phys.* **2020** (2020), no. 12, article no. 66, 45 pp. Zbl 1457.81095 MR 4239386

[3] H. F. Blichfeldt, The minimum values of positive quadratic forms in six, seven and eight variables. *Math. Z.* **39** (1935), no. 1, 1–15 Zbl 0009.24403 MR 1545485

[4] A. Bondarenko, D. Radchenko, and K. Seip, Fourier interpolation with zeros of zeta and L-functions. *Constr. Approx.* **57** (2023), no. 2, 405–461 Zbl 1528.11076 MR 4577389

[5] J. W. S. Cassels, On a problem of Rankin about the Epstein zeta-function. *Proc. Glasg. Math. Assoc.* **4** (1959), no. 2, 73–80 Zbl 0103.27602 MR 0117193

[6] H. Cohn and N. Elkies, New upper bounds on sphere packings I. *Ann. of Math. (2)* **157** (2003), no. 2, 689–714 Zbl 1041.52011 MR 1973059

[7] H. Cohn and A. Kumar, Universally optimal distribution of points on spheres. *J. Amer. Math. Soc.* **20** (2007), no. 1, 99–148 Zbl 1198.52009 MR 2257398

[8] H. Cohn and A. Kumar, Optimality and uniqueness of the Leech lattice among lattices. *Ann. of Math. (2)* **170** (2009), no. 3, 1003–1050 Zbl 1213.11144 MR 2600869

[9] H. Cohn, A. Kumar, S. Miller, D. Radchenko, and M. Viazovska, Universal optimality of the E_8 and Leech lattices and interpolation formulas. *Ann. of Math. (2)* **196** (2022), no. 3, 983–1082 Zbl 1509.11061 MR 4502595

[10] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. Viazovska, The sphere packing problem in dimension 24. *Ann. of Math. (2)* **185** (2017), no. 3, 1017–1033 Zbl 1370.52037 MR 3664817

[11] J. H. Conway and N. J. A. Sloane, What are all the best sphere packings in low dimensions? *Discrete Comput. Geom.* **13** (1995), no. 3–4, 383–403 Zbl 0844.52013 MR 1318784

[12] P. H. Diananda, Notes on two lemmas concerning the Epstein zeta-function. *Proc. Glasg. Math. Assoc.* **6** (1964), no. 4, 202–204 Zbl 0128.04501 MR 0168537

[13] W. Duke, Ö. Imamoglu, and Á. Tóth, Cycle integrals of the j-function and mock modular forms. *Ann. of Math. (2)* **173** (2011), no. 2, 947–981 Zbl 1270.11044 MR 2776366

[14] V. Ennola, On a problem about the Epstein zeta-function. *Proc. Cambridge Philos. Soc.* **60** (1964), no. 4, 855–875 Zbl 0146.05504 MR 0168535

[15] M. Gerbelli-Gauthier and A. Venkatesh, Cohomology of Fuchsian groups and Fourier interpolation. *Essent. Number Theory* **4** (2025), no. 2, 217–249 Zbl 08085866 MR 4959855

[16] T. C. Hales, A proof of the Kepler conjecture. *Ann. of Math. (2)* **162** (2005), no. 3, 1065–1185 Zbl 1096.52010 MR 2179728

[17] M. I. Knopp, Some new results on the Eichler cohomology of automorphic forms. *Bull. Amer. Math. Soc.* **80** (1974), no. 4, 607–632 Zbl 0292.10022 MR 0344454

[18] A. Korkine and G. Zolotareff, Sur les formes quadratiques positives quaternaires. *Math. Ann.* **5** (1872), no. 4, 581–583 JFM 04.0081.03 MR 1509795

[19] A. Korkine and G. Zolotareff, Sur les formes quadratiques positives. *Math. Ann.* **11** (1877), no. 2, 242–292 JFM 09.0139.01 MR 1509914

[20] A. Kulikov, F. Nazarov, and M. Sodin, Fourier uniqueness and non-uniqueness pairs. *J. Math. Phys. Anal. Geom.* **21** (2025), no. 1, 84–130 Zbl 1565.42003 MR 4866154

[21] H. L. Montgomery, Minimal theta functions. *Glasg. Math. J.* **30** (1988), no. 1, 75–85 Zbl 0639.10017 MR 0925561

[22] D. Radchenko and M. Stoller, Fourier non-uniqueness sets from totally real number fields. *Comment. Math. Helv.* **97** (2022), no. 3, 513–553 Zbl 1497.11259 MR 4468993

[23] D. Radchenko and M. Viazovska, Fourier interpolation on the real line. *Publ. Math. Inst. Hautes Études Sci.* **129** (2019), no. 1, 51–81 Zbl 1455.11075 MR 3949027

[24] J. P. G. Ramos and M. Sousa, Fourier uniqueness pairs of powers of integers. *J. Eur. Math. Soc. (JEMS)* **24** (2022), no. 12, 4327–4351 Zbl 1504.42014 MR 4493626

[25] J. P. G. Ramos and M. Sousa, Perturbed interpolation formulae and applications. *Anal. PDE* **16** (2023), no. 10, 2327–2384 Zbl 1537.41001 MR 4678143

[26] J. P. G. Ramos and M. Stoller, Perturbed Fourier uniqueness and interpolation results in higher dimensions. *J. Funct. Anal.* **282** (2022), no. 12, article no. 109448, 34 pp. Zbl 1486.42001 MR 4403065

[27] R. A. Rankin, A minimum problem for the Epstein zeta-function. *Proc. Glasg. Math. Assoc.* **1** (1953), no. 4, 149–158 Zbl 0052.28005 MR 0059300

[28] T. Tao, An uncertainty principle for cyclic groups of prime order. *Math. Res. Lett.* **12** (2005), no. 1, 121–127 Zbl 1080.42002 MR 2122735

[29] W. J. Ventevogel and B. R. A. Nijboer, On the configuration of systems of interacting particles with minimum potential energy per particle. *Phys. A* **98** (1979), no. 1–2, 274–288 MR 0546896

[30] M. S. Viazovska, The sphere packing problem in dimension 8. *Ann. of Math. (2)* **185** (2017), no. 3, 991–1015 Zbl 1373.52025 MR 3664816

Danylo Radchenko
Laboratoire Alexander Grothendieck, Institut des Hautes Études Scientifiques, CNRS, 35 route de Chartres, 91440 Bures-sur-Yvette, France; radchenko@ihes.fr

Other prize lectures

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/41

Learning physics simulations for improving design processes in the industry

Fabien Casenave

Abstract. This article is the written version of the Felix Klein prize lecture, given on 16 July 2024, at the 9th European Congress of Mathematics in Seville. In the same fashion as the lecture, general ideas and results are given, and references are provided for the reader interested in details and rigorous developments.

In an industrial group like Safran, numerical simulation of physics phenomena is used in the majority of the design processes of its manufactured products. At Safran's corporate research center, we develop technologies to improve these processes by deriving fast and reliable surrogate models for various domains of physics.

We address two technologies developed in recent years. First, a physical reduced-order modeling method for non-linear structural mechanics and thermal analysis. This method is applied to the lifetime computation of high-pressure turbine blades and the fast heat analysis of high-pressure compressors. Second, the learning of physics simulations under non-parameterized geometrical variability using classical machine learning tools, thanks to non-linear deterministic dimensionality reduction (involving morphing and principal component analysis) and Gaussian process regression, with predictive uncertainties. We also present some of our contributions to the open-source and open-data community.

1. Introduction: Surrogate modeling in the industry

Consider an operator F modeling a large and complex numerical problem or an expensive experiment. Design processes in the industry often require evaluating such expensive operators a large number of times, in so-called many-query tasks. Examples of such tasks are parametric exploration, optimization, uncertainty quantification or calibration. Without any modification or simplification, they are intractable and cannot be applied to real-life industrial configurations.

To give an example of a many-query task, consider the optimization of the shape of an aircraft wing, in order to maximize its finesse under a minimum thickness constraint; see Figure 1. When using an iterative algorithm to compute an approximated

Mathematics Subject Classification 2020: 41A46 (primary); 65N99 (secondary).
Keywords: reduced-order modeling, machine learning, industrial applications.

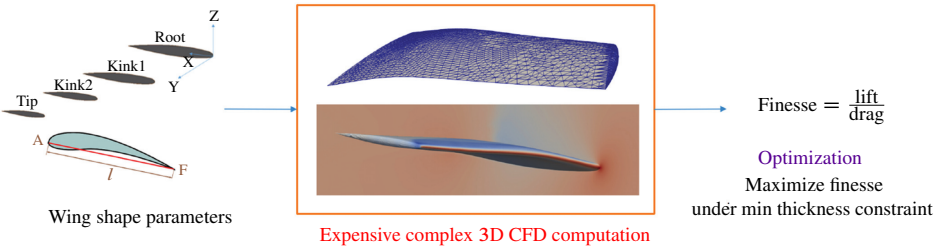


Figure 1. Optimization of the shape of an aircraft wing.

solution of this optimization problem, we need to run the complex fluid solver a large number of times.

A classical way of recovering tractable tasks is by replacing the costly operator F by a surrogate f. The behavior of F is learned by extracting a trend from data. A training database is built by choosing training inputs from a design of experiments (DoE) algorithm and computing the corresponding outputs by evaluating the operator F. The surrogate is constructed in the form of a regressor trained on the obtained input/output couples in the training stage. This surrogate should be accurate enough to replace, in the exploitation stage, the operator F in the considered many-query task and fast enough to make it tractable on an industrial configuration.

A favorable situation concerns tabular problems, where input and output are vectors in fixed low-dimensional vector spaces: $\mathbb{R}^d \ni x \mapsto y = F(x) \in \mathbb{R}^p$, with $d, p \in \mathbb{N}^*$. Many classical methods are available for low-dimensional input-output regression problems, such as linear or polynomial regression, nearest neighbors, random forests, support vector regression or Gaussian process regression.

In this work, we are interested in learning physics problems, which differ from tabular problems in the sense that physical settings involve complex heterogeneous objects of possible variable high dimension, such as meshes, boundary conditions, fields or material models. It should be noted that a physics setting differs from the inputs of a learning problem. The inputs constitute the subset of the physical setting that varies in the considered many-query task. Likewise, the outputs to be predicted by the surrogate model are the subset of the solution that will be used to make an informed decision after the many-query task has ended.

To give an example, consider the AirfRANS physics learning problem [3]; two samples in the training dataset are illustrated in Figure 2. In this learning problem, the physics setting is constituted of the mesh of the fluid domain, the boundary conditions, the Reynolds-averaged Navier–Stokes equations and the $k - \omega$ turbulence model; see [3] for more details. The inputs of the learning problem are the mesh, the angle of attack and the inlet velocity, while the outputs are the velocity, pressure and turbulent velocity fields. The chosen turbulence model is part of the physics setting,

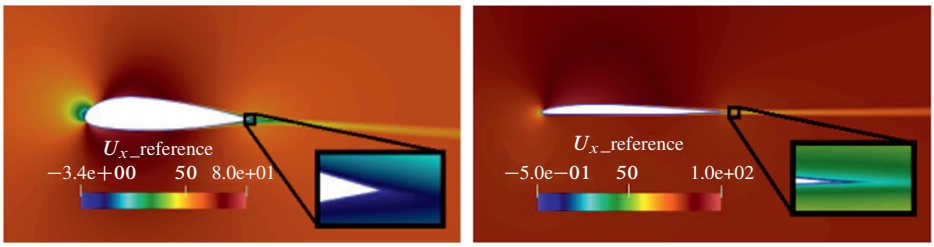


Figure 2. Component-x of the velocity fields of two samples in the AirfRANS dataset.

but being fixed by the samples of the training dataset, it is not part of the inputs of the learning problem.

The advantages of considering a physics learning problem instead of a tabular problem with only scalar inputs and outputs are the possibility to change the shape outside the training parameterization, for instance, in exemption studies, and the possibility to change the post-treatments of the predicted solution fields, after the learning has ended.

In what follows, we start by presenting some physical reduced-order modeling technologies developed at Safran in Section 2. Then, we introduce a machine learning method for dealing with non-parameterized geometrical variability in Section 3. Finally, we present some of our initiatives in the open-source and open-data communities in Section 4.

2. Physical reduced-order modeling

Physical reduced-order modeling (ROM), also called projection-based or *a posteriori* reduced-order modeling, is a class of surrogate modeling methods that relies on the resolution of the underlying physics equations in the exploitation stage.

2.1. Linear data compression

In the training stage, we are given inputs generated using a DoE algorithm and corresponding outputs computed by evaluating the operator F. A linear dimensionality reduction algorithm is then applied to the outputs to construct a reduced-order basis. The exploitation stage consists of applying the same algorithms as the ones used when solving the high-dimensional problems involved in evaluating F, but restricting the solution to the subspace spanned by the reduced-order basis. To give an example, assume that evaluating the operator F consists of solving a costly high-dimensional initial boundary value problem. A numerical solution can be calculated by deriving a variational formulation, choosing a time discretization scheme (for instance, the implicit Euler), a space discretization scheme (for instance, a finite element basis)

and computing at each time step the solution to the obtained system of non-linear equations (using, for instance, Newton's algorithm). In this case, a physical ROM of this problem can be derived by replacing the finite element basis by our computed reduced-order basis. We notice that to keep using the Galerkin method, we need to express the reduced solution as a linear combination of the elements of the reduced-order basis, in the same fashion as the reference solution is expressed as a linear combination of the finite element basis. Many variants have been proposed to take into account other types of problems and numerical simulation methods involved in the operator F evaluations. Other variants in the training stage have also been proposed; for instance, the reduced basis method [22] does not rely on a design of experiments, but rather it starts with a single input/output couple and enriches the data by selecting, in a greedy manner, the configuration where the current reduced-order model makes the largest error. Such enrichment strategies rely on fast and accurate error bounds [7] and can serve in the exploitation stage to provide accuracy assessments and trigger enrichment steps.

In physical ROM methods, the speedup comes from the fact that the cardinality of the reduced-order basis is much smaller than that of the finite element basis, leading to much smaller linear systems to solve in Newton's algorithm of the exploitation stage. However, constructing these small linear systems can be very costly, in particular when numerical integration over the geometrical domain (after discretization: the mesh) is required, which is generally the case with non-linear problems. Being able to assemble the reduced operators involved in Newton's algorithm as fast as possible is critical to obtain practical speedup. Precomputations and constructions of approximations (called hyperreduction) are carried out in the training stage in an operator compression step; see [6, Section 3.2] for more details and [8, Section 1] for references on hyperreduction methods.

We focus on non-linear structural mechanics and non-linear transient heat problems; see [6] and [8], respectively, for details on the considered formulations. In both cases, we rely on the snapshot proper orthogonal decomposition (POD) for the dimensionality reduction step, since it can be easily parallelized when domain decomposition is used. For the operator compression step, we use a localized version of the empirical cubature method (ECM) by deriving an approximate quadrature formula independently for each subdomain. The ECM aims to replace the expensive high-dimensional quadrature schemes (that cannot be precomputed) used to numerically integrate various quantities over the mesh when evaluating the operator F by reduced schemes requiring much fewer quadrature points. The optimization problem to solve consists of minimizing the number of non-zero elements of the vector of quadrature weights in a set of admissible solutions. The latter is the set of vectors with positive components that approximate all the integrals computed when evaluating F on the training inputs, up to a given relative accuracy. This problem being NP-hard, it

is replaced by a tractable greedy approach called non-negative orthogonal matching pursuit algorithm. The positivity of the weights ensures that the reduced operators inherit the spectral properties of F, namely in our context, that the reduced tangent operator in Newton's algorithm in the exploitation stage is guaranteed to be symmetric positive definite. Details on our two physics of interest are given in [6, Section 4.3] and [8, Section 3.2]. In the former, we illustrate the ability of our physical ROM to accelerate the computation of the lifetime prediction of 3D elastoviscoplastic industrial high-pressure turbine blades from a few weeks to a few hours (see Figure 3) and in the latter, to accelerate the computation of transient heat analysis of high-pressure compressors with non-linear radiation boundary conditions and non-linear heat capacity and conductivity, with error compatible with industrial exploitation.

In structural mechanics applications, the quantities of interest are usually dual quantities, such as the plastic strain or the von Mises stress. Reconstruction of dual quantities over the complete mesh can be carried out using the Gappy-POD. Moreover, enforcing the reduced quadrature scheme to include well-chosen quadrature points can ensure the Gappy-POD reconstruction is well posed; see [6, Proposition 1]. In [5], we noticed in our numerical experiments that the prediction error on dual quantities is well correlated to the ROM-Gappy-POD residual, defined as the error on the reduced quadrature points between the online computation of the corresponding quantity by the constitutive law solver and the Gappy-POD reconstruction. Hence, we introduce an error indicator as a regressor trained to predict the error from the

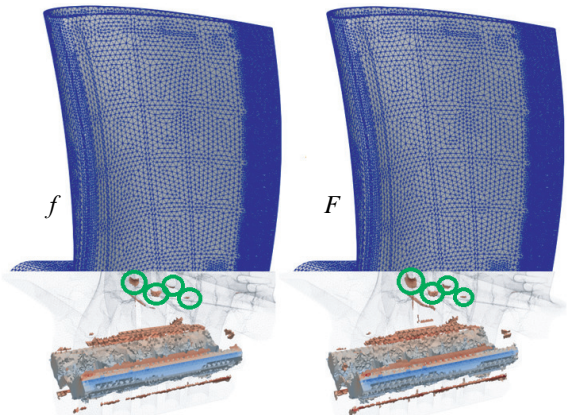


Figure 3. Comparison between the physical ROM f and the reference operator F: largest stress tensor eigenvalue in the areas where the von Mises criterion is large, at max rotation of the 100th cycle. Here, the ROM is trained using the outputs from 1 cyclic loading, and the following 99 cycles are used in the exploitation stage to compute a fast prediction. The four areas circled in green correspond to the locations where the internal cooling channels make a U-turn in the blade: the physical ROM was able to identify them correctly.

ROM-Gappy-POD residual. We illustrate in [5] that this error indicator succeeds in maintaining the error below a chosen threshold by triggering enrichment steps.

Linear dimensionality reduction, or data compression, enables the use of the Galerkin method in the exploitation stage, but this restriction may lead to situations where the cardinality of the reduced-order basis is too large to obtain meaningful speedup factors, even when hyperreduction is used. These situations of poorly reducible problems are the subject of the following section.

2.2. Piecewise linear data compression

The reducibility of the regression problem of approximating F by f using linear dimensionality reduction is evaluated by inspecting the rate of decay of the Kolmogorov N-width of the set of output solutions with respect to the dimension N of an optimal approximation vector space. The Kolmogorov N-width is the worst approximation error of the considered set of outputs, made by the best N-dimensional subspace. When the rate of decay is very fast, the approaches of the previous section, involving linear data compression, can be successful. Otherwise, the reduced-order basis has too many elements, and the reduced problems become slow to solve: the problem is poorly reducible. This is the context of the Ph.D. work by Thomas Daniel [11].

Various techniques have been proposed in the literature to address this situation; see the references in [11]. Here, we focus on the construction of a dictionary of local physical ROMs, using piecewise linear data compression on the outputs; see Figure 4: the outputs are grouped in clusters. Consider again the problem of lifetime prediction of 3D elastoviscoplastic industrial high-pressure turbine blades, and now suppose that the temperature loading field is an input. To be able to use a dictionary of local ROMs, we must know in advance in which cluster of the outputs the solution is located in

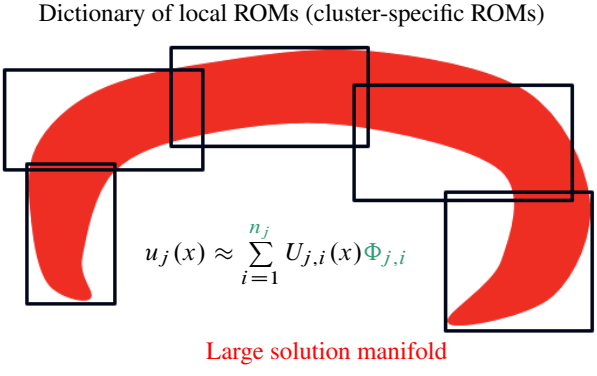


Figure 4. Piecewise linear data compression.

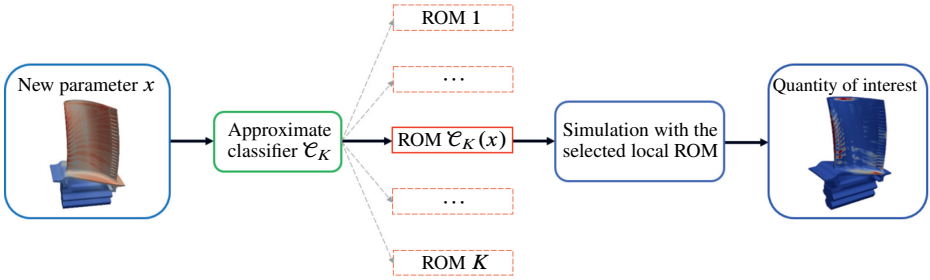


Figure 5. Local ROM recommendation.

order to use the adequate local ROM. To gain this insight, we train a classifier to recommend the right local ROM from the 3D input temperature field, in the same fashion as classifiers in computer graphics that can recognize an object type on a picture; see Figure 5. Such workflows have been carried out in [13]. Contrary to image classifiers, we are in a situation where we have very few inputs, which are of a large dimension. To improve the performance of the training of our classifier, we developed dedicated feature selection and data augmentation techniques in [14].

Partitioning of the outputs has been investigated in the literature using the L^2 norm in the ambient solution space. We notice that the L^2 norm does not necessarily give low-dimensional approximation subspaces; see Figure 6. It appears preferable to group the outputs with respect to their relative orientation instead of their magnitudes. In [12, Property 4.13], we show that the partitions of the outputs minimizing the k-medoid cost function with the sine dissimilarity are exactly the minimizers of the sum of a local (to the partition element) variant of the Kolmogorov 1-width weighted by the volume of the partition element. This result does involve the relative angles between outputs and has an important practical consequence: it provides an algorithm to partition the outputs in an optimal way for our task at hand.

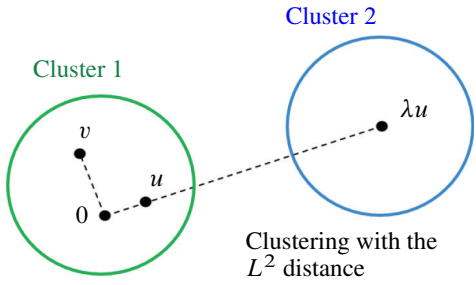


Figure 6. For two clusters, using the L^2 norm leads to cluster 1 having two independent directions, as many as a global ROM. However, considering a cluster with v and another one with u and λu leads to two local reduced-order bases of cardinal 1.



Figure 7. Some output fields for the 2D advection problem [12].

Consider a 2D advection problem, where Gaussians of two amplitudes (small: 0.1 and large: 1) and various vertical positions ξ_2^0 are translated from the left to the right; see Figure 7. For each value ξ_2^0, both small and large magnitudes are considered as inputs.

Figure 8 illustrates multidimensional scaling representations of the clustering of the 2D advection problem outputs with 5 clusters using the L^2 norm and the sine dissimilarity. We notice that with the L^2 norm, all the outputs corresponding to the small amplitude 0.1 are tightly packed at the center of the figure, meaning that they are likely to be affected by the same cluster (cluster 5 in the figure). By construction, the set of all small magnitude outputs contains all the independent directions of the outputs; hence the Kolmogorov N-width of this cluster is not smaller than that of the set of all outputs. On the contrary, for the sine dissimilarity on the right-hand side of Figure 8, three connected components appear, corresponding to the three values

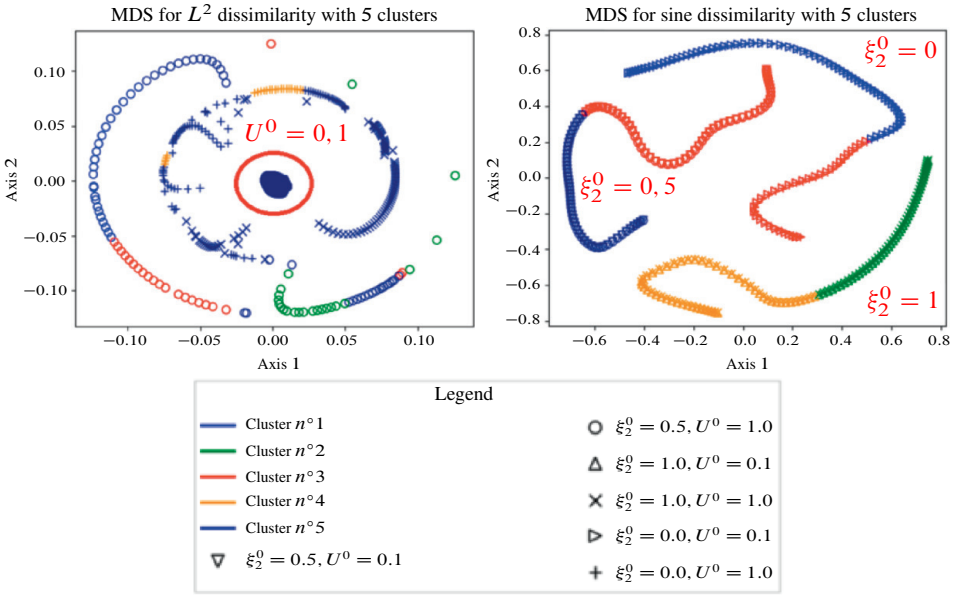


Figure 8. Multidimensional scaling representations of the clustering of the 2D advection problem outputs with 5 clusters using the L^2 norm (left) and the sine dissimilarity (right) [12].

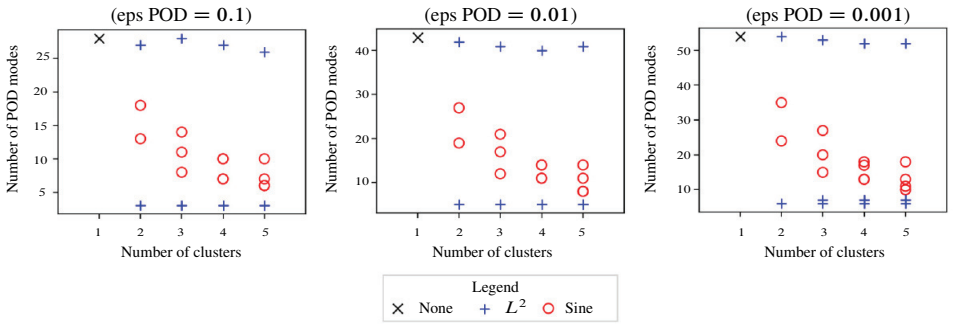


Figure 9. Number of POD modes for each local basis with respect to the number of clusters for various accuracy criteria of the POD truncation applied to the 2D advection problem [12].

of ξ_2^0. At each position are located two outputs, corresponding to the same values of ξ_2^0 and ξ_1 (horizontal component), for both small and large magnitudes: it means that two outputs varying only by their magnitude have zero sine dissimilarity.

In Figure 9, we investigate the ability of the clusterings using the L^2 norm and the sine dissimilarity to construct local ROMs of small dimensions. Using the L^2 norm, we notice that one of the clusters has the same dimension as a global ROM: it is the one containing all the small magnitude outputs. This means that regardless of the chosen number of clusters, one of them is as poorly reducible as the set of all the outputs. On the contrary, using the sine dissimilarity, the dimension of all the local ROMs decreases as the number of clusters increases: it becomes profitable to cluster the outputs since we succeed in reducing the dimension of the ROMs. This illustrates that the choice of the dissimilarity used in the clustering of the outputs is very important and that the sine one is effective in this situation.

In [15], a dictionary of local ROMs has been constructed using the techniques developed in this section to quantify the uncertainty of the plastic strain and the von Mises stress in areas of interest with respect to the input temperature loading of a high-pressure turbine blade. First, we choose some parameterization of the variability of the input temperature field and construct two DoEs, one MaxProj of size 80 and one Sobol' of size 120. Then, we evaluate the operator F for these 200 configurations (7 days and 9 hours on 48 computer cores). We cluster the outputs on the MaxProj DoE using the sine dissimilarity with the PAM k-medoid clustering algorithm to obtain a partition of size 2 of the outputs (5 minutes). We construct the 2 corresponding local ROMs using the POD-ECM technique presented in Section 2.1 (5 hours on 24 computer cores). We label the 120 outputs of the Sobol' DoE by computing the sine dissimilarity between each output and each of the two previously identified medoids (5 minutes). We apply our feature selection algorithm (16 minutes on 280 computer cores) and train our classifier: a logistic regression with elastic net

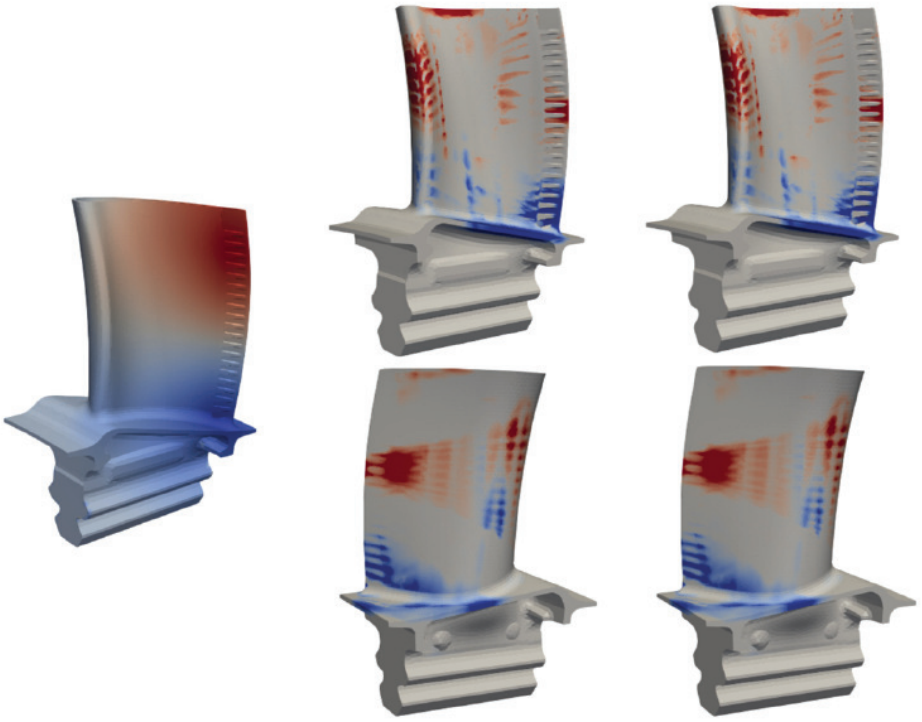


Figure 10. Comparison between high-fidelity predictions (middle column) and the dictionary of ROMs' predictions (right-hand column). The field on the left represents the difference between the current temperature field (belonging to cluster 1) and the reference one. The other fields correspond to the increments of the plastic strain with respect to its reference state obtained with the reference temperature field [15].

regularization (1 minute). At this stage, we have constructed a ROM dictionary of 2 local ROMs and trained the classifier to be able to recommend the local ROM to use in the exploitation stage, from the input temperature loading (see Figure 5); our surrogate f is available. Finally, probability density estimates of the plastic strain and the von Mises stress in areas of interest of the high-pressure turbine blade are computed from 1008 Monte Carlo evaluations of f, with relative errors of 1–2% and a speedup of 600; see Figure 10.

2.3. Limitations of physical reduced-order modeling

Although physical ROMs have been successfully derived for many equations and configurations, two important limitations remain. First, non-parameterized geometrical variations: when two outputs are supported on different meshes, comparing them (to estimate correlations) is not direct. Second, global non-linear data compression,

which would enable the derivation of a single model for poorly reducible problems, prevents us from using the Galerkin method in the exploitation stage.

Various authors have succeeded in tackling each of these limitations, while still constructing a surrogate f that relies on solving the physics equations. However, this comes with the price of complex algorithms, in particular for maintaining practical speedups: to cite a few, geometrical variations [2, 19] and non-linear data compression [1, 20].

In the following section, we propose a surrogate f that relies on non-linear data compression and deals with non-parameterized geometrical variations, but no longer relies on solving the physics equations in the exploitation stage.

3. Machine learning for learning physical problems under non-parameterized geometrical variability

3.1. Mesh morphing Gaussian process

In our recent work presented in [10], we introduce a technique named mesh morphing Gaussian process (MMGP). MMGP integrates four key components: (i) mesh morphing, (ii) finite element interpolation, (iii) dimensionality reduction and (iv) Gaussian process regression. This combination is used to learn solutions to PDEs that involve geometric variations not parameterized explicitly.

Figure 11 demonstrates the inference process for predicting a field of interest using this method. The figure should be interpreted from left to right: on the left side, various geometries corresponding to different samples are depicted, while on the right side, the target fields to be predicted on these geometries are shown.

When dealing with non-parameterized meshes, it is necessary to perform a shape embedding process, converting the meshes into learnable objects. In the MMGP

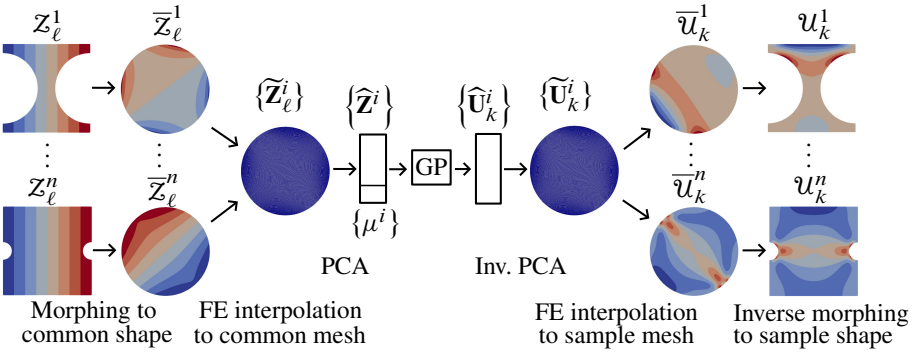


Figure 11. Illustration of the MMGP inference workflow for the prediction of an output field of interest [10].

approach, we utilize the coordinates of the mesh vertices as continuous fields. The left column of Figure 11 illustrates the continuous field of the x-component of the coordinates, naturally showing vertical iso-values. Next, we employ a deterministic morphing process to transform the input meshes into meshes of a chosen common shape, the unit disk in the case of Figure 11. At this stage, depicted in the second column of the figure, each sample is converted into a distinct mesh of the unit disk. This transformation uniquely modifies the coordinate fields for each sample based on the input mesh shapes. We then select a mesh of the common shape and project the coordinate fields from each morphed mesh onto this common mesh using finite element interpolation. Consequently, the coordinate fields of all samples are now represented on the same mesh, thus having a uniform size. This allows us to apply classical dimensionality reduction techniques, such as principal component analysis (PCA), to generate low-dimensional vectors that represent our meshes. If the learning problem includes scalar inputs, we append them at the end of these vectors. It is important to note that our shape embedding relies on the morphing of the coordinate fields, which are transformed according to the input shapes.

Similarly, for the output fields, a field embedding process is required to convert the variable-sized output fields into learnable objects. We apply the same transformations used for the coordinate fields: morphing onto a common shape, finite element interpolation onto a common mesh and PCA compression. This results in low-dimensional vectors that represent our fields of interest.

By applying these deterministic pre-treatments, we reduce the physics learning problem to a low-dimensional input-output tabular regression problem. The complete pre-treatments, including the morphing, should be regarded as the dimensionality reduction step, which is then possibly highly non-linear. The deterministic nature of these pre-treatments is crucial as it simplifies the machine learning stage, removing the complexity of embedding large and variable-dimensional objects. We rely on Gaussian process regression due to its accuracy and its capability to estimate predictive uncertainties effectively.

We refer to [10] for numerical experiments, where MMGP has been found to provide more accurate solutions than recent deep learning technologies (MeshGraphNet and U-net graph convolutional neural network), while being very cheap to train.

3.2. ML4PhySim competition

We have taken part in a competition [17] organized by IRT SystemX and other partners, using our MMGP technology as our entry. This competition consists of learning to predict the pressure, velocity and turbulent viscosity fields on the AirfRANS dataset [3], presented in Section 1.

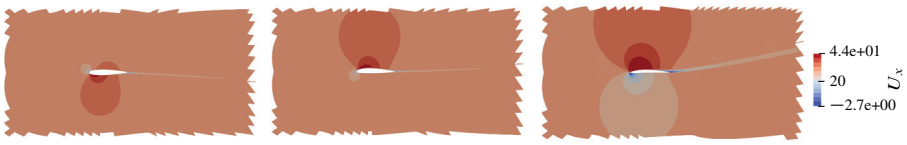


Figure 12. Illustration of the horizontal velocity for three samples of the AirfRANS dataset: the external boundaries are very irregular.

MMGP, being in its early development stages, still needs some dedicated treatment, especially in the morphing step. To derive an effective morphing, we first extend all the 2D meshes up to a common bounding box, since the external boundaries are very irregular; see Figure 12.

For the common shape, we select the geometry of the first pre-treated mesh from the training set. We employ a radial basis function (RBF) morphing algorithm, which involves transforming specific control points and using them to compute the transformation for all other points. The control points chosen for this process include the intrado and extrado of the airfoil, the points at the center of the wake, and the points on the external boundary. Figure 13 illustrates the morphing process applied to the second mesh in the training set. All fields, including input coordinates and output fields, are now projected onto the common mesh, which is the pre-treated and morphed mesh of the first sample of the training set. For dimensionality reduction, we utilize the snapshot proper orthogonal decomposition (POD), as detailed in [10, Annex C]. Objects reconstructed from a reduced representation obtained through snapshot POD are expressed as linear combinations of the compressed samples, preserving all zero linear relationships in the reduced objects. In the AirfRANS dataset, the velocity field satisfies a zero boundary condition at the airfoil surface. Since the competition scoring function involves scalar quantities derived from integrating output fields at the

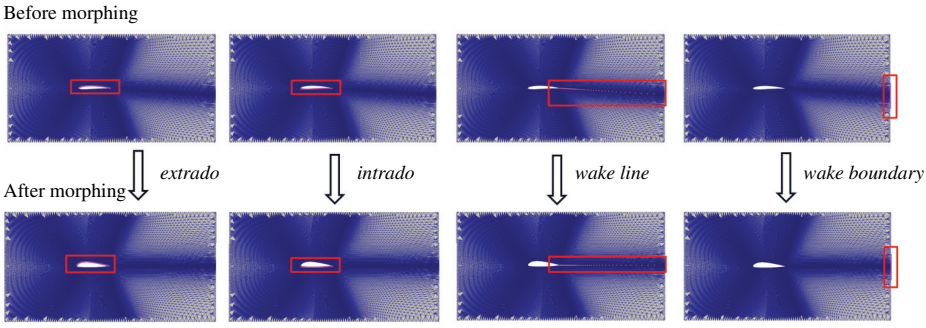


Figure 13. Morphing of the second mesh of the training set.

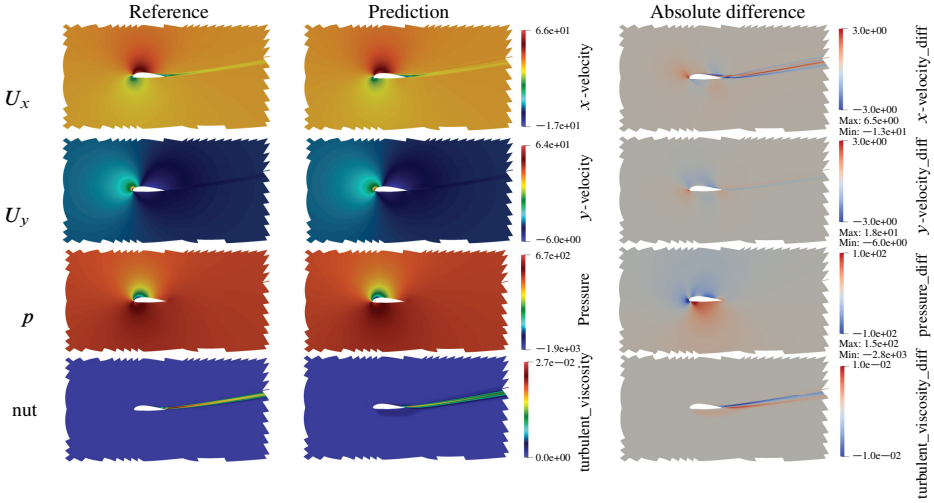


Figure 14. Illustration of the result on a test sample.

airfoil surface, maintaining this zero boundary condition is crucial. By using snapshot POD with MMGP, all predicted fields are expressed as linear combinations of velocity fields, which helps ensure that the zero boundary condition is respected.

We have completed the preprocessing of all meshes, including morphing them, projecting the fields onto a common mesh, and reducing the dimensionality of both the input coordinate fields and output fields. For the competition, we trained separate Gaussian processes for each generalized coordinate associated with the snapshot POD of each output field of interest. The Gaussian process kernel incorporates a constant term combined with an RBF kernel and a white noise term.

Some results are illustrated in Figure 14.

The score is a scalar between 0 and 100 that takes into account the speedup of the surrogate model and the accuracy of the output fields, as well as scalars obtained by integrating these fields (lift and drag coefficients, and their Spearman's rank correlation coefficients), on two different test sets, including one being out of distribution. In Figure 15, the final leaderboard of the competition is presented. Our method ranked first and won the competition, despite involving very simple and classical tools, while our competitors used various state-of-the-art deep learning methods.

3.3. Remarks and outlook

MMGP offers a method for approximating both field and scalar outputs of interest related to complex physics problems involving non-parameterized geometrical variability. The approach introduces two key innovations: firstly, the combination of mesh morphing with finite element interpolation, and secondly, a shape embedding process that uses dimensional reduction of coordinates, treated as continuous fields over the

		Criteria category																			
		ML related (40%)						Physics (30%)				OOD generalization (30%)									
	Method	Accuracy (75%)					Speedup (25%)	Physical criteria				OOD accuracy (42%)					OOD physics (33%)				Speedup (25%)	Global score (%)
Rank		$\overline{u}_x$	$\overline{u}_y$	$\overline{p}$	$\overline{\nu}_t$	$\overline{p}_s$		C_D	C_L	ρ_D	ρ_L	$\overline{u}_x$	$\overline{u}_y$	$\overline{p}$	$\overline{\nu}_t$	$\overline{p}_s$	C_D	C_L	ρ_D	ρ_L		
1	MMGP	🟢	🟢	🟢	🟢	🟢	27.40	🟢	🟢	🟠	🟢	🟢	🟢	🟠	🟢	🟠	🟢	🟢	🟠	🟢	28.08	81.29
2	GNN-FC	🟢	🟢	🔴	🟢	🟠	570.77	🟢	🟠	🟠	🟢	🟠	🟢	🔴	🟠	🔴	🟢	🟠	🟠	🟠	572.3	66.81
3	MINR	🟢	🟢	🟠	🟢	🟠	518.58	🟠	🟠	🔴	🟠	🟢	🟢	🔴	🟢	🔴	🟠	🟠	🔴	🟠	519.21	58.37
4	Bi-Trans	🟢	🟢	🔴	🟢	🔴	552.97	🟠	🟠	🔴	🟢	🟠	🟠	🔴	🟠	🔴	🟠	🔴	🔴	🟠	556.46	51.24
5	NeurEco	🟢	🟢	🟠	🟢	🔴	44.93	🟠	🟠	🔴	🟠	🟢	🟢	🔴	🟢	🔴	🟠	🟠	🔴	🟠	44.78	50.72

Figure 15. ML4PhySim competition final leaderboard [17].

geometric support. This approach eliminates the need for the machine learning task to manage variable-sized samples or to learn implicit shape embeddings. By reducing the dimensions of inputs and outputs, Gaussian process regression can be efficiently applied to smaller objects.

MMGP has several significant advantages. It is capable of handling very large meshes with ease, can be trained on a CPU, is interpretable, demonstrates high accuracy in our numerical experiments and the ML4PhySim competition, and provides readily available predictive uncertainties. We believe that in the context of industrial part design, the configurations typically have relatively low intrinsic dimensions. Therefore, small machine learning models, if properly reparameterized in a deterministic manner (using morphing in the case of MMGP), can be highly efficient. However, MMGP does have some limitations. The morphing process must be customized for each new use case; there are assumptions regarding fixed topology, and the speedup is limited due to the need for computing morphing and finite element interpolation during the inference stage.

As future work, we plan to work on the previously mentioned limitations. In [18], we recently proposed a generic method for computing the morphing of the shape embedding that relies on successive solutions of elasticity problems. Moreover, an efficient procedure allows us to quickly approximate the morphing in the exploitation stage. Next, we plan to derive an optimal notion of morphing, in the sense that it aims to minimize the number of PCA modes, leading to a regression problem in the smallest possible dimension.

4. Open-source and open-data: Playing a role in the community

Since the inauguration of the Safran corporate research center in 2015, we have introduced internal processes to publish our research codes as open-source and the non-confidential datasets that we produced as open-data.

We promote accessibility, quality and robustness of our libraries by resorting to DevOps tools and workflows: collaborative development is enforced using git and code reviews, automatic testing and internal deployment pipelines and extensive documentation. We choose permissive licenses for our codes (MIT or BSD 3-Clause when possible) and our datasets (CC-BY-SA), and our open-source codes are published on GitLab, documentation is deployed on Read the Docs and packages are distributed on conda-forge, as much as possible.

4.1. Open-source codes

Recently, we developed a datamodel called PLAID (Physics-Learning AI Data model) and published an implementation of this datamodel in the form of a library

of the same name. We aim to formalize physics learning problems, starting with the datasets, which involve complex and heterogeneous objects, as introduced in Section 1. If practitioners share the same standard for producing their datasets and training their models, it would favor productive sharing of data and results and streamlined workflows. The library PLAID heavily relies on the CGNS [16] standard and comes with many functions to handle the data.

The codes developed for the methods presented in the previous sections have been published as open-source; see Figure 16. First, we mention Muscat [4, 30], a library including a set of powerful tools to work on meshes in the context of finite element computations, in particular for pre- and post-treatments. Many readers and writers are available, enabling the handling of data computed by various commercial and open-source codes. We aim to promote PLAID [31] at the base of our open-source developments as our standard for physics learning problems. The physical ROM technologies detailed in Section 2 rely on two libraries: mordicus [21], which we co-developed within the MOR_DICUS consortium, and genericROM [26], fully developed by Safran. The implementation of genericROM is non-intrusive, parallel with distributed memory, and can deal with non-parameterized variability. We refer to [25] for details on these notions and a description of mordicus and genericROM. The code for the MMGP method, presented in Section 3, is also available as open-source [28]. Finally, we mentioned tabular learning problems in Section 1, for which powerful and classical methods exist. Our team at Safran has developed an open-source platform called Lagun [27], where workflows involving design of experiments, surrogate modeling, sensitivity analysis, uncertainty quantification, optimization and calibration can be carried out for these tabular problems.

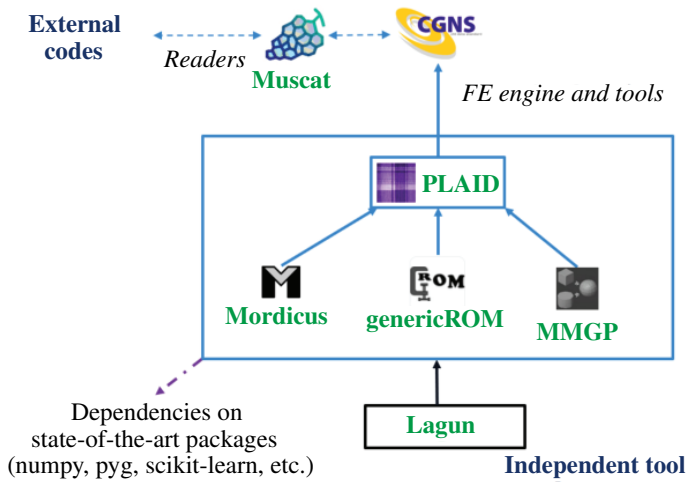


Figure 16. Open-source libraries involved in this article.

4.2. Open-data datasets

We are in the process of distributing the non-confidential datasets on which we develop our physics learning methods as open-data in PLAID format for the community to use. We use the sharing services of https://zenodo.org and https://huggingface.co: the three datasets used in [10] to compare MMGP with state-of-the-art deep learning methods are available on Zenodo [9, 23, 24] and on Hugging Face.

Hugging Face provides hosting services for machine learning applications called Spaces. We have distributed a demo [29] of our method MMGP applied to the physics learning problem in the ML4PhySim competition presented in Section 3.2; see Figure 17. In this demo, the user chooses a sample ID, which triggers the computation of the MMGP surrogate f, which involves the morphing, the finite element interpolation, the projection of the coordinate fields onto the snapshot POD modes, the inference of the trained Gaussian process regressor and the inverse snapshot POD and finite element interpolation. The user can then choose an output field to plot, which shows the prediction, reference and the error. We mention that this demo

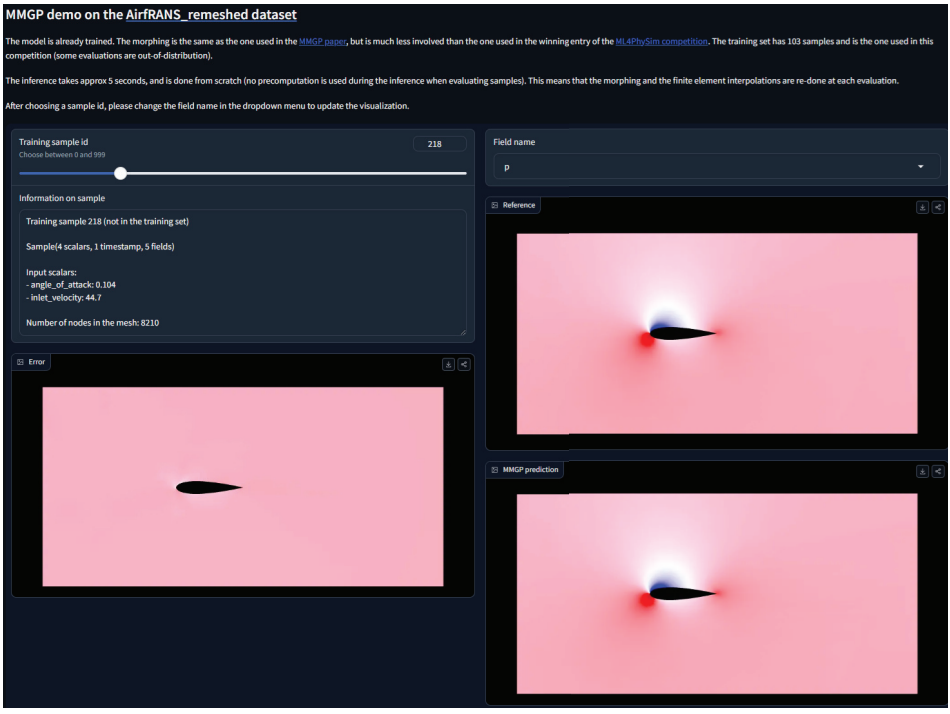


Figure 17. Hugging Face Space hosting a demo of our method MMGP applied to the physics learning problem in the ML4PhySim competition.

involves a remeshed dataset, where the meshes have been coarsened from approximately 180, 000 nodes down to approximately 8000 nodes, for the demo to be in real time.

References

[1] N. Akkari, F. Casenave, E. Hachem, and D. Ryckelynck, A Bayesian nonlinear reduced order modeling using variational autoencoders. *Fluids* **7** (2022), no. 10, article no. 334, 24 pp.

[2] N. Akkari, F. Casenave, D. Ryckelynck, and C. Rey, An updated Gappy-POD to capture non-parameterized geometrical variation in fluid dynamics problems. *Adv. Model. Simul. Eng. Sci.* **9** (2022), no. 1, article no. 3, 34 pp.

[3] F. Bonnet, J. A. Mazari, P. Cinnella, and P. Gallinari, AirfRANS: high fidelity computational fluid dynamics dataset for approximating Reynolds-Averaged Navier–Stokes solutions. In *Proceedings of the 36th conference on Neural Information Processing Systems (NeurIPS 2022) track on datasets and benchmarks, November 28–December 9, 2022, New Orleans, LA, USA*, pp. 23463–23478, Adv. Neural Inform. Process. Syst. 35, Neural Information Processing Systems Foundation (NeurIPS), San Diego, CA, USA, 2022

[4] F. Bordeu, F. Casenave, and J. Cortial, BasicTools: a numerical simulation toolbox. *J. Open Source Softw.* **8** (2023), no. 86, article no. 5142, 4 pp.

[5] F. Casenave and N. Akkari, An error indicator-based adaptive reduced order model for nonlinear structural mechanics—application to high-pressure turbine blades. *Math. Comput. Appl.* **24** (2019), no. 2, article no. 41, 28 pp. MR 3977955

[6] F. Casenave, N. Akkari, F. Bordeu, C. Rey, and D. Ryckelynck, A nonintrusive distributed reduced-order modeling framework for nonlinear structural mechanics—application to elastoviscoplastic computations. *Internat. J. Numer. Methods Engrg.* **121** (2020), no. 1, 32–53 Zbl 1548.74772 MR 4047864

[7] F. Casenave, A. Ern, and T. Lelièvre, Accurate and online-efficient evaluation of the *a posteriori* error bound in the reduced basis method. *ESAIM Math. Model. Numer. Anal.* **48** (2014), no. 1, 207–229 Zbl 1288.65157 MR 3177842

[8] F. Casenave, A. Gariah, C. Rey, and F. Feyel, A nonintrusive reduced order model for nonlinear transient thermal problems with nonparametrized variability. *Adv. Model. Simul. Eng. Sci.* **7** (2020), no. 1, article no. 22, 19 pp.

[9] F. Casenave, X. Roynard, and B. Staber, *Tensile2d: 2D quasistatic non-linear structural mechanics solutions, under geometrical variations* [Data set]. Zenodo, 2023, https://zenodo.org/records/14840177 visited on 8 May 2026

[10] F. Casenave, B. Staber, and X. Roynard, MMGP: a Mesh Morphing Gaussian Process-based machine learning method for regression of physical problems under non-parametrized geometrical variability. In *Proceedings of the 37th conference on Neural Information Processing Systems (NeurIPS 2023), December 10–16, 2023, New Orleans, LA, USA*, pp. 43972–43999, Adv. Neural Inform. Process. Syst. 36, Neural Information Processing Systems Foundation (NeurIPS), San Diego, CA, USA, 2023

[11] T. Daniel, *Machine learning for nonlinear model order reduction*. Ph.D. thesis, University of Paris Sciences and Letters, 2021

[12] T. Daniel, F. Casenave, N. Akkari, A. Ketata, and D. Ryckelynck, Physics-informed cluster analysis and a priori efficiency criterion for the construction of local reduced-order bases. *J. Comput. Phys.* **458** (2022), article no. 111120, 32 pp. Zbl 1560.65714 MR 4393883

[13] T. Daniel, F. Casenave, N. Akkari, and D. Ryckelynck, Model order reduction assisted by deep neural networks (ROM-net). *Adv. Model. Simul. Eng. Sci.* **7** (2020), no. 1, article no. 16, 27 pp.

[14] T. Daniel, F. Casenave, N. Akkari, and D. Ryckelynck, Data augmentation and feature selection for automatic model recommendation in computational physics. *Math. Comput. Appl.* **26** (2021), no. 1, article no. 17, 25 pp.

[15] T. Daniel, F. Casenave, N. Akkari, D. Ryckelynck, and C. Rey, Uncertainty quantification for industrial numerical simulation using dictionaries of reduced order models. *Mech. Ind.* **23** (2022), article no. 3, 20 pp.

[16] CFD General Notation System (CGNS), https://cgns.github.io visited on 8 May 2026

[17] IRT SystemX, Ansys, NVIDIA, Exaion, and RTE, ML4PhySim challenge. https://www.codabench.org/competitions/1534 visited on 8 May 2026

[18] A. Kabalan, F. Casenave, F. Bordeu, V. Ehrlacher, and A. Ern, Elasticity-based morphing technique and application to reduced-order modeling. *Appl. Math. Model.* **141** (2025), article no. 115929, 25 pp. Zbl 1573.65052 MR 4851782

[19] E. N. Karatzas, F. Ballarin, and G. Rozza, Projection-based reduced order models for a cut finite element method in parametrized domains. *Comput. Math. Appl.* **79** (2020), no. 3, 833–851 Zbl 1443.65348 MR 4049381

[20] K. Lee and K. T. Carlberg, Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *J. Comput. Phys.* **404** (2020), article no. 108973, 32 pp. Zbl 1454.65184 MR 4043884

[21] MOR_DICUS consortium, Mordicus. Code: https://gitlab.com/mor_dicus/mordicus; doc: https://mordicus.readthedocs.io visited on 8 May 2026

[22] A. T. Patera and G. Rozza, *Reduced basis approximation and a posteriori error estimation for parametrized partial differential equations*. MIT Pappalardo Grad. Monogr. Mech. Eng., Massachusetts Institute of Technology, Cambridge, MA, 2007

[23] X. Roynard, F. Casenave, and B. Staber, *Rotor37: a 3D CFD RANS dataset, under geometrical variations of a compressor blade* [Data set]. Zenodo, 2023, https://zenodo.org/records/14840190 visited on 8 May 2026

[24] X. Roynard, F. Casenave, and B. Staber, *AirfRANS_original* [Data set]. Zenodo, 2024, https://zenodo.org/records/14840387 visited on 8 May 2026

[25] D. Ryckelynck, F. Casenave, and N. Akkari, Resources: software and tutorials. In *Manifold learning. Model reduction in engineering*, pp. 53–69, Springer, Cham, 2024

[26] Safran, genericROM. Code: https://gitlab.com/drti/genericrom; doc: https://genericrom.readthedocs.io visited on 8 May 2026

[27] Safran, Lagun. Code: https://gitlab.com/drti/lagun; doc: https://lagun.readthedocs.io visited on 8 May 2026

[28] Safran, MMGP. Code: https://gitlab.com/drti/mmgp; doc: https://mmgp.readthedocs.io visited on 8 May 2026

[29] Safran, MMGP demo. https://huggingface.co/spaces/fabiencasenave/MMGP_demo visited on 8 May 2026

[30] Safran, Muscat. Code: https://gitlab.com/drti/muscat; doc: https://muscat.readthedocs.io visited on 8 May 2026

[31] Safran, PLAID. Code: https://github.com/PLAID-lib/plaid; doc: https://plaid-lib.readthedocs.io visited on 8 May 2026

Fabien Casenave
Digital Sciences and Technologies Department, Safran Tech, Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France; fabien.casenave@safrangroup.com

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/42

On the 125th anniversary of his birth: My "encounters" with Otto Neugebauer

Reinhard Siegmund-Schultze

Abstract. Otto Neugebauer was arguably the most influential historian of mathematics in the 20th century. His main merit was the interpretation of the mathematical content in Old Babylonian texts (around 2000–1600 BCE) and in Neo-Babylonian astronomical texts (around 600–0 BCE) which have come down to us as cuneiform clay tablets. The lecture focused on Neugebauer's other activities for mathematics, highlighting their astonishing breadth and effectiveness. These include his role as "architect" of the new building for the Mathematical Institute in Göttingen in the 1920s, his work as a founder and director of two leading mathematical abstracting journals in the 1930s and 1940s, his experiences as a non-Jewish refugee from Nazi Germany, and his appearance as a keen and often sarcastic observer of political events. The paper has some overlap with my article (2016) but adds new literature and a discussion of the transition of Neugebauer's research interests to mathematical astronomy, as revealed in his talk at the ICM in Oslo 1936. In the final section, some work on Babylonian astronomy is mentioned which continues and draws on Neugebauer's research.

Dedicated to my partner and fellow historian June Barrow-Green (London), who has improved the content and language of my manuscripts for nearly two decades and has thus played a significant role in my winning the Otto Neugebauer Prize.

Introduction

This paper is a reworking of my Neugebauer prize lecture from July 2024 in Sevilla. I never met Neugebauer (Figure 1), not least because of travel restrictions which affected me as an East German before 1990, the year when Neugebauer died in American exile. However, I have worked as a historian on important contributions

Mathematics Subject Classification 2020: 01A17 (primary); 01A70, 01A74 (secondary).
Keywords: Babylonian mathematics and astronomy, Otto Neugebauer, *Zentralblatt*, emigration from Nazi Germany.

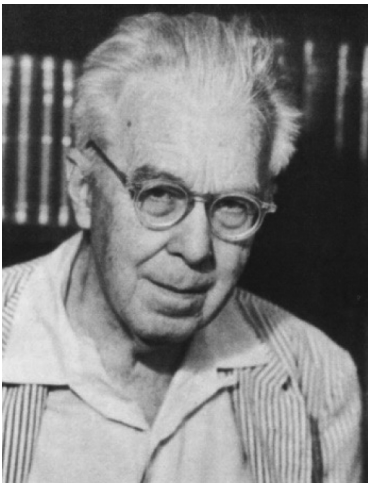


Figure 1. Otto Neugebauer (1899–1990) from ca. 1975 [36, pp. 139–140]. Courtesy of *Centaurus*; used with permission.

of Neugebauer's to the community of mathematicians, in particular his organizational work, above all in mathematical reviewing. I doubt that a personal meeting with Neugebauer would have been very useful for shedding light on these questions. In his summarizing book on the "History of Ancient Mathematical Astronomy" (HAMA), the 76-year-old Neugebauer wrote categorically, apparently reacting to methodological criticism of his work by a young generation of historians:

> *I have tried to come as close as possible to the astronomical problems themselves without hiding my ignorantia behind the smoke-screen of sociological, biographical and bibliographical irrelevancies.* [23, p. vii]

My historical "encounters" with Neugebauer, however, show that Neugebauer on various occasions had to take a political and philosophical stance. Moreover, even in his work on Babylonian mathematics and astronomy, Neugebauer revealed occasionally political and philosophical positions, as will be shown below. I am not a specialist in ancient Babylonian mathematics which would require, among other things, knowledge of the old Sumerian and Akkadian languages.[1] I will therefore restrict my remarks about Neugebauer's most important accomplishments between 1935 and 1975 (MKT, UAA, MCT, ACT, and HAMA; see references below) to some more general and cursory ones (in Sections 2, 6–8, and 11).

For more details on Neugebauer's life and work, the reader is referred to the biography by Neugebauer's student, Noel M. Swerdlow, and to the collection edited by

[1] For more details, see [6, 9, 23, 28].

Alexander Jones et al. in 2016.[2] Neugebauer's archive is kept at the Shelby White and Leon Levy Archives Center of the Institute for Advanced Study (IAS) in Princeton. It contains manuscripts but almost no correspondence.[3] An important source is the Nachlass of Neugebauer's friend, Erich Bessel-Hagen (1898–1946), which is located at the University Library in Bonn [11]. Bessel-Hagen's archive contains 160 letters between him and Neugebauer. It was used for the first time in [33].

1. Start as a mathematician

Neugebauer was born in Innsbruck (Austria) in 1899. He studied physics and mathematics in Graz, München, and Göttingen. When he arrived in Göttingen in 1922, physics and its geometrical aspects (relativity theory, B. Riemann, H. Weyl) were among his foremost interests, not history. However, in this context, he had already shown some interest in the history of non-Euclidean geometry and in Hilbert's axiomatics, as his early manuscripts and lecture notes in Princeton show [29]. Given existing biographical accounts, it seems unlikely that he engaged in a systematic study of ancient cultures before he studied physics and mathematics.

At first, the young Neugebauer was inspired by the enthusiasm and the brilliance of the mathematicians in Göttingen, including David Hilbert, Emmy Noether, and Edmund Landau. The mathematician Richard Courant (1888–1972) in Göttingen became his mentor and life-long friend. Courant was the successor to the great Göttingen scholar Felix Klein (1849–1925), who had always advocated the equal cultivation of pure and applied mathematics. After Klein's death in 1925, Courant involved his protégé in the publication of Klein's "Lectures on the development of mathematics in the 19th century" [10]. One reads in the preface by the editors Courant and Neugebauer from 1926:

> *At a time when also within science the practitioners are much too impressed by the present and tend to look at this as in unnatural enlargement and exaggerated importance, Klein's work enables the opening of the eyes of many for the connections and lines of development of our science in its totality.* [10, p. v]

Occasionally, Neugebauer shared some of the hubris of the Göttingen mathematicians, among whom it was fashionable to belittle those considered dilettantes. But at the same time, he strongly felt the inadequacy of his own mathematical talent, particularly at the beginning of his career, when he was still looking for a topic to work

[2] See [9, 26, 29, 33, 35, 36].

[3] Neugebauer destroyed most of his private correspondence [26, p. 274].

on. On 21 April 1924, Neugebauer wrote to Bessel-Hagen, who at the time was also a student in Göttingen:

> *The constant and constantly growing doubt about my mathematical talent tends to paralyze my energy.* [33, p. 67]

Around this time, Neugebauer felt the need to change his environment, so he went to Harald Bohr (1887–1951) in Copenhagen. Bohr, the brother of the famous physicist Niels Bohr, happened to be a good friend of Richard Courant. He was a specialist in almost periodic functions, and Neugebauer became his co-author in the paper [2] which would remain Neugebauer's only mathematical publication.

2. Start as a historian of mathematics: Egyptian fractions (1926) and beginnings of research on Babylonian mathematics (1927–1929)

When Neugebauer was in Copenhagen, Bohr invited him to write a review of Thomas Eric Peet's edition of the Rhind Mathematical Papyrus, a major source on ancient Egyptian mathematics [29]. This review, written in German, appeared in the Danish journal *Matematisk Tidsskrift* and was Neugebauer's first published article [12]. Neugebauer's review sparked the work that was to form the basis of his doctoral dissertation in Göttingen on Egyptian fractions [13].[4] Had Neugebauer been a student of Hilbert or Landau, his chances of completing his studies with a dissertation on Egyptian fractions would have been very scant. Luckily, however, he had Courant on his side, someone who valued human character above formal requirements [29, p. 30]. As late as 1963, in his address on Courant's 75th birthday [22], one senses the relief Neugebauer must have experienced when he finally found a field of study which was far removed from the work of the leading and awe-inspiring Göttingen mathematicians, but still tolerated and acknowledged by them, his research in the history of ancient mathematics and astronomy.

Very shortly after his PhD, Neugebauer became interested in Babylonian texts, of which there were far more extant than the few Egyptian sources. As in the case of the Egyptian texts, careful philological work had been done on Babylonian texts before 1900, though not always by scholars with sufficient mathematical knowledge. As early as 1927, Neugebauer obtained his habilitation (teaching permit) in Göttingen with a treatise on the origins of the Babylonian sexagesimal system [14].[5] Høyrup reports on the immense surprise among historians and mathematicians when, 2 years

[4] In [29, p. 32], details of Neugebauer's doctoral examination are discussed.
[5] Details in [29].

later, Neugebauer found in an Old Babylonian text the solution of second-degree equations for the partition of trapezoids [6, 15].

Neugebauer occasionally used his teaching license to give historical lectures, but he had to earn his living by giving lectures on descriptive geometry, etc. But even more than in historical research or mathematical lectures, Neugebauer was involved in improving the infrastructure of mathematics.

3. Rockefeller and the Mathematical Institute in Göttingen

After the Great War, Courant was heavily involved in several activities aimed at "modernizing" the infrastructure of mathematics in Göttingen and throughout Germany, both in cooperation with Ferdinand Springer (focusing on publications) and together with the American Rockefeller philanthropy. The latter activities culminated in the erection of a new mathematics building in Göttingen in 1929 [30]. As Courant's biographer Constance Reid remarked:

> *For this work Courant turned to Neugebauer, whose artistic gifts and knowledge of engineering, his draughtsmanship from his studies in Graz, and his awareness of the needs of mathematicians made him the perfect choice. Courant and Neugebauer were eleven years apart in age and completely different characters.* [27, p. 108]

"Courant's word was maybe," as the mathematician Hans Lewy told Reid, "Neugebauer's was yes, yes, no, no." But the two were fast becoming an administrative team, with the purposeful disorganization of Courant balanced by the efficiency of Neugebauer.

Working behind the scenes, Neugebauer helped draw up new blueprints for a T-shaped building designed with an eye toward functionality and comfort. This design called for two large lecture halls, special rooms for drawing and applied mathematics, a colloquium room for a more intimate atmosphere, and a spacious *Lesezimmer* (reading room) that would house the institute's large and rapidly growing library [16]. Construction went on for 2.5 years before the new quarters could open in November 1929. In 1930, one Rockefeller officer supervising the Göttingen project wrote back to his superiors in the USA:

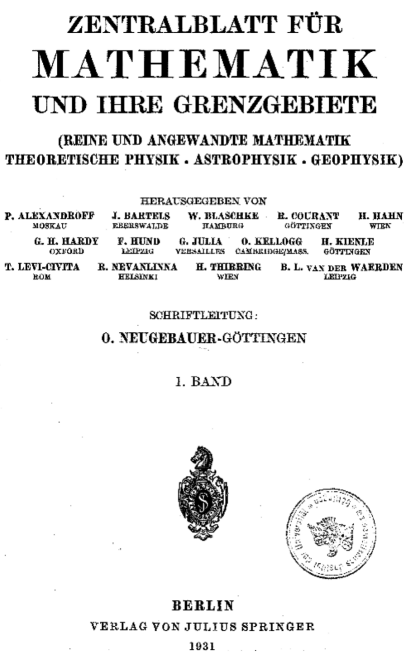
> *Neugebauer is a very able fellow, has had real charge of the details of the building of the new institute here, forceful, pleasing to meet, efficient, and probably worthwhile if his program is at all good.* [33, p. 73]

In fact, Neugebauer did not get Rockefeller support for his historical work before 1933, in particular no travel stipend. Only much later as a refugee in Copenhagen was he supported by Rockefeller and other foundations, and then mostly for his editorial work in mathematical reviewing.

4. Mathematical reviewing: *Zentralblatt* (1931) and its conflict with the old *Jahrbuch*

Another dimension of Neugebauer's work for the wider community of mathematicians was his activities in mathematical publishing. This happened in collaboration with the Springer publishers in Berlin, who had taken much responsibility for German mathematics after the Great War. Springer, with whom Courant was closely connected through personal friendship, had published Neugebauer's PhD thesis on Egyptian fractions as a book in 1926 [13]. Springer also paved the way for an outlet which would publish many historical papers of Neugebauer's, in particular MKT and UAA (see references below), in the years to come: the short-lived *Quellen und Studien zur Geschichte der Mathematik, Astronomie und Physik* (Sources and Studies on the History of Mathematics, Astronomy and Physics). The journal was launched in 1929 and ended about a decade later. It was edited by Neugebauer, Julius Stenzel, and Otto Toeplitz [29, p. 26] and assisted by Neugebauer's friend Bessel-Hagen.

But Neugebauer's cooperation with Springer went far beyond his own historical interests. Neugebauer's and Courant's foundation of the abstracting journal *Zentralblatt für Mathematik und ihre Grenzgebiete* (Central Journal for Mathematics and its Bordering Subjects) in 1931 took place against the backdrop of institutional and conceptual conflicts within the German mathematical community, in particular between Göttingen and Berlin. These conflicts were now perpetuated by the competition between the new *Zentralblatt* and the older *Jahrbuch über die Fortschritte der Mathematik* (Yearbook on the Progress of Mathematics) from the 19th century. Problems of the modernity of mathematical production, of the relation between the national and the international, and of defining and reviewing mathematical progress were invoked here. The *Jahrbuch* was strongly supported by the future extreme Nazi, the prominent function theorist Ludwig Bieberbach in Berlin. It was published in the name of the Prussian Academy of Sciences and put much emphasis on its systematic, year-wise reviewing, on the historical perspective on mathematics, on the thoroughness which lay in its slowness, and on the mental discipline which allegedly came with insisting on German as the principal language of reviewing. As managing editor (Schriftleiter) of the young *Zentralblatt*, Neugebauer had to – somewhat ironically – support the priority of the research interests of modern mathematicians over historical


ZENTRALBLATT FÜR
MATHEMATIK
UND IHRE GRENZGEBIETE
(REINE UND ANGEWANDTE MATHEMATIK
THEORETISCHE PHYSIK . ASTROPHYSIK . GEOPHYSIK)
HERAUSGEGEBEN VON
P. ALEXANDROFF MOSKAU · J. BARTELS EBERSWALDE · W. BLASCHKE HAMBURG · R. COURANT GÖTTINGEN · H. HAHN WIEN
G. H. HARDY OXFORD · F. HUND LEIPZIG · G. JULIA VERSAILLES · O. KELLOGG CAMBRIDGE/MASS. · H. KIENLE GÖTTINGEN
T. LEVI-CIVITA ROM · R. NEVANLINNA HELSINKI · H. THIRRING WIEN · B. L. VAN DER WAERDEN LEIPZIG
SCHRIFTLEITUNG:
O. NEUGEBAUER-GÖTTINGEN
1. BAND
BERLIN
VERLAG VON JULIUS SPRINGER
1931


ERGEBNISSE DER MATHEMATIK
UND IHRER GRENZGEBIETE
HERAUSGEGEBEN VON DER SCHRIFTLEITUNG
DES
„ZENTRALBLATT FÜR MATHEMATIK"
ZWEITER BAND
3
GRUNDBEGRIFFE DER
WAHRSCHEINLICHKEITS-
RECHNUNG
VON
A. KOLMOGOROFF
BERLIN
VERLAG VON JULIUS SPRINGER
1933

Figure 2. First title page of *Zentralblatt* (1931) and title page of A. N. Kolmogorov's *Ergebnisse* report (1933), also edited by Neugebauer.

reflection. As noted above, he and Courant, in their preface to Klein's historical lectures, had warned against the "unnatural enlargement and exaggerated importance" of the present.

Neugebauer's duties went beyond the *Zentralblatt* itself, because he also served as an editor of Springer's new *Ergebnisse* (results) series, where, for instance, A. N. Kolmogorov's pioneering "Foundations of the Theory of Probability" appeared in the original German in 1933 (Figure 2).

Similar to Neugebauer's obligation to prioritize mathematical over historical lectures in Göttingen, Neugebauer's work in mathematical publishing was something of a "bread and butter job" as he rather drastically described it later. In a job interview in February 1939 at Brown University, Neugebauer allegedly replied to the question whether he would be willing to work permanently as managing editor of the *American Mathematical Reviews*, then in the process of being founded:

> *Effectively, he said eight years on the* Zentralblatt *were enough for a man who preferred not to have a life sentence to abstract journal editing.* [33, p. 93]

5. The Nazis looming and Neugebauer's emigration to Copenhagen

Before the German Nazis came to power in 1933, Neugebauer saw the "writing on the wall" with the rise of extremism, particularly from the right. One finds a rare outburst of political anxiety on the part of Neugebauer at the end of one of his letters to Bessel-Hagen on the eve of the 1930 elections for the German Reichstag (Figure 3). The elections in September 1930 led to a sensational gain in votes for the Nazi Party (NSDAP), thus confirming Neugebauer's fears.

Philip Davis of Brown University, who knew him well, reports the following remark by Neugebauer from a conversation which testifies to his skepticism vis-à-vis biographical and sociological research on the history of mathematics:

> *If you never heard the sound of Nazi boots below you in the street you cannot understand the history of the period.* [4, p. 130]

When the Nazis came to power on 30 January 1933, they soon introduced laws to dismiss those they defined as "Jews" and those they found "politically unreliable," in particular Social Democrats and Communists. As a non-Jew and a politically not very active person, Neugebauer was not immediately threatened, but almost all of his friends such as Noether and Courant were thrown out [31]. He and the mathematician Kurt Friedrichs then mounted a campaign to gain signatures for a petition to the Prussian Ministry of Education aimed at reinstating Courant as director of the Göttingen Institute of Mathematics. After much effort, including many refusals, they came up with 28 distinguished "Aryan names," including Artin, Blaschke, Carathéodory, Hasse, Heisenberg, Herglotz, Hilbert, von Laue, Mie, Planck, Prandtl, Schrödinger, Sommerfeld, van der Waerden, and Weyl [27, pp. 151–152], [29, p. 41].



Figure 3. Political drawing from Neugebauer's letter of 20 August 1930 to Bessel-Hagen alluding to tactical collaboration between Nazis and Communists. The text reads "Hand in Hand fürs Vatterland!" thereby mocking the sanctified notion of "Vaterland" (fatherland) by misspelling. Courtesy of University Library Bonn, Nachlass E. Bessel-Hagen. See also [33, p. 75].

This initiative led nowhere under the circumstances of increasing political intimidation. After a visit to Copenhagen, Courant wrote on 12 September 1933 to Neugebauer, advising him in clandestine language (because mail was monitored) to leave Germany:

> *Generally, Harald [Bohr] tends to emphasize the transformation theory instead of fixed point theorems. That method is also more rewarding for your work.* [33, p. 78]

The insecure situation for *Zentralblatt* and the lack of funds for his historical work contributed to Neugebauer's increasing problems in Göttingen. Neugebauer's letters to foreigners asking them to protest against Nazi attacks on *Zentralblatt* made him even more suspicious in the eyes of the regime.

Finally, Neugebauer obtained a position in Copenhagen, financed by Danish foundations and by Rockefeller. He moved the editorial office of *Zentralblatt* to Copenhagen in 1934.

6. Neugebauer on race and racism in science and mathematics, in particular at the Oslo International Congress of Mathematicians (ICM) 1936

In Copenhagen, Neugebauer continued publishing on Old Babylonian mathematics,[6] and, somewhat later, on Neo-Babylonian mathematical astronomy.[7] At the same time, he continued – with the assistance of his wife – mathematical reviewing, involving an increasing number of non-German referees. As a rule, Ludwig Bieberbach's infamous racist publications were quoted only by their titles in *Zentralblatt* but otherwise ignored. No line of commentary or quote is, for instance, given in *Zentralblatt* for Bieberbach's embarrassing "Stilarten mathematischen Schaffens [Styles of Mathematical Creation]" (1934) [1]. The parallel review of the same paper in the *Jahrbuch* by the Berlin topologist Georg Feigl reproduces a half-page quote from Bieberbach's paper but gives no commentary either.

Note that Neugebauer, in common with most of his contemporaries before WWII, occasionally used the notion of "race" in an intuitive sense and did not doubt the existence of different styles of science in different "races."[8] But he differed from racists

[6] In particular, MKT (1935–1937). A good introduction to this line of research is by Neugebauer himself [19], properly written in Danish, the language of his country at the time.

[7] In particular, UAA (1938). Good introductions by Neugebauer are [20, 21].

[8] After the war, the unscientific notion of a "race" became discredited due to the Nazis and the Holocaust.

such as Bieberbach by denying priority for one "mathematical style" over others and by his belief in the mutual fertilization of different approaches to mathematics.[9] Indeed, in July 1933, half a year after the Nazis had come to power and probably as an expression of his dissent from rising racism, Neugebauer had commented in a book review about the positive effects of an "intermingling of cultures and races (in the cuneiform cultures) as opposed to the narrowly restricted Egypt development [17, p. 563]."[10] Implicitly, Neugebauer opposed a biologistic interpretation of talents within certain "races" – as the Nazis preferred to do – by pointing to the "complicated historical processes" which resulted in different scientific approaches. According to Neugebauer, simple biological or ethnical explanations would amount to replacing history by reference to a "whim of creation" (Laune der Schöpfung). This transpires rather clearly from Neugebauer's plenary talk in German at the International Congress for Mathematicians (ICM) in Oslo 1936 on "Greek Mathematics and its relationship to pre-Greek mathematics." Here, Neugebauer says:

> *It might seem obvious to try to explain this contrast between the numerical methods of the Babylonians and the geometric models of the Greeks by saying that the typically intuitive talent of the Greeks, their sense of geometric form, is revealed here in contrast to the way of thinking of the oriental cultures. It seems to me, however, that such a way of looking at things corresponds roughly to the procedure of solving difficulties in a physical theory by simply invoking a whim of creation. To be satisfied with such 'solutions' is to renounce the foundation of our entire scientific methodology. Everything I have to say in the following is intended to show you that it was complicated historical processes that brought about these apparently so clear final results.* [20, p. 160]

No doubt Neugebauer saw the specific societal conditions in classical Greece, including philosophy, as a determining factor for the "intuitive talents" of their mathematicians. But he denied Greek mathematics to be the towering point of reference and stressed the equal value of cultures and the benefit of various styles of mathematics for the development of mathematics as a whole. This could at the same time be used as an argument against racist theories of separation between peoples, such as in Bieberbach's *Deutsche Mathematik*. In an interview which Neugebauer gave on

[9] This insistence was in the tradition of Felix Klein [37] and of his successor Courant, although other self-proclaimed followers of Klein, such as Bieberbach, tried to insinuate racist feelings into Klein.

[10] This was a review in *Die Naturwissenschaften* of Johannes Tropfke's "History of Elementary Mathematics."

14 July 1936, during the ICM, to the Norwegian daily *Arbeiderbladet*, his insistence on the mingling of cultures for making scientific progress again became clear. Answering the question: "Why is it that the Babylonians were mathematically so talented?" Neugebauer replied:

> *Probably due to the special blend of different types of people down there. The Babylonians built on the Sumerian culture, which the English have investigated very extensively through excavations at Ur.* [5, p. 250]

It is hard *not* to read this as an indirect commentary on Bieberbach!

7. Shifting from Old Babylonian mathematics to Neo-Babylonian mathematical astronomy

The title of Neugebauer's plenary in Oslo "Greek Mathematics and its relationship to pre-Greek mathematics" is somewhat misleading. It does not reveal that at Oslo Neugebauer gave a first report on the mathematical methods of Neo-Babylonian *astronomy.* The talk was published a year later in the Proceedings of the ICM.[11] There Neugebauer explained in some detail how he made sense of the masses of data contained in the cuneiform tablets, registering periodic movements within the Babylonian zodiac "coordinate system" (angular distances, locations below and above the ecliptic, times, even velocities) of planets, the Moon, and the Sun, only few of them observed, most of them calculated. Neugebauer explained what he meant by the "contrast between the numerical methods of the Babylonians and the geometric models of the Greeks" (see quote above):

> *Babylonian astronomy describes periodic processes in a first approximation by 'linear zig-zag functions' [lineare Zackenfunktionen], i.e. by first-order difference series that oscillate back and forth between two extreme values. However, it is often necessary to describe the course of a function more precisely than by linear interpolation. Second-order difference series are then formed by summing up first-order series and putting them together with suitable signs (labeled 'tab' or 'lai' ...) ... The numerical methods described here ... are characteristic of Babylonian astronomy as a whole. In complete contrast to it is Greek astronomy, which uses the Babylonian results in all numerical determinations of constants, but has abandoned the purely numerical methods of the Babylonians to describe the movements as a whole and*

[11] Neugebauer's German lectures "Vorlesungen über babylonische Astronomie" (1936) were apparently never published; cf. [35, p. 305].

1948] MATHEMATICAL METHODS IN ANCIENT ASTRONOMY 1021

[Linear interpolation]

then leads to a continuous function $f(n)$ whose graph looks like Fig. 3. We shall call such functions "linear zigzag functions." The period of the function $f(x)$ may be called P and it is obvious that

$$P = 2\Delta/d$$

where Δ is the amplitude and d the slope of the increasing branch.

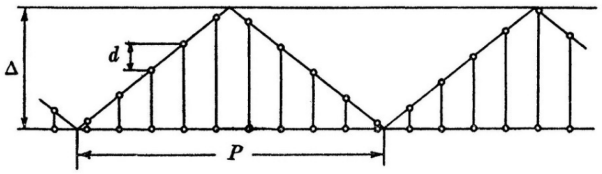


FIG. 3

Figure 4. Neugebauer's linear zigzag function for the interpretation of periodic celestial motion (From [21, p. 1021], emphasis by this author.) (Neugebauer): Courtesy of AMS; used with permission.

> *instead operates with kinematic models, namely with the well-known excenter and epicycle models.* [20, pp. 157–160]

The Greek models are clearly exemplified in the *Almagest* of Ptolemy (around 150 CE), whom Neugebauer admired immensely as a leading applied mathematician of antiquity. Ptolemy used some of the observational data of the Babylonian astronomers and the sexagesimal number system (based on powers of 60), traces of which are still in use today.

A publication in English by Neugebauer 12 years after his Oslo talk in the *Bulletin of the American Mathematical Society* from which Figure 4 is taken gives a similar but more detailed introduction to Babylonian mathematical astronomy, which is still very readable. Here, one finds an allusion to modern "harmonic analysis":

> *To use modern terminology the Babylonian procedure is very close to harmonic analysis. . . . We do not know what were the Babylonian concepts about the physical structure of the universe. It seems safe to say that whatever concepts might have existed they were not directly reflected in the mathematical methods for the lunar and planetary movements.* [21, p. 1020]

The reference to "harmonic analysis" on the part of the Neo-Babylonians might suggest that Neugebauer had in mind his joint publication with Harald Bohr, "On linear differential equations with constant coefficients and almost periodic right-hand

side" (1926) [2]. Of course, this publication was much more sophisticated mathematically and a piece of "pure mathematics." A comparison with the applied mathematics in Babylonian astronomy might therefore be far-fetched unless one understands it in the restricted sense that Neugebauer's mind throughout his historical research was always primarily mathematical.

As Swerdlow and Rowe have convincingly argued, Neugebauer found the rich textual sources for Mesopotamian mathematics and mathematical astronomy far more important than anything he could ever have written about Greek mathematics, let alone Egyptian mathematics [29]. From 1938 onward, he published in his journal *Quellen und Studien* exclusively on his new topic, Babylonian mathematical astronomy.

Neugebauer's comparison of Old Babylonian mathematics with classical Greek geometry was the second, largely independent theme of his Oslo talk.[12] The same year, 1936, he published "Zur geometrischen Algebra" [18]. According to the thought-provoking historiographical concept of "geometric algebra," Greek geometry can be considered to some extent as an analogue (or even as a reformulation) of algebraic reasoning existing either informally in Greek mathematics or earlier. A strong version of this concept was the largely undocumented claim – at least temporarily supported by Neugebauer himself – that there must have existed a direct historical link to older Babylonian algebraic thinking [7]. This claim was criticized by a new generation of historians from the 1960s.[13] Whatever the merits of the discussion of "geometric algebra" may have been, there is little doubt that it cannot be reduced to an ahistorical insinuation by mathematicians projecting the modern standards of mathematics into the past. The discussion was also broadening historical perspectives, and it followed partly political and philosophical motivations in a manner similar to Neugebauer's interpretation of mathematical astronomy, namely as a move against the one-sided and uncritical admiration of the Greeks as the most important representatives of "ancient science and mathematics."

Summing up Neugebauer's talk in Oslo: He stressed the independent importance of – on the one hand – Old Babylonian mathematics compared to later Greek geometry and – on the other hand – of Neo-Babylonian mathematical astronomy compared to contemporary and to somewhat later Greek astronomy. Indeed, he would never again – at least in the titles of his publications – speak about "pre-Greek mathematics" [36, p. 148], [29, p. 47].

[12] The discussion in [5, pp. 180–183] of Neugebauer's talk in Oslo undervalues mathematical astronomy and focuses too strongly on "geometric algebra."

[13] See remarks below in Section 11.

8. Growing international appreciation for Neugebauer's historical work

During his time in Copenhagen, Neugebauer's international reputation as a historian of mathematics and astronomy grew. On 6 September 1933, the historian of science George Sarton, who edited the journal *Isis* from his modest position at Harvard University, wrote a letter to the director of the IAS in Princeton, Abraham Flexner:

> *As compared with Neugebauer I am only a dilettante. He works in the front-trenches while I amuse myself wayback in the rear – praising the ones, blaming the others; saying this ought to be done, etc. – and doing very little myself.* [33, p. 79]

Flexner was obviously impressed and wrote to the leading American mathematician at his Institute, Oswald Veblen, on 8 September 1933:

> *My disposition would be to invite Neugebauer, because he would bring to this country something absolutely new, namely the historical and humanistic side of mathematics.* [33, p. 79]

After a conversation with Niels Bohr that took place near New York City on 14 February 1937, the Rockefeller functionary and mathematician Warren Weaver wrote in his diary:

> *B. talks with great enthusiasm and considerable feeling concerning the fundamental contribution which Neugebauer is making to our knowledge of ancient civilizations … There seems to be no question that this is the most significant contribution that has ever been made to the history of mathematics and thus directly to our knowledge of the historical development of science in general.* [33, p. 90]

Despite this broad appreciation for Neugebauer as a historian, the Americans, including the Rockefeller philanthropy, found it advisable not to invite (or as sometimes was said "import") Neugebauer to the USA but rather support him at his place of exile in Copenhagen. The main reason was the overwhelming interest of the American community of mathematicians in Neugebauer's work for mathematical reviewing (*Zentralblatt*) which in 1937 still seemed to be better placed in Europe.

9. Political crises and Neugebauer's second emigration, from Denmark to the USA

In the middle of the 1930s, the political situation in Europe became ever more nervous and dangerous. From the ICM in Oslo, the Italians had been absent due to sanctions against Italy's war in Ethiopia, and the Russians were forbidden to attend by Stalin and the Politburo [5]. On 14 March 1937, Neugebauer wrote the following letter from Copenhagen to Courant in New York about the intrusion of Stalinism into mathematics:

> *Dear Courant,*
>
> *You will certainly be interested to learn that Kolmogoroff and Khintchine had big scandals in Russia due to their Ergebnisse-reports, published in Germany. As a matter of fact, in Russia there is flourishing now the same idiotic nationalism as in the Third Reich. Of course you should not write about these things to Russia, but because of Yellow Books you should be interested to know. For instance, I do not believe that one of the two would now be able to write a Yellow Book without danger.*
>
> *Many cordial greetings in haste, Your O.N.* [33, pp. 82–83].[14]

The situation became unbearable for Neugebauer in 1938, after racist legislation had been promulgated in Italy with consequences for the Italian co-editor of *Zentralblatt*, the differential geometer Tullio Levi-Civita. After P. S. Aleksandrov had been forced by the Russian politicians to withdraw from the editorial board in May 1938, Levi-Civita was excluded by Springer in October that same year, even without informing the managing editor Neugebauer. In November 1938 – about the time of the anti-Semitic November pogrom in Germany (Kristallnacht) – Neugebauer sent a printed postcard to all the reviewers of *Zentralblatt*:

> *Since one of the editors of the* Zentralblatt für Mathematik *has been canceled without communicating with him, with me or with the other editors, since further it has been demanded of me to consider other than purely objective points of view in the distribution of reviews, I have resigned the editorship of the* Zentralblatt. *I have to thank all my contributors most warmly for their*

[14] In 1933, A. N. Kolmogorov and A. Khinchin had published reports on probability theory with Springer's German *Ergebnisse* (see Figure 2, right side). Both had agreements with Springer and Courant for textbooks on real function theory and probability theory in the prestigious Springer *Grundlehren* series, which appeared in yellow covers. These plans failed under the given political circumstances.

> *many years of distinguished service and, above all, for the understanding with which they have accommodated themselves to the demands, not always convenient, that had to be placed upon them. O. Neugebauer.* [33, p. 86]

After Neugebauer's retreat, Springer continued the publication of *Zentralblatt* in Berlin. The journal now came under even closer political scrutiny from the Nazis and had to coordinate its reviews with the old *Jahrbuch*. As far as Neugebauer was concerned, he severed his connections with Springer totally. In a letter from 9 November 1938 to Neugebauer, Springer had expressed a lack of understanding for Neugebauer's withdrawal from *Zentralblatt.* The publisher then still hoped for a continuation of *Quellen und Studien* with Neugebauer as editor. However, on 3 December 1938, Neugebauer received a letter from Oskar Becker, the philosopher and historian of mathematics in Bonn. In it, Becker demanded the resignation of Toeplitz from *Quellen und Studien* because of his Jewish descent, and he proposed Bessel-Hagen as a replacement. Neugebauer responded with the announcement of his own resignation and proposed Becker and Bessel-Hagen as the new editors:

> *I have inferred from your letter that also scientists in Germany consider the fact of extensive pogroms a sufficient reason for making it impossible for a scholar of outstanding merit even to work scientifically. . . . You are writing that you as a National Socialist apparently have an opinion differing from mine, in spite of your personal respect for Mr. Toeplitz. I can only reply that I am not in the happy possession of any 'Weltanschauung' and I am therefore in need to consider in each case individually, what to do, without being able to retire to a previously given dogma.* [33, p. 87]

As a result, *Quellen und Studien* collapsed. Toeplitz emigrated to Palestine in 1939, and Stenzel's Jewish widow went to the USA in 1939. Neugebauer found a new home in Providence, Rhode Island, where he helped start *Mathematical Reviews* and built a department for the History of Mathematics at Brown University.

The protocol of Neugebauer's 18 February 1939 job interview in Providence, quoted above, shows that he was anxious that the responsibilities for the new reviewing journal should be distributed on broader shoulders than had been the case with *Zentralblatt* in Copenhagen [33, p. 92]. Nevertheless, Neugebauer kept busy with *Mathematical Reviews* in the first 2 or 3 years of its existence. In 1943, the immigrant from Germany and Sweden, the probabilist Willy Feller, 7 years Neugebauer's junior, became managing editor.

10. After the War: Increasing political disillusionment and return to historical research

In a letter to a UNESCO representative, who had offered help in reviving the historical journal *Eudemus*, Neugebauer wrote in December 1947:

> *I confess that I belong to the dying-out tribe of scholars who deeply dislike organized work. . . . I know that more of my remaining lifetime is necessary to complete the studies which interest me deeply. Thus as far as I am concerned I shall spend all my efforts in the continuation of my own research program and the only hope I have is that I am as little as possible disturbed by outside influences.* [33, p. 96]

Of course, Neugebauer had done his fair share of work for the community in the past. In all likelihood, he had become disillusioned by the various instances of political interference with his work. Fortunately, the "remaining lifetime" for the then 48-year-old Neugebauer would amount to over four decades with many fundamental publications to come, among them arguably the most important being the three-volume "Astronomical Cuneiform Texts" (ACT 1955).[15]

By the same token, he kept his skeptical attitude vis-à-vis sociological and biographical work, at least with respect to the history of ancient science. A particularly telling episode is connected with Neugebauer's sharp criticism of a review written by the Belgian theoretical physicist Leon Rosenfeld for the Danish historical journal *Centaurus*. The review was of Neugebauer's revised edition of H.G. Zeuthen's *Forelæsninger over Mathematikens Historie: Oldtiden* from 1893. Rosenfeld praised this book as a "great work," but then added the following remarks:

> *Even the very best human endeavour, however, has its weaknesses. [. . .] I could not imagine those old mathematicians as men of flesh and blood. [. . .] Zeuthen was obviously too much of a mathematician and too little of a genuine historian to pay proper attention to this human side of the question.* [33, p. 100]

In a letter from 17 October 1950, Neugebauer protested to the editor of *Centaurus*, Mogens Pihl. Although his protest ultimately went unpublished, it deserves to be quoted almost in its entirety:

> *I may say that the only flaw [in the present issue of the journal] I found is the review of Rosenfeld, which I can only consider as monumental nonsense. He*

[15] This was originally published in London and later taken over by Springer, whose relationship with Neugebauer was apparently not yet fully recovered in 1955.

> *has obviously not yet understood that historical research is a science and not poetry. He wants to learn something about the motives of men of whom we know hardly more than their time of life. Not to say that I seriously doubt that we know the motives of any scholar, even of our contemporary colleagues. I hope Rosenfeld's lack of insight into the methodology in our field will not exercise too much influence on the editorial policy.* [33, p. 100]

One can safely assume that the Marxist physicist Rosenfeld was not as naive as his review might appear at first glance. What he probably wanted to express was a desire to learn more about the societal conditions under which ancient mathematics was produced, in particular the philosophical connections as well as possible applications. With the memories of World War II still vivid, and in view of the developing atomic race, Rosenfeld probably also wanted to make another point – a bit mistaken for the topic in question – namely to stress the responsibility of scientists for their work. One may assume that Neugebauer, the liberal-minded refugee from Germany, felt similarly disillusioned in the post-war environment. As to social factors, Neugebauer had implicitly acknowledged their importance in his polemics against the "typically intuitive talent (implicitly 'Aryan style') of the Greeks" (see above). But while Rosenfeld reacted with political activity and increased historical interest, the disillusioned historian, writing from the McCarthy-stricken USA,[16] only felt a sense of embarrassment. As a result, he argued in the extreme opposite direction compared to Rosenfeld, denying that one could make judgments about motives "even of our contemporary colleagues." On 19 December 1950, he wrote to Pihl in Copenhagen a follow-up letter:

> *The only reasonable thing to do is to shoot oneself as soon as possible and not to wait for the politicians, the militaries, and the physicists.* [33, p. 101]

In order to better understand Neugebauer's political and philosophical views and his disillusionment after the war, one must certainly take into account his general attitude toward Germany and especially the insufficient and uncritical manner in which the Germans went about "coping with the past." During his entire emigration, Neugebauer had followed political events in Nazi Germany closely, particularly the

[16] Jens Høyrup allowed me to use the following anecdote which he has from a Danish colleague, himself a rather unpolitical person: "He told me that he had visited Neugebauer during the McCarthy period. He spoke, scandalized, about the treatment given to Dirk Struik [a prominent Dutch-American differential geometer, communist and historian of mathematics who was under indictment at the time for un-American activities]. Neugebauer got up, and closed the door, and said 'I have once been forced to move from one continent to another for political reasons. I do not want to repeat that."' (Email from Jens Høyrup to me, 27 October 2024).

political behavior of the scholars who remained behind. We have already noted his critical and somewhat cynical response to Becker. Even regarding his old friend, the Dutch mathematician Bartel L. van der Waerden (1903–1996), who had once attended Neugebauer's lectures in Göttingen on the history of ancient mathematics[17] but who later stayed in Germany (at Leipzig) in spite of opportunities abroad, Neugebauer grew increasingly critical. In 1940, he wrote to Bessel-Hagen: "What you write about Bartel is all well but for reasons I cannot explain here I don't give a damn for him." In August 1945, Neugebauer wrote to his friend, the topologist Heinz Hopf in Zürich, about van der Waerden, who at the time had political problems due to remaining in Germany even as a Dutch citizen [32]:

> *I do not mind his remaining a German professor until the end. I do mind his remaining German professor at the beginning. However, I feel very differently than the Lord and I do not intend to do anything positive or negative. I know that Bessel-Hagen behaved excellently and I would like very much to hear about him as soon as he is again localized.* [33, p. 97]

To van der Waerden himself, Neugebauer had written a week before: "What has happened to the *Zentralblatt*? I hope it is completely ruined." When van der Waerden complained that Neugebauer continued to write personal letters in English, not using his mother tongue even though the war had ended, Neugebauer offered the following rejoinder, which, however, probably expressed only half of the truth: "I must remark that the language I use in my letters does not depend on my mother but on my secretary." [33, p. 97] On 15 August 1946, Neugebauer wrote to his old friend, the mathematician Franz Rellich (1906–1950), who happened to be van der Waerden's brother-in-law. Rellich was known to have behaved decently during the Nazi years and had been appointed director of the Mathematical Institute in Göttingen in 1946. Neugebauer wrote concerning *Mathematical Reviews*:

> *This letter is actually intended to ask you for your advice in a delicate matter. We would like to extend our staff of reviewers to include Germans, but we must avoid inviting people who are politically 'belastet' [incriminated].*[18]

[17] Van der Waerden is known, besides his *Moderne Algebra* (1930–1931) and his book on mathematical statistics (1957), also for his historical books *Science Awakening* (Ontwakende Wetenschap) I (1950) and II (1965), the latter [38] on Ancient Mathematical Astronomy.

[18] Niedersächsische Staats- und Universitätsbibliothek Göttingen, Handschriftenabteilung, Cod_Ms_F_Rellich_1_61. Rellich replied on 19 September 1946 with an interesting list of 17 German mathematicians. Only two of them (G. Grüss and W. Magnus) had been professors before 1945, and the two were known as dissidents, as many others on the list. Only five of the

Neugebauer kept reservations against Germany in the years to come which might have been partly because the University of Göttingen did not do much to acknowledge the reasons for his emigration in 1934. In November 1945 – before Rellich's appointment – the university authorities had submitted to the British Military Government a list of possible candidates for professorships among former emigrants. From this list, Neugebauer, who had been proposed by the British, was excluded because:

> *He went in 1933 to Copenhagen, without political pressure having been exerted on him, that means voluntarily. There he accepted the chair for history of mathematics which already had been prepared for him.* [33, p. 98]

The document has an additional commentary in pencil about Neugebauer: "Out of the question, went voluntarily." So the proposal of the British was refused by the University of Göttingen. Ten years later, Neugebauer got compensation, but only for the level of "assistant professor."

11. Later historical work connecting to Neugebauer

As indicated above, Neugebauer's approach to ancient Babylonian mathematics and astronomy was partly influenced by his perspective as a modern mathematician trained in Göttingen. Other more competent mathematicians such as van der Waerden and Peter J. Huber (born 1934) have followed Neugebauer's historical interests in their own work, contributing important details.[19] Huber, who was van der Waerden's student in Zürich in the 1950s [3] and collaborated in the latter's book on the *Birth of Mathematical Astronomy* which was originally published in 1965 [38], said much later (2010) to a colleague:

> *For me, all this was an amusing hobby. I felt that history of science was not a serious profession, especially not for young people (and apart from that, the job situation predictably would be difficult).* [35, p. 326]

In his "History of Ancient Mathematical Astronomy" (HAMA), Neugebauer praised particularly Huber's article of 1957 on the daily movement of Jupiter in

17 have entries in the biographical website of St. Andrews University today. Although Rellich may have found most established mathematicians to be unlikely referees for the American journal, the list seems to indicate – by their exclusion – a strong involvement of many of them with the Nazis.

[19] Høyrup tells me that Huber is still active in the field: "These later years he has made very interesting work on Old Babylonian linguistics (that is, the linguistics of the ancient scribes)." (Email, 27 October 2024).

Babylonian texts [8] and the "excellent discussion" of linear interpolation methods there [23, p. 412]. Huber went on with a successful career in mathematical statistics,[20] influenced in this field also by van der Waerden.

Research in the mathematical aspects of Babylonian astronomy continues. In 2016, some recent progress was reported in the widely read journal *Science.* It relates to two major discoveries of Neugebauer's on Old Babylonian texts, published in 1929, and on the Neo-Babylonian mathematical astronomy in his Oslo talk of 1936, which have been mentioned above [15, 20]. The author of the *Science* article was the Dutch historian M. Ossendrijver. He wrote about his recent discovery of a Babylonian cuneiform tablet from the last half-millenium BCE:

> *The idea of computing a body's displacement as an area in time-velocity space is usually traced back to 14th-century Europe. I show that in four ancient Babylonian cuneiform tablets, Jupiter's displacement along the ecliptic is computed as the area of a trapezoidal figure obtained by drawing its daily displacement against time. This interpretation is prompted by a newly discovered tablet on which the same computation is presented in an equivalent arithmetical formulation. The tablets date from 350 to 50 BCE. The trapezoid procedures offer the first evidence for the use of geometrical methods in Babylonian mathematical astronomy, which was thus far viewed as operating exclusively with arithmetical concepts.* [24, p. 482]

Ossendrijver explains Jupiter's displacement further for the case of a time period of constant deceleration (linear decrease of velocity) of the planet, where equal areas mean equal distances traveled (see Figure 5, right):

> *The trapezoid for days 0 to 60 is divided into two smaller trapezoids of equal area. In order to achieve this, the Babylonian astronomers applied a partition procedure that is well-attested in Old Babylonian (2000 to 1800 BCE) mathematics.* [24, p. 484]

The author makes clear that the trapezoid in the picture is an interpretation of the text, which contains no geometrical figures. Ossendrijver also explains that "geometrical" here has a different, more abstract meaning compared to the "geometrical models" of Greek astronomers which were discussed above.

> *The Babylonian trapezoid procedures are geometrical in a different sense than the methods of the mentioned Greek astronomers, since the geometrical figures describe configurations not in physical space but in an abstract mathematical space defined by time and velocity (daily displacement).* [24, p. 484]

[20] Among other things, Huber is a recognized specialist in the field of "robust statistics."

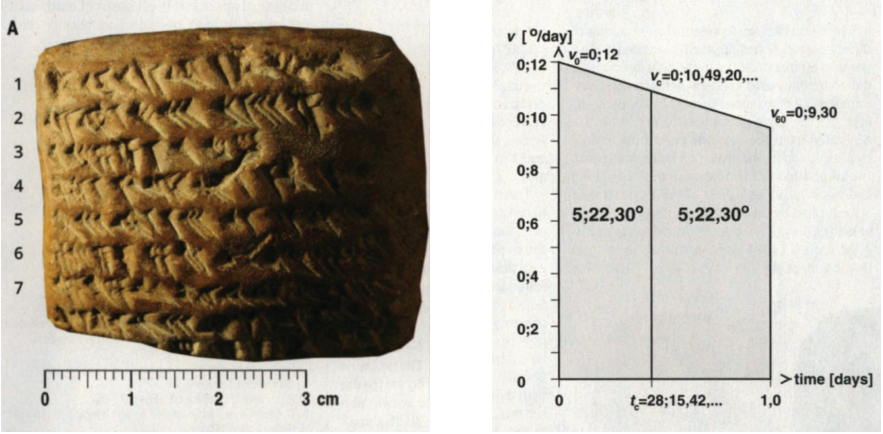


Figure 5. Two figures from Ossendrijver's article in *Science* (2016) [24]. Left: the "newly discovered" Babylonian text. Right: Ossendrijver's interpretation. Courtesy of AAS; used with permission.

In another article of the same year, Ossendrijver points out that historians have enlarged their perspective from a predominantly mathematical standpoint of Neugebauer's to a more general historical one:

> *Subsequent research on Babylonian mathematical astronomy has proceeded from the advanced stage where Neugebauer left it, . . . the aim has shifted from understanding Babylonian mathematical astronomy in our own terms to interpreting its algorithms and the underlying mathematical and astronomical concepts in Babylonian terms.* [25, pp. 340–341]

This broader historical view is sometimes interpreted as criticism of Neugebauer's historical methodology, in particular connected to his support for the "geometrical algebra" hypothesis, as indicated above [7].[21] Nevertheless, all critical remarks do not concern the core of Neugebauer's positive historical discoveries and his

[21] According to Høyrup, "Neugebauer is the... indubitable source for van der Waerden's view of Greek 'geometric algebra' as a translation of Babylonian algebraic knowledge into geometric language necessitated by the discovery of incommensurability" [7, p. 143]. The same author quotes Neugebauer's admission from 1969 that "simplicity of a historical explanation" (assuming the translation hypothesis) is "by no means equivalent with historical proof" [7, p. 140]. Høyrup also disagrees with some "idiosyncratic" claims of Neugebauer's with respect to the "algebraic character of Babylonian mathematics" [7, p. 138]. I find some ahistorical insinuation of algebraic thinking in Neugebauer's early work on Egyptian mathematics as well [34]. In recent years, the heuristic potential of the "geometric algebra" hypothesis has been increasingly acknowledged [7].

groundbreaking work as editor and interpreter of Babylonian texts. In any case, Neugebauer's linguistic expertise and historical sense prevented him from falling into the traps of naive historical views sometimes entertained and exaggerated by research mathematicians.[22] Progress in the historiography of Old Babylonian mathematics and Neo-Babylonian astronomy has supplemented Neugebauer's results on many fronts. Among these investigations figure those on the geometrical background of Babylonian numerical algorithms, on the social, in particular bureaucratic and educational environment, and on linguistic nuances.

It is possible that Neugebauer found some of this broader work as belonging to a "smokescreen of sociological, biographical, and bibliographical irrelevancies" as he said in 1975. I believe, however, to have shown that Neugebauer himself could not avoid considering the social and philosophical environment and that he had to acknowledge the impact of "complicated historical processes" (as he called them) on the shaping of mathematical and astronomical theories. His claim to be "not in the happy possession of any Weltanschauung" was formulated in an extreme situation of political pressure in the 1930s, and his later attack on sociological and related investigations was apparently influenced by increasing personal political disillusionment, as reported by colleagues [4].

Acknowledgments. The archives containing the estates of Neugebauer (Princeton) and Bessel-Hagen (Bonn) as well as the Manuscript Division in Göttingen are thanked for giving access to their sources.

I thank June Barrow-Green (London) for advice with respect to language and content, Guillermo P. Curbera (Sevilla) for kind help with the manuscript, and Jens Høyrup (Copenhagen) for his expertise in Babylonian mathematics. All remaining errors are mine.

Neugebauer's most important publications

MKT Mathematische Keilschrift-Texte (1935/37 in *Quellen und Studien*)

UAA Untersuchungen zur Antiken Astronomie (1938 in *Quellen und Studien*)

MCT Mathematical Cuneiform Texts (1945, completion of MKT in English, with A. Sachs)

ACT Astronomical Cuneiform Texts (1955, completion of UAA in English)

[22] Van der Waerden, who had to rely on Neugebauer's philological expertise, went on to publish speculations about the existence of a "Stone Age Mathematics," which were largely dismissed by professional historians.

HAMA History of Ancient Mathematical Astronomy (1975, see specific literature below)

References

[1] L. Bieberbach, Stilarten mathematischen Schaffens. *Sitzungsber. Preuß. Akad. Wiss., Phys.-Math. Kl.* (1934), 351–360 Zbl 0009.38801

[2] H. Bohr and O. Neugebauer, Über lineare Differentialgleichungen mit konstanten Koeffizienten und fastperiodischer rechter Seite. *Nachr. Gesellschaft Wiss. Göttingen Math.-Phys. Kl.* (1926), 8–22 JFM 52.0464.01

[3] A. Buja and H. R. Künsch, A conversation with Peter Huber. *Statist. Sci.* **23** (2008), no. 1, 120–135 Zbl 1327.62011 MR 2523944

[4] P. J. Davis, Otto Neugebauer: reminiscences and appreciation. *Amer. Math. Monthly* **101** (1994), no. 2, 129–131 Zbl 0795.01024 MR 1259824

[5] C. D. Hollings and R. Siegmund-Schultze, *Meeting under the integral sign? The Oslo Congress of Mathematicians on the Eve of the Second World War*. Hist. Math. 44, American Mathematical Society, Providence, RI, 2020, 338 pp. Zbl 1495.01004 MR 4249622

[6] J. Høyrup, *Lengths, widths, surfaces. A portrait of old Babylonian algebra and its Kin*. Sources Stud. Hist. Math. Phys. Sci., Springer, New York, 2002, 462 pp. Zbl 0999.01001 MR 1873291

[7] J. Høyrup, What is "geometric algebra", and what has it been in historiography? *AIMS Math.* **2** (2017), no. 1, 128–160 Zbl 1431.01057

[8] P. Huber, Zur täglichen Bewegung des Jupiter nach babylonischen Texten. *Z. Assyriologie Vorderasiatische Archäologie* **52** (1957), no. Jahresband, 265–303

[9] A. Jones, C. Proust, and J. M. Steele (eds.), *A mathematician's journeys. Otto Neugebauer and modern transformations of ancient science*. Archimedes 45, Springer, Cham, 2016, 342 pp. Zbl 1335.01010 MR 3410193

[10] F. Klein, *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert. Teil I*. Für den Druck bearbeitet von R. Courant und O. Neugebauer. Springer, Berlin, 1926 JFM 52.0022.05 MR 0529278

[11] E. Neuenschwander, Der Nachlass von Erich Bessel-Hagen im Archiv der Universität Bonn. *Historia Math.* **20** (1993), no. 4, 382–414 Zbl 0792.01026 MR 1248472

[12] O. Neugebauer, Litteraturanmeldelse (T. E. Peet, The Rhind Mathematical Papyrus). *Mat. Tidsskr. A* (1925), 66–70

[13] O. Neugebauer, *Die Grundlagen der ägyptischen Bruchrechnung*. Springer, Berlin, 1926, 45 pp. JFM 52.0004.02

[14] O. Neugebauer, Zur Entstehung des Sexagesimalsystems. *Abh. Akad. Wiss. Göttingen Neue Folge* **13** (1927), 1–55 JFM 53.0003.04

[15] O. Neugebauer, Zur Geschichte der babylonischen Mathematik. In O. Neugebauer, J. Stenzel, and O. Toeplitz (eds.), *Quellen und Studien zur Geschichte der Mathematik, Astronomie und Physik. Abteilung B: Studien / Band 1 / Heft 1*, pp. 67–80, Springer, Berlin, 1929 JFM 55.0593.02

[16] O. Neugebauer, Das Mathematische Institut der Universität Göttingen. *Die Naturwissenschaften* **18** (1930), 1–4 JFM 56.0081.01

[17] O. Neugebauer, Review of J. Tropfke, Geschichte der Elementarmathematik, Zweiter Band, 3. Auflage, 1933. *Die Naturwissenschaften* **21** (1933), 563–564

[18] O. Neugebauer, Zur geometrischen Algebra. *Quell. Stud. Gesch. Math. Astron. Phys. B* **3** (1936), 245–259 Zbl 0013.33702

[19] O. Neugebauer, Matematiske Kileskrifttekster. *Mat. Tidsskr. B* (1936), 1–9 JFM 62.0005.05

[20] O. Neugebauer, Über griechische Mathematik und ihr Verhältnis zur vorgriechischen. In *Comptes rendus du Congrès International des Mathématiciens, Oslo 1936, Vol. 1*, pp. 157–170, A.W. Broggers Boktrykkeri, Oslo, Norway, 1937 Zbl 63.0799.01

[21] O. Neugebauer, Mathematical methods in ancient astronomy. *Bull. Amer. Math. Soc.* **54** (1948), 1013–1041 Zbl 0033.24101 MR 0028217

[22] O. Neugebauer, *Reminiscences on the Göttingen Mathematical Institute on the Occasion of R. Courant's 75th Birthday*, Otto Neugebauer Papers, Box 14, IAS, 1963

[23] O. Neugebauer, *A history of ancient mathematical astronomy*. Sources Stud. Hist. Math. Phys. Sci. 1, Springer, Berlin, 1975, 1456 pp. Zbl 0323.01002 MR 0465672

[24] M. Ossendrijver, Ancient Babylonian astronomers calculated Jupiter's position from the area under a time-velocity graph. *Science* **351** (2016), no. 6272, 482–484 Zbl 1355.01005 MR 3467973

[25] M. Ossendrijver, Translating Babylonian mathematical astronomy: Neugebauer and beyond. In A. Jones, C. Proust, and J. Steele (eds.), *A mathematician's journeys. Otto Neugebauer and modern transformations of ancient science*, pp. 333–342, Archimedes 45, Springer, Cham, 2016 Zbl 1342.01047 MR 3496903

[26] L. Pyenson, Inventory as a route to understanding: Sarton, Neugebauer, and sources. *Hist. Sci.* **33** (1995), no. 3, 253–282 MR 1356858

[27] C. Reid, *Courant in Göttingen and New York. The story of an improbable mathematician*. Springer, New York, 1976, 314 pp. Zbl 0353.01010 MR 0532647

[28] E. Robson, *Mesopotamian mathematics, 2100–1600 BC. Technical constants in bureaucracy and education*. Oxf. Ed. Cuneif. Texts 14, The Clarendon Press, Oxford, 1999, 334 pp. Zbl 0963.01001 MR 1735671

[29] D. E. Rowe, From Graz to Göttingen: Neugebauer's early intellectual journey. In A. Jones, C. Proust, and J. Steele (eds.), *A mathematician's journeys. Otto Neugebauer and modern transformations of ancient science*, pp. 1–59, Archimedes 45, Springer, Cham, 2016 Zbl 1342.01037 MR 3496893

[30] R. Siegmund-Schultze, *Rockefeller and the internationalization of mathematics between the two world wars. Document and studies for the social history of mathematics in the 20th century*. Sci. Networks Hist. Stud. 25, Birkhäuser, Basel, 2001, 341 pp. Zbl 0981.01013 MR 1826906

[31] R. Siegmund-Schultze, *Mathematicians fleeing from Nazi Germany. Individual fates and global impact*. Princeton University Press, Princeton, 2009, 471 pp. Zbl 1179.01024 MR 2522825

[32] R. Siegmund-Schultze, Bartel Leendert van der Waerden (1903–1996) im Dritten Reich: Moderne Algebra im Dienste des Anti-Modernismus? In D. Hoffmann and M. Walker (eds.), *„Fremde" Wissenschaftler im Dritten Reich. Die Debye-Affäre im Kontext*, pp. 200–229, Göttingen, Wallstein, 2011

[33] R. Siegmund-Schultze, "Not in possession of any Weltanschauung": Otto Neugebauer's flight from Nazi Germany and his search for objectivity in mathematics, in reviewing, and in history. In A. Jones, C. Proust, and J. Steele (eds.), *A mathematician's journeys. Otto Neugebauer and modern transformations of ancient science*, pp. 61–106, Archimedes 45, Springer, Cham, 2016 Zbl 1341.01049 MR 3496894

[34] R. Siegmund-Schultze, Another look at the two Egyptian pyramid volume 'formulas' of 1850 BCE. *Br. J. Hist. Math.* **37** (2022), no. 3, 171–178 Zbl 1521.01002 MR 4516926

[35] J. M. Steele, Neugebauer's *Astronomical Cuneiform Texts* and its reception. In A. Jones, C. Proust, and J. Steele (eds.), *A mathematician's journeys. Otto Neugebauer and modern transformations of ancient science*, pp. 303–332, Archimedes 45, Springer, Cham, 2016 Zbl 1335.01078 MR 3496902

[36] N. M. Swerdlow, Otto E. Neugebauer (26 May 1899–19 February 1990). *Proc. Amer. Philos. Soc.* **137** (1993), no. 1, 139–165 MR 3363428

[37] R. Tobies, *Felix Klein. Visions for mathematics, applications, and education*. Vita Math. 20, Birkhäuser, Cham, 2021, 677 pp. Zbl 1480.01002 MR 4292586

[38] B. L. Van der Waerden, *Science awakening. II. The birth of astronomy* (with contributions by Peter Huber). Noordhoff, Leyden, 1974, 347 pp. (Dutch original 1965)

Reinhard Siegmund-Schultze
Department of Mathematics, University of Agder, Kristiansand, Norway;
reinhard.siegmund-schultze@uia.no

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/43

Integrable PDEs for random growth

Jeremy Quastel

Abstract. We describe the KPZ equation and explore the KPZ fixed point, its connection to the Kadomtsev–Petviashvili equation and how the stochastic integrability arises from a microscopic model, the polynuclear growth model (PNG).

1. The KPZ universality class

What is the universal equation for random growth of a one-dimensional interface? Physicists will certainly answer that it is the Kardar–Parisi–Zhang (KPZ) equation [17]: A height function $h(t, x)$, $x \in \mathbb{R}$, $t \geq 0$, is growing according to the stochastic partial differential equation

$$\partial_t h = \lambda(\partial_x h)^2 + \nu \partial_x^2 h + \sqrt{D}\xi, \tag{1.1}$$

where $\xi(t, x)$ is *space-time white noise*, i.e., the delta correlated Gaussian field,

$$\langle \xi(t, x), \xi(s, y) \rangle = \delta(t - s)\delta(x - y).$$

This equation was introduced to model stochastic growth processes such as *ballistic deposition*, a continuous time surface growth model on an integer lattice $h(x, t) \in \mathbb{Z}$, $x \in \mathbb{Z}$, with the following stochastic rule. Attached to each x there is an independent Poisson process. At each jump time of the Poisson process, a particle falls at x and either (1) stacks on top of the particles already there or (2) sticks to one of the neighbouring stacks, whichever is larger. So the rule is that

$$h(x) \mapsto \max\{h(x - 1), h(x) + 1, h(x + 1)\}$$

independently at each site x at rate 1. The non-linear term $(\partial_x h)^2$ in the KPZ equation is supposed to model the fact that there is a *lateral* part of the growth – the sticky particles – so the radial part of the growth should depend non-linearly on the slope of the interface. The parameter λ reflects the strength of this non-linearity. Similarly, ν and D are the strengths of diffusivity and noise.

Mathematics Subject Classification 2020: 60K35 (primary); 82C31 (secondary).
Keywords: integrable systems, random growth, KPZ universality, stochastic PDEs.

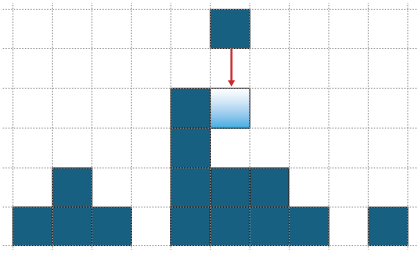


Figure 1. Ballistic deposition.

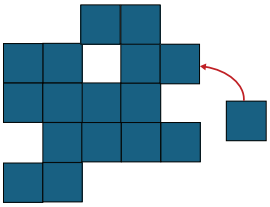


Figure 2. Eden model.

Ballistic deposition (see Figure 1) in turn is a toy model for the *Eden model* (see Figure 2), a growth model on $\mathbb{Z}^2$: Start a growing cluster with a single site. At each time step, look at all sites adjacent to the cluster and choose one uniformly at random to add to the cluster.

The Eden model can be thought of as a model for a fire front or a growing bacterial colony, etc. There is also a class of models like it, called *first passage percolation*. The same sort of thing should happen whenever we have a stable phase invading an unstable phase in the presence of randomness in two dimensions. An earlier way to model such things was through a stochastic reaction-diffusion equation for $\varphi(t, x_1, x_2)$:

$$\partial_t \varphi = \Delta \varphi + \xi V'(\varphi),$$

where ξ is some sort of noise and V is a potential with one stable and one unstable equilibrium.

These latter models are undirected, and their exact location of the front is a bit unclear, though one could certainly introduce some definition. If one tried to choose parameters λ, ν, D to approximate the Eden model by the KPZ equation (1.1), one would find oneself stuck quite quickly. Models like Ballistic deposition (see Figure 1), which are height functions over a one-dimensional substrate, are called *directed*. All the models that we can analyse are of directed type, though for most of them, like Ballistic deposition, we cannot say much.

The phenomenon we want to study is that fluctuations are observed to be of size $t^{1/3}$ at time t and live on spatial scales of size $t^{2/3}$. Furthermore, in the corresponding large space-time rescaling, one is supposed to see certain universal stochastic processes.

The KPZ equation has an advantage over the others mentioned so far, in that it has an explicit invariant measure. Two-sided Brownian motion with variance $\frac{D}{(2\nu)}$ plus any drift bx are invariant, and there are no others. Note that it does not depend on λ, and for $\lambda = 0$, it is easy to prove. For $\lambda \neq 0$, it is quite miraculous. The earliest physics arguments of the invariance [22] were very non-rigorous, but it can be seen from approximating models that preserve random walk interfaces [6]. Only recently, a proof in the continuum has become available [12].

One can devise all kinds of growth models, but one has to be extremely lucky to have a tractable invariant measure. This appears to be a necessary, but not sufficient, condition to make any sort of analytical progress. For example, there is a spatial discretization of KPZ which has Gaussian random walks as invariant measure, but nobody has been able to prove much about its large-scale behaviour. On the other hand, recently highly non-trivial invariant measures for the KPZ equation in an interval or half-line with Neumann boundary conditions have been discovered [5, 9]. So there are examples where the invariant measure is not simple. And in fact, we only know them by approximation from microscopic models; there is no proof of their invariance at the continuum level yet.

If we start the KPZ equation with a nice smooth height function $h_0(x)$, at a very small time $t > 0$, it will look like a locally Brownian version of h_0. For example, the quadratic variation in an interval $[a, b]$ will instantly jump from 0 at time 0 to $\frac{D}{2\nu}|b - a|$ for $t > 0$. This means that the non-linear term $(\partial_x h)^2$ in (1.1) cannot make classical sense, and some renormalization needs to be done to it. This was most spectacularly achieved in [14], but an easy way out [6] is to just define (1.1) via the Hopf–Cole transformation

$$h(t, x) = \frac{\nu}{\lambda} \log Z(t, x), \quad \partial_t Z = \nu \partial_x^2 - \beta Z, \tag{1.2}$$

with $\beta = -\frac{\sqrt{D}\lambda}{\nu}$. This is called the *Hopf–Cole solution.* The multiplicative stochastic heat equation (1.2) turns out to be a rare non-linear (in the noise ξ) stochastic partial differential equation which can be handled by a naive generalization of the Itô theory of stochastic differential equations. On the other hand, for stability and convergence, one often needs a direct theory of (1.1) which has been supplied by Hairer and others [13, 15]. The multiplicative stochastic heat equation has a Feynman–Kac representation as a Brownian expectation:

$$Z(t, x) = E[e^{-\beta \int_0^t \xi(t-s, B_s + x) ds} Z(0, B_t + x)]. \tag{1.3}$$

The space-time white noise is far too irregular for this to make classical sense, but if one expands the exponential as a Taylor series and replaces each resulting multiple integral $\int_{[0,t]^n} ds_1 \cdots ds_n$ by its ordered version $n! \int_{0<s_1<\cdots<s_n<t} ds_1 \cdots ds_n$, the result is a convergent series of multiple stochastic integrals for the Itô solution of (1.2). So we can just define the right-hand side of (1.3) to mean that. The Feynman–Kac representation (1.3) provides a link to *directed polymers in a random environment* where we assign probabilities to random walk paths X_n on $\mathbb{Z}$ according to

$$Z_\beta^{-1} e^{-\beta \sum_{i=1}^n \xi_{i,X_i}},$$

where $\xi_{i,j}$ are now independent identically distributed random variables, and Z_β is the normalization that makes it a probability measure. The *free energy* $h = \log Z$ can be thought of as a discrete KPZ and is expected in great generality to have the same large-scale behaviour. Moreover, β is thought of as an inverse temperature: As $\beta \searrow 0$ (high temperature), the polymer is free; as $\beta \nearrow \infty$ (low temperature), it is an energy-optimizing path, called *last passage percolation*. In $1+1$ dimensions at any finite temperature, the paths on large scale are localized around favourite paths. All of them have lateral fluctuations of size $t^{2/3}$, as opposed to $t^{1/2}$ at $\beta = 0$. But it is only in special last passage cases that one can do full computations.

In order to understand the $t^{1/3}$ height fluctuations from a scaling perspective, we rescale (1.1). Let us use ε for the relation between small and large spatial scales. We want to choose exponents χ and z in order to have an interesting limit as $\varepsilon \searrow 0$ of

$$h_\varepsilon(t,x) = \varepsilon^\chi (h(\varepsilon^{-z} t, \varepsilon^{-1} x) - c\varepsilon^{-z} t). \tag{1.4}$$

The fact that Brownian motion is invariant forces

$$\chi = \frac{1}{2}$$

since Brownian motion is the natural fixed point of $h(x) \mapsto \varepsilon^{1/2} h(\varepsilon^{-1} x)$. Since white noise scales as the square root of the Dirac delta, we obtain after a little work

$$\partial_t h_\varepsilon = \varepsilon^{\frac{3}{2}-z} \lambda (\partial_x h_\varepsilon)^2 - c_\varepsilon + \varepsilon^{2-z} \nu \partial_x^2 h_\varepsilon + \varepsilon^{1-\frac{1}{2}z} \sqrt{D}\xi$$

which forces [10, 23]

$$z = \frac{3}{2}. \tag{1.5}$$

So the height fluctuations $h \sim \varepsilon^{-\chi} = (\varepsilon^{-z})^{1/3}$, and this is just another way of saying that the fluctuations are of size $t^{1/3}$. It also identifies the scale on which we expect to see non-trivial spatial correlations as $\varepsilon^{-1} \sim t^{2/3}$. On the other hand, to prove scaling (1.5) requires sophisticated arguments [4].

Scaling (1.4) shows that the KPZ equation is *not scaling invariant*. It is not our universal model for $1+1$-dimensional random growth. It is more of a nice continuum model of the same sort as ballistic aggregation, but with the advantage that the invariant measure is explicit.

Let us take $\lambda = \nu = \frac{1}{4}$ and $D = 1$ which is a kind of standard case. So we are solving the 1:2:3 scaled KPZ equation

$$\partial_t h_\varepsilon = \frac{1}{4}(\partial_x h_\varepsilon)^2 - c_\varepsilon + \varepsilon^{\frac{1}{2}} \frac{1}{4} \partial_x^2 h_\varepsilon + \varepsilon^{\frac{1}{4}} \xi \tag{1.6}$$

and asking what the limit looks like as $\varepsilon \to 0$ if we have initial data correspondingly rescaling as $h_\varepsilon(0, x) \to \mathfrak{h}_0(x)$. The constant c_ε is going to infinity, but one could simply work at the level of the derivative $u_\varepsilon = \partial_x h_\varepsilon$ which satisfies the rescaled stochastic Burgers equation

$$\partial_t u_\varepsilon = \frac{1}{4} \partial_x u_\varepsilon^2 + \varepsilon^{\frac{1}{2}} \frac{1}{4} \partial_x^2 u_\varepsilon + \varepsilon^{\frac{1}{4}} \partial_x \xi, \tag{1.7}$$

and then the constant just is not there anymore. So one might guess the limiting evolution would be the Burgers equation

$$\partial_t u = \frac{1}{4} \partial_x u^2.$$

This equation is terribly ill-posed. Usually, in a class in partial differential equations, we are told the true solution is the inviscid solution, which means to add $\nu \partial_x^2 u$ to the right-hand side and then let $\nu \searrow 0$. But this cannot be the answer for us because white noise is invariant for (1.7) all the way along. But one can compute the inviscid solution starting with white and it is not white noise. In fact, a remnant of the noise survives, and one has the following result.

Theorem 1.1. *The limit of* (1.6) *as* $\varepsilon \searrow 0$ *is a Markov process of* $\mathfrak{h}(t, x)$ *called* the KPZ fixed point. *It satisfies a Hopf–Lax-type variational formula, with a non-trivial stationary forcing noise, the* Airy sheet *(see* (3.8)*): For fixed* $t > 0$,

$$\mathfrak{h}(t, x; \mathfrak{h}_0) = \sup_{y \in \mathbb{R}} \left\{ t^{\frac{1}{3}} \widehat{\mathcal{A}}(t^{-\frac{2}{3}} x, t^{-\frac{2}{3}} y) - \frac{1}{t}(x - y)^2 + \mathfrak{h}_0(y) \right\}. \tag{1.8}$$

Its one-dimensional distributions' $F(t, x, r) = \mathrm{Prob}(\mathfrak{h}(t, x) \le r)$ *logarithmic derivative* $\phi = \partial_r^2 \log F$ *solves the* Kadomtsev–Petviashvili equation

$$\partial_t \phi + \frac{1}{2} \partial_r \phi^2 + \frac{1}{12} \partial_r^3 \phi + \frac{1}{4} \partial_r^{-1} \partial_x^2 \phi = 0. \tag{1.9}$$

It is worth comparing the KP equation (1.9) with the KPZ equation (1.6). One thinks intuitively that the quadratic non-linearity in KP must have something to do with the quadratic non-linearity in KPZ. But they are not even in the same variable. As far as we know, there is no heuristic or physical argument for the appearance of (1.9). The KP equation is 1:2:3 scaling invariant, and it is a big surprise that it arises as a universal equation for random growth. The proof combines many results and is essentially a long algebraic computation combined with various asymptotic analyses. What happens is that there are certain stochastically integrable models in the class – we will describe one here, the polynuclear growth model. Scaling these, we discover the 1:2:3 scaling invariant Markov process at the centre of the class, the KPZ fixed point, which is also stochastically integrable and whose finite-dimensional distributions, given by appropriate Fredholm determinants, turn out to satisfy KP equations.

2. Stochastic integrability

The route to (1.9) started a quarter century ago with a groundbreaking result of Baik–Deift–Johansson [3] on the fluctuations of the longest increasing subsequence of a random permutation. Let σ be chosen uniformly from S_N, the set of permutations of $\{1, \dots, N\}$, and L_N be the length of the longest increasing subsequence $\sigma(i_1) < \cdots < \sigma(i_L), i_1 < \cdots < i_L$. In undergraduate combinatorics classes, we prove that once N is greater than n^2, there is always either an increasing or decreasing subsequence of length $n+1$, so it is reasonable to conjecture, as Ulam did, that $L_N \sim CN^{1/2}$. This was proved by Hammersley in 1972. In 1976, Vershik–Kerov and Logan–Shepp independently showed that $C = 2$. The fluctuations are much more refined. Baik–Deift–Johansson discovered in 1999 that

$$\lim_{N\to\infty} \text{Prob}\left(\frac{L_N - 2N^{\frac{1}{2}}}{N^{\frac{1}{6}}} \le r\right) = F_{\text{GUE}}(r), \quad \text{where } F_{\text{GUE}}(r) := e^{-\partial^{-2}q^2(r)} \tag{2.1}$$

is the Tracy–Widom GUE distribution (often called F_2). Here, ∂^{-2} denotes the second anti-derivative, and q is the *Hastings–McLeod* [16] solution (i.e., $q(r) \sim 0$ as $r \to \infty$, and $q(r) \sim |r/2|^{1/2}$, implying $D^{-2}q^2 \sim r^3/2$ as $r \to -\infty$) of the Painlevé II equation

$$q'' = rq + 2q^3,$$

which already hints that there are integrable systems behind the scenes. It had been discovered a decade earlier as the limiting distribution of the top eigenvalue of a matrix from the Gaussian unitary ensemble. The connection between random matrices and longest increasing subsequences was murky at best. And what do longest increasing subsequences have to do with random growth models?

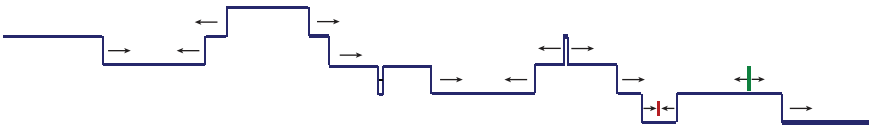


Figure 3. In the PNG dynamics, flat islands with integer heights expand deterministically in both directions with speed 1, merging when they collide (red), and with new islands (green) appearing above existing ones at the times of a space-time Poisson process with rate 2.

In fact, the longest increasing subsequence turns out to be a special case of a classical random growth model called PNG (polynuclear growth model). This is a Markov process whose state space consists of height functions $h : \mathbb{R} \to \mathbb{Z} \cup \{-\infty\}$. So they are piecewise constant, taking only integer jumps at discrete positions. We also (for good reasons) allow h to take the value $-\infty$. Like all these growth models, the natural topology on the state space is local Hausdorff convergence of hypographs, with the range of the heights compactified at $-\infty$. The PNG dynamics consists of two parts running simultaneously: (1) nucleations: at random times and random spatial location x, $h \mapsto h + 1_x$, i.e., a tiny island of height 1 is nucleated on top of the existing interface. This occurs as a Poisson point process in space-time with rate 2 (the 2 is arbitrary, but people learned over time that this avoids strings of $2^{1/3}$ in the exact formulas); (2) a deterministic dynamics $h \mapsto \sup_{|y-x|\le t} h(y)$, i.e., islands spread out laterally at speed 1 and merge on contact (see Figure 3).

Almost equivalently, the model can be thought of as a collection of one-dimensional kinks (down steps) and antikinks (up steps) moving on $\mathbb{R}$. They move ballistically at speed 1: kinks to the right, antikinks to the left. They annihilate upon collision and are also produced in pairs, a kink at x^+ and an antikink at x^-, according to a rate 2 space-time Poisson process. The height $h(t, 0)$ is just its value $h(0, 0)$ at time 0, plus the number of kinks and antikinks which have crossed the spatial point 0 up to time t.

The *narrow wedge* at $y \in \mathbb{R}$ is the (legal) height function $\mathfrak{d}_y(y) = 0$ and $\mathfrak{d}_y(x) = -\infty$ for $x \neq y$. If we start PNG with $\mathfrak{d}_y$, we could call the resulting solution at time t, as a function of x, $\mathcal{A}_t(x, y)$. The PNG dynamics, under the natural coupling where one starts with two or more different initial conditions and runs them using the same Poisson noise, preserves the max operation

$$h(t, x; \max\{h_1, h_2\}) = \max\{h(t, x; h_1), h(t, x; h_2)\},$$

where $h(t, x; h_0)$ indicates the initial data is h_0. Hence one has the variational formula

$$h(t, x; h_0) = \sup_y \{\mathcal{A}_t(x, y) + h_0(y)\}. \tag{2.2}$$

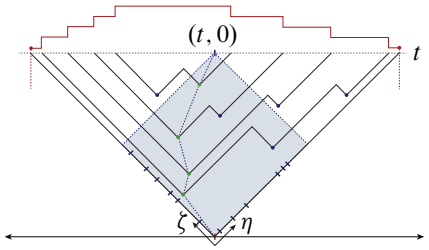


Figure 4. Longest increasing subsequence in PNG with narrow wedge initial data. From $\eta = t + x$ to $\zeta = t - x$, order of points in box corresponds to a permutation.

Clearly, from the variational formula, the narrow wedge solutions play a special role. If we start with one of them, say, $\mathfrak{d}_0$, one can see after a little thought that the height $h(t, 0)$ at position 0 at time t depends only on the Poisson points in the space-time 45° rotated square with vertices $(0, 0)$, $\left(\frac{t}{2}, \frac{t}{2}\right)$, $\left(\frac{t}{2}, -\frac{t}{2}\right)$ and $(t, 0)$, and in fact the height is the largest number of such points picked up by a path from $(0, 0)$ to $(t, 0)$, piecewise linear between the Poisson points, with slopes always ≤ 1 in absolute value. A collection of N points in a box generically give you a permutation in S_N: the way the order in one coordinate is transformed into the order in the other coordinate. The points picked up by one of our paths correspond to an increasing subsequence of the permutation. A collection of N points uniformly and independently chosen in a box correspond to a uniformly random permutation in S_N. The rate 2 Poisson points in our diamond of volume $\frac{1}{2}t^2$ correspond to choosing N points with $P(N = k) = \frac{t^{2k}}{k!}e^{-t^2}$ and then placing them uniformly and independently (see Figure 4). Hence $h(t, 0; \mathfrak{d}_0)$ is the length of the longest increasing subsequence of a random uniformly chosen permutation from S_N, where N has the Poisson distribution with rate t^2. In this way, the longest increasing subsequence problem can be seen as a special initial data for our particular growth model.

KPZ growth models like PNG do not really have an invariant measure because there is a non-trivial height shift. But we can talk about invariant measures modulo the absolute height. For PNG, these consist of continuous time random walks $h(x)$ which jump a unit step up at rate ρ and a unit step down at rate ρ^{-1}. The average slope $\rho - \rho^{-1}$ can be thought of as the average density of antikinks minus kinks. It is globally preserved, since they are created and annihilated in kink-antikink pairs, and parametrizes the set of invariant measures.

We will now show that PNG is a *stochastic integrable system*, by which we mean that its Markov transition probabilities are in-principle computable in somewhat the same way that one solves classical completely integrable systems.

From the variational formula (2.2), we obtain *skew time reversal invariance*

$$P(h(t, x; g) \leq -f) = P(h(t, x; f) \leq -g).$$

It is called that because you can think of it saying that the dynamics is preserved by $t \mapsto -t$, $h \mapsto -h$. There is another way to think about it. Let $\mathbf{1}_{f>g}$ be the indicator that $f(x) > g(x)$ for all $x \in \mathbb{R}$. Let L be the infinitesimal generator of the PNG process and L^* the infinitesimal generator of the "backwards" process $-h$. As the notation indicates, this is the adjoint of L with respect to the invariant measure, and if we start in such a measure, it is literally the process viewed backwards in time. Then the infinitesimal version of skew time reversal invariance is

$$L_f \mathbf{1}_{f>g} = L_g^* \mathbf{1}_{f>g},$$

where the subscripts in L_f and L_g^* indicate which variable the generators are acting on. Of course, this also says $L_f(\mathbf{1} - \mathbf{1}_{f>g}) = L_g^*(\mathbf{1} - \mathbf{1}_{f>g})$ and

$$\mathbf{1} - \mathbf{1}_{f>g} = g \text{ hits the hypograph of } h,$$

where the hypograph of $h = \{(x, y) \in \mathbb{R}^2 : y \le h(x)\}$ is closed if and only if h is upper semicontinuous. So it suggests considering the *hit kernel*

$$P_{a,b}^{\mathrm{hit}(h)}(u, v) = P(g \text{ hits the hypograph of } h \text{ on } [a, b] \text{ and } g(b) = v \mid g(a) = u),$$

where on the right-hand side, the probability measure on g is the invariant measure, i.e., the continuous time random walk $g(x)$ which jumps up at rate ρ and down at rate ρ^{-1}. It turns out not to matter what ρ is, so we take it to be 1 to be concrete. The generator of this random walk is then the discrete Laplacian

$$\Delta f(x) = f(x + 1) - 2f(x) + f(x - 1).$$

Let $E_{a,u;b,v}$ denote expectation with respect to random path g in $[a, b]$ with $g(a) = u$ and $g(b) = v$. Then,

$$L_h P_{a,b}^{\mathrm{hit}(h)}(u, v) = E_{a,u;b,v}[L_h(\mathbf{1} - \mathbf{1}_{g>h})] = E_{a,u;b,v}[L_g^*(\mathbf{1} - \mathbf{1}_{g>h})].$$

Here, L^* has the same invariant measures as L, so if it were not for the pinning at either end, the last term would vanish. If one looks at why such models can have explicit invariant measures, it is always because an integration by parts produces a telescoping series. So it is not hard to guess that with the pinned measure, the telescoping series will produce boundary terms. For PNG, they can be computed explicitly, and the result is

$$E_{a,u;b,v}[L_g^*(\mathbf{1} - \mathbf{1}_{g>h})] = 2[P_{a,b}^{\mathrm{hit}(h)}, \nabla].$$

Here, ∇ is the symmetric difference operator:

$$\nabla f(u) = \frac{1}{2}(f(u + 1) - f(u - 1)),$$

and $[A, B] = AB - BA$ is the *commutator* of A and B. If we define, for $a < b$,

$$T^h_{a,b} = e^{a\Delta} P^{\mathrm{hit}(h)}_{a,b} e^{-b\Delta},$$

then because Δ is the generator of the walk

$$\partial_t T^h_{x-t,x+t} = 0.$$

But this means that if we define, for $x, r \in \mathbb{R}$ and $t \geq 0$, with $\tau_r f(u) = f(u + r)$,

$$K(t, x, r; h) = \tau_r e^{-2t\nabla - x\Delta} T^h_{x-t,x+t} e^{2t\nabla + x\Delta} \tau_{-r}, \tag{2.3}$$

we have

$$(\partial_t - L)K = 0$$

as an operator, or we can think of them as integral kernels, and so a vast collection of solutions depending on parameters $u_1, u_2 \in \mathbb{Z}$.

Next we use determinants to build transition probabilities out of them. The PNG generator L is a sum of two pieces, a differential operator corresponding to the deterministic part of the dynamics with the kinks going one way and the anti-kinks going the other way, mutually annihilating on contact, and a creation operator $2\int dy(f(h + 1_y) - f(h))$. It is not hard to check from the definition that $K(h + 1_y) - K(h)$ is rank one. But then by linearity and the action of differential operators and rank one perturbations on determinants,

$$(\partial_t - L)\det(I - K) = \det(I - K)\,\mathrm{tr}((I - K)^{-1}(\partial_t - L)K),$$

so since $(\partial_t - L)K = 0$, we have

$$(\partial_t - L)\det(I - K) = 0$$

as well. In order to get the determinant to produce enough functions at $t = 0$ to identify the Markov transition probabilities, we have to extend it in the following way. For $x_1 < \cdots < x_n$, the *extended kernel* acting on the n-fold product of $\ell^2(\mathbb{Z}_{>0})$ whose i, jth matrix element is, with the notation $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{r} = (r_1, \ldots, r_n)$,

$$K^{\mathrm{ext}}_{ij}(t, \vec{x}, \vec{r}, h) = \tau_{r_i}(-1_{x_i < x_j} e^{(x_j - x_i)\Delta} + e^{-2t\nabla - x_i\Delta} T^h_{x_i - t, x_j + t} e^{2t\nabla - x_j\Delta})\tau_{-r_j}. \tag{2.4}$$

Here, we sometimes use $P^{\mathrm{hit}(h)}_{a,b}$ with $a > b$ which we just set to 0. Now let

$$U(t, h) = \det(I - K^{\mathrm{ext}})_{(\ell^2(\mathbb{Z}_{>0}))^n}.$$

The same argument shows we also have $(\partial_t - L)U = 0$, and setting $t = 0$ in (2.4), we get an upper triangular matrix with diagonal entries $\tau_{r_i} P^{\mathrm{hit}(h)}_{x_i,x_i} \tau_{-r_i}(u, v) =$

$1_{u=v\le h(x_i)-r_i}$. The determinant is just the product of the determinants of these diagonal entries, each one on $\ell^2(\mathbb{Z}>0)$ giving the indicator that $h(x_i)-r_i\ge 0$, i.e., we get $\prod_{i=1}^n 1_{r_i\le h(x_i)}$. Hence

$$U(0,h)=1_{h(x_i)\le r_i,\,i=1,\dots,n}.$$

As usual, finite-dimensional distributions determine the spatial stochastic process. In other words, we have solved the Kolmogorov equation for general h!

Theorem 2.1. *The finite-dimensional distributions of the transition probabilities of PNG are given by*

$$P_h(h(t,x_i)\le r_i, i=1,\dots,n)=\det(I-K^{\mathrm{ext}}(t,\vec{x},\vec{r},h))_{(\ell^2(\mathbb{Z}_{>0}))^n}.$$

Proof. We have shown they satisfy the Kolmogorov backward equation with the required initial data. So all we need is uniqueness of that equation. There is no general theorem of uniqueness for the Kolmogorov backward equation, but there is a general argument that uniqueness of the backward equation follows if we have existence of the adjoint equation. But here the adjoint is just the same process upside down, so our derivation of transition probabilities gives us the existence there as well, and therefore the uniqueness of our equation. ∎

We emphasize that this is a very unusual situation. Usually, in Markov process theory, the transition probabilities are defined as solutions of the Kolmogorov equation, but we do not expect to be able to write down the solutions.

Let us take as an example the narrow wedge initial data $\mathfrak{d}_0$, which as we saw leads to a Poissonized version of the longest increasing subsequence. Since the random walk g can only hit the hypograph of $\mathfrak{d}_0$ at the origin, for $a<0<b$,

$$T^{\mathfrak{d}_0}_{a,b}=e^{a\Delta}P^{\mathrm{hit}(\mathfrak{d}_0)}_{a,b}e^{-b\Delta}=\mathbf{1}_{\mathbb{Z}_{\le 0}}.$$

One can readily check that

$$e^{2t\nabla+x\Delta}(u_1,u_2)=e^{-2x}\left(\frac{t-x}{t+x}\right)^{\frac{u_2-u_1}{2}}J_{u_2-u_1}\big(2\sqrt{t^2-x^2}\big),$$

where J_n are Bessel functions of the first kind. If we set $\sqrt{t^2-x^2}=1$ and only worry about one-point distribution, we get

$$\begin{aligned}F_r&:=P(h(1,0;\mathfrak{d}_0)\le r)=\det(I-\tau_r B\tau_{-r})_{\ell^2(\mathbb{Z}_{>0})}\\&=e^{-s^2}\det(I_{i-j}(2s))_{i,j=0,\dots,r-1},\end{aligned}$$

where

$$B(u,v) = \frac{J_{u-1}(2)J_v(2) - J_u(2)J_{v-1}(2)}{u-v}$$

is the *discrete Bessel kernel* and I_n are modified Bessel functions of the first kind. The second equality is from [8]. The last is Gessel's formula [11] which eventually led to the Tracy–Widom asymptotics for the longest increasing subsequence (2.1).

The Toeplitz determinants $\det(I_{i-j}(2s))_{i,j=0,\dots,r-1}$ are intimately related to the orthogonal polynomials on the unit circle associated with the weight $\varphi_s(\zeta) = e^{s(\zeta+\zeta^{-1})}$. In the Szegő three-term recurrence for such orthogonal polynomials on the circle with respect to a general weight [27], only one coefficient α_n depends on the choice of weight. Thus, the α_ns, known as the *Verblunsky coefficients*, "encode" the weight and can be thought of as the analogue of reflection coefficients in this context. In the present case, they are related to F_r through

$$\alpha_r^2 = 1 - \frac{F_r F_{r+2}}{F_{r-1}^2}.$$

They are known ([7] or see [32, §3]) to satisfy the discrete Painlevé II equation

$$-s(1-\alpha_r^2)(\alpha_{r+1} + \alpha_{r-1}) = (r+1)\alpha_r$$

as well as the Ablowitz–Ladik lattice [1]

$$\partial_s \alpha_r = (1-\alpha_r^2)(\alpha_{r+1} - \alpha_{r-1}).$$

The Ablowitz–Ladik hierarchy is a reduction of the 2D Toda hierarchy [28], suggesting the generalization to arbitrary initial data. Setting $\zeta = \frac{1}{2}(t - \overline{x})$ and $\eta = \frac{1}{2}(t + \overline{x})$, where $\overline{x} = x_1 + \cdots + x_n$, $\partial_{\overline{x}} = \partial_{x_1} + \cdots + \partial_{x_n}$ and $r \pm 1 = (r_1, \dots, r_n) \pm (1, \dots, 1)$, it is not hard to check that the PNG kernel K from (2.4) satisfies for $u, v > 0$, and $r_i > \underline{r} = \sup_{|y-x|\le t} h_0(y)$, the height achieved at (t, x) without nucleations, the following:

(1) $K_r(u,v) = K_{r-1}(u+1, v+1)$.

(2) $\partial_\eta K_r(u,v) = K_{r-1}(u+1,v) - K_r(u+1,v)$.

(3) $\partial_\zeta K_r(u,v) = K_{r-1}(u,v+1) - K_r(u,v+1)$.

Then it is a general fact [19] that for $r > \underline{r}$, the $n \times n$ matrix

$$Q_r = (I - K_r)^{-1}(1,1)$$

satisfies the *non-Abelian* 2D *Toda equation*

$$\partial_\zeta(\partial_\eta Q_r Q_r^{-1}) + Q_r Q_{r-1}^{-1} - Q_{r+1} Q_r^{-1} = 0. \tag{2.5}$$

Theorem 2.2 ([19]). *Let h_0 be any deterministic initial data for PNG. The matrix function Q_r is related to the n point distributions of the PNG height at time $t > 0$ by*

$$\frac{F(t, x_1, \dots, x_n, r_1 + 1, \dots, r_n + 1)}{F(t, x_1, \dots, x_n, r_1, \dots, r_n)} = \det Q_r, \tag{2.6}$$

where $F_r(t, x_1, \dots, x_n, r_1, \dots, r_n) := \mathrm{Prob}_{h_0}(h(t, x_i) \le r_i,\ i = 1, \dots, n)$ and for $r_i > \underline{r}(t, x_i)$, they satisfy the non-Abelian 2D *Toda equation* (2.5).

The non-Abelian 2D Toda equation forms a well-known integrable system [20]. Lax and Zakharov–Shabat (zero-curvature) forms of the equation can be found in [31, §3.1] (see also [28]). They arise from compatibility of the equations for vectors Φ_r [21]

$$\partial_\eta \Phi_r = V_r \Phi_r + \lambda_r \Phi_{r-1}, \quad \partial_t \Phi_r = U_r \Phi_{r+1}$$

with

$$U_r = Q_r Q_{r-1}^{-1}, \quad V_r = -\partial_\eta Q_r Q_r^{-1}.$$

The non-Abelian 2D Toda equation is sometimes written in terms of U_r and V_r as

$$\partial_\eta U_r + V_r U_r - U_r V_{r-1} = 0, \quad \partial_\xi V_r + U_{r+1} - U_r = 0.$$

When $n = 1$, there is no need to take the determinant in (2.6), and the previous theorem reduces via $U_r = \frac{F_{r+1} F_{r-1}}{F_r^2}$ to the following more accessible statement.

Corollary 2.3. *Let h_0 be deterministic. Then for $t > 0$ and $r > \underline{r} = \sup_{|y-x| \le t} h_0(y)$, the one-point distributions of PNG, $F_r(t, x) := \mathrm{Prob}_{h_0}(h(t, x) \le r)$ satisfy the scalar* 2D *Toda equation*

$$\frac{1}{4}(\partial_t^2 - \partial_x^2) \log F_r = \frac{F_{r+1} F_{r-1}}{F_r^2} - 1. \tag{2.7}$$

In the case of flat initial data, $h_0 \equiv 0$, F_r is independent of x, the x derivative drops out of (2.7) and taking $g_r = \log F_r - \log F_{r-1}$ and $t \mapsto 2t$ we obtain the classic Toda lattice,

$$\ddot{g}_r = e^{g_{r+1} - g_r} - e^{g_r - g_{r-1}}.$$

3. The KPZ fixed point

Since we have a general formula for the transition probabilities of a model in the KPZ class, if we zoom out in the 1:2:3 scaling

$$\varepsilon^{\frac{1}{2}} h(\varepsilon^{-\frac{3}{2}} t, \varepsilon^{-1} x) - C_\varepsilon t,$$

we can expect to obtain the transition probabilities of the Markov process which is a fixed point of this rescaling. It was done first for TASEP [18, 26], but the rescaling is most transparent for the PNG formulas. The *strong KPZ universality conjecture* says that we should get the same answer for any model in the class. Of course, this can alternatively be thought of as the definition of the KPZ universality class. Certainly, we do get the same answer for several models which can be solved explicitly like this, and for which the limit has been computed.

3.1. Brownian scattering transform

Let us see how it all works with the PNG kernel (2.3). First of all, note that the underlying space $\mathbb{Z}_{>0}$ in which the Fredholm determinant is taken has the physical meaning of height. So it should be rescaled to $\varepsilon^{1/2}\mathbb{Z}_{>0}$ which is approximating $\mathbb{R}_+$. The operators ∇ and Δ act on functions on this space and are now of order $\varepsilon^{1/2}$ and ε, respectively. We can expand

$$\nabla = \varepsilon^{\frac{1}{2}}\partial + \varepsilon^{\frac{3}{2}}\frac{1}{6}\partial^3 + o(\varepsilon^{\frac{3}{2}}), \quad \Delta = \varepsilon\partial^2 + o(\varepsilon).$$

The operator ∂ acts on functions on $\mathbb{R}$. Now our formula for PNG rescales as

$$\begin{aligned} &P_{\varepsilon^{1/2}h_0(\varepsilon^{-1}x)}(\varepsilon^{\frac{1}{2}}h(\varepsilon^{-\frac{3}{2}}x_i, \varepsilon^{-1}t) - C_\varepsilon t \le r_i) \\ &\quad = \det(I - \varepsilon^{-1}K(\varepsilon^{-\frac{3}{2}}t, \varepsilon^{-1}\vec{x}, \varepsilon^{-\frac{1}{2}}(\vec{r} + C_\varepsilon t), \varepsilon^{\frac{1}{2}}h_0(\varepsilon^{-1}x)))_{\ell^2(\varepsilon^{1/2}\mathbb{Z}_{>0})}. \end{aligned}$$

The ε^{-1} in front of the K is because we want the measure on $\varepsilon^{-1/2}\mathbb{Z}_{>0}$ to give mass $\varepsilon^{1/2}$ to each site so that it approximates Lebesgue measure on $[0, \infty)$. We are assuming $\mathfrak{h}_0(x) = \lim_{\varepsilon\searrow 0}\varepsilon^{1/2}h_0(\varepsilon^{-1}x)$ and look at the three terms in the second part of the kernel (2.4) and their approximations up to terms vanishing in ε:

$$\underbrace{\tau_{\varepsilon^{-1/2}(r_i+C_\varepsilon t)}e^{-2\varepsilon^{-\frac{3}{2}}t\nabla - \varepsilon^{-1}x_i\Delta}}_{\tau_{r_i+C_\varepsilon t}\ e^{-2\varepsilon^{-1}t\partial - (t/3)\partial^3 - x_i\partial^2}} \underbrace{T^{\varepsilon^{\frac{1}{2}}h_0(\varepsilon^{-1}\cdot)}_{\varepsilon^{-1}x-\varepsilon^{-3/2}t, \varepsilon^{-1}x+\varepsilon^{-3/2}t}}_{\mathrm{T}^{\mathfrak{h}_0}} \cdot \underbrace{e^{2\varepsilon^{-\frac{3}{2}}t\nabla + \varepsilon^{-1}x_j\Delta}\tau_{-\varepsilon^{-1/2}(r_j+C_\varepsilon t)}}_{e^{2\varepsilon^{-1}t\partial + (t/3)\partial^3 + x_j\partial^2}\tau_{-r_j - C_\varepsilon t}} \cdot$$

In the first and third terms, there is a huge shift $\tau_{r_i+C_\varepsilon t}$ followed by $e^{-2\varepsilon^{-1}t\partial} = e^{-2\varepsilon^{-1}t\partial} = \tau_{-2\varepsilon^{-1}t}$ since we are on $\mathbb{R}$. So we take $C_\varepsilon = 2\varepsilon^{-1}$. In the middle one, the operator converges to what we call the *Brownian scattering transform* $\mathrm{T}^{\mathfrak{h}_0}$ of $\mathfrak{h}_0$. To summarize, we get in the 1:2:3 limit

$$K = e^{-\frac{t}{3}\partial^3 - x_i\partial^2 + r_i\partial}\mathrm{T}^{\mathfrak{h}_0}e^{\frac{t}{3}\partial^3 + x_j\partial^2 - r_j\partial}. \tag{3.1}$$

The presentation is a bit informal, and $\mathrm{T}^{\mathfrak{h}_0}$ is not really defined by itself. For $t \neq 0$, the operators on the outside are convolution operators with a completely explicit kernel in terms of the Airy function Ai:

$$e^{\frac{t}{3}\partial^3 + x\partial^2}(u_1, u_2) = t^{-\frac{1}{3}} e^{\frac{2x^3}{3t^2} - \frac{(u_1-u_2)x}{t}} \operatorname{Ai}(-t^{-\frac{1}{3}}(u_1 - u_2) + t^{-\frac{4}{3}}x^2). \tag{3.2}$$

Although it is not immediate, at $t = 0$ it is not too hard to check that it converges to the usual heat kernel

$$e^{x\partial^2}(u_1, u_2) = \frac{1}{\sqrt{4\pi x}} e^{-\frac{(u_2-u_1)^2}{4x}}.$$

As for $\mathrm{T}^{\mathfrak{h}_0}$, we are asymptotically computing the probability for a Brownian motion (of variance 2) to hit $\mathfrak{h}$ (see Figure 5) but surrounded by backward (!) heat equations

$$\mathrm{T}^{\mathfrak{h}} = \lim_{L\to\infty} e^{-L\partial^2} \mathrm{P}^{\mathrm{hit}(\mathfrak{h})}_{-L,L} e^{-L\partial^2}. \tag{3.3}$$

We do not really mean to make sense of (3.3) directly. What we really mean is that for any $t \neq 0$,

$$e^{-\frac{t}{3}\partial^3} \mathrm{T}^{\mathfrak{h}} e^{\frac{t}{3}\partial^3} = \lim_{L\to\infty} e^{-\frac{t}{3}\partial^3 - L\partial^2} \mathrm{P}^{\mathrm{hit}(\mathfrak{h})}_{-L,L} e^{\frac{t}{3}\partial^3 - L\partial^2}$$

makes sense (as a trace class operator on $L^2([r, \infty))$) and (3.1) is not really a product of operators but an informal way to write the family in $t, \vec{x}, \vec{r}$ of integral kernels $K(t, \vec{x}, \vec{r}, u_1, u_2)$ in u_1, u_2 which solve

$$\begin{gathered} \partial_t K = -\frac{1}{3}(\partial^3_{u_1} + \partial^3_{u_2})\, K, \quad \partial_{x_i} K = -\partial^2_{u_1} K, \quad \partial_{x_j} K = \partial^2_{u_2} K, \\ \partial_{r_i} K = \partial_{u_1} K, \quad \partial_{r_j} K = \partial_{u_2} K. \end{gathered} \tag{3.4}$$

When $\mathfrak{h}$ has compact support, which in this world means that it is equal to $-\infty$ outside a box $[-\ell, \ell]$, suppose $L \gg \ell$, and we want to compute

$$\lim_{L\to\infty} e^{-\frac{t}{3}\partial^3 - x_i\partial^2 + r_i\partial} e^{-L\partial^2} P^{\mathrm{hit}(\mathfrak{h})}_{-L,L} e^{-L\partial^2} e^{\frac{t}{3}\partial^3 + x_j\partial^2 - r_j\partial}. \tag{3.5}$$

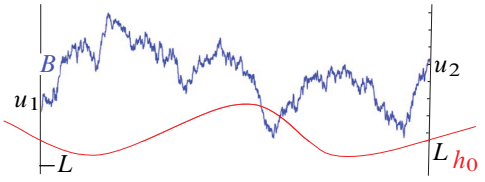


Figure 5. The quantity $P^{\mathrm{hit}(\mathfrak{h})}_{-L,L}(u_1, u_2)$ is the probability density that a Brownian motion $B(x)$ is at u_2 at time L given that it starts at u_1 at time $-L$ and that it hits the hypograph of $\mathfrak{h}$ along the way.

But the Brownian semigroup is $e^{t\partial^2}$, so since the Brownian particle is free on the intervals $[-L,-\ell)$ and $(\ell,L]$, we can write

$$P^{\mathrm{hit}(\mathfrak{h})}_{-L,L} = e^{(L-\ell)\partial^2} P^{\mathrm{hit}(\mathfrak{h})}_{-\ell,\ell} e^{(L-\ell)\partial^2},$$

and (3.5) becomes, even without taking the limit, completely well defined:

$$e^{-\frac{t}{3}\partial^3+(-x_i-\ell)\partial^2+r_i\partial} P^{\mathrm{hit}(\mathfrak{h})}_{-\ell,\ell} e^{\frac{t}{3}\partial^3+(x_j-\ell)\partial^2-r_j\partial}. \tag{3.6}$$

For example, in the case of one narrow wedge at 0, $P^{\mathrm{hit}(\mathfrak{h})}_{-\ell,\ell}$ is just $e^{\ell\partial^2}\mathbf{1}_{(-\infty,0]}e^{\ell\partial^2}$, and after adding in the term $-e^{(x_j-x_i)\partial^2+(r_i-r_j)\partial}1_{x_i<x_j}$, the limit of the first term in (2.4), one obtains from (3.2) the *extended Airy kernel*.

One could also, say, start with two narrow wedges at ± 1, in which case $P^{\mathrm{hit}(\mathfrak{h})}_{-\ell,\ell}$ is given by

$$e^{(\ell-1)\partial^2}\mathbf{1}_{(-\infty,0]}e^{2\partial^2}\mathbf{1}_{(-\infty,0]}e^{(\ell-1)\partial^2}.$$

This approach leads to a practical formula for any compactly supported data.

For noncompactly supported data, it looks like one has to take a limit. For example, for flat initial data $\mathfrak{h}_0 = 0$, one might take approximations which are 0 on $[-N,N]$ and $-\infty$ on $[-N,N]^C$, compute the scattering operator by the reflection principle and send $N\to\infty$. One finds, much to one's satisfaction, that the limiting operator T^0 is exactly the reflection operator $\varrho f(u) = f(-u)$. Then one can check by complex analysis that $e^{-(1/3)\partial^3}\varrho e^{(1/3)\partial^3}(u,v) = 2^{-1/3}\,\mathrm{Ai}(2^{-1/3}(u+v))$, the kernel for $F_{\mathrm{GOE}}(2^{-1/3}r)$; the rescaling by $2^{-1/3}$ represents the slight difference between the one-point distribution, at time 1 of the KPZ fixed point starting from $\mathfrak{h}_0\equiv 0$, and the limiting fluctuations of the top eigenvalue of a standard $N\times N$ matrix from the Gaussian orthogonal ensemble.

Writing the operator as a product such as in (3.1) can be misleading, especially as these operator products are on $L^2(\mathbb{R})$ while the Fredholm determinants are being taken on $L^2([0,\infty))$. Otherwise, the pre- and postfactors $e^{-(t/3)\partial^3}Te^{(t/3)\partial^3}$ would just be a conjugation and the determinant would not describe any non-trivial dynamics.

For general noncompactly supported data, one might wonder if a limiting operation always works. Actually, there is a trick to avoid taking any limit at all, which we call *splitting*. Recall that the hitting is being done from above, i.e., hitting $\mathfrak{h}$ means entering its hypograph. We can write, with the obvious notation

$$P^{\mathrm{hit}(\mathfrak{h})}_{-\ell,\ell} = e^{2\ell\partial^2} - P^{\mathrm{no\,hit}(\mathfrak{h})}_{-\ell,\ell}$$

and slip the two terms into (3.6). By the Markov property of Brownian motion,

$$P^{\mathrm{no\,hit}(\mathfrak{h})}_{-\ell,\ell}(u,v) = \int_{\mathfrak{h}(0)}^{\infty} dz\, P^{\mathrm{no\,hit}(\mathfrak{h})}_{-\ell,0}(u,z)\, P^{\mathrm{no\,hit}(\mathfrak{h})}_{0,\ell}(z,v).$$

Now one can go back and write $P^{\text{no hit}(\mathfrak{h})}_{-\ell,0} = e^{\ell\partial^2} - P^{\text{hit}(\mathfrak{h})}_{-\ell,0}$ and $P^{\text{no hit}(\mathfrak{h})}_{0,\ell} = e^{\ell\partial^2} - P^{\text{hit}(\mathfrak{h})}_{0,\ell}$. One of these sees $\mathfrak{h}^+$ which is just the function $\mathfrak{h}$ on $[0,\infty)$, and the other sees $\mathfrak{h}^-$ which is $\mathfrak{h}(-x)$, $x \in [0,\infty)$. Let $\tau_{\mathfrak{h}^+}$ and $\tau_{\mathfrak{h}^-}$ be the hitting times of $\mathfrak{h}^\pm$ of a Brownian motion starting at z. Since there has to be a first time it hits, after which we do not care, we can write

$$P^{\text{hit}(\mathfrak{h})}_{0,\ell}(z,v) = \int_0^\ell P_z(\tau_{\mathfrak{h}^+} \in dx) e^{(\ell-x)\partial^2}(B(x),v),$$

where B is the hitting Brownian motion. But now when we put it back into (3.6), all the ℓs cancel, and we can let $\ell \to \infty$ free of charge. Putting it all together, we get the extended kernel

$$\begin{aligned} \mathfrak{K}^{\text{ext}}_{ij}(t,\vec{x},\vec{r},\mathfrak{h}) = &-e^{(x_j-x_i)\partial^2+(r_i-r_j)\partial}1_{x_i<x_j} \\ &+ (S_{t,x_i,r_i} - S^{\mathfrak{h}^-}_{t,x_i,r_i})^*(S_{t,-x_j,r_j} - S^{\mathfrak{h}^+}_{t,-x_j,r_j}), \end{aligned} \tag{3.7}$$

where

$$S^{\mathfrak{h}}_{t,x,r}(z,v) = E_z[S_{t,x-\tau_{\mathfrak{h}},r}(B(\tau),v)], \quad S_{t,x,r} = e^{\frac{t}{3}\partial^3 + x\partial^2 - r\partial}.$$

One needs an upper bound on $\mathfrak{h}_0$ for it to make sense. For example, if $\mathfrak{h}_0(x) = ax^2$, it is not hard to see (for example, from the variational formula (1.8)) that the fixed point explodes to $+\infty$ at time $t = a^{-1}$. Height functions should naturally be upper semicontinuous, and a nice class of such functions is given by those that are bounded by $C_1 + C_2|x|$ for some $C_1, C_2 < \infty$. No lower bound is needed, and $\mathfrak{K}^{\text{ext}}$ given by (3.7) is a nice continuous map from this space equipped with the local Hausdorff topology into trace class operators [18].

The KPZ fixed point is the 1:2:3 scaling invariant Markov process of height functions which sits in the middle of the class, much as Brownian motion sits in the middle of the universality class of stochastic processes without long-range correlations which scale to it in the diffusive 1:2 scaling (see Figure 6). It is defined through its transition probabilities, whose finite-dimensional distributions are given by

$$P_{\mathfrak{h}_0}(\mathfrak{h}(t,x_1) \le r_1, \dots, \mathfrak{h}(t,x_n) \le r_n) = \det(I - \mathfrak{K}^{\text{ext}}(t,\vec{x},\vec{r},\mathfrak{h}_0))_{L^2([0,\infty))^n}.$$

It is proved in [18] that these are indeed transition probabilities and satisfy the Chapman–Kolmogorov equations. So the stochastic integrability is $\mathfrak{h}_0 \mapsto$ Brownian scattering transform $\mapsto$ linear evolution in height/space/time $\overset{\det}{\mapsto}$ transition probs, analogous to scattering/linear evolution/inverse scattering for classical integrable PDE.

The following basic properties all follow from the formula; they can also be proved by approximation from appropriate models:

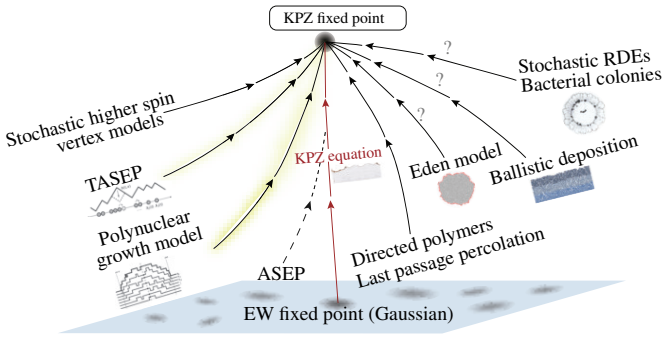


Figure 6. Schematic of the space of models with EW and KPZ fixed points under *zooming out*.

(1) *(1:2:3 invariance)* $\varepsilon^{1/2}\mathfrak{h}(\varepsilon^{-3/2}t, \varepsilon^{-1}x; \varepsilon^{-1/2}\mathfrak{h}_0(\varepsilon x)) \stackrel{\text{dist}}{=} \mathfrak{h}(t, x; \mathfrak{h}_0)$.

(2) *(Invariance of Brownian motion)* If $B(x)$ is a two-sided Brownian motion with diffusion coefficient 2, then for each $t > 0$, $\mathfrak{h}(t, x; B) - \mathfrak{h}(t, 0; B)$ is also two-sided Brownian motion in x with diffusion coefficient 2.

(3) *(Stationarity in space)* $\mathfrak{h}(t, x + u; \mathfrak{h}_0(x - u)) \stackrel{\text{dist}}{=} \mathfrak{h}(t, x; \mathfrak{h}_0)$.

(4) *(Reflection invariance)* $\mathfrak{h}(t, -x; \mathfrak{h}_0(-x)) \stackrel{\text{dist}}{=} \mathfrak{h}(t, x; \mathfrak{h}_0)$.

(5) *(Affine invariance)* $\mathfrak{h}(t, x; \mathfrak{h}_0(x) + a + cx) \stackrel{\text{dist}}{=} \mathfrak{h}(t, x + \frac{1}{2}ct; \mathfrak{h}_0(x)) + a + cx + \frac{1}{4}c^2t$.

(6) *(Skew time reversibility)* $P(\mathfrak{h}(t, x; \mathfrak{g}) \le -\mathfrak{f}(x), x \in \mathbb{R}) = P(\mathfrak{h}(t, x; \mathfrak{f}) \le -\mathfrak{g}(x), x \in \mathbb{R})$.

If we start from a narrow wedge at 0, the spatial process is called the *parabolic Airy process*

$$\mathcal{A}(x) = \mathfrak{h}(1, x; \mathfrak{d}_0).$$

If we add a parabola $\widehat{\mathcal{A}}(x) = \mathcal{A}(x) + x^2$, it turns out to be stationary in x; for example, for each x, $\widehat{\mathcal{A}}(x) \stackrel{\text{dist}}{=} F_{\text{GUE}}(x)$. You get the same thing if you take the large N limit of the top line of Dyson Brownian motion version of GUE, i.e., the eigenvalues of a dynamical $N \times N$ Hermitian matrix whose upper diagonal entries are independent stationary Ornstein–Uhlenbeck processes. From last passage models like PNG, the KPZ fixed point inherits the *preservation of max* property: For any $\mathfrak{f}_1, \mathfrak{f}_2$,

$$\mathfrak{h}(t, x; \mathfrak{f}_1 \vee \mathfrak{f}_2) \stackrel{\text{dist}}{=} \mathfrak{h}(t, x; \mathfrak{f}_1) \vee \mathfrak{h}(t, x; \mathfrak{f}_2).$$

The *parabolic Airy sheet* is the two-parameter process

$$\mathcal{A}(x, y) = \mathfrak{h}(1, y; \mathfrak{d}_x), \tag{3.8}$$

and its stationary version is $\widehat{\mathcal{A}}(x, y) = \mathcal{A}(x, y) - (x - y)^2$. Using stationarity in space, $\mathfrak{h}(1, y; \mathfrak{d}_x) \overset{\text{dist}}{=} \mathfrak{h}(1, y - x; \mathfrak{d}_0)$, and by reflection invariance, $\mathfrak{h}(1, y - x; \mathfrak{d}_0) \overset{\text{dist}}{=} \mathfrak{h}(1, x - y; \mathfrak{d}_0)$. Therefore,

$$\mathcal{A}(x, y) \overset{\text{dist}}{=} \mathcal{A}(y, x),$$

and the Airy sheet is an Airy process in each of its coordinates.

From the max property, the KPZ fixed point satisfies the variational formula (1.8). This is the closest thing we have to a stochastic equation for the KPZ fixed point, and it is incomplete because we do not exactly know what the joint distributions of $\mathcal{A}(x, y)$ are. Of course, many things about it can be understood by approximation from various models.

4. The KPZ fixed point and integrable PDE

Let $F(t, x, r)$ be the one-point distribution of PNG. We now perform the same scaling on it as a solution of the 2D Toda lattice. So we are interested in the limit $\mathfrak{F} = \lim_{\varepsilon \searrow 0} \mathfrak{F}_\varepsilon$, where the latter is defined by

$$F(t, x, r) = \mathfrak{F}_\varepsilon(\varepsilon^{\frac{3}{2}} t, \varepsilon x, \varepsilon^{\frac{1}{2}}(r - 2t)).$$

We expand $\mathfrak{F}_\varepsilon$ in Taylor series even though it really lives on $\varepsilon^{1/2}\mathbb{Z}$. Under this scaling, (2.7) becomes

$$\frac{1}{4}(\varepsilon^3 \partial_t^2 - \varepsilon^2 4 \partial_{\text{tr}}^2 + \varepsilon 4 \partial_r^2 - \varepsilon^2 \partial_x^2) \log \mathfrak{F} = \frac{\mathfrak{F}(\cdot + \varepsilon^{\frac{1}{2}})\mathfrak{F}(\cdot - \varepsilon^{\frac{1}{2}})}{\mathfrak{F}^2} - 1, \tag{4.1}$$

where for notational simplicity we removed the ε from the subscript in $\mathfrak{F}_\varepsilon$. Expanding (4.1) out and multiplying by $\mathfrak{F}^2$, we get

$$\varepsilon^2 \Big[\mathfrak{F} \partial_{\text{tr}}^2 \mathfrak{F} - \partial_t \mathfrak{F} \, \partial_r \mathfrak{F} + \frac{1}{4} \mathfrak{F} \, \partial_x^2 \mathfrak{F} - \frac{1}{4} (\partial_x \mathfrak{F})^2 + \frac{1}{12} \mathfrak{F} \, \partial_r^4 \mathfrak{F} - \frac{1}{3} \partial_r \mathfrak{F} \, \partial_r^3 \mathfrak{F} + \frac{1}{4} (\partial_r^2 \mathfrak{F})^2 \Big] + o(\varepsilon^2) = 0.$$

The vanishing of the square brackets constitutes the Hirota form of the Kadomtsev–Petviashvili equation KP-II.

Let $\mathfrak{K} = I - \mathfrak{K}^{\text{ext}}$, the fixed point kernel from (3.7), and $Q = \mathfrak{K}\mathfrak{K}^{\text{ext}}(0, 0)$ which is an $n \times n$ matrix-valued function of $t, x_1, \dots, x_n, r_1, \dots, r_n$, and the initial height profile $\mathfrak{h}_0$. Let $\mathcal{D}_r = \partial_{r_1} + \cdots + \partial_{r_n}$, $\mathcal{D}_x = \partial_{x_1} + \cdots + \partial_{x_n}$. One can check that for $\mathfrak{F} := P_{\mathfrak{h}_0}(\mathfrak{h}(t, x_1) \le r_1, \dots, \mathfrak{h}(t, x_n) \le r_n)$,

$$\mathcal{D}_r \log \mathfrak{F} = \operatorname{tr} Q,$$

in other words, the n-dimensional distributions can be reconstructed from Q. One can prove [24] the following as a general consequence for Fredholm determinants of kernels satisfying evolution equations (3.4).

Theorem 4.1. *The function Q and its derivative $q = \mathcal{D}_r Q$ solve the matrix Kadomtsev–Petviashvili (KP) equation*

$$\partial_t q + \frac{1}{2}\mathcal{D}_r q^2 + \frac{1}{12}\mathcal{D}_r^3 q + \frac{1}{4}\mathcal{D}_x^2 Q + \frac{1}{2}[q, \mathcal{D}_x Q] = 0,$$

where $[A, B] = AB - BA$.

In particular, the one-point marginals of the KPZ fixed point (i.e., in the case $n = 1$), $\phi = \partial_r^2 \log \mathfrak{F}$, satisfy the scalar KP-II equation

$$\partial_t \phi + \frac{1}{2}\partial_r \phi^2 + \frac{1}{12}\partial_r^3 \phi + \frac{1}{4}\partial_r^{-1}\partial_x^2 \phi = 0. \tag{4.2}$$

Classical examples include the following.

Example 4.2. For narrow wedge initial data $\mathfrak{h}_0 = \mathfrak{d}_0$, the distribution of $\mathfrak{h}(t, x) + \frac{x^2}{t}$ does not depend on x and should be the Tracy–Widom GUE scaled by $t^{1/3}$. So we look for a self-similar solution to KP (4.2) of the form

$$\phi^{\mathrm{nw}}(t, x, r) = t^{-\frac{2}{3}}\psi^{\mathrm{nw}}(t^{-\frac{1}{3}}r + t^{-\frac{4}{3}}x^2),$$

which gives the ODE

$$(\psi^{\mathrm{nw}})''' + 12\psi^{\mathrm{nw}}(\psi^{\mathrm{nw}})' - 4r(\psi^{\mathrm{nw}})' - 2\psi^{\mathrm{nw}} = 0. \tag{4.3}$$

The transformation $\psi^{\mathrm{nw}} = -q^2$ takes (4.3) into Painlevé II:

$$q'' = rq + 2q^3. \tag{4.4}$$

As $r \to -\infty$, the solution is approximately $\phi^{\mathrm{nw}}(t, x, r) \sim -\left(\frac{r}{2t} + \frac{x^2}{2t^2}\right)$, picking out the Hastings–McLeod solution $q(r) \sim -\operatorname{Ai}(r)$ as $r \to \infty$. Thus, we recover $F(t, x, r) = F_{\mathrm{GUE}}(t^{-1/3}r + t^{-4/3}x^2)$,

$$F_{\mathrm{GUE}}(s) = e^{-\partial^{-2}q^2} = \exp\left\{-\int_s^\infty du\,(u - s)q^2(u)\right\}.$$

Example 4.3. For flat initial data $\mathfrak{h}_0(x) \equiv 0$, there is no x dependence and KP-II (4.2) reduces to KdV. We look for a self-similar solution of KdV of the form

$$\phi^{\mathrm{fl}}(t, r) = \left(\frac{t}{4}\right)^{-\frac{2}{3}}\psi^{\mathrm{fl}}\left(\left(\frac{t}{4}\right)^{-\frac{1}{3}}\right)$$

which reduces it to the ODE

$$(\psi^{\mathrm{fl}})''' + 12(\psi^{\mathrm{fl}})'\psi^{\mathrm{fl}} - r(\psi^{\mathrm{fl}})' - 2\psi^{\mathrm{fl}} = 0.$$

Miura's transform

$$\psi^{\mathrm{fl}} = \frac{1}{2}(q' - q^2)$$

brings this to Painlevé II (4.4), with the same behaviour as $r \to -\infty$ as in the previous example. So we recover $F(t, x, r) = F_{\mathrm{GOE}}(4^{1/3}t^{-1/3}r)$,

$$F_{\mathrm{GOE}}(r) = e^{-\partial^{-2}\frac{1}{2}(q'+q^2)} = \exp\left\{-\frac{1}{2}\int_r^\infty du\, q(u)\right\} F_{\mathrm{GUE}}(r)^{\frac{1}{2}}.$$

These distributions arose earlier as the limits of rescaled top eigenvalues from, respectively, the Gaussian unitary and Gaussian orthogonal ensembles. But unlike other limit distributions in probability, they have no scaling invariance. KP is invariant under

$$\phi(t, x, r) \mapsto \alpha^{-2}\phi(\alpha^{-3}t, \alpha^{-2}x, \alpha^{-1}r), \quad \mathfrak{h}_0(x) \mapsto \alpha^{-1}\mathfrak{h}_0(\alpha^2 x).$$

This is not surprising; we essentially derived it by taking this scaling. So now we have embedded the Tracy–Widom distributions into a system which does have scaling invariance, i.e., we have enlarged our perspective enough to see the underlying invariance principle.

This begs the question: Where do the KP equations come from, and what do they mean? One of the key characteristics of KP is solitons. But our solutions have no solitons at all. Universal non-linear partial differential equations like KP and KdV, or the non-linear Schrödinger equation, are generally obtained from more complicated equations by rescalings, asymptotic expansions and/or ansatz of special forms. Unlike other basic PDEs like the heat and wave equations, Burgers or the Euler equations, they are not derived from first- principle physical mechanisms, but only as model equations governing the non-linear effects in certain asymptotic regimes. It is all "this is what you get in the limit" as opposed to "the stuff is doing this". For example, KP is usually derived from water wave equations in the limit of long waves with small amplitude and slow dependence on the transverse coordinate. KdV is derived from one-way travelling waves in Fermi–Pasta–Ulam, etc. Since these model equations appear as limits of so many equations, they have an opportunity to inherit integrability in the scaling limit from even just one integrable example within their domain of attraction. So it is not surprising that many of them are integrable.

Our case is unusual. There is no "more complicated equation" reducing to KP in the scaling limit. And the KP equations are discovered not through the scaling limit, but through the exact formulas themselves, which are then recognized as examples of inverse scattering solutions of KP. None of this was very easy to predict.

One could perhaps think of the full evolution equation, for, say, PNG, as the more complicated equation. But this just leads to a bigger mystery. It is certainly not true that one had any reason to expect that one-point functions for PNG satisfy a closed equation. In general, in a many-body problem like this, one expects to write down an infinite hierarchy of equations, for example, for n particle density functions, the so-called BBGKY hierarchy. So one thing to keep in mind is that the biggest surprise is not that one-point distributions of the KPZ fixed point are governed by KP, but that they are governed by any closed equation at all.

The KPZ equation itself is not stochastically integrable, though it seems to have partial solvability. It can be obtained in the limit of other models as a weakly asymmetric or weak noise limit in which the asymmetry or noise is scaled down as we scale the solution with diffusive scaling. Some of the approximating models are themselves partially solvable. For example, for the asymmetric exclusion process, formulas for narrow wedge and step Bernoulli initial data were obtained in the breakthrough works [29, 30]. From the formulas, one has convergence to the KPZ fixed point for these special initial data. For general initial data, it is a very recent result [2]. An earlier attempt [25] contains a mistake in a shift argument used to identify the limiting fluctuations (however, it does prove tightness and even works for non-nearest neighbour exclusion height functions in the 1:2:3 scaling, a model that is not even partially solvable). That the KPZ equation converges to the KPZ fixed point, i.e., it is a member of the universality class bearing its name, is proved in [33, 34]. The key challenge for the future is to devise methods to prove the convergence to the KPZ fixed point for the larger class of non-solvable KPZ models.

Funding. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] M. Adler and P. van Moerbeke, Integrals over classical groups, random permutations, Toda and Toeplitz lattices. *Comm. Pure Appl. Math.* **54** (2001), no. 2, 153–205 Zbl 1086.34545 MR 1794352

[2] A. Aggarwal, I. Corwin, and M. Hegde, KPZ fixed point convergence of the ASEP and stochastic six-vertex models. 2024, arXiv:2412.18117v1

[3] J. Baik, P. Deift, and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.* **12** (1999), no. 4, 1119–1178 Zbl 0932.05001 MR 1682248

[4] M. Balázs, J. Quastel, and T. Seppäläinen, Fluctuation exponent of the KPZ/stochastic Burgers equation. *J. Amer. Math. Soc.* **24** (2011), no. 3, 683–708 Zbl 1227.60083 MR 2784327

[5] G. Barraquand and P. Le Doussal, Steady state of the KPZ equation on an interval and Liouville quantum mechanics. *Europhysics Lett.* **137** (2022), no. 6, article no. 61003, 8 pp.

[6] L. Bertini and G. Giacomin, Stochastic Burgers and KPZ equations from particle systems. *Comm. Math. Phys.* **183** (1997), no. 3, 571–607 Zbl 0874.60059 MR 1462228

[7] A. Borodin, Discrete gap probabilities and discrete Painlevé equations. *Duke Math. J.* **117** (2003), no. 3, 489–542 Zbl 1034.39013 MR 1979052

[8] A. Borodin and A. Okounkov, A Fredholm determinant formula for Toeplitz determinants. *Integral Equations Operator Theory* **37** (2000), no. 4, 386–396 Zbl 0970.47014 MR 1780118

[9] I. Corwin, Some recent progress on the stationary measure for the open KPZ equation. In E. Basor, A. Böttcher, T. Ehrhardt, and C. A. Tracy (eds.), *Toeplitz operators and random matrices. In Memory of Harold Widom*, pp. 321–360, Oper. Theory Adv. Appl. 289, Birkhäuser, Cham, 2022 Zbl 07827908 MR 4573955

[10] D. Forster, D. R. Nelson, and M. J. Stephen, Large-distance and long-time properties of a randomly stirred fluid. *Phys. Rev. A (3)* **16** (1977), no. 2, 732–749 MR 0459274

[11] I. M. Gessel, Symmetric functions and P-recursiveness. *J. Combin. Theory Ser. A* **53** (1990), no. 2, 257–285 Zbl 0704.05001 MR 1041448

[12] Y. Gu and J. Quastel, Integration by parts and invariant measure for KPZ. *Trans. Amer. Math. Soc.* **379** (2026), no. 2, 1435–1452 Zbl 08149756 MR 5019586

[13] M. Gubinelli, P. Imkeller, and N. Perkowski, Paracontrolled distributions and singular PDEs. *Forum Math. Pi* **3** (2015), article no. e6, 75 pp. Zbl 1333.60149 MR 3406823

[14] M. Hairer, Solving the KPZ equation. *Ann. of Math. (2)* **178** (2013), no. 2, 559–664 Zbl 1281.60060 MR 3071506

[15] M. Hairer, A theory of regularity structures. *Invent. Math.* **198** (2014), no. 2, 269–504 Zbl 1332.60093 MR 3274562

[16] S. P. Hastings and J. B. McLeod, A boundary value problem associated with the second Painlevé transcendent and the Korteweg–de Vries equation. *Arch. Ration. Mech. Anal.* **73** (1980), no. 1, 31–51 Zbl 0426.34019 MR 0555581

[17] M. Kardar, G. Parisi, and Y.-C. Zhang, Dynamic scaling of growing interfaces. *Phys. Rev. Lett.* **56** (1986), no. 9, 889–892 Zbl 1101.82329

[18] K. Matetski, J. Quastel, and D. Remenik, The KPZ fixed point. *Acta Math.* **227** (2021), no. 1, 115–203 Zbl 1505.82041 MR 4346267

[19] K. Matetski, J. Quastel, and D. Remenik, Polynuclear growth and the Toda lattice. *J. Eur. Math. Soc. (JEMS)* (2025), DOI 10.4171/JEMS/1558

[20] A. V. Mikhailov, The reduction problem and the inverse scattering method. *Phys. D* **3** (1981), no. 1–2, 73–117 Zbl 1194.37113

[21] J. J. C. Nimmo and R. Willox, Darboux transformations for the two-dimensional Toda system. *Proc. Roy. Soc. A* **453** (1997), no. 1967, 2497–2525 Zbl 0890.35147 MR 1490247

[22] G. Parisi, On the replica approach to random directed polymers in two dimensions. *J. Phys. France* **51** (1990), no. 15, 1595–1606 MR 1063802

[23] Y. Pomeau and P. Résibois, Time dependent correlation functions and mode-mode coupling theories. *Phys. Rep.* **19** (1975), no. 2, 63–139

[24] J. Quastel and D. Remenik, KP governs random growth off a 1-dimensional substrate. *Forum Math. Pi* **10** (2022), article no. e10, 26 pp. Zbl 1502.60161 MR 4413350

[25] J. Quastel and S. Sarkar, Convergence of exclusion processes and the KPZ equation to the KPZ fixed point. *J. Amer. Math. Soc.* **36** (2023), no. 1, 251–289 Zbl 1520.60063 MR 4495842

[26] D. Remenik, Integrable fluctuations in the KPZ universality class. In D. Beliaev and S. Smirnov (eds.), *Proceedings of the International Congress of Mathematicians (ICM), Vol. 6. Sections 12–14*, pp. 4426–4450, EMS Press, Berlin, 2023 Zbl 1533.60182 MR 4680409

[27] G. Szegő, *Orthogonal polynomials*. 4th edn., Amer. Math. Soc. Colloq. Publ. 23, American Mathematical Society, Providence, RI, 1975, 432 pp. Zbl 0305.42011 MR 0372517

[28] K. Takasaki, Toda hierarchies and their applications. *J. Phys. A* **51** (2018), no. 20, article no. 203001, 35 pp. Zbl 1406.37053 MR 3803581

[29] C. A. Tracy and H. Widom, A Fredholm determinant representation in ASEP. *J. Stat. Phys.* **132** (2008), no. 2, 291–300 Zbl 1144.82045 MR 2415104

[30] C. A. Tracy and H. Widom, Asymptotics in ASEP with step initial condition. *Comm. Math. Phys.* **290** (2009), no. 1, 129–154 Zbl 1184.60036 MR 2520510

[31] K. Ueno and K. Takasaki, Toda lattice hierarchy. In K. Okamoto (ed.), *Group representations and systems of differential equations (Tokyo, 1982)*, pp. 1–95, Adv. Stud. Pure Math. 4, North-Holland, Amsterdam, 1984 Zbl 0577.58020 MR 0810623

[32] W. Van Assche, *Orthogonal polynomials and Painlevé equations*. Austral. Math. Soc. Lect. Ser. 27, Cambridge University Press, Cambridge, 2017, 179 pp. Zbl 1387.33001 MR 3729446

[33] B. Virag, The heat and the landscape I. 2020, arXiv:2008.07241v1

[34] X. Wu, The KPZ equation and the directed landscape. *Ann. Probab.* **54** (2026), no. 3, 1212–1257 Zbl 08200773 MR 5061894

Jeremy Quastel
Department of Mathematics, University of Toronto, 40 St. George Street, Toronto, ON, M5S 1L2, Canada; quastel@math.toronto.edu

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/44

MUMPS: MUltifrontal Massively Parallel Solver for the direct solution of sparse linear equations

Patrick R. Amestoy, Jean-Yves L'Excellent, Theo Mary, and Chiara Puglisi

Abstract. The scientific library MUMPS (MUltifrontal Massively Parallel Solver) solves large systems of sparse linear equations using a direct method. It is an open-source software library distributed under the CeCILL-C license which can be downloaded from the MUMPS solver website.

Direct methods are widely appreciated for their numerical robustness, reliability, and ease of use. They are commonly used either to compute the solution or as a preconditioner within iterative methods. However, they are also known for their high computational complexity.

In this communication, we briefly describe the main features of the MUMPS solver. We then explain how data sparsity with block low-rank (BLR) approximations and mixed precision can be used to reduce the complexity of direct methods. We conclude by showing results on large real-world applications (matrices of order several hundred million) and describing ongoing research work.

1. Introduction

In many areas of numerical simulation (such as geophysics, structural mechanics, economy, chemistry, and health), the solution of sparse linear systems is a critical and expensive part in terms of calculation time and memory footprint. In fact, the initial problem, often non-linear, is commonly approximated by a succession of linear problems which, depending on the formulation and the modeling carried out, can be excessively difficult to solve numerically. Actual problems provided by our users and industrial customers, whether they come from the study of the aging of nuclear power plants, the study of noise inside the passenger compartment of a car, or the study of the deformation and wear of a tire, are numerically challenging. It should be noted

Mathematics Subject Classification 2020: 65F05 (primary); 65F50, 65F55, 65Y05 (secondary).
Keywords: sparse linear algebra, direct methods, low-rank approximation, mixed-precision computation, parallel computation.

that a numerically difficult problem will often lead to longer calculation times, higher memory usage, and significant power consumption.

Today, applications require solving linear systems of increasingly large sizes (several hundred million unknowns), allowing complex problems to be addressed while maintaining performance and numerical stability. In this context, it is necessary to design linear algebra libraries that know how to exploit the computers, regardless of the size of the problem, the architecture, and the power of the computer (from a laptop with a few cores to supercomputers with hundreds of thousands of heterogeneous computing units).

The `MUMPS` library solves large sparse linear systems of equations, $AX = B$, using direct methods based on the multifrontal method [22, 23]. The multifrontal method [10] organizes the operations that take place during the factorization of sparse matrices in such a way that the entire factorization is performed through partial factorizations of a sequence of dense and small submatrices (*frontal matrices*). It is guided by a tree, the so-called *elimination tree* [32], that represents the dependencies between those partial factorizations.

The matrix A is a sparse square matrix, and X and B are vectors or matrices whose sparsity can also be exploited. Please note that the solution X is, in general, not sparse even if A and B are sparse. The sparsity of X comes from the fact that in many applications, the user is only interested in part of the solution. There are two main classes of methods to solve such problems:

- *direct methods* [21], based on a factorization of A (see [27]), for example, in the form LU, LDL^T, where L and U are (respectively, lower and upper) triangular matrices and D is a diagonal (or block diagonal) matrix when Gaussian elimination is used to factor the matrix, or in the form QR, where Q is an orthogonal matrix and R an upper triangular matrix in case of orthogonal factorization;
- *iterative methods*, which build a series of iterates converging to the solution and often rely on a preconditioner for which a direct method may also play a role.

Direct methods, in general, are more costly (in terms of number of operations and memory usage) than iterative methods but are more numerically robust.

Among direct methods, the multifrontal methods compute a sequence of partial factorizations of smaller dense frontal matrices following the elimination tree. Multifrontal methods, like many direct methods, use an approach in three phases to solve $AX = B$: analysis, factorization, and solve.

(1) Analysis: The undirected graph associated with the matrix A is analyzed. The undirected graph of A is defined on the symmetrized matrix $A + A^T$; it has as many nodes as columns in the matrix and an edge between nodes i and j if and only if a_{ij} is non-zero. A permutation (*ordering*) that reduces the number of operations during the factorization is calculated. The analysis also

predicts the data structures for the factorization. This requires carrying out a symbolic factorization to determine the frontal matrices, to calculate the task dependency tree, and to partially assign tasks to processors.

(2) Factorization: The permuted matrix is factorized in the form LU or LDL^T, depending on the properties of matrix A. Numerical pivoting based on partial pivoting is commonly used (see [29, Theorem 9.5]) to preserve the numerical stability of LU factorization. Note that due to partial pivoting, the data structures predicted during the analysis can be modified during the factorization phase. The factorization is generally the most computationally intensive part.

(3) Solve: Triangular systems are solved to compute a solution. Post-processing can also be applied, such as iterative refinement and/or error analysis.

The MUMPS library implements robust and efficient parallel algorithms, taking into account the following objectives and constraints:

(1) Matrix factorizations are more stable when numerical pivoting is allowed. Efficient implementations are challenging in a parallel distributed environment. One difficulty with sparse matrices is that numerical pivoting results in a modification during the factorization phase of the task dependency graph computed during the analysis phase. Furthermore, the stability of a pivot must be checked against other entries in the matrix, which is not trivial because it often requires costly interprocessor communications.

(2) Large-scale high-performance computers generally have their memory physically distributed. Therefore, the message-passing paradigm (MPI) is used in combination with the shared-memory paradigm (OpenMP) to address efficiently computing nodes that share memory between a large number of cores. Exploiting GPU accelerators is an ongoing work.

(3) Architecture evolutions make it very hard or even impossible to perfectly predict the behavior of current computing platforms, with several levels of caches, high-performance redundant networks with a complex topology, some degree of heterogeneity between processors or at the network level, varying load in possibly multi-user environments, etc. Furthermore, the characteristics of the graph of tasks used in MUMPS are not fully predictable when numerical pivoting is allowed (see previous point). Therefore, adaptive approaches are used, with so-called *dynamic scheduling*. Static information is used to reduce the space to be explored by the dynamic schedulers.

(4) There are often time constraints in numerical simulation. Therefore, it is critical that the algorithms reduce the time for solution (or *makespan*) as much as possible on a given number of processors.

(5) If the problem is large, special attention must be paid to memory which, if not correctly used, is a very strong limiting factor for parallel direct methods. The order in which the tasks are scheduled has a strong impact not only on the makespan but also on memory, and unfortunately, more parallelism often leads to more global memory used. Therefore, it is crucial to take these two conflicting criteria into account when designing scheduling strategies for the computational tasks.

With these objectives in mind, in the `MUMPS` solver, a fully asynchronous and distributed scheduling approach [13] to process the graph of tasks during the factorization has been designed, and several types of pivoting strategies have been implemented.

In recent years, the `MUMPS` group has developed a new low-rank method called block low-rank (BLR) [1, 11] to exploit data sparsity. For a wide range of problems arising from elliptical PDEs, we have proved in [8] that we can reduce the complexity of sparse solvers in terms of the number of floating-point operations and memory usage of the numerical phases and in particular of the factorization phase. In [30], the BLR method was proved to be backward stable. This means that the computed solution is the exact solution of a nearby problem with a small perturbation of the input data (in our case, the matrix A and the right-hand side B). We will describe in Section 2 the low-rank approximation approaches, introducing the block low-rank method.

Moreover, low-rank approximations exhibit remarkable properties that allow for the use of mixed-precision arithmetic without loss of accuracy. For any unitarily invariant norm, the optimal low-rank approximation is the truncated singular value decomposition (SVD), a result known as the Eckart–Young theorem [24]. During BLR compression, an SVD decomposition of the block to be compressed is computed, and we have shown in [2] that the singular vectors associated with small singular values can be safely stored in lower precisions while preserving high accuracy overall. This will be described in Section 3.

In Section 4, we present some promising results on real-world applications combining BLR compression and mixed-precision algorithms. Tracks for future research are mentioned as concluding remarks in Section 5.

2. Low-rank approximations

Motivations. In many applications (such as those coming from the discretization of elliptic partial differential equations), the matrix A has been shown to have a low-rank property: conveniently defined off-diagonal blocks of its Schur complements can be approximated by low-rank products [16]. Several formats have been

proposed to exploit this property, mainly differing on whether they use a strong or weak admissibility condition, whether they have a nested basis property, and whether they use a hierarchy of levels. The most general of the hierarchical formats is the $\mathcal{H}$-matrix format [16, 17, 28], which is non-nested and strongly admissible. The $\mathcal{H}^2$-matrix format [17] is its nested counterpart. HODLR matrices [14] are based on the weak admissibility condition, and HSS [19, 37] and the closely related HBS [26] additionally possess nested bases. These low-rank formats can be exploited within dense or sparse solvers to provide a substantial reduction of their complexity, which is quasilinear even for dense solvers.

In the MUMPS solver, we exploit a different low-rank format: the block low-rank (BLR) format [1]. Unlike hierarchical formats, the BLR format is based on a flat, non-hierarchical blocking of the matrix. We have shown in [8] that despite being non-hierarchical, the BLR format can still achieve a substantial complexity reduction. Moreover, when applied to sparse solvers, the complexity reductions from exploiting sparsity of the matrix A and data sparsity thanks to BLR approximations compound, leading to complexities as low as $O(n \log n)$ for storage and $O(n^{4/3})$ for floating-point operations (for 3D problems with constant ranks), which are not so far from the $O(n)$ hierarchical ones. Moreover, the constant prefactor hidden in the $O(\cdot)$ notation is higher for hierarchical formats, and hierarchical formats can also be more complex to implement, especially in parallel, fully featured, general-purpose sparse solvers. For these reasons, the BLR format has been preferred in most sparse direct solvers, including MUMPS and PaStiX [36]; STRUMPACK [25] adopts a composite approach [20] which combines the BLR format with hierarchical ones.

The block low-rank format. A block low-rank (BLR) representation $\widetilde{A}$ of a dense matrix A has the $p \times p$ block form

$$\widetilde{A} = \begin{bmatrix} A_{11} & \widetilde{A}_{12} & \cdots & \widetilde{A}_{1p} \\ \widetilde{A}_{21} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ \widetilde{A}_{p1} & \cdots & \cdots & A_{pp} \end{bmatrix}, \tag{2.1}$$

where off-diagonal blocks A_{ij} of size $n_i \times n_j$ are approximated by low-rank matrices $\widetilde{A}_{ij} = X_{ij} Y_{ij}^T$ of rank k_{ij}, where $X_{ij} \in \mathbb{R}^{n_i \times k_{ij}}$ and $Y_{ij} \in \mathbb{R}^{n_j \times k_{ij}}$. If the ranks k_{ij} are chosen such that $\|\widetilde{A}_{ij} - A_{ij}\| \leq \varepsilon \|A\|$, we obtain a BLR approximation satisfying a relative accuracy of order ε, $\|\widetilde{A} - A\| \leq c\varepsilon \|A\|$, where c is a constant that depends on the choice of norm (for example, for the Frobenius norm, $c = p$). If the rank k_{ij} is small enough so that $k_{ij}(n_i + n_j) < n_i n_j$, storing the low-rank factors X_{ij} and Y_{ij} requires fewer entries than storing the full block A_{ij}, and A_{ij} is referred to as a low-rank block. Otherwise, it is more efficient to keep A_{ij} as is, and it is referred to as

a full-rank block. The BLR representation is not only useful for reducing storage; it can also be exploited to reduce the cost of computations, in particular that of LU factorization. As explained above and proved in [8], BLR compression allows for an asymptotic complexity reduction, which means that the larger the matrix, the greater the storage gain.

The $\widetilde{A}_{ij}$ approximation of each block can be computed in different ways, such as an SVD computation or a truncated QR orthogonal factorization with column pivoting [18]. A truncated QR corresponds to a QR factorization with pivoting which is truncated as soon as the norm of the trailing submatrix falls below the prescribed threshold $\varepsilon\|A\|$.

This ε parameter, called the low-rank threshold, defines the accuracy of the low-rank approximations. An important question is how the use of BLR compression affects the stability of LU factorization. Indeed, while the threshold ε straightforwardly controls the blockwise truncation errors, the question of how these errors propagate within the LU factorization is far from obvious. It has been proved in [30] that the ε parameter can be used to control the backward error of the final solution. Hence the BLR LU factorization is backward stable with respect to ε.

Because the multifrontal method relies on dense factorizations, the BLR approximations can be easily incorporated into the multifrontal factorization by representing the frontal matrices as BLR matrices as defined by equation (2.1).

High-performance implementation. While the reduced complexity of BLR LU factorization should in principle lead to a significantly faster execution time, translating these theoretical gains into actual speedups is quite challenging, especially in a parallel computing environment. One of the main reasons is that BLR computations tend to be memory (or communication) bound. Indeed, while BLR approximations reduce both flops and memory costs, the former are reduced by a much larger factor than the latter. Moreover, BLR approximations not only require partitioning the matrix into smaller blocks but also replace these blocks by low-rank products of even smaller dimensions. The granularity, i.e., the size of a task in terms of number of floating-point operations, of BLR computational tasks is therefore radically smaller than that of standard dense linear algebra.

In order to achieve efficiency, it is therefore necessary to rethink standard algorithms to reduce the amount of data movement or communications. `MUMPS` is based on such *communication-avoiding* implementations of the BLR LU factorization described in [11]. They are based on using a specific order of computations (left-looking factorization) that minimizes accesses to uncompressed data and on grouping low-rank blocks in order to increase their granularity and hence the compute intensity of the factorization. Our performance experiments with `MUMPS` presented in [11] confirm that these optimizations lead to significant speedups. In a more recent work [3],

we have also extended these ideas to accelerate the BLR triangular solution step with many right-hand sides by minimizing accesses to the uncompressed right-hand sides.

3. Mixed-precision computations

The increasing support for lower-precision arithmetics in hardware provides new opportunities for high-performance scientific computing. However, even though low-precision arithmetics can provide significant speed, communication, and energy benefits, its use in scientific computing poses the challenge of preserving the accuracy and stability of the computation.

To address this issue, a variety of mixed-precision algorithms that combine low and high precisions have emerged.

Adaptive-precision BLR. Block low-rank approximations allow for the use of mixed-precision arithmetic without loss of accuracy. During BLR compression, an SVD decomposition of the block to be compressed is computed. We have shown in [2] that the singular vectors associated with small singular values can be safely stored in lower precisions while preserving high accuracy overall. Given p available precision formats with unit roundoffs $u_1 \leq \varepsilon < u_2 < \cdots < u_p$, for any off-diagonal low-rank block $\widetilde{A}_{ij} = U\Sigma V^T$, this leads to an adaptive-precision representation of the form

$$\widehat{A}_{ij} = [\widehat{U}_1 \cdots \widehat{U}_p] \begin{bmatrix} \widehat{\Sigma}_1 & & \\ & \ddots & \\ & & \widehat{\Sigma}_p \end{bmatrix} [\widehat{V}_1 \cdots \widehat{V}_p]^T,$$

where $\widehat{U}_k$, $\widehat{\Sigma}_k$, and $\widehat{V}_k$ are stored in precision u_k. According to the theory developed in [2], the singular values should be partitioned such that $\|\Sigma_k\| \leq \varepsilon \|A\|/u_k$. With this criterion, we have $\|\widehat{A}_{ij} - \widetilde{A}_{ij}\| \leq (2p-1)\varepsilon\|A\|$, and so a preserved accuracy of order ε is guaranteed.

For typical BLR matrices, the singular values of the low-rank blocks decay rapidly. As a result, a large part of the coefficients and operations can be switched to lower-precision arithmetics. This leads to significant storage savings as we will illustrate in the next section.

Iterative refinement and MUMPS as a preconditioner. Another possible usage of MUMPS in mixed precision is to compute the LU factorization in a lower precision, use this lower-precision LU factorization to obtain an approximate initial solution, and use an iterative solver in higher precision to refine the accuracy of the solution. In particular, we have shown in [6] that, as long as the problem is not too ill-conditioned, LU-based iterative refinement can deliver full double-precision accuracy even with

a single-precision factorization, reducing the time and memory costs of the solver by a factor of 2×. Moreover, this iterative refinement can also be combined with a BLR factorization (with potentially very coarse accuracy ε) for even greater time and memory savings.

In order to handle (potentially very) ill-conditioned problems, LU-based refinement can be replaced by more robust variants based on Krylov iterative solvers such as GMRES [7]. The low-precision LU factors computed by `MUMPS` can then be used as a robust preconditioner.

4. Numerical results

We have been able to successfully use the mixed-precision block low-rank feature of `MUMPS` [4, 33] in the context of collaborations with the WIND project on full-waveform inversion applications (see [33]) and with TotalEnergies-Inria Makutu on hybridizable discontinuous Galerkin discretization for 3D wave propagation (see [4, 5]).

We focus here on results obtained with full-waveform inversion of the Gorgon model with the WIND project (see [33, 34]).

High-resolution imaging has many applications in civil engineering, oil exploration, earthquake seismology, medical imaging, non-destructive control testing, etc. The so-called full-waveform inversion (FWI) has become the baseline imaging method in many application domains with the development of new acquisition technologies and the continuous advances in high-performance computing. FWI relies on simple principles: FWI aims at estimating the constitutive parameters of a medium contained in the coefficients of the wave equation by minimizing the distance between waves recorded by an array of sensors and the numerically simulated counterparts. As such, it is a partial differential equation (PDE)-constrained optimization problem, which is solved with iterative gradient-based methods. At each iteration, building the gradient of the misfit function with the adjoint-state method requires computing the incident wavefields triggered by each source of the experiment and the so-called adjoint wavefields by propagating the data residuals backward in time from the sensors. The wave equation is classically solved with numerical methods such as finite difference or finite element schemes to exploit the full information content in the data. Moreover, the wave equation can be solved either in the time-space or in the frequency-space domain. In the former case, solving the wave equation is an initial condition evolution problem, which can be solved with explicit (matrix-free) schemes. In the latter case, solving the time-harmonic wave equation for each source of the experiment is a boundary value problem requiring the solution of a large and sparse complex-valued linear system with multiple sparse right-hand sides. Operto et al. [35] have developed a frequency-domain full-waveform inversion (FWI) for

LU size (TBytes)			Flops		Time (seconds) BLR+Mixed		Scaled Residual
FR	BLR	+Mixed	FR	BLR+Mixed	Analysis +Factorization	Solve	BLR+Mixed
73	34	26	2.6×10^{18}	0.5×10^{18}	5946	27	7×10^{-4}

Table 1. Results for the Gorgon model, reservoir 23 km × 11 km × 6.5 km, grid size 15 m, Helmholtz equation, 25 Hz: FR (full-rank) with single-precision fp32; BLR (block low-rank) with $\varepsilon_{\text{BLR}} = 10^{-5}$; mixed-precision BLR: 3 precisions (32 bits, 24 bits, 16 bits) for storage using 48,000 cores (500 MPI × 96 threads/MPI). Solve is time to compute the solution with one right-hand side.

geophysical applications to image a geological medium containing several tens to hundreds of millions of unknowns after discretization, while the seismic experiment involves hundreds to thousands of seismic sources (right-hand sides). They implement FWI in the frequency domain, where the Helmholtz equation is solved with a finite difference method and the MUMPS solver.

In [33], we show how the memory and computational time of the LU factorization can be mitigated with adaptive-precision BLR approximations. In Table 1, the performance of MUMPS is illustrated with an industrial seabed case study from the Northwest Shelf of Western Australia: the Gorgon gas field (see [34]). The inversion is designed as a hierarchical multiscale approach proceeding across increasing frequencies between 1.7 Hz and 25 Hz. For each frequency, the discretization of the computational domain is matched to the frequency to optimize the size of the linear system to solve. The largest matrix to be processed at 25 Hz contains 531 million degrees of freedom. The number of floating-point operations (flops) for a full-rank (FR) LU factorization is 2.6×10^{18}, and the estimated memory necessary for the LU factors is equal to 73 TBytes. We remark that the FR factorization could not be carried out due to the large memory necessary for the computations. The results show that the use of BLR approximations reduces the size of the LU factors by a factor of about 2×, and using mixed precision further increases this factor to 3×. These results have been achieved on the supercomputer Adastra at CINES (Montpellier, France) on 48,000 cores (500 MPI × 96 threads/MPI). It is interesting to mention that since hundreds to thousands of right-hand sides need to be processed, the FWI simulation will be dominated by the time for solution which becomes more critical than the time for factorization.

5. Concluding remarks

A notable feature of the adaptive-precision BLR representation is that it can exploit an arbitrary number of precision formats. In the current version of MUMPS, we use seven

precision formats – including custom formats with truncated mantissa – that maximize the storage reduction. Additionally, blocks in BLR fronts that are not admissible to BLR compression may be stored in their full-rank format in low precision if their norm is small. In order to maintain efficiency with these custom formats which can be critical for the solution phase, we have developed a memory accessor approach [9] that implements an efficient, lightweight conversion from custom to standard formats. The efficiency of the approach relies on cache-aware mechanisms. When a computation is executed using (non-standard) low-precision operands, the operation is performed if such an operand and a working-precision copy of it can coexist in the cache. If this is not possible, the operation is split up into smaller blocks.

In the last few years, nodes of parallel computers have become very heterogeneous, leading to extremely powerful nodes equipped with multicore CPUs together with accelerators such as GPUs. Using accelerators efficiently, in the context of a general-purpose sparse direct solver, is very challenging. Our approach is based on offloading BLAS-3 kernels using XKBlas [31] from Inria. In collaboration with Inria, we are studying computational nodes with unified memory platforms. The unified memory architectures significantly reduce memory transfer overheads between the CPU and the GPU, enabling efficient offloading of smaller BLAS-3 kernels to the GPU and offering great opportunities to cover a wider range of real-world applications.

As mentioned, the asymptotic complexity of the flat BLR format, although already quite low, is however not optimal. Some more compact alternative formats, achieving intermediate complexities between BLR and fully hierarchical formats, have been theoretically proposed: the BLR^2 [15] and the multilevel BLR (MBLR) [12] formats. In collaboration with EDF (Electricité de France) company, we are currently assessing their practical potential for further improving the complexity of `MUMPS`.

Acknowledgments. The `MUMPS` project is a joint effort by several researchers and institutions (CERFACS, CNRS, ENS Lyon, Toulouse INP, Inria, Sorbonne University, and the University of Bordeaux) spanning more than 25 years of research and development. Today, the `MUMPS` solver is developed and maintained by Mumps Technologies, an enterprise founded in 2019. We are grateful to Stéphane Operto for his careful proofreading of the section on full-waveform inversion.

Funding. This work was carried out in the context of two PhD theses supported by EDF and is supported by a consortium of industrial and academic users.

References

[1] P. Amestoy, C. Ashcraft, O. Boiteau, A. Buttari, J.-Y. L'Excellent, and C. Weisbecker, Improving multifrontal methods by means of block low-rank representations. *SIAM J. Sci. Comput.* **37** (2015), no. 3, A1451–A1474 Zbl 1314.05111 MR 3355774

[2] P. Amestoy, O. Boiteau, A. Buttari, M. Gerest, F. Jézéquel, J.-Y. L'Excellent, and T. Mary, Mixed precision low-rank approximations and their application to block low-rank LU factorization. *IMA J. Numer. Anal.* **43** (2023), no. 4, 2198–2227 Zbl 1525.65038 MR 4621842

[3] P. Amestoy, O. Boiteau, A. Buttari, M. Gerest, F. Jézéquel, J.-Y. L'Excellent, and T. Mary, Communication avoiding block low-rank parallel multifrontal triangular solve with many right-hand sides. *SIAM J. Matrix Anal. Appl.* **45** (2024), no. 1, 148–166 Zbl 1535.65052 MR 4688196

[4] P. Amestoy, A. Buttari, F. Faucher, M. Gerest, J.-Y. L'Excellent, and T. Mary, Mixed precision sparse direct solver applied to 3D wave propagation. In *Proceedings of the 8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2022), Oslo, Norway*, 2022

[5] P. Amestoy, A. Buttari, F. Faucher, J.-Y. L'Excellent, and T. Mary, Application of the sparse direct solver MUMPS to Hybridizable Discontinuous Galerking (HDG) discretization for wave modeling. In *ECCOMAS Congress 2024, Lisbon, Portugal*, 2024

[6] P. Amestoy, A. Buttari, N. J. Higham, J.-Y. L'Excellent, T. Mary, and B. Vieublé, Combining sparse approximate factorizations with mixed-precision iterative refinement. *ACM Trans. Math. Software* **49** (2023), no. 1, article no. 4, 29 pp. Zbl 07908566 MR 4567890

[7] P. Amestoy, A. Buttari, N. J. Higham, J.-Y. L'Excellent, T. Mary, and B. Vieublé, Five-precision GMRES-based iterative refinement. *SIAM J. Matrix Anal. Appl.* **45** (2024), no. 1, 529–552 Zbl 1545.65124 MR 4704629

[8] P. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary, On the complexity of the block low-rank multifrontal factorization. *SIAM J. Sci. Comput.* **39** (2017), no. 4, A1710–A1740 Zbl 1372.65089 MR 3691219

[9] P. Amestoy, A. Jego, J.-Y. L'Excellent, T. Mary, and G. Pichon, BLAS-based mixed precision block memory accessor with applications to sparse direct solvers. *ACM Trans. Math. Softw.* **52** (2026), no. 2, article no. 10, 29 pp.

[10] P. Amestoy, A. Buttari, I. Duff, A. Guermouche, J.-Y. L'Excellent, and B. Uçar, Multifrontal method. In D. Padua (ed.), *Encyclopedia of parallel computing*, pp. 1209–1216, Springer, 2011

[11] P. R. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. Mary, Performance and scalability of the block low-rank multifrontal factorization on multicore architectures. *ACM Trans. Math. Software* **45** (2019), no. 1, article no. 2, 26 pp. Zbl 1471.65025 MR 3939092

[12] P. R. Amestoy, A. Buttari, J.-Y. L'Excellent, and T. A. Mary, Bridging the gap between flat and hierarchical low-rank matrix formats: the multilevel block low-rank format. *SIAM J. Sci. Comput.* **41** (2019), no. 3, A1414–A1442 Zbl 1434.65051 MR 3945811

[13] P. R. Amestoy, I. S. Duff, J.-Y. L'Excellent, and J. Koster, A fully asynchronous multifrontal solver using distributed dynamic scheduling. *SIAM J. Matrix Anal. Appl.* **23** (2001), no. 1, 15–41 Zbl 0992.65018 MR 1856597

[14] A. Aminfar, S. Ambikasaran, and E. Darve, A fast block low-rank dense solver with applications to finite-element matrices. *J. Comput. Phys.* **304** (2016), 170–188 Zbl 1349.65595 MR 3422408

[15] C. Ashcraft, A. Buttari, and T. Mary, Block low-rank matrices with shared bases: potential and limitations of the BLR^2 format. *SIAM J. Matrix Anal. Appl.* **42** (2021), no. 2, 990–1010 Zbl 07365307 MR 4275498

[16] M. Bebendorf, *Hierarchical matrices. A means to efficiently solve elliptic boundary value problems*. Lect. Notes Comput. Sci. Eng. 63, Springer, Berlin, 2008, 290 pp. Zbl 1151.65090 MR 2451321

[17] S. Börm, L. Grasedyck, and W. Hackbusch, Introduction to hierarchical matrices with applications. *Eng. Anal. Bound. Elem.* **27** (2003), no. 5, 405–422 Zbl 1035.65042

[18] P. Businger and G. H. Golub, Linear least squares solutions by householder transformations. *Numer. Math.* **7** (1965), no. 3, 269–276 Zbl 0142.11503 MR 0176590

[19] S. Chandrasekaran, M. Gu, and T. Pals, A fast ULV decomposition solver for hierarchically semiseparable representations. *SIAM J. Matrix Anal. Appl.* **28** (2006), no. 3, 603–622 Zbl 1120.65031 MR 2262971

[20] L. Claus, P. Ghysels, Y. Liu, T. A. Nhan, R. Thirumalaisamy, A. P. S. Bhalla, and S. Li, Sparse approximate multifrontal factorization with composite compression methods. *ACM Trans. Math. Software* **49** (2023), no. 3, article no. 24, 28 pp. Zbl 07910039 MR 4658761

[21] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct methods for sparse matrices*. Monographs on Numerical Analysis, The Clarendon Press, Oxford University Press, New York, 1986, 341 pp. Zbl 0604.65011 MR 0892734

[22] I. S. Duff and J. K. Reid, The multifrontal solution of indefinite sparse symmetric linear equations. *ACM Trans. Math. Software* **9** (1983), no. 3, 302–325 Zbl 0515.65022 MR 0791968

[23] I. S. Duff and J. K. Reid, The multifrontal solution of unsymmetric sets of linear equations. *SIAM J. Sci. Statist. Comput.* **5** (1984), no. 3, 633–641 Zbl 0557.65017 MR 0754490

[24] C. Eckard and G. Young, The approximation of one matrix by another of lower rank. *Psychometrica* **1** (1936), no. 3, 211–218 JFM 62.1075.02

[25] P. Ghysels, X. S. Li, F.-H. Rouet, S. Williams, and A. Napov, An efficient multicore implementation of a novel HSS-structured multifrontal solver using randomized sampling. *SIAM J. Sci. Comput.* **38** (2016), no. 5, S358–S384 Zbl 1352.65092 MR 3565567

[26] A. Gillman, P. M. Young, and P.-G. Martinsson, A direct solver with $O(N)$ complexity for integral equations on one-dimensional domains. *Front. Math. China* **7** (2012), no. 2, 217–247 Zbl 1262.65198 MR 2897703

[27] G. H. Golub and C. F. Van Loan, *Matrix computations*. 4th edn., Johns Hopkins Stud. Hist. Math., Johns Hopkins University Press, Baltimore, MD, 2013, 756 pp. Zbl 1268.65037 MR 3024913

[28] W. Hackbusch, A sparse matrix arithmetic based on $\mathcal{H}$-matrices. Part I: Introduction to $\mathcal{H}$-matrices. *Computing* **62** (1999), no. 2, 89–108 Zbl 0927.65063 MR 1694265

[29] N. J. Higham, *Accuracy and stability of numerical algorithms*. 2nd edn., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002, 680 pp. Zbl 1011.65010 MR 1927606

[30] N. J. Higham and T. Mary, Solving block low-rank linear systems by LU factorization is numerically stable. *IMA J. Numer. Anal.* **42** (2022), no. 2, 951–980 Zbl 1511.65024 MR 4410733

[31] J. V. F. Lima, T. Gautier, V. Danjean, B. Raffin, and N. Maillard, Design and analysis of scheduling strategies for multi-CPU and multi-GPU architectures. *Parallel Comput.* **44** (2015), 37–52 MR 3326470

[32] J. W. H. Liu, The role of elimination trees in sparse factorization. *SIAM J. Matrix Anal. Appl.* **11** (1990), no. 1, 134–172 Zbl 0697.65013 MR 1032223

[33] S. Operto, P. Amestoy, H. Aghamiry, S. Beller, A. Buttari, L. Combe, V. Dolean, M. Gerest, G. Guo, P. Jolivet, J.-Y. L'Excellent, F. Mamfoumbi, T. Mary, C. Puglisi, A. Ribodetti, and P.-H. Tournier, Is 3D frequency-domain FWI of full-azimuth/long-offset OBN data feasible? The Gorgon data FWI case study. *The Leading Edge* **42** (2023), no. 3, 173–183

[34] S. Operto, P. Amestoy, S. Beller, A. Buttari, L. Combe, F. Faucher, P. Jolivet, J.-Y. L'Excellent, T. Mary, and S. Migeon, Pushing the limits of 3D frequency-domain FWI with the 2015/2016 OBN Gorgon dataset. In *85th EAGE Annual Conference & Exhibition, Oslo*, 10–13 June 2024

[35] S. Operto, A. Miniussi, R. Brossier, L. Combe, L. Métivier, V. Monteiller, A. Ribodetti, and J. Virieux, Efficient 3-D frequency-domain mono-parameter full-waveform inversion of ocean-bottom cable data: application to Valhall in the visco-acoustic vertical transverse isotropic approximation. *Geophys. Journal Int.* **202** (2015), no. 2, 1362–1391

[36] G. Pichon, E. Darve, M. Faverge, P. Ramet, and J. Roman, Sparse supernodal solver using block low-rank compression: design, performance and analysis. *J. Comput. Sci.* **27** (2018), 255–270 MR 3834258

[37] J. Xia, S. Chandrasekaran, M. Gu, and X. S. Li, Fast algorithms for hierarchically semiseparable matrices. *Numer. Linear Algebra Appl.* **17** (2010), no. 6, 953–976 Zbl 1240.65087 MR 2759603

Patrick R. Amestoy
Mumps Technologies SAS, 193 Rue Marcel Mérieux, 69007 Lyon, France;
patrick.amestoy@mumps-tech.com

Jean-Yves L'Excellent
Mumps Technologies SAS, 193 Rue Marcel Mérieux, 69007 Lyon, France;
jean-yves.l.excellent@mumps-tech.com

Theo Mary
Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75005 Paris, France; theo.mary@lip6.fr

Chiara Puglisi
Mumps Technologies SAS, 193 Rue Marcel Mérieux, 69007 Lyon, France;
chiara.puglisi@mumps-tech.com

Public lecture

© 2026 EMS Press
This work is licensed under a CC BY 4.0 license
DOI 10.4171/9ECM/45

Tales of (mathematical) paradise

José Ferreirós

Abstract. Around 100 years ago, Hilbert stated his famous words that "nobody shall expel us" from the Cantorian paradise. The theory of sets had emerged rather suddenly, exactly 150 years ago, with a surprising paper of Cantor that has a peculiar inner story. In this history talk, I presented three vignettes from the early history of set theory. They represent three key moments: 1874, marking the birth of transfinite cardinalities under funny circumstances; 1905, a highly polemical moment when the schism between constructivists and "formalists" emerged; and 1925, the high time of the foundational debate between critics (Weyl, Brouwer, Skolem) and partisans of the higher infinite. I will not follow the obvious chronology; the narrative will be like those movies that include flashbacks.

> *"Aus dem Paradies, das Cantor uns geschaffen, soll uns niemand vertreiben können."* (Hilbert, in Münster on 4 June 1925, during a meeting in memory of Weierstraß.)

Almost a century ago, David Hilbert was saying his famous words, "nobody shall expel us" from the Cantorian paradise. Some of his contemporaries believed that the purported "paradise" was completely artificial. We shall revisit some facets of this debate in this history talk. I will be presenting three scenes or vignettes from the early history of set theory, covering a period of 50 years. They represent three key moments: 1874, marking the birth of transfinite cardinalities under peculiar circumstances; 1905, a highly polemical moment when the divide between constructivists and so-called "formalists" emerged; and 1925, the high time of the foundational debate between critics (Weyl, Brouwer, Skolem) and partisans of the higher infinite.

I intend to provide you with some interesting personal details concerning the real people involved in the story (mathematics is a human endeavor!), and also some details that have to do with methodological subtleties of this story. In this way, we will come to understand some reasons why the problems of set theory ended up splitting the mathematical community into parties that disagreed.

Mathematics Subject Classification 2020: 01A60 (primary); 01A72 (secondary).
Keywords: set theory, Hilbert, Cantor, Dedekind, constructivism, formalism, foundational debate, Borel, Rey Pastor, König.

I hope you will excuse me if I do not follow the obvious chronology. Let me begin a century ago, with Hilbert and his famous conference "On the Infinite." Then we will move back in time to 1874.

1. 1925: Hilbert in paradise?

In 1925, the foundational debate between intuitionism and formalism was at its peak. This began to get heated in 1921 with Weyl's intervention; reactions to intuitionism started to gain momentum, with a peak around 1925–1926.[1]

In 1924 and 1925, there was quite a lot of activity, with the brilliant intuitionist L. E. J. Brouwer very actively publishing papers,[2] Hilbert delivering his speech on the infinite, Weyl publishing a couple of papers, plus several other lesser figures – Fraenkel, Gonseth, Kolmogorov, Lietzmann, etc.

Historians have remarked that the tone of the debate was very political and allegorical from the beginning: Brouwer was called "the revolution," Weyl and Brouwer a "repeated attempt of a putsch," their views compared with Kronecker's (vs. the good Cantor, who is the point of comparison for Hilbert); and then Hilbert refusing to abandon the Paradise (a rebellious Adam!), "in the beginning was the sign," and so on.[3]

But besides the personal and political undertones, there were clear foundational logical topics, concerning the justification for actual infinity, the status of pure existence principles like AC, or the need to restrict methods to constructive procedures. This was the context for Hilbert's famous talk "On the Infinite" in 1925.

That is well remembered as the paper in which he included the famous words: "From the Paradise which Cantor created for us, no one shall expel us." Once again, as he had for many years, he stood forcefully on the side of set theory and so-called "classical" mathematics. The work is also remembered as a detailed and subtle presentation of the Hilbert program for metamathematics, whose goal was to provide a consistency proof for the core mathematical theories (Peano arithmetic, real analysis, set theory).

---

[1] See Hesseling (2003).

[2] Brouwer in the *Math. Ann.* (1925), *Crelle* (1924), and *Jb.DMV* (1925). Weyl in *Math. Zeitschr.* (1924) and a popular presentation in the philosophical journal *Symposion*.

[3] The whole thing would explode in 1928, with the "battle of frogs and mice," only partially due to foundational reasons – perhaps more due to the Bologna ICM in September 1928. Brouwer promoted a boycott, and there was an increasing "incompatibility of our views on fundamental matters" (D. Rowe and H. Mehrtens have written on this topic; see also van Dalen 1990).

This is not the place to go into the details of the Hilbert program, but I would like to revisit the main philosophical lessons that he tried to establish. These are somewhat ambivalent and thus surprising.

Hilbert begins with a lengthy discussion of the infinite in natural science. He argues in detail that neither in the direction of the infinitely small nor in the infinitely large does one find infinity. So reality is finite, the infinite is not real in the sense of *wirklich*, actual in the physical world. "Yet it could very well be the case that the infinite has a well-justified place *in our thinking* and plays the role of an indispensable notion."

But Hilbert seems to be of two minds here: on the one hand, he insists on the whole Cantorian universe, for set theory gives us "the deepest insight into the nature of the infinite"; this is the Romantic side, full of enthusiasm for the transfinite numbers, for the richness of mathematics. Cantor's "theory of *transfinite numbers* appears to me to be the most admirable flower of the mathematical intellect and in general one of the highest achievements of purely rational human activity." But there is also a critical side, and he ends up denying the infinite a substantive role in our thinking.

In the middle of the talk, we find the famous passage discussing "a completely satisfactory way of avoiding the paradoxes without betraying our science." The right attitude and viewpoints are these:

(1) We will carefully investigate and promote fruitful definitions [*Begriffsbildungen*] and deductive methods, wherever there is any hope of salvage. We will nurse them, strengthen them, and make them useful. No one shall drive us out of the paradise which Cantor created for us.

(2) We must establish throughout mathematics the same certitude for our deductions as exists in ordinary elementary number theory, which no one doubts and where contradictions and paradoxes arise only through our own carelessness.
Obviously, these goals can be attained only after we have fully elucidated *the nature of the infinite*.

But surprisingly, the Göttingen mathematician ends up denying the infinite a substantive role and relegating it to the status of a regulative idea in a Kantian sense:

> *The final result then is: nowhere is the infinite realized; it is neither present in Nature nor admissible as a foundation in our rational thinking – a remarkable harmony between being and thought.*[4] *... The right to operate with the infinite*

[4]This idea runs contrary to Cantor's philosophical position regarding the infinite, and we can be certain that he would have been unhappy with Hilbert's views.

> *can be secured only by means of the finite* [i.e., justifying the procedure by proving its consistency].

Here, we find the critical side of Hilbert's, far from the above-mentioned Romanticism, where he is trying to accommodate the thought that it is all "ideal elements" and a *manner of speech* (*façon de parler*) which boils down to finite procedures. Thus, he ends by saying:

> *The role that remains to the infinite is, rather, merely that of an Idea – if, in accordance with Kant's words, we understand by an Idea a concept of reason that transcends all experience and through which the concrete is completed so as to form a totality – an idea, moreover, in which we may have unhesitating confidence within the framework furnished by my theory.* (Hilbert 1925, p. 392)

This last is merely a hope, which was later contradicted by Gödel's arch-famous results.

We see that Hilbert was to some extent ambivalent, of two hearts concerning the higher infinite. He stands forcefully on the side of set theory, but his critical analysis of the foundational situation inclines him toward skepticism, or at least to a moderate standpoint. That was also the situation for many other mathematicians at the time.

Indeed, in the 1920s, there were important attempts to ground the Zermelo–Fraenkel axiom system ZFC, but also to restrict it, to the despair of Zermelo. Few people were as optimistic as Hilbert; many expressed an attitude of *neutrality* toward the actual infinite – this applies even to great experts like Sierpiński and Luzin – and some leading mathematicians rejected the Cantorian paradise.

Let me give just two examples. Émile Borel had been important by introducing Cantor's work in France, but after the discussion on the AC and the well-ordering theorem, in 1905, he turned into a critic of set theory (even rejecting the notion of $\aleph_1$). And he was highly influential. We will come back to this in the last section.

Another example is the leading mathematician in Spain and Argentina at the time, Julio Rey Pastor (1888–1962). His book *La matemática superior* was written in 1916 under the influence of a stay at Göttingen shortly before. But in the second edition, in 1951, Rey Pastor distances himself from his "youthful enthusiasm" for set theory. In 1916, he had explained that mathematics would be defined in the future as "the *Science of Sets*." In the later edition, he changed the first chapter of the book and wrote:

> *At the beginning of the 20th century, ephemeral golden age of Cantorism, one considered the transfinite Arithmetic and Geometry of Cantor as the indispensable basis of Analysis; the new treatises devoted a long initial chapter to them. The successes of the French school (Borel, Baire, Lebesgue, Fréchet)*

> *confirmed this prestige, while the criticism of Borel and Poincaré moderated the excessive enthusiasm, making clear that the whole of classical Analysis can be rigorously built without the actual infinite but with some Cantorian concepts that are indispensable to promote the theory of real functions... (accumulation points, neighborhoods,...). The rise of the unexpected paradoxes, early in the century, cooled down the enthusiasm of which this author was infected, giving the transfinite undue prominence in these conferences.* (Rey Pastor 1951, pp. 20–21, my translation)

According to him, mathematics in the 19th century, while going from the concrete to the abstract, had never lost sight of reality and the natural sciences. Meanwhile, it is "the characteristic of our times: arbitrary abstract constructions, arbitrary postulates and arbitrary functions; a sum of arbitrariness that horrified Poincaré" (Rey Pastor 1951, p. 7).

Similar second thoughts about set theory can be found in many relevant people: Poincaré, Lebesgue, Baire, Weyl, Skolem, van der Waerden, etc. As I have explained elsewhere, in the interwar period, an "atmosphere of insecurity" reigned. But let us now move back in time and consider the beginnings.

2. 1874: Birth in sin

Set theory was born in 1874 with the publication of Cantor's paper "On a property of the collection of all real algebraic numbers."[5] It was born in sin, as we shall see.

The contents of this elegant 1874 paper took form in the correspondence between Cantor and Dedekind, November–December 1873. The work presents two theorems, neither of them trivial at the time: a proof of the denumerability of the algebraic numbers, and a proof of the non-denumerability of $\mathbb{R}$. Those results provide an "essential distinction" between two kinds of infinite sets and a proof of the existence of transcendental (non-algebraic) numbers.

Here are some inconvenient facts we know:[6]

(1) The first theorem was proven by Dedekind in a letter on 30 November.

(2) The title highlights the first theorem and Dedekind's main working field, i.e., algebraic number theory.

[5]Cantor (1874). This assertion of course is somewhat exaggerated. Set theory was already beginning some years before (see Ferreirós 2024). But in this 1874 publication, it was shown, for the first time, that infinite sets with different "sizes" exist. In this sense, it was the beginning of *transfinite* set theory.

[6]From the published correspondence, Cantor and Dedekind (1937), also available in English.

(3) Notes are preserved in which Dedekind expresses surprise at how Cantor used his letters (30 November and 8 December) "almost literally" in print, *Crelle's Journal*, vol. 77.

(4) The idea of the second proof was certainly Cantor's, but Dedekind simplified it very much in one of his letters (8 December) and modified it somewhat.[7]

(5) All was somehow related to the "Berlin circumstances" (which means, to Weierstraß and Kronecker), but Dedekind never received more details – they "never discussed the paper" later on.[8]

Indeed, Weierstraß had a great interest in the first theorem (algebraic numbers ordered into a simple countable sequence) and encouraged Cantor to publish.

On pondering these facts, one is led to wonder – a text written by whom, and who should be credited with the theorems? I suggest it should have been a joint paper. And then, there is the matter of the title! Dedekind had just begun to make his mark (1871) in algebraic number theory, indeed offering a kind of redefinition of higher number theory. The combination of the use of his letters *verbatim*, and the insertion of that title, must have been mortifying.

Instead of presenting my own evaluation of the situation, let me just cite what other experts think. S. Müller-Stach, who has written about Dedekind's life and work (2024), speaks of Cantor's "unthoughtful publication" of some common ideas "without mentioning Dedekind, thus violating scientific standards boldly." He remarks that a letter from Dedekind's nephew to Aunt Julie Dedekind from 1887 proves that this was a well-known issue in the entire Dedekind family. This may sound like a mere anecdote, but it is indeed a very telling one.

Oliver Deiser, who is much more favorable to Cantor,[9] writes: "As a compromise, one might in the future call the result of countability of the algebraic numbers, Theorem of Cantor–Dedekind. The rest of the complex may remain with Cantor." Cantor reacted to Dedekind's proof of denumerability with the comment that his method was similar to how he proved the denumerability of tuples of integers. That is the basis for Deiser's claim that the theorem is due to both. However, Cantor's method cannot be extended to the case of the algebraic numbers unless one employs the key concept of the "height" of a polynomial, due to Dedekind (see Ferreirós 1993, p. 358).

[7] It seems this version went verbatim into the paper, except for five words (hence the "almost"). In another letter, Cantor acknowledged that Dedekind's "comments and expressions" were "so valuable to me" that he "thought it appropriate to adopt" them (!).

[8] The strange circumstances of the 1874 paper may have involved a certain quarrel between Kronecker (and Kummer) and Dedekind – a typical priority issue, both being experts in the theory of algebraic numbers. Compare Dedekind's *Werke*, vol. 3, 481.

[9] See Deiser's website https://www.aleph1.info/?call=Puc&permalink=cd1 (and search for Die Veröffentlichung von Cantor 1874).

In any event, the personalities and attitudes of both mathematicians were quite different and apparently sometimes incompatible.

But there is a *second reason* to talk about sinful birth. This is because of the deceitful clarity and perspicuity involved in the definitions of the real numbers that Cantor and Dedekind put forward (1872). They seemed to provide a perfectly precise characterization of the real-number continuum, but neither of them was fully aware of the import of their actual presuppositions. What I mean is the (implicit) powerset principle and the vague notion of arbitrary subsets.

The 1872 papers rely, respectively, on Dedekind cuts on $\mathbb{Q}$ and on fundamental sequences of rational numbers (a.k.a. Cauchy sequences). In order to obtain the complete system of the real numbers, this means relying on *arbitrary* sequences of rationals with the Cauchy property (Cantor), or equivalently, on arbitrary cuts on $\mathbb{Q}$ (Dedekind). They assume given a *totality of all* such arbitrary infinitary constructs, considered to be extensionally definite (not merely intensionally defined).[10]

In the technical terminology of set theory, Dedekind and Cantor were implicitly (and unclearly) adopting the *quasi-combinatorial* conception of sets. That they were so doing was only gradually clarified, especially through the criticism leveled against their work by constructivist mathematicians (Kronecker, Poincaré, Borel, Baire, Brouwer, Weyl, Skolem, and others). One can see how the critics did a great service to set-theoretic mathematics by helping to clarify its foundations.

Let me put it differently, to be more clear: it was explicit already (as of 1872) that these definitions relied on *actually infinite* sets, but it was left obscure the way in which they presupposed *arbitrary* or "*random*" sets (not merely "lawful" sets, given by a definite law) – and moreover an alleged totality of all such.

This situation was acknowledged years later, e.g., by Hilbert [11], when he mentioned "the doubts which have been raised against the existence of the totality of all real numbers" based on the notion that "by the set of real numbers" we have "to imagine, say, the totality of all possible laws according to which the elements of a fundamental sequence can proceed." The situation is even worse, since we assume that there are sequences that do not accord with *any* law.

[10]This distinction is due to Weyl; see the detailed explanation in Crosilla and Linnebo (2024). The idea is this: a condition is *intensionally* definite if, given some object, either the condition applies to that object or it does not; on the other hand, a condition or collection can be definite in the sense that the totality of all its instances or members has been circumscribed – this is *extensional* definiteness.

3. 1904–1905: Open polemics

In the decade of 1900, there were big polemical debates concerning the strange or surprising features of set theory: its strongly non-constructive orientation, its acceptance of "arbitrary" sets, and AC (the axiom of choice). The polemics was mostly about the question which methodology was admissible, especially in the context of the study of infinite sets – hence also about the acceptance or rejection of some fundamental axioms.

Leaving behind early ideas about logicism, this was mostly a debate between modern "formalists" (those in favor of the free methods and fruitful conceptual developments typical of set theory) and critical constructivists (critics of the "arbitrariness" of set theory, partisans of more restrictive methods).

The context included several things. First, from 1899 to 1903, the set-theoretic paradoxes were "discussed in many meetings" (Bernstein), and perhaps more widely after the publication of books by Russell and Frege in 1903.[11] But second, and even more important, there was a heated debate about the Continuum Problem at the 1904 ICM. This was Hilbert's [10] problem no. 1: Cantor's continuum hypothesis, CH, and the well-ordering theorem.

At the 1904 ICM in Heidelberg, Gyula König proved that the continuum is no aleph, in the presence of both Cantor and Hilbert. This caused a sensation; it is reported that the other sessions of the congress were rescheduled so that all participants could attend König's presentation. We have a moving description of Cantor's reaction:

> *Then Cantor took the word in a state of deep excitement. He expressed his gratefulness to God for having allowed him to live and see this refutation of his errors. The newspapers informed of König's talk. The Grand Duke of Baden himself asked to be informed of these sensational events through Felix Klein.* (Kowalewski 1950, pp. 198–203)

Unfortunately, one cannot always trust reminiscences.[12] A real eyewitness of the events gave a totally different account: "It is characteristic of Cantor that, from the beginning, he considered König's result as invalid, despite his exact demonstrations.

[11] As Cantor insisted (letters to Hilbert and Dedekind, 1899), there were reasons against the notions of a "collection of all" ordinals, of "all" cardinals or alephs, and even an actual infinity of "all" sets. All of this, made more complex due to many other "logical" paradoxes (Russell 1903, Richard 1905, etc.), later distinguished as semantic paradoxes.

[12] According to the Acts of the ICM, after König's talk *Zum Kontinuum-Problem* Cantor, Hilbert and Schoenflies took the word. Kowalewski was not present at the 1904 ICM.

[He rejoiced in saying that he did not mistrust the König, but only His Minister]." (Schoenflies 1922, p. 100)

In reaction to König's proof, a group of mathematicians including Cantor, Hilbert and Zermelo joined to discuss the situation and clarify it. If true, König's result would have destroyed the edifice of Cantor's transfinite set theory. Later in 1904, contributions due to Hausdorff and Zermelo again transformed the situation. Hausdorff clarified the issue with König's theorem, showing that his result had to be restricted: the fault was not due to König himself (all his original arguments were valid) but to a theorem which had been incorrectly formulated in Bernstein's dissertation;[13] ironically, Bernstein was a doctoral student of Cantor and Hilbert. Then, in September 1904, Zermelo sent his proof of the well-ordering theorem based on the axiom of choice, AC, in the form of a letter. It was immediately published in the *Mathematische Annalen*, whose editors were Klein and Hilbert.

Notice that the well-ordering theorem implies that every set has an aleph as its cardinal number. It is a very significant strengthening of the Cantorian edifice. But Zermelo's argument raised many doubts. Immediately, there came reactions to be published in the *Math. Annalen* (see vols. 60 and 61). A huge debate followed, a surprisingly intense series of arguments and counter-arguments, with authors from all the main European countries: France, Germany, Britain, Italy, with the "five letters" exchanged in France,[14] plus contributions due to Peano, Poincaré, Hardy, Jourdain, Russell...

There was confusion because of numerous sources of doubt: not only the new axiom AC, which was very explicitly non-constructive. Even those willing to accept it believed that there could be other problems with the proof, e.g., that it might depend on the unacceptable "set" of all ordinals. This was the case of Bernstein. Zermelo clarified the situation in 1908, with a new proof of the theorem, an explicit defense of the axiom, and the presentation of a whole axiom system for set theory.

It is instructive to consider that Hilbert, in his famous talk on Mathematical Problems [10], was asking for an *effective* well-ordering of the reals, a constructive solution to the problem of well-ordering $\mathbb{R}$:

> *The question now arises whether the totality of all [real] numbers may not be arranged in another manner so that every subset may have a first element, i.e., whether the continuum cannot be considered as a well-ordered set—a*

[13] A result which is only valid for alephs with a limit ordinal subindex, not for successor ordinals – hence, valid for $\aleph_\omega$ but not for $\aleph_1, \aleph_2, \ldots$

[14] Published in the *Bull. Société Mathématique de France*; see Borel et al. (1905). The *Math. Annalen* vol. 60(2) featured four papers discussing Zermelo's short piece (counting up to seven with vol. 61).

> *question which Cantor thinks must be answered in the affirmative. It appears to me most desirable to obtain a direct proof of this remarkable statement of Cantor's, say, by actually giving an arrangement of numbers such that in every subsystem a first number can be pointed out.*

Nobody has found such a constructive solution, and it is consistent with ZFC that there is none. What Zermelo provided, instead, was a totally non-constructive solution of the problem.

Many other people were also requesting an effective, constructive solution. For Émile Borel, the result of Zermelo was merely *conditional*: if you can define a choice function (on the powerset of *S*), *then* you can well-order the set (*S*); otherwise, your hands are empty. Borel continued to debate the problem for years, and he was highly influential. In his papers and talks, he came to the point of rejecting the notion of $\aleph_1$ as ill-defined.[15]

But notice that the axiom AC is perfectly natural, provided that we understand set theory to deal with *arbitrary infinite sets*. It applies to "every set for which the totality of subsets, etc. is meaningful [*einen Sinn hat*]," as Zermelo said. However, does the idea of a completely arbitrary function make actual sense? That is the problem which led to a split among mathematicians.

The debate would continue for many years to come, even if Hilbert's side was winning in practice. To give one further example, let me quote a paper by O. Perron in 1911:

> *Today there is a branch of mathematics, in which the viewpoints are very divided; some take for correct, what the others reject. It is the so-called theory of sets, in which the security of mathematical deduction seems to have been completely lost.*
>
> *The worst contrapositions arise in the question of the well-ordering of the continuum. Where one half regards the proof as absolutely free of difficulties,... the other half declares it all false.* (Perron 1911)

One can give further examples of critics of set theory among relevant mathematicians in all generations from 1904 until 2024. What will be the future of set theory? Despite enormous technical progress in the field during the last century, some reasons for skepticism have been present, most obviously the Continuum Problem (whose "resolution" or lack thereof may be construed as a reason for distrust).

[15] This situation was even interpreted as a *refutation* of the AC and the well-ordering theorem – assuming impossible an explicit well-ordering of $\mathbb{R}$.

Was the Cantorian paradise, after all, just an artificial paradise?[16]

References

[1] E. Borel et al., Cinq lettres sur la theorie des ensembles. In his *1950, Leçons sur la théorie des fonctions*, 4th edn. Gauthier-Villars, Paris, 1905; 1st edn., 1898 (English translation as appendix to G. H. Moore, *Zermelo's Axiom of Choice*. Springer, Berlin, 1982)

[2] G. Cantor, Ueber eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen. *J. Reine Angew. Math.* **1874** (1874), no. 77, 258–262 (Also in his *Abhandlungen*, 115–118. English translation in Ewald 1996, Vol. 2) MR 1579605

[3] G. Cantor and R. Dedekind, In E. Noether and J. Cavaillès (eds.), *Briefwechsel Cantor–Dedekind*. Hermann, Paris, 1937, 61 pp. (English translation in Ewald 1996, Vol. 2) Zbl 0017.38502

[4] L. Crosilla and Ø. Linnebo, Definiteness in early set theory. *J. Philos. Math.* **1** (2024), 43–62

[5] O. Deiser, Die Veröffentlichung von Cantor 1874 (search for the Anhänge). In *Der Briefwechsel zwischen Cantor und Dedekind* (online). 2020, https://www.aleph1.info/?call=Puc&permalink=cd1 visited on 8 May 2026

[6] W. Ewald, *From Kant to Hilbert. A source book in the foundations of mathematics, Vol. II*. Oxford University Press, Oxford, 1996 Zbl 0859.01002 MR 1465678

[7] J. Ferreirós, On the relations between Georg Cantor and Richard Dedekind. *Historia Math.* **20** (1993), no. 4, 343–363 Zbl 0792.01031 MR 1248470

[8] J. Ferreirós, The early development of set theory. In E. N. Zalta and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, Winter 2024 edn., https://plato.stanford.edu/archives/win2024/entries/settheory-early/ visited on 8 May 2026

[9] D. E. Hesseling, *Gnomes in the Fog. The reception of Brouwer's intuitionism in the 1920s*. Sci. Networks Hist. Stud. 28, Birkhäuser, Basel, 2003, 448 pp. Zbl 1037.01008 MR 2023449

[10] D. Hilbert, Mathematische probleme. In his *Werke (1932/35)* **3** (1900), 290–339 (English translation in *Bull. Amer. Math. Soc.* **8** (1902), 437–479)

[11] D. Hilbert, Über den Zahlbegriff. *Jahresber. Dtsch. Math.-Ver.* **8** (1900), no. 1, 180–194 (English translation in Ewald 1996, Vol. 2) JFM 31.0165.02

[12] D. Hilbert, Über das Unendliche. *Math. Ann.* **95** (1926), no. 1, 161–190 (English translation in J. van Heijenoort, *From Frege to Gödel*, pp. 367–392, Harvard University Press, 1967) JFM 51.0044.02 MR 1512272

[16] As Dieudonné wrote; see P. Dugac, *Richard Dedekind et les fondements des mathématiques* (Paris, Vrin, 1976), p. 11.

[13] G. Kowalewski, *Bestand und Wandel*. R. Oldenbourg, München, 1950, 309 pp. Zbl 0037.00203

[14] S. Müller-Stach, *Richard Dedekind. What are and what should the numbers be? Continuity and irrational numbers*. Class. Texts Sci., Springer, Berlin, 2024, 236 pp. Zbl 1560.01051 MR 4881998

[15] O. Perron, Über Wahrheit und Irrtum in der Mathematik. *Jahresber. Dtsch. Math.-Ver.* **20** (1911), 196–211 JFM 42.0073.03

[16] J. Rey Pastor, *La matemática superior*. Iberoamericana, Buenos Aires, 1951; 1st edn., 1916

[17] A. Schoenflies, Zur Erinnerung an Georg Cantor. *Jahresber. Dtsch. Math.-Ver.* **31** (1922), 97–106 JFM 48.0019.02

[18] D. van Dalen, The war of the frogs and the mice, or the crisis of the *Mathematische Annalen*. *Math. Intelligencer* **12** (1990), no. 4, 17–31 Zbl 0723.01006 MR 1076531

[19] E. Zermelo, Beweis, daß jede Menge wohlgeordnet werden kann. *Math. Ann.* **59** (1904), no. 4, 514–516 (in Z. Werke, Vol. 1 (2010), 80–119. English translation in van Heijenoort, op. cit. (1967). See also the 1908 papers in *Werke*, Vol. 1, 120–230) JFM 35.0088.03 MR 1511281

José Ferreirós
Department of Logic and Philosophy of Science, Universidad de Sevilla, C/ Camilo J. Cela s/n, 41018 Sevilla, Spain; josef@us.es

European Congress of Mathematics

Sevilla, 15–19 July, 2024

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), Berlin (2016), and Portorož (2021), the Ninth European Congress of Mathematics (9ECM) took place in Sevilla, Spain, July 15–19, 2024, with about 1400 participants from all over the world.

Ten plenary and thirty-one 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, four more prizes were awarded: The Felix Klein Prize for a remarkable solution of an industrial problem, the Otto Neugebauer Prize for a highly original and influential piece of work in the history of mathematics. The Paul Lévy Prize in Probability and the EMS/ECMI Lánczos Prize in Mathematical Software were awarded for the first time. The program was complemented by five public lectures, several exhibitions, sixty-four minisymposia, and eighteen contributed sessions, with about 950 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, ten prize, twenty-seven invited lectures, and a public lecture, all of which were delivered during the congress.

https://ems.press

ISBN 978-3-98547-105-8





EMS PRESS