

Proceedings of the  
**International Congress of  
Mathematicians**

Seoul 2014





SEOUL ICM 2014  
INTERNATIONAL  
CONGRESS OF  
MATHEMATICIANS

# Proceedings of the International Congress of Mathematicians

Seoul 2014

**VOLUME I**  
**Plenary Lectures  
and Ceremonies**

**Editors**

Sun Young Jang  
Young Rock Kim  
Dae-Woong Lee  
Ikkwon Yie

Editors

Sun Young Jang, University of Ulsan  
Young Rock Kim, Hankuk University of Foreign Studies  
Dae-Woong Lee, Chonbuk National University  
Ikkwon Yie, Inha University

Technical Editors

Young Rock Kim, The Korean  $\text{\TeX}$  Society  
Hyun Woo Kwon, The Korean  $\text{\TeX}$  Society

Proceedings of the International Congress of Mathematicians  
August 13–21, 2014, Seoul, Korea

Published by

KYUNG MOON SA Co. Ltd.  
174, Wausan-ro Mapo-gu Seoul, Korea  
Tel: +82-2-332-2004 Fax: +82-2-336-5193  
E-mail: [kyungmoon@kyungmoon.com](mailto:kyungmoon@kyungmoon.com)  
Homepage: [www.kyungmoon.com](http://www.kyungmoon.com)

© 2014 by SEOUL ICM 2014 Organizing Committee

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

This work was supported by the Korean Federation of Science and Technology Societies Grant funded by the Korean Government.

ISBN 978-89-6105-804-9  
ISBN 978-89-6105-803-2 (set)

Printed in Korea

# Preface

The Proceedings of SEOUL ICM 2014 consist of four volumes. In compliance with precedents, the Organizing Committee published both electronic and traditional print versions of the Proceedings for easier access and dissemination. Volumes II through IV were published before the Congress, electronic versions of which were stored into USBs and distributed to all registered participants. A policy to allow unlimited access to the electronic versions of all four volumes for non-commercial use was implemented. They are available at [www.icm2014.org/en/vod/proceedings](http://www.icm2014.org/en/vod/proceedings) and [www.mathunion.org/ICM](http://www.mathunion.org/ICM).

This volume is divided into five main parts. The first part consists of the speeches delivered during the opening ceremony of the Congress, including the presentations of the Fields Medals, the Rolf Nevanlinna Prize, the Carl Friedrich Gauss Prize and the Chern Medal Award, as well as the speeches and presentation of the Leelavati Prize made during the closing ceremony. The second part contains the laudations on the work of the awardees. The third part compiles the articles of the plenary speakers while the fourth part compiles those of the awardees and the Emmy Noether lecturer. The last part gathers the articles of the panel discussions organized at the Congress. Volumes II through IV contains the articles of the invited speakers categorized into 19 sections.

On behalf of the Organizing Committee, we would like to take this opportunity to express our sincere gratitude to all authors for graciously providing their valuable articles and to Kyungmoonsa for their devotion in the publications of this Congress. Lastly, we would like to thank the technical editors, whose committed support and endeavor were essential in ensuring the quality of the presentation of these publications.

Sun Young Jang  
Young Rock Kim  
Dae-Woong Lee  
Ikkwon Yie

## **Editors' Note**

The articles of Manjul Bhargava and Martin Hairer, who were a plenary speaker and an invited speaker, respectively, before the announcement of the 2014 Fields Medals, have been relocated to the Special Lectures section of this volume. However, the article of Martin Hairer also appears on pages 49 through 73 of Volume IV as the publications of the invited lectures were completed prior to the Congress. Additionally, the following speaker and panels did not opt to publish an article in this publication: Artur Avila (2014 Fields Medalist), panel on Future of Publishing, and panel on R&D Policy.

# Contents

<b>Preface</b>	v
<b>Past Congresses</b>	1
<b>Organization of the Congress</b>	4
<b>Committees of the Congress</b>	10
<b>Other Collaborators of ICM 2014</b>	19
<b>List of Sponsors</b>	20
<b>Opening Ceremony</b>	21
<b>Closing Ceremony</b>	36

## The Work of the Winners of the Fields Medal, the Nevanlinna Prize, the Gauss Prize, and the Chern Medal

<b>Étienne Ghys</b>	
The work of Artur Avila	47
<b>Benedict H. Gross</b>	
The work of Manjul Bhargava	56
<b>Ofer Zeitouni</b>	
The work of Martin Hairer	65
<b>Curtis T. McMullen</b>	
The work of Maryam Mirzakhani	73
<b>Sanjeev Arora</b>	
The work of Subhash Khot	81
<b>Ron Fedkiw, Jean-Michel Morel, Guillermo Sapiro, Chi-Wang Shu, and Wotao Yin</b>	
The work of Stanley Osher	90
<b>Mark L. Green</b>	
The work of Phillip Griffiths	114

## Plenary Lectures

<b>Ian Agol</b>	
Virtual properties of 3-manifolds	141

<b>James Arthur</b>	
<i>L</i> -functions and automorphic representations	171
<b>Alexei Borodin</b>	
Integrable probability	199
<b>Franco Brezzi</b>	
The Great Beauty of VEMs	217
<b>Emmanuel J. Candès</b>	
Mathematics of sparsity (and a few other things)	235
<b>Demetrios Christodoulou</b>	
Hyperbolic P.D.E. and Lorentzian geometry	259
<b>Fernando Codá Marques</b>	
Minimal surfaces: variational theory and applications	283
<b>Alan Frieze</b>	
Random Structures and Algorithms	311
<b>Ben Green</b>	
Approximate algebraic structure	341
<b>Jun-Muk Hwang</b>	
Mori geometry meets Cartan geometry: Varieties of minimal rational tangents	369
<b>János Kollár</b>	
The structure of algebraic varieties	395
<b>Jean-François Le Gall</b>	
Random geometry on the sphere	421
<b>Mikhail Lyubich</b>	
Analytic low-dimensional dynamics: From dimension one to two	443
<b>Frank Merle</b>	
Asymptotics for critical nonlinear dispersive equations	475
<b>Takuro Mochizuki</b>	
Wild harmonic bundles and twistor $\mathcal{D}$ -modules	499
<b>Benoît Perthame</b>	
Some mathematical aspects of tumor growth and therapy	529
<b>Jonathan Pila</b>	
O-minimality and Diophantine geometry	547
<b>Vojtěch Rödl</b>	
Quasi-randomness and the regularity method in hypergraphs	573
<b>Vera Serganova</b>	
Finite dimensional representations of algebraic supergroups	603

## Special Lectures

<b>Georgia Benkart</b> ( <i>Emmy Noether Lecture</i> )	
Connecting the McKay correspondence and Schur-Weyl duality	633
<b>Manjul Bhargava</b> ( <i>Fields Medal</i> )	
Rational points on elliptic and hyperelliptic curves	657
<b>Martin Hairer</b> ( <i>Fields Medal</i> )	
Singular stochastic PDEs	685
<b>Subhash Khot</b> ( <i>Nevanlinna Prize</i> )	
Hardness of Approximation	711
<b>Adrián Paenza</b> ( <i>Leelavati Prize</i> )	
I want to play with mathematics	729

## Panel Discussions

<b>Deborah Ball, Bill Barton*</b> , <b>Jean-Marie Laborde</b> , and <b>Man Keung Siu</b>	
How should we teach mathematics better?	739
<b>James H. Davenport</b>	
Mathematical Massive Open Online Courses (MOOCs): Report of a Panel Discussion	743
<b>Carla Cederbaum, Alicia Dickenstein, Gert-Martin Greuel*</b> , <b>David Grünberg, Hyungju Park, and Cédric Villani</b>	
IMAGINARY PANEL: Math communication for the future – A Vision Slam	755
<b>Peter J. Olver</b>	
The World Digital Mathematics Library: Report of a Panel Discussion	773
<b>Jean-Pierre Bourguignon, Ingrid Daubechies, Myung-Hwan Kim, and Youngah Park*</b>	
Why STEM (Science, Technology, Engineering and Mathematics)?	787
<b>Eduardo Colli, Fidel R. Nemenzo, Konrad Polthier, and Christiane Rousseau*</b>	
Mathematics is everywhere	799
<b>Other Activities</b>	813
<b>List of Participants</b>	814
<b>Participants by Country</b>	865
<b>Author Index</b>	867



# Past Congresses

1897	Zürich	1962	Stockholm
1900	Paris	1966	Moscow
1904	Heidelberg	1970	Nice
1908	Rome	1974	Vancouver
1912	Cambridge, UK	1978	Helsinki
1920	Strasbourg	1982	Warsaw (held in 1983)
1924	Toronto	1986	Berkeley
1928	Bologna	1990	Kyoto
1932	Zürich	1994	Zürich
1936	Oslo	1998	Berlin
1950	Cambridge, USA	2002	Beijing
1954	Amsterdam	2006	Madrid
1958	Edinburgh	2010	Hyderabad



2014 Seoul

# Past Winners of the Fields Medal, the Nevanlinna and the Gauss Prizes, the Chern Medal, and the Leelavati Prize

## Fields Medal

1936	Lars V. Ahlfors Jesse Douglas	1986	Simon K. Donaldson Gerd Faltings
1950	Laurent Schwartz Atle Selberg		Michael H. Freedman
1954	Kunihiko Kodaira Jean-Pierre Serre	1990	Vladimir G. Drinfeld Vaughan F. R. Jones Shigefumi Mori Edward Witten
1958	Klaus F. Roth René Thom	1994	Jean Bourgain Pierre-Louis Lions Jean-Christophe Yoccoz Efim Zelmanov
1962	Lars Hörmander John W. Milnor	1998	Richard E. Borcherds William T. Gowers Maxim Kontsevich Curtis T. McMullen
1966	Michael F. Atiyah Paul J. Cohen Alexander Grothendieck Stephen Smale	2002	Laurent Lafforgue Vladimir Voevodsky
1970	Alan Baker Heisuke Hironaka Sergeĭ P. Novikov John G. Thompson	2006	Andrei Okounkov *Grigori Perelman Terence Tao Wendelin Werner
1974	Enrico Bombieri David B. Mumford	2010	Elon Lindenstrauss Ngô Bào Châu Stanislav Smirnov Cédric Villani
1978	Pierre R. Deligne Charles L. Fefferman Gregori A. Margulis Daniel G. Quillen		
1982	Alain Connes William P. Thurston Shing-Tung Yau		

### **Rolf Nevanlinna Prize**

1982 Robert E. Tarjan  
1986 Leslie G. Valiant  
1990 Alexander A. Razborov  
1994 Avi Wigderson

1998 Peter W. Shor  
2002 Madhu Sudan  
2006 Jon Kleinberg  
2010 Daniel Spielman

### **Carl Friedrich Gauss Prize**

2006 Kiyosi Itô

2010 Yves Meyer

### **Chern Medal Award**

2010 Louis Nirenberg

### **Leelavati Prize**

2010 Simon Singh

# Organization of the Congress

## Hyungju Park, Chairman of the Seoul ICM 2014 Organizing Committee

### Korea's ICM Bidding

The first mathematician from Korea to attend the Congress made it to Helsinki International Congress of Mathematicians (ICM) in 1978 with a help of an IMU travel grant program for developing countries. It could be said that the mathematical research in Korea until the early 1980s was rather isolated and sporadic at best. Korean mathematical community began its globalization efforts by joining IMU in 1981 as a Group I member. In the mid-1980s, Korea endeavored to jump start its mathematical research, inviting renowned mathematicians from abroad to deliver lectures and providing young Korean mathematicians with opportunities to glimpse at mainstream mathematics. Visible spinoffs occurred, including modernization of academic curriculums and diversification of research centers. In light of such advancement, Korea became a Group II member of IMU in 1993. In the 1990s, Korea made quite a progress in improving its mathematical research both in quantity and quality, partly aided by the influx of talented Korean mathematicians educated abroad who returned to Korea after obtaining their degrees. The establishment of the first research institute in Korea devoted to mathematics and theoretical physics, Korea Institute for Advanced Study (KIAS), in 1996 provided the infrastructure for further development.

A dramatic display of the progress of Korean mathematics was made at Madrid ICM in 2006, to where three mathematicians, Jun-Muk Hwang, Jeong Han Kim and Yong-Geun Oh, were invited as first Korean ICM invited speakers. This ignited a festivity among the Korean mathematical community. Inspired by the evidence of the momentum in Korean mathematics, Korean Mathematical Society (KMS) applied for a raise in its IMU group level and was announced as a Group IV member in 2007. In the history of IMU, this still remains as the only instance in which the group level of a member country was raised by two in one shot. Highly motivated by this series of developments and to continue the momentum, KMS decided to place a bid to host ICM 2014. In June of 2007, KMS launched an ICM Bidding Committee of twelve members and appointed Hyungju Park as its Chair. Subsequently, an ICM Bidding Advisory Committee of twelve members was also launched. David Eisenbud and Efim Zelmanov of USA, Masaki Kashiwara of Japan and Yang Lo of China served as advisory committee members. The committee proceeded with full-fledged preparatory work and invested considerable amount of efforts and energy to establish bidding strategies and write up a bidding proposal. In the early stage of preparation, the committee worked hard to convince the Korean government on the significance of hosting the Congress in Korea. Many effortful visits were made to the Ministry of Science and Technology, the Ministry of Planning and Budget, and the National Assembly, that is, the Korean parliament.

On August 27, 2007, the Ministry of Science and Technology notified the bidding committee of the positive evaluation result on the viability of hosting ICM in Korea. The result was sent for a financial feasibility assessment to the Ministry of Planning and Budget, whose positive review was then sent to the Prime Minister's Office that convened its 45th Interna-

tional Event Review Committee Meeting in January 2008. This committee made an official decision to declare the ICM bidding as a matter of national importance. With an active support from the government, the bidding committee was able to secure 200 million Korean won (about 180 thousand US dollars) and 150 million Korean won (about 135 thousand US dollars) in years 2008 and 2009, respectively, for its bidding efforts.

Four cities, Seoul, Busan, Daegu and Jeju, expressed their interest in hosting the Congress and submitted proposals to the committee by October 2007. The committee organized a site evaluation subcommittee and visited the convention centers in those cities. After the visits and several meetings, the committee chose Korea's capital city, Seoul, as its host city candidate. A matter of urgency at the moment was to secure the fund for a travel grant program to invite mathematicians from developing countries. To that end, Korean mathematical community began its strenuous fundraising efforts and an ambitious plan to invite 1,000 mathematicians from developing countries began to materialize.

Brazil, Canada and Korea ended up submitting bidding proposals to IMU in November 2008. Korea's proposal was centered on the theme, "Dreams and Hopes for Late Starters". Evidence of the growth of Korean mathematical research during the recent half-century was demonstrated through various statistical data. Korea, which started as a developing country with poor research environment, had achieved an unparalleled growth. This led to the conclusion that Korea's hosting of ICM will convey a strong message of hope to those countries still struggling. Korea emphasized its genuine concern and interest in aiding underprivileged mathematicians and proposed a daring plan to invite 1,000 of those mathematicians.

The committee considered it important to secure a letter of support from the President of Korea. Eventually, the President's Office expressed great interest and enthusiastically provided the needed letter. Letters of support from the Minister of Education, Science and Technology, the Minister of Culture, Sports and Tourism, the Mayor of Seoul and the Vice Minister of Foreign Affairs were also received.

An IMU site evaluation committee visited Korea from the 23rd till the 26th of February, 2009. In writing up the bidding proposal and welcoming the IMU team, the bidding committee established as one of its primary arguments the high respect for scholarship deeply rooted in Korean culture. As a more practical measure to convey the firm governmental support, the Prime Minister, the Minister of Education, Science and Technology and the Mayor of Seoul personally met with the IMU team. The team also visited Samsung Electronics and experienced Korea's state-of-the-art IT technology and was provided with a collection of more than one hundred ICM-related news articles by Korean press along with their English summaries.

During the bidding process, the following twelve countries voluntarily sent letters of support for Korea: Japan, China, India, Hong Kong, Vietnam, Singapore, Malaysia, Thailand, Cambodia, Philippines, Tunisia and Morocco. IMU held its annual Executive Committee (EC) meeting in Fuzhou, China, during April 18 and 19, 2009. There, IMU decided to recommend Seoul as the hosting city for ICM 2014. IMU EC's recommendation was formally approved at the 2010 IMU General Assembly (GA) in Bangalore, India, held from August 16 to 17, 2010.

## **Preparation**

At a plenary meeting on November 15, 2013, the National Assembly of Korea overwhelmingly adopted a resolution in support of Seoul ICM. The resolution states that the National Assembly

- is dedicated to supporting Seoul ICM to hold a successful Congress,
- will make special efforts to encourage every citizen in all fields, including the government, industry, press and academia, to participate in and to have interest in the Congress,
- is aware that the development and popularization of fundamental science, including mathematics, is of significant importance to the national competitiveness, and
- urges the government to provide full support for Seoul ICM.

In celebration of the adoption of the parliamentary resolution, an ICM forum on “Beyond and After Seoul ICM” was held on December 10th at the National Assembly, attended by two legislators, a delegation from the National Assembly Standing Committee, a Deputy Minister of Science, ICT and Future Planning, along with over 120 mathematicians.

Additionally, the Korean government declared the year 2014 as the Korean Mathematical Year and an official proclamation ceremony was held on January 13, 2014. With over 100 representatives from the science community, the Minister of Science, ICT and Future Planning and the Minister of Education addressed the audience at the ceremony. Prior to the ceremony, a forum on “Mathematics, a Key Player of Creative Economy” was held to kick-start the activities planned for the year, at which Maria J. Esteban delivered a keynote lecture on “Linkage between Mathematics and Industry”, followed by two invited lectures and a panel discussion.

## Hosting of the Congress

The 17th IMU General Assembly was held from August 10 to 11, 2014, in the historic city of Gyeongju located in southern Korea. The history of Gyeongju dates back to ancient times, when the city was the capital of the Silla Dynasty (57 BC–AD 935) that reigned for the longest period in the history of Korea. The entire city has been designated as a UNESCO World Cultural Heritage under the name, Gyeongju Historic Areas. During the assembly at such historically rich city, Rio de Janeiro, Brazil, was announced as the next ICM host city.

The Seoul ICM was held for nine days from the 13th through the 21st of August, 2014, at COEX in Seoul. Seoul, the capital of Korea, is the center of Korean culture and education as well as politics and economics. As a city with a history of more than 600 years, Seoul is unique in that historical sites and modern cultural facilities coexist in harmony. It is always bustling with colorful events, performances and reenactments of traditional activities.

The Congress venue, COEX, is a business and cultural hub located in the heart of Seoul’s business district. It is a convention and exhibition center as well as a popular entertainment attraction for both domestic and foreign visitors. Asia’s largest underground mall, three five-star hotels, a department store, a subway station, an airport terminal, to name a few, are all located at COEX.

With a total of 5,217 registrants from 122 countries, the Seoul ICM set a new record for highest numbers of participants and countries in the history of ICM. The Congress was a scholarly festival of academic presentations and discussions and a feast of diverse cultural programs for adolescents and the general public. Also, as the year’s Fields Medalists included the very first female winner, an unprecedented and exceptional scene was presented during the awarding ceremony as the host (the President of IMU), the awardee (the President of Korea) and the awardee were all females. Various public programs, held during the days of the Congress with an aim of popularizing mathematics, were participated by 21,227 adolescents

and the general public. The record of a total of 27,359 participants at Seoul ICM will be remembered for many years ahead.

It is a tradition of the Congress for international conferences in various mathematics fields to hold before and after the Congress days in the hosting and neighboring countries. For the case of Seoul ICM, a total of 51 satellite conferences, 35 in Korea and 16 in neighboring countries, took place.

The opening ceremony on the first day of the Congress began with an opening address and a welcoming address by Hyungju Park, the Chairman of the Seoul ICM Organizing Committee, and Ingrid Daubechies, the President of IMU. Park Geun-hye, the President of Korea, attended the ceremony. After awarding the Fields Medals, the Rolf Nevanlinna Prize, the Carl Friedrich Gauss Prize and the Chern Medal Award, she delivered a congratulatory address, welcoming foreign visitors from all over the world and emphasizing the importance of mathematics as a basis of advancement for the history of mankind. Over 4,000 domestic and foreign figures from the mathematical community, the industry and the media attended the ceremony, which was also broadcasted live through EBS (Educational Broadcasting System) TV station. Numerous press personnel also visited to cover the ceremony. There were more than 1,500 domestic media coverages during the Congress, reflecting the excitement and the heated interest of the general public.

The EBS TV station worked closely with the organizing committee to bring a successful hosting and expanding the base of mathematics. To this end, EBS designated an 18-day-long period from August 4th through 21st as the Math Popularization Week and broadcasted various mathematics-related contents such as mathematical documentaries. EBS also broadcasted the opening ceremony and the public lectures of the Congress.

Daily newspapers, Math&Presso, were produced jointly with Korea Joongang Daily and distributed during the days of the Congress, which included interviews of award winners and special lecturers, and coverages on the cultural events that took place on every corner of the venue. The organizing committee also collaborated with Springer to publish the Seoul Intelligencer and distributed to all registered participants. This magazine contains informative articles and several scientific papers, and was well received by its readers.

The organizing committee made special efforts to make all lectures available for watching on the internet. VODs of official events such as the ceremonies and academic lectures can be watched from the official website of Seoul ICM at [www.icm2014.org](http://www.icm2014.org) and from a Youtube channel at [www.youtube.com/user/ICM2014VOD](http://www.youtube.com/user/ICM2014VOD). Photographs taken during the Congress can also be found from the Seoul ICM website and its Facebook webpage at [www.facebook.com/SEOULICM2014](http://www.facebook.com/SEOULICM2014).

Four award winner lectures, five laudations, two award lectures, four special lectures, 19 plenary lectures 177 invited lectures, 646 short communications and 388 poster presentations were made as the scientific program of the Congress. This includes the Emmy Noether Lecture by Georgia Benkart and the Abel Lecture by John Milnor. As a result, a total of 1,245 research results were introduced and participants were provided with an opportunity to peek into the recent advancements in mathematics. In addition, eight invited panels were organized during the Congress on the following topics:

- Why STEM?
- How should we teach better?
- Mathematics is everywhere
- R&D policy
- IMAGINARY Panel: Math communication for the future – a Vision Slam

- Mathematical Massive Open Online Courses
- Future of Publishing
- World Digital Mathematics Library

The Seoul ICM was positively evaluated not only on the quantity and quality of its scientific programs and on its smooth management, but also on its role in extending support for developing countries and reinforcing mathematical popularization. The Korean mathematical community has continuously expressed and emphasized the solidarity and support for mathematicians in developing countries striving amidst poor research environment. As a result of such efforts, the NANUM program invited 1,000 mathematicians from developing countries, of whom 664 participated in the Congress from 85 countries in South America, Southeast Asia, Eastern Europe and Africa. The financial support was granted in three categories: 45% senior mathematicians, 45% junior mathematicians and 10% advanced graduate students.

Upon this opportunity, IMU also organized and hosted the MENAO (Mathematics in Emerging Nation: Achievements and Opportunities) Symposium at the Congress venue on the day before the opening ceremony to discuss the measures in further supporting mathematics in developing countries. During this symposium, efforts to prepare various devices for shared growth of developing countries began in earnest by establishing a research fellowship for graduate students and raising funds.

An important feature of Seoul ICM compared to its precedents is that it not only served as an academic event for scholars but that an extended participation of the general public was drawn through various cultural programs.

In the evening of the opening ceremony, a public lecture was delivered by the honorary president of the Renaissance Technologies, James Simons, who is renowned for integrating mathematical theories into analyzing the stock market. Simons is also well-known for contributing a large portion of his assets to the advancement of science and education. During his public lecture in front of an audience of 5,000 on the very special role mathematics played in his life, Simons confidently assured that mathematics will be the greatest weapon in the scientific, technological and economic advancement of the future.

The IMAGINARY exhibition, jointly offered by the National Institute for Mathematical Sciences (NIMS) of Korea and Mathematisches Forschungsinstitut Oberwolfach of Germany, introduced and provided opportunity to experience contents-centered mathematical concepts to visitors through the application of state-of-the-art hardware such as touch panels.

The Bridges conference is the largest-scale mathematics-based international conference integrating mathematics and the arts. Bridges took place at Gwacheon National Science Museum, a subway-ride away from COEX, during the Congress days. It held various events and was visited by more than fifty thousand people.

On August 19th, a Math Movie Screening, cosponsored by the organizing committee and the French Embassy in Korea, played a French documentary film titled, *How I Came to Hate Maths (Comment J'ai Détesté Les Maths)*. Cédric Villani personally attended the event to introduce the film and to hold a Q&A session with the audience after the screening along with two renowned mathematicians, Jean-Pierre Bourguignon and Gert-Martin Greuel. This was an opportunity to present to the public the evidently expanding influence of mathematics. On the same day, Baduk (Go) events, including simultaneous games with professional players, drew intense interests from mathematicians and the general public.

For the logo of Seoul ICM, the organizing committee conducted a nationwide call for its design. Selected after several rounds of review, the logo comprises two golden spirals that

grow and expand at the rate of golden ratio, symbolizing the main theme of the Congress and depicting the Tae-Geuk of the Korean national flag.

A mathematical calendar was produced and distributed through Seoul ICM website, which integrated math formulas and facts with each day in the calendar. The calendar contains various contents from simple ones for elementary school students to challenging ones for experienced mathematicians, from common arithmetic operations to remarkably complex expressions, and from witty mathematical jokes to rather serious facts. The main schedule of the Congress was also marked in the calendar to aid the participants. The organizing committee also worked collaboratively with the Korean government to issue Seoul ICM commemorative stamps, which features the Pythagorean Theorem, Euler's theorem giving necessary and sufficient conditions of a graph having an Eulerian tour, and the Pascal's Triangle.

A very notable component of Seoul ICM is its corps of volunteers. The organizing committee regarded ICM volunteers as potential leaders of future mathematics and designed the volunteer program as an educational process rather than merely a source of manpower. The committee decided to rule out minors and limited volunteers to college and graduate students. In fact, there were strong interests from high school students, whose parents expressed dissatisfaction to the committee during the selection process. The committee received more than 700 volunteer applications. While the majority was mathematics majors, there were also students from other diverse disciplines. Eventually, after application screenings and two interviews, 280 volunteers were selected. A training program consisting of three rounds of training courses was mandated. The program not only consisted of logistic training matters, but also included lectures by renowned mathematicians on the importance of modern mathematics. Almost all volunteers remained onsite until the end of the Congress without dropouts and the volunteers' blog was filled with touching stories of memorable moments. Some of these students are already discussing how to save money to attend the next Rio de Janeiro ICM. We believe this enthusiasm is no less than any success we may have accomplished and advise future congress organizers to plan well ahead to run an energetic and successful volunteer program.



# Committees of the Congress

## I. Local Organizing Committees

### Executive Organizing Committee

Hyungju Park, POSTECH & NIMS, *Chair*  
Hyang-Sook Lee, Ewha Womans University, *Senior Vice Chair*  
Dongsu Kim, KAIST, *Vice Chair*  
Kyewon Koh Park, Ajou University, *Vice Chair*  
Jaeduck Jang, Hankuk University of Foreign Studies  
Sun Young Jang, University of Ulsan  
Hyeonbae Kang, Inha University  
JongHae Keum, KIAS  
Jeong Han Kim, KIAS  
Seonja Kim, Chungwoon University  
June-Yub Lee, Ewha Womans University  
Jungseob Lee, Ajou University

### Organizing Committee

#### IMU General Assembly Committee

Jeong Han Kim, KIAS, *Chair*  
Yun Sung Choi, POSTECH  
Mihyun Kang, Technische Universitat Graz, Austria  
Young Soo Kwon, Yeungnam University  
Yong Hoon Lee, Pusan National University  
Yongdo Lim, Sungkyunkwan University

#### Local Program Committee

JongHae Keum, KIAS, *Chair*  
YoungJu Choie, POSTECH  
Bokhee Im, Chonnam National University  
Sung-Eun Koh, Konkuk University  
Keonhee Lee, Chungnam National University  
Seok-Zun Song, Jeju National University

#### Planning and Finance Committee

Jungseob Lee, Ajou University, *Chair*  
Jaeduck Jang, Hankuk University of Foreign Studies  
Jung-Rye Lee, Daejin University  
Hye Sook Park, Seowon University

**International Exchanges Committee**

Dongsu Kim, KAIST, *Chair*  
June Gi Kim, Kangwon National University  
Sung-A Kim, Dongguk University  
Minkyu Kwak, Chonnam National University  
Yongnam Lee, KAIST  
Kyewon Koh Park, Ajou University  
Heesung Shin, Inha University

**Public and Media Relations Committee**

Hyang-Sook Lee, Ewha Womans University, *Chair*  
Jun-Muk Hwang, KIAS  
Intae Jeon, The Catholic University of Korea  
Suh-Ryung Kim, Seoul National University  
Chang-Ock Lee, KAIST  
Jongwoo Lee, Kwangwoon University  
Yongjin Song, Inha University

**Cultural Activities Committee**

Seunghun Yi, Youngdong University, *Acting Chair*  
Sunah Kim, Chosun University  
June Bok Lee, Yonsei University  
Sang-Gu Lee, Sungkyunkwan University  
Poo-Sung Park, Kyungnam University

**Web & Electronic Communications Committee**

June-Yub Lee, Ewha Womans University, *Chair*  
Jin-Hwan Cho, University of Suwon  
Sang-il Oum, KAIST  
Jae-Suk Park, POSTECH  
Seonhwa Kim, IBS-CGP

**Publications Committee**

Sun Young Jang, University of Ulsan, *Chair*  
Young Rock Kim, Hankuk University of Foreign Studies  
Dae-Woong Lee, Chonbuk National University  
Ikkwon Yie, Inha University

**Parallel Scientific Activities Committee**

Hyeonbae Kang, Inha University, *Chair*  
Sunghan Bae, KAIST  
Minhyong Kim, University of Oxford  
Woo Young Lee, Seoul National University

Yong-Geun Oh, IBS-CGP & POSTECH  
Jongil Park, Seoul National University

### **Social Activities and Logistics Committee**

Seonja Kim, Chungwoon University, *Chair*  
Youngook Choi, Yeungnam University  
Soon-Yi Kang, Kangwon National University  
Hyun Seok Kim, Sogang University  
Ki-Heon Yun, Sungshin Women's University

### **Ex officio**

Gwang Hui Kim, Kangnam University  
Do Sang Kim, Pukyong National University  
Joongul Lee, Hongik University  
Ke-Seung Lee, Korea University  
Nany Lee, University of Seoul  
Jaebum Sohn, Yonsei University

### **Advisory Committee**

*Co-Chairs*  
Dohan Kim, Seoul National University  
Kyung Chan Min, Yonsei University

## **II. IMU Committees for ICM 2014**

### **Program Committee**

Carlos Kenig, University of Chicago, USA, *Chair*  
Erwin Bolthausen, University of Zürich, Switzerland  
Sun-Yung Alice Chang, Princeton University, USA  
Wellington de Melo, IMPA, Brazil  
Hélène Esnault, Universität Berlin, Germany  
Timothy Gowers, Royal Society and University of Cambridge, UK  
Ravindran Kannan, Microsoft Research Labs. India, India  
JongHae Keum, KIAS, Republic of Korea  
Claude Le Bris, des Ponts & Inria, France  
Alex Lubotzky, Hebrew University of Jerusalem, Israel  
Jaroslav Nesetril, IUUK MFF Charles University in Prague, Czech Republic  
Andrei Okounkov, Columbia University, USA

### **Sectional Panels of the Program Committee**

#### **1. Logic and Foundations**

Boris Zilber, University of Oxford, UK, *Chair*  
Alexander Kechris, California Institute of Technology, USA

Menachem Magidor, Hebrew University of Jerusalem, Israel  
Toniann Pitassi, University of Toronto, Canada  
Theodore Slaman, University of California Berkeley, USA

## 2. Algebra

Efim Zelmanov, University of California San Diego, USA, *Chair*  
Lars Hesselholt, Nagoya University and University of Copenhagen, Japan and Denmark  
Gopal Prasad, University of Michigan, USA  
Mark Sapir, Vanderbilt University, USA  
Agata Smoktunowicz, University of Edinburgh, UK  
Bemd Sturmfels, University of California Berkeley, USA

## 3. Number Theory

Chandrashekar Khare, University of California Los Angeles, USA, *Chair*  
John Friedlander, University of Toronto, Canada  
Roger Heath-Brown, Mathematical Institute Oxford University, UK  
Kazuya Kato, University of Chicago, USA  
Mark Kisin, Harvard University, USA  
Báo Châu Ngô, University of Chicago, USA  
Peter Sarnak, Princeton University, USA  
Freydoon Shahidi, Purdue University, USA  
Gisbert Wüstholtz, ETH Zürich, Switzerland

## 4. Algebraic and Complex Geometry

Yujiro Kawamata, University of Tokyo, Japan, *Chair*  
Arnaud Beauville, Université de Nice, France  
Jean-Pierre Demailly, Institut Fourier, France  
Christopher Hacon, University of Utah, USA  
Jun-Muk Hwang, KIAS, Republic of Korea  
Dmitry Kaledin, Steklov Mathematical Institute, Russia  
Richard Thomas, Imperial College London, UK

## 5. Geometry

Simon Donaldson, Imperial College London, UK, *Chair*  
Fernando Coda Marques, IMPA, Brazil  
Matthew Gursky, University of Notre Dame, USA  
Gerhard Huisken, Universität Tübingen, Germany  
Sergei Olegovich Ivanov, St. Petersburg State University, Russia  
Bruce Kleiner, New York University, USA  
Yiming Long, Nankai University, P.R. China  
Dusa McDuff, Barnard College and Columbia University, USA  
Frank Pacard, École Polytechnique, France  
Toshikazu Sunada, Meiji University, Japan

## 6. Topology

Michael Hopkins, Harvard University, USA, *Chair*  
Martin Bridson, Oxford University, UK  
Ronald Fintushel, Michigan State University, USA  
Marc Levine, Universität Duisburg-Essen, Germany  
John Luecke, University of Texas at Austin, USA  
Wolfgang Lueck, University of Bonn, Germany  
Yair Minsky, Yale University, USA  
Kaoru Ono, RIMS Kyoto University, Japan  
Ulrike Tillmann, Oxford University, UK

## 7. Lie Theory and Generalizations

David Kazhdan, Hebrew University of Jerusalem, Israel, *Chair*  
Michel Brion, Institut Fourier, France  
Marc Burger, ETH Zürich, Switzerland  
Maryam Mirzakhani, Stanford University, USA  
Hee Oh, Yale University, USA  
Eric Opdam, University of Amsterdam, Netherlands  
Peter Schneider, Universität Münster, Germany  
Wolfgang Soergel, Universität Freiburg, Germany

## 8. Analysis and Applications

Jean Bourgain, Institute for Advanced Study, USA, *Chair*  
Luigi Ambrosio, Scuola Normale Superiore, Italy  
Sergei Konyagin, Steklov Mathematical Institute, Russia  
Pekka Koskela, University of Jyväskylä, Finland  
Ngaiming Mok, University of Hong Kong, Hong Kong  
Assaf Naor, Princeton University, USA  
Tatiana Toro, University of Washington, USA  
Dan-Virgil Voiculescu, University of California Berkeley, USA  
Horng-Tzer Yau, Harvard University, USA

## 9. Dynamical Systems and ODE

Michael Benedicks, KTH Royal Institute of Technology, Sweden, *Chair*  
Dmitry Dolgopyat, University of Maryland, USA  
Manfred Einsiedler, ETH Zürich, Switzerland  
Sergei Kuksin, CNRS and Université Paris Diderot, France  
Leonid Polterovich, Tel Aviv University, Israel  
Enrique Pujals, IMPA, Brazil  
Mitsuhiro Shishikura, Kyoto University, Japan  
Amie Wilkinson, University of Chicago, USA  
Jean-Christophe Yoccoz, Collège de France, France

**10. PDE**

Luis Caffarelli, University of Texas at Austin, USA, *Chair*  
Manuel del Pino, FCFM University of Chile, Chile  
Sergiu Klainerman, Princeton University, USA  
Tai-Ping Liu, Academia Sinica, Taiwan  
Frank Merle, Université de Cergy-Pontoise, France  
Sylvia Serfaty, Université Pierre et Marie Curie, France  
Sijue Wu, University of Michigan, USA

**11. Mathematical Physics**

Giovanni Felder, ETH Zürich, Switzerland, *Chair*  
Joel Feldman, University of British Columbia, Canada  
Michio Jimbo, Rikkyo University, Japan  
Nikita Nekrasov, Simons Center for Geometry and Physics, USA  
Igor Rodnianski, Princeton University, USA  
Herbert Spohn, TU Munich, Germany  
Leon Takhtajan, Brook University, USA  
Craig Tracy, University of California Davis, USA  
Pavel Wiegmann, University of Chicago, USA

**12. Probability and Statistics**

Jean-François Le Gall, Université Paris-Sud, France, *Chair*  
Itai Benjamini, Weizmann Institute, Israel  
Maury Bramson, University of Minnesota, USA  
David Donoho, Stanford University, USA  
Alice Guionnet, MIT and CNRS, USA and France  
Shigeo Kusuoka, University of Tokyo, Japan  
Gregory Lawler, University of Chicago, USA  
Sara van de Geer, ETH Zürich, Switzerland

**13. Combinatorics**

Alexander Schrijver, University of Amsterdam and CWI, Netherlands, *Chair*  
Mireille Bousquet-Mélou, CNRS Université de Bordeaux, France  
Gil Kalai, University of Jerusalem, Israel  
Jeong Han Kim, KIAS, Republic of Korea  
Serge Lando, Institute of System Research, Russia  
Jiri Matousek, Charles University Prague and ETH Zürich, Czech Republic and Switzerland  
Cheryl E. Praeger, University of Western Australia, Australia  
Oliver Riordan, University of Oxford, UK  
Benny Sudakov, ETH Zürich, Switzerland  
Carsten Thomassen, Technical University of Denmark, Denmark

**14. Mathematical Aspects of Computer Science**

Madhu Sudan, Microsoft Research, USA, *Chair*

Manindra Agrawal, IIT Kanpur, India

Irit Dinur, Weizmann Institute, Israel

Mark Jerrum, University of London, UK

Alexander Razborov, University of Chicago and Steklov Mathematical Institute, USA

Daniel Spielman, Yale University, USA

**15. Numerical Analysis and Scientific Computing**

Wolfgang Hackbusch, Max Planck Institute for Mathematics in the Sciences, Germany, *Chair*

Zhiming Chen, AMSS Chinese Academy of Sciences, P.R. China

Thomas Yizhao Hou, California Institute of Technology, USA

Rolf Jeltsch, ETH Zürich, Switzerland

Yvon Maday, Labo J.-L. Lions University Pierre et Marie Curie, France

Philip Protter, Columbia University, USA

Endre Süli, University of Oxford, UK

Mary Fanett Wheeler, University of Texas at Austin, USA

**16. Control Theory and Optimization**

Jean-Michel Coron, Université Pierre et Marie Curie, France, *Chair*

Mihai Anitescu, Argonne National Laboratory, USA

Vivek Borkar, Indian Institute of Technology Bombay, India

William Cook, University of Waterloo, Canada

Rekha Thomas, University of Washington, USA

Margaret Wright, Courant Institute of Mathematical Sciences NYU, USA

Xu Zhang, Sichuan University, P.R. China

**17. Mathematics in Science and Technology**

Emmanuel Candes, Stanford University, USA, *Chair*

Mitchell Luskin, University of Minnesota, USA

Vincent Caselles, Universitat Pompeu Fabra, Spain

Olivier Faugeras, Inria, France

John Ball, University of Oxford, UK

Maria J. Esteban, CNRS & University Paris-Dauphine, France

Peter Markowich, University of Cambridge, UK

Hans Föllmer, Humboldt University of Berlin, Germany

Bernhard Schölkopf, Max Planck Institute for Intelligent Systems, Germany

Zuwei Shen, National University of Singapore, Singapore

**18. Mathematics Education and Popularization of Mathematics**

Mina Teicher, Bar-Ilan University and NYU, USA, *Chair*

Yuriko Y Baldin, UFSCar, Brazil

Keith Devlin, Stanford University, USA

Celia Hoyles, University of London, UK

Gabriele Kaiser, Universität Hamburg, Germany  
Oh Nam Kwon, Seoul National University, Republic of Korea  
Frederick K.S. Leung, University of Hong Kong, Hong Kong  
Tomas Recio, Universidad de Cantabria, Spain

### **19. History of Mathematics**

Karine Chemla, CNRS, France, *Chair*  
Leo Corry, Tel Aviv University, Israel  
Moritz Epple, Goethe University Frankfurt am Main, Germany  
Niccolò Guicciardini, University of Bergamo, Italy  
Jan Pieter Hogendijk, University Utrecht, Netherlands  
Tinne Hoff Kjeldsen, RUC, Denmark  
Edith Sylla, North Carolina State University, USA

### **Fields Medal Committee for 2014**

Ingrid Daubechies, Duke University, USA, *Chair*  
Luigi Ambrosio, Scuola Normale Superiore, Italy  
David Eisenbud, MSRI and University of California Berkeley, USA  
Kenji Fukaya, State University of New York, USA  
Étienne Ghys, CNRS-ENS Lyon, France  
Benedict Gross, Harvard University, USA  
Frances Kirwan, Oxford University, UK  
János Kollár, Princeton University, USA  
Maxim Kontsevich, Institut des Hautes Études Scientifiques, France  
Michael Struwe, ETH Zürich, Switzerland  
Ofer Zeitouni, Weizmann Institute of Science, Israel  
Günter M. Ziegler, Freie Universität Berlin, Germany

### **Rolf Nevanlinna Prize Committee for 2014**

Avi Wigderson, Institute for Advanced Study, USA, *Chair*  
Thierry Coquand, Gothenburg University, Sweden  
Yurii Nesterov, Université Catholique de Louvain, Belgium  
Jaikumar Radhakrishnan, Tata Institute of Fundamental Research, India  
Éva Tardos, Cornell University, USA  
Leslie Valiant, Harvard University, USA

### **Carl Friedrich Gauss Prize Committee for 2014**

Alfio Quarteroni, Ecole Polytechnique Fédérale de Lausanne, Switzerland, *Chair*  
Weinan E., Peking University and Princeton University, P.R. China and USA  
Barbara Keyfitz, Ohio State University, USA  
Aad van der Vaart, Leiden University, Netherlands  
Andrés Weintraub, University of Chile, Chile

**Chern Medal Award Committee for 2014**

Robert Bryant, Duke University, USA, *Chair*  
Kazuo Murota, University of Tokyo, Japan  
Felix Otto, Max Planck Institute for Mathematics in the Sciences, Germany  
Alain-Sol Sznitman, ETH Zürich, Switzerland  
Claire Voisin, CNRS, France

**Leelavati Prize Committee for 2014**

David Mumford, Brown University, USA, *Chair*  
Oh Nam Kwon, Seoul National University, Republic of Korea  
Guillermo Martínez, University of Buenos Aires, Argentina  
Madabusi Santanam Raghunathan, Indian Institute of Technology Bombay, India  
Srinivasa Varadhan, New York University, USA

**Emmy Noether Lecture Committee for 2014**

Christiane Rousseau, University of Montreal, Canada, *Chair*  
Maria J. Esteban, CNRS & University Paris-Dauphine, France  
Raman Parimala, Emory University, USA  
Claudia Sagastizabal, IMPA, Brazil  
Anatoly Vershik, Steklov Mathematical Institute, Russia

**Travel Grants Committee for 2014**

Not appointed since Korea has offered the NANUM travel grants program.

# Other Collaborators of ICM 2014

## Secretariat

Heeyeun Choi (Director), Yoonkyeng Jang, Yesel Jun, Hyejin Kim

## Organizing Agency

Heeun Cho, Bannie Kim (CEO), Dohyun Kim, Ellen Eunsuk Lee (Project Manager)

## Publishing Office

Jong Hwa Park, Su Youn Park



Volunteers of Seoul ICM

# List of Sponsors

## Host

- International Mathematical Union (IMU)

## Partner

- Korean Mathematical Society

## Supported by

- Ministry of Science, ICT and Future Planning
- Seoul Metropolitan Government
- Korea Institute for Advanced Study (KIAS)
- National Institute for Mathematical Sciences (NIMS)
- Gyeongsangbuk-do
- Gyeongju City
- Korean Federation of Science and Technology Societies (KOFST)
- Korea Tourism Organization

## Sponsored by

- Samsung Electronics
- Kyungmoonsa
- Posco Group
- Naver Cultural Foundation
- Sempio
- Naver
- Dong A Science

## Media Sponsor

- Korean Education Broadcasting System (EBS)

## Supporter

- Hankook Ilbo

# Ceremonies



# Opening Ceremony (13 August 2014)

## Opening Address

### Hyungju Park, Chairman of the Seoul ICM 2014 Organizing Committee

President Park Geun-hye, Minister Choi Yanghee, Ambassadors, Members of the National Assembly, Excellencies, Distinguished guests, Ladies and gentlemen, Dear friends and fellow mathematicians,

On behalf the organizers of Seoul ICM, I am truly excited to welcome you from around the world to this Congress.

More than 125 countries are represented in this congress, and even more if we include the fifty one satellite conferences. I sincerely thank the International Mathematical Union for the help and support it provided during the past years, which saved us from many mistakes and pitfalls. And my heartfelt congratulations go to the prize winners to be announced today.

During the many years of preparations for this congress, the level of support from the government and corporations of Korea has been phenomenal. The law-making body of Korea, the National Assembly, adopted a resolution in support of Seoul ICM in November of 2013 and the Korean government declared the year 2014 as the Korean Mathematical Year in order to maximize the impact of Seoul ICM.

Several prominent corporations made considerable contributions to this Congress underscoring the growing importance of mathematics in the society. This experience of working together with many faces of our society will certainly help to open a new era of expanded roles of mathematics in the 21st century.

With an illiteracy rate close to zero, the education of children is often the highest priority for Korean families. This high regard for education and scholarship explains the steady influx of gifted students into the mathematics profession. It undoubtedly contributed to the rapid economic development of the country.

Our NANUM program required focused and concerted efforts of the Korean math community. It is our wish that the participants of this congress take the ICM excitement back home, further extending the positive impacts of the Congress to future generations in their respective countries.

This Congress also put much emphasis on public outreach programs. The public lectures by James Simons and by the Leelavati prize winner, the Baduk match (go game) against renowned masters, and the math movie projection event, to name a few, were made possible by the efforts of our outreach team. These efforts will undoubtedly contribute to making mathematics an essential part of mass culture of our times.



I hope that you enjoy and are rejuvenated by the exciting mathematical lectures and by the company of colleagues from afar. I hope you will also be able to savor some of the fine attractions that our country offers.

Now, this is the opening day. May the excitement last for the remaining days. Again, welcome.

## Welcoming Address

### Ingrid Daubechies, President of the International Mathematical Union



President Park Geun-hye, Minister Choi Yanghee, distinguished guests, Prize Winners and families, everybody who is participating and attending here:

In the opening images you saw scenes from past ICMs. We were in Hyderabad four years ago, and before that, in Beijing and Madrid. At the recent General Assembly, which was held just few days ago in beautiful Gyeongju, it was voted that the next ICM will be in Rio in Brazil. But you will hear much more about that in the Closing Ceremony. Now we are here, this year, in Seoul for ICM 2014 in beautiful Korea. During the opening film, you already were introduced to Korea's history, its culture, its beauty, its serenity. During the next few days, we will of course enjoy hearing about the recent advances in mathematics. We will celebrate the IMU Prize Winners. We will revel in being again the company of old

friends and in making new acquaintances, making new contacts, maybe laying the base for new collaborations in this international mathematical community. I hope that you will also find some time to enjoy Korea and its beautiful culture, its wonderful nature and its fabulous food. (I'm sure I've gained already several pounds in the last couple of days!) But above all, I think you will find that all this is made possible by the very smooth organization and very hard and sustained work of the Local Organizing Committee. I have had many occasions already, in the preparations to the ICM, to witness their dedication and efforts, and we will continue to do so during the whole ICM. I would like to thank all of them on your behalf, for this work.

To end this welcome address to the ICM, I want to express again my hope that you will enjoy the remainder of this Opening Ceremony and of the whole Conference. Thank you.

## Awards Ceremony

### Ingrid Daubechies

Dear respected participants of the Congress, I am greatly honored to host this memorable event, in my capacity as the President of IMU.

Today's Awards Ceremony at Seoul ICM has a new component compared to past ICMs. Each of the Fields Medalist and the Rolf Nevanlinna Prize winner, who have to be under 40, will be introduced to you by a short movie. These videos are the result of a collaboration of the IMU and the Simons Foundation; the IMU is grateful to the Simons Foundation for having accepted to fund and produce these movies. After each movie, the laureate will step forward

to be acknowledged before we proceed to the next movie. The Fields Medalist movies will be shown in the alphabetical order of the last names of the laureates; they will be followed by the movie for the Rolf Nevanlinna Prize winner. After the five movies are concluded, we will proceed to the actual handing out of these five IMU awards.

Fields Medals have been awarded by the IMU since 1936. They recognize outstanding mathematical achievement for existing work and for the promise of future achievement. From the start, they were meant for young mathematicians. The rule is now that to be eligible to receive a Fields Medal, mathematicians must have their 40th birthday after January 1st of the year in which the ICM is held. Let us now meet this ICM's Fields Medalists!

The four 2014 Fields Medalists are, in alphabetical order, Artur Avila, Manjul Bhargava, Martin Hairer and Maryam Mirzakhani.

We will now proceed to the Nevanlinna Prize. The Rolf Nevanlinna Prize has been awarded at every ICM since 1982. It was established by the IMU together with the Finnish Academy of Sciences. It recognizes outstanding contributions in mathematical aspects of information sciences. It is subject to the same age limit as the Fields Medal: to be eligible for the Nevanlinna Prize, the 40th birthday of the winner must be after January 1st of the ICM year.

The 2014 Nevanlinna Prize Winner is Subhash Khot.

At this time, the medals to the Fields Medalists and the Nevanlinna Prize winner will be given to them by the President of the Republic of Korea. We start with the Fields Medals. As before, we will follow the alphabetical order of the laureates' last names. Professor Myung-Hwan Kim, the President of the Korean Mathematical Society, and Professor Sug Woo Shin, Associate Professor of UC Berkeley and Assistant Professor of MIT, will assist the Awards Ceremony. As this ceremony proceeds, I will read the citations for each Medalist.

Artur Avila is awarded a Fields Medal for his profound contributions to dynamical systems theory, which have changed the face of the field, using the powerful idea of renormalization as a unifying principle.

Manjul Bhargava is awarded a Fields Medal for developing powerful new methods in the geometry of numbers, which he applied to count rings of small rank and to bound the average rank of elliptic curves.

Martin Hairer is awarded a Fields Medal for his outstanding contributions to the theory of stochastic partial differential equations, and in particular for the creation of a theory of regularity structures for such equations.

Maryam Mirzakhani is awarded the Fields Medal for her outstanding contributions to the dynamics and geometry of Riemann surfaces and their moduli spaces.

For the Nevanlinna Prize, Professor Pertti Mattila, the representative of the Finnish Academy, will join us on the podium; while he assists with the award ceremony proper, I will read you the citation:

Subhash Khot is awarded the Nevanlinna Prize for his prescient definition of the "Unique Games" problem, and leading the effort to understand its complexity and its pivotal role in the study of efficient approximation of optimization problems; his work has led to breakthroughs in algorithmic design and approximation hardness, and to new exciting interactions between computational complexity, analysis and geometry.

We will now announce the next two IMU Prize winners. The winners of both the Gauss Prize and the Chern Medal Award will be called to the stage and acknowledged first. We start with the first established of these two prizes, namely the Carl Friedrich Gauss Prize.

The Gauss Prize has been awarded at the ICM for the first time in 2006 and is now awarded at every ICM. The Prize was established by the IMU and the German Mathematical Society.

It honors a scientist whose mathematical research has had an impact outside mathematics—either in technology, in business, or simply in people’s everyday lives. The Gauss Prize will be presented by Professor Alfio Quarteroni, the Chair of the Carl Friedrich Gauss Prize Committee for 2014, who is now joining us on the podium. Professor Jürg Kramer, the President of the German Mathematical Society will represent the German Mathematical Society.

### **Alfio Quarteroni, Chair of the Carl Friedrich Gauss Prize Committee for 2014**

Stanley Osher is awarded the Gauss Prize for his influential contributions to several fields in applied mathematics, and for his far-ranging inventions that have changed our conception of physical, perceptual, and mathematical concepts, giving us new tools to apprehend the world.

### **Ingrid Daubechies**

It is now the turn of the Chern Medal Award. The Chern Medal Award was awarded for the first time at the 2010 ICM and is now awarded at every ICM. It was established by the IMU and the Chern Medal Foundation in cooperation with the Simons Foundation. It is awarded to an individual whose accomplishments warrant the highest level of recognition for outstanding achievements in the fields of mathematics. The Chern Medal Foundation and the Simons Foundation will be represented by Trustee May Chu of the Chern Medal Foundation and President James Simons of the Simons Foundation, who are now joining us on the podium.

I present the 2014 Chern Award winner, Phillip Griffiths. The chair of the Chern Medal Award committee, Robert Bryant, asked me to read the citation to you.

Phillip Griffiths is awarded the 2014 Chern Medal for his groundbreaking and transformative development of transcendental methods in complex geometry, particularly his seminal work in Hodge theory and periods of algebraic varieties.

At this time, we are ready for the awards to be handed to the Gauss and Chern Award winners. This will be done by President Park Geun-hye of the Republic of South Korea. The 2014 Carl Friedrich Gauss Prize winner, Stanley Osher!

For the Chern Medal Award, we will be joined by May Chu, the Trustee of the Chern Medal Foundation, and President Jim Simons of the Simons Foundation.

We will now proceed with the Chern Medal Award.

The 2014 Chern Medalist, Phillip Griffiths!

## Congratulatory Address

**Park Geun-hye, President of Republic of Korea**



Honored mathematicians, respected Ingrid Daubechies, President of IMU (International Mathematical Union), distinguished guests from home and abroad, and ladies and gentlemen, I am highly delighted that the International Congress of Mathematicians, which boasts more than one hundred years of history and tradition, is being held in Seoul.

Today more than 4,000 mathematicians from approximately 120 countries and plenty of young people dreaming to create a better future through mathematics have joined us here. I would like to extend my sincere welcome to all of them. Let me first express my sincere congratulations to seven mathematicians, who have been awarded the Fields Medals, the Rolf Nevanlinna Prize, the Carl Friedrich Gauss Prize, and the Chern Medal Award. In particular, I highly honor and admire the great spirit of challenge and passion of Dr. Maryam Mirzakhani, the first female to be awarded the Fields Medal in its history.

Ladies and gentlemen, the study of mathematics enjoys the longest history within academia and its magnificent legacy has been with us throughout the entire history of humanity. From ancient times when humanity still lived without letters people started to calculate and measure. Indeed, mathematics transformed the life of humanity as universal language, going beyond regions and nations, and serving as the basis of human logical thinking.

Even in this modern era, mathematics is still a critical foundation that stands at the center of the development of advanced science and technology and changes in our lives. Without mathematics, it would have been impossible to develop digital technology, which played a critical role in bringing about the ICT revolution. Without mathematics, we would now live in a world without our favorite movies and animations produced by computer graphics. By applying mathematical models in finance and analysis of Big Data, new services and markets are created. As we can see from these examples, mathematics allows us to solve problems with new methods and principles and creates much higher added values by converging with various fields, such as science and technology, industry, and culture and art. I firmly believe that the development of humanity in the future is closely intertwined with mathematics.

The world has now entered an era of creativity and innovation where a single individual's outstanding creation and ideas can move the entire world. In this sense, creative, logical and rational thinking that we acquire through mathematics is one of the most critical qualifications for our future leaders. That is why I sincerely hope that mathematics will develop not only as a pure academic subject for mathematicians but also as enjoyable and understandable learning

for the general public and our youngsters who will be leading the future. I would like to ask the honored mathematicians gathered here to inspire our young generation to enjoy mathematics and grow up as creative and talented individuals with a sense of creativity and rationality who ultimately contribute to the future of humanity.

Honored mathematicians, Korea achieved remarkable economic growth within a short period of time and a parallel advance was achieved in the study of mathematics in spite of a late start. Korea first joined the International Mathematical Union (IMU) in 1981 as a member country of the lowest ranking group I. Later in 1993, its status was elevated to group II and in 2007, it was promoted to group IV, climbing two rungs of the ladder at once. Koreans are deeply grateful to the world's mathematical community who have cordially welcomed Korean mathematicians into their midst. Under the name of "NANUM," we invited approximately 1,000 mathematicians from developing countries to share the dreams and hopes that Koreans have enjoyed. Korea will be more than willing to contribute to co-prosperity of the entire humanity by sharing our experience and know-how with the rest of the world in various sectors, including the economy.

Respected mathematicians, to my knowledge, more than 1,200 papers have been released through this Congress and various public lectures will be delivered by renowned mathematicians in tandem with financial investors. We count on your dedication and commitment to expand the academic foundation of mathematics through deeper and broader discussion and to ultimately contribute to the advancement of humanity. It is my sincere expectation that many students and public who find mathematics very difficult can have an opportunity to develop genuine interest in mathematics and find out the unique charm and joy that only mathematics can give.

Encircled by beautiful traces of our long history, Seoul is a city with numerous ancient wonders, such as our royal palaces, but also a modern city with state-of-the-art in industry and culture. Please enjoy the beautiful cultural heritage and vitality Korea offers and fill your hearts with precious and wonderful memories during your stay.

Thank you.

## **Announcements**

### **Ingrid Daubechies**

We have one more announcement to make, one more IMU prize to announce, namely the Leelavati Prize. The Leelavati Prize was awarded for the first time during the Closing Ceremony of the ICM in 2010. It was established by the IMU and the Indian government; it is now funded by Infosys, as a permanent IMU prize to be awarded at every ICM. It is awarded for outstanding contributions to increase the public awareness of mathematics as an intellectual discipline and the crucial role it plays in diverse human endeavors.

The 2014 Leelavati Prize will be awarded to Adrián Paenza for his decisive contributions to changing the mind of a whole country, namely Argentina, about the way it perceives mathematics in daily life and, in particular, for his books, his TV programs, and his unique gift of enthusiasm and passion in communicating the beauty of mathematics.

The prize itself will be awarded only at the Closing Ceremony; you will have your own chance of witnessing his enthusiasm and his passion during the public lecture he will give, in this very hall, at 8 in the evening on August 20, the last full day of the Conference.

Next, I have an extra surprise for you. Although all the IMU prizes have been awarded or announced, there is an extra announcement to be made and an extra award to be given.

The Chern Medal Award consists of several components. Two of these are a Medal and an award for the recipient, which Phillip Griffiths, the 2014 Chern Medal Award Winner, received just minutes ago from the hands of President of the Republic of Korea. But there is a third component, which is a check of 250,000 dollars that the Chern Medal Award Winner can direct to a charitable organization of his choice. I learned, with great delight, that Phillip Griffiths, has designated the African Mathematics Millennium Science Initiative, or AMMSI, as the designee of this award. AMMSI is an organization close to the heart and the concerns of the Commission for Developing Countries (CDC) of the IMU, it was featured yesterday at the one-day symposium Mathematics in Emerging Nations: Achievements and Opportunities (MENAO), organized by the IMU. AMMSI is affiliated with several CDC initiatives that help developing countries reach higher level of mathematics and sustain and foster their mathematical communities. I would like to invite May Chu and Jim Simons, who are representing the Chern Medal Foundation and the Simons Foundation, respectively, the Chern Medal Award Winner for 2014, Phillip Griffiths, and Wandera Ogana, the director of AMMSI, to join me on the stage, please.

Phillip Griffiths is known for not only his outstanding mathematics and his brilliant and effective service to the mathematical community; in recent years he has also been very active in fostering mathematics in developing countries, especially in Africa. I am sure that this award is appreciated not only by AMMSI but by every one of us who strives to work towards bringing mathematics everywhere. Thank you so much, Professor Griffiths.

IMU has become more active in the last few decades in fostering the growth of advanced mathematical education and research in Emerging Nations. The NANUM initiative, cited several times already and in particular by the President of Korea, was a wonderful initiative of the Korean Local Organizing Committee, as well as the whole Korean mathematical community, that fit beautifully in this whole framework, thereby freeing up CDC resources. Typically the CDC sets aside money every year, prior to every ICM, to bring mathematicians from developing countries and fund their participation in the ICM.

With part of these freed up resources, CDC organized the one-day MENAO symposium yesterday. The symposium made the case, through presentations of economics expert in these matters, as well as through case studies (Korea was one among those) and, examples, that promoting advanced mathematical development in a country benefits its economic development.

MENAO also showcased many further opportunities, where, with modest funding, a great impact can be obtained. We had presentations from a whole alphabet-soup of organizations, each of which does fantastic things with fairly modest means, reaching many mathematicians. Just to give a few examples: CIMPA organizes fantastic summer schools in all developing countries in the world; UMALCA, the Latin American and Caribbean association, coordinates immense efforts from the more developed countries in that region to help the less developed; TWAS, the World Academy of Sciences, strives to find fellowships for graduate students everywhere in the developing world; CANP, an initiative of ICMI, the instruction branch of the IMU, helps build capacity for mathematical education in less developed regions by establishing network programs. So many more initiatives were showcased at MENAO, among which AMMSI, which we just saw honored by the Chern Medal Award Winner.

In connection with all this, CDC is launching the Adopt-a-Mathematics-Graduate-Student initiative, which we hope will interest mathematicians in developed country interested in mentoring and helping support a student in a developing country. If this is your case, CDC is working on a framework to match you, one-on-one with such a student. You can find a preliminary description on the Friends of the IMU webpage; more will come on the CDC webpage soon.

At MENAO, we couldn't yet announce the fantastic gift directed by 2014 Chern Medal Award Winner to AMMSI, because the Chern Award winner hadn't been announced publicly. But we had yet another fantastic announcement, which I am happy to also broadcast to all of you here. Namely, the five inaugural winners of the Breakthrough Prize in Mathematics, Simon Donaldson, Maxim Kontsevich, Jacob Lurie, Terence Tao and Richard Taylor, have let me know just a few days ago that they will each donate 100,000 dollars, for a total of 500,000 dollars, to the IMU CDC to endow a fund that will award Breakout Graduate Fellowships to math grad students from and in the developing world.

The IMU is profoundly grateful for the support from Phil Griffiths and from the five Breakthrough Prize Winners to graduate education in the developing world. It also hopes that this generous and shining example by top leaders of our mathematical community, who believe in the small and collective efforts that we make and the impact these have, will inspire others, both within and from outside the mathematical research world. Should you already want to emulate them in a small way, you can do so right during ICM by participating in DonAuction, a fundraising initiative that will last only for the period of the ICM. Check it out at [www.donauction.org](http://www.donauction.org) or at the IMAGINARY stand in the ICM Exhibition space.

IMU's most important business listed at the top of its charge is the organization of our quadrennial International Congress. During the Congress, we will hear Plenary Lectures by outstanding mathematicians, who have been asked make them accessible to a wide range of mathematicians present here. We also will have Invited Lectures in many different directions. This ICM will have a record number of cross-listed talks, in the different sectional meetings, illustrating the large extent to which different subfields within mathematics are cross-fertilizing and influencing each other, a wonderful development. We will, of course, all celebrate our Prize Winners, who will give their own lectures. We will have public evening lectures, the first one tonight by Jim Simons and the last one on August 20th by the Leelavati Prize Winner, Adrián Paenza. Apart from all this, I hope that you will also enjoy some of the outreach activities, and visit the ICM exhibition space. And if you get a chance, visit the Bridges 2014 conference, held in parallel to ICM, just a subway ride away in Gwacheon Museum organized by the Bridges Organization of Math and Art.

I hope you will fully enjoy the ICM, the core of the Conference itself, the many other activities that the Local Organizing Committee has organized around it and your stay in the city of Seoul. Thank you so much for coming.

## IMU Status Report

### Martin Grötschel, Secretary of the International Mathematical Union



The exquisite glamour and the particular thrill of this opening ceremony are almost over. It is now time for the “boring stuff”. The IMU Secretary is supposed to report, in the last presentation of this event, a glimpse of the “State of the Union”.

Before doing this, let me mention that there has been very hard work behind the shine that you have seen and the lightness and friendliness that you have experienced in Seoul so far.

I chaired of the organizing committee of ICM 1998, and I remember the effort involved well.

Having been in close contact with the ICM 2014 organizing committee, I do know what Hyungju Park and his team have suffered in the last years, in particular in the recent months and days. Let us not forget, the team consists of volunteers and they do all that in their free time with great energy and outstanding enthusiasm. Please give an extra applause to the colleagues involved in the organization of this great congress and all the additional activities associated with it.

My job is to report to you about those people and organizations who have been working in the last years behind the scenes for IMU and the mathematical community in general. I want to point out that it has been a pleasure to collaborate with all the colleagues. Consensus could always be reached, even despite initial dissents, since everyone served a joint good cause: to promote and foster mathematics. Having been active for IMU for the last 20 years and finishing my term as IMU Secretary at the end of this year, I can honestly state that I have enjoyed all the work, that I feel happy to be a mathematician and to belong to this wonderful community.

The role of the IMU Secretary is not to provide you with a vision of mathematics and tell you how I or how IMU thinks the future of mathematics is going to be. Many lectures at this congress will take care of that. I will tell you about some details of our work so that you know what IMU has done in the recent four years. It will be brief, and I will highlight only a few memorable topics.

The ICM-related committees are of particular importance. You have seen the winners of the IMU awards a few moments ago and you have certainly studied the list of invited ICM lecturers. Many colleagues were involved in choosing them. The IMU Executive Committee (EC) set up the ICM 2014 Program Committee and one committee for each of the IMU prizes: the Fields Medals, Nevanlinna Prize, Gauss Prize, Chern Medal Award and the Leelavati Prize as well as for the ICM Emmy Noether Lecture.

The *Program Committee* (PC) is responsible for the scientific program of ICMs; the 2014 PC was chaired by Carlos Kenig and had eleven further members: Erwin Bolthausen, Alice Chang, Welington de Melo, Hélène Esnault, Tim Gowers, Ravi Kannan, Jong Hae Keum, Claude Le Bris, Alex Lubotzky, Jarik Nesetril, Andrei Okunkow. The PC was supported by nineteen section panels. They jointly succeeded in coming up with an outstanding list of speakers and in reaching a reasonable balance of regions, gender and mathematical fields. A rough count shows that the number of PC and panel members is about the same as the number of invited speakers, i.e. one PC/panel member chose one speaker - a truly significant selection effort.

The *Fields Medal Committee* consisted of Luigi Ambrosio, David Eisenbud, Kenji Fukaya, Étienne Ghys, Benedict Gross, Frances Kirwan, János Kollár, Maxim Kontsevich, Michael Struwe, Ofer Zeituni, Günter M. Ziegler and was chaired by IMU President Ingrid Daubechies. This committee is fully responsible for the choice of the award winners. The same rule holds for the other IMU prizes. The EC does not interfere; all IMU prize selection committees work autonomously—their choice is IMU's choice. I think we all agree that great choices were made.

The *Selection Committee for the Nevanlinna Prize* was chaired by Avi Wigderson and had Thierry Coquand, Yuri Nesterov, Jaikumar Radhakrishnan, Éva Tardos, and Leslie Valiant as additional members. The *Gauss Prize Selection Committee* consisted of Weinan E, Barbara Keyfitz, Andrés Weintraub, Aad van der Vaart and Alfio Quarteroni as chair. Robert Bryant chaired the *Chern Medal Award Selection Committee* which had Kazuo Murota, Felix Otto, Alain-Sol Sznitman and Claire Voisin as additional members. Finally, David Mumford chaired the *Leelavati Selection Committee* and was supported by Oh Nam Kwon, Guillermo

Martínez, M.S. Raghunathan, and Srinivasa Varadhan. The task of the Leelavati Selection Committee is particularly difficult since it has to search the world with its many different languages and cultural habits to find a person that contributed significantly to the public awareness of mathematics as an intellectual discipline and of the crucial role it plays in diverse human endeavors. A wonderful Argentinian prize winner was detected. Please, attend his lecture on August 20.

The responsibility for all IMU activities rests with the IMU Executive Committee (EC), which consists of a President, a Secretary, two Vice Presidents, six Members-at-Large, and the Past President, who has no voting rights. The EC is elected by IMU's General Assembly (GA), which appoints the EC for a four-year term. The GA is the "international parliament of mathematics" and consists of the delegates of all members of IMU. Each member is represented by a number of delegates that depends on the membership group (1 to 5) it adheres to.

The GA met on August 10 and 11, 2014 in Gyeongju and decided on the IMU leadership for the term 2015–2018. Shigefumi Mori will be the next President, Helge Holden the IMU Secretary, and Alicia Dickenstein and Vaughan Jones the new Vice Presidents. The Members-at-Large will be Benedict Gross, Hyungju Park, Christiane Rousseau, Vasudevan Srinivas, John Toland and Wendelin Werner. Shigefumi Mori, a former Fields Medalist, is the first IMU President from Asia ever. VP Alicia Dickenstein is the first mathematician from Argentina joining the EC and Hyungju Park the first Korean member of the EC. With Vaughan Jones and Wendelin Werner the EC has three Fields Medalists as its members: the highest EC Fields Medal density ever. Together with Ingrid Daubechies, these eleven colleagues will work hard in the next four years to promote, encourage and support many international mathematical activities.

There are subtle issues that the IMU usually does not mention in public. IMU often receives requests to help mathematicians who have been imprisoned for political reasons (not for crimes) or who have been treated unfairly. Advice is requested from political institutions intending to advance mathematics. Mathematical institutes in danger of getting shut down or in financial trouble ask for support in their struggle to survive, and we seek for donors and sponsors for mathematical activities. IMU representatives work behind the scenes and try their best to help mathematicians wherever possible. If you have crucial difficulties and feel that international assistance might be of advantage, just send a message to the IMU Secretary.

Significant work is done in IMU's Commissions and Committees. The largest commission is the *International Commission for Mathematical Instruction* (ICMI) that, founded in 1908, is actually older than IMU itself. ICMI has a wide range of activities. I want to mention just one which I consider to be of particular importance: the Capacity and Network Project (CANP). CANP aims to enhance mathematics education at all levels in developing countries making their people capable of meeting the challenges these countries face. CANP helps develop the educational capacity of those responsible for mathematics teachers and creates sustained and effective regional networks of teachers, mathematics educators and mathematicians, also linking them to international support. The major activity of a CANP project is a two-week workshop for about forty participants, half of them coming from the host country and half from regional neighbors, primarily aimed at mathematics teacher educators. CANP workshops have been held in Mali with participants from Sub-Saharan Africa in September 2011, in Costa Rica in August 2012 with Central American and Caribbean participants, in Cambodia in 2013, and the next one will be held in Tanzania in September 2014. Financial support came primarily from IMU, contributions from UNESCO, ICSU, ICIAM, RECSAM and SEAMS are acknowledged.

The Internet and the World Wide Web have transformed mathematical communication at least as much as the introduction of journals. This transformation and many of the commercial pressures affect mathematicians in many ways. The IMU EC formed the *Committee on Electronic Information and Communication* (CEIC) in 1998 to watch these developments, advise the EC, through it the IMU and mathematicians generally about these trends and best ways to adapt to these changes. I want to mention here only three panels organized by CEIC at this congress that represent typical aspects of CEIC activity. Lead by the CEIC chair Peter Olver, CEIC members and invited experts will discuss the perspectives of “Mathematical Massive Open Online Courses”, provide their views of “The Future of Mathematical Publishing” and describe efforts that may lead to “The World Digital Mathematics Library” that we are all dreaming of, namely an electronic repository that makes the mathematical literature of all time online available for everyone everywhere in the world free of charge. Please, stay informed by attending these panel discussions.

The Commission for Developing Countries (CDC) has been mentioned so often so far at this opening ceremony that I will not add much to it. It is impossible, though, not to praise in this context the NANUM project that is an initiative of our Korean colleagues to invite about 1,000 mathematicians from the developing world to ICM 2014. I hope that the NANUM grantees will seize the unique chance to meet and network here with colleagues from the world over. The MENAO (Mathematics in Emerging Nations: Achievements and Opportunities) Symposium that CDC launched yesterday capitalized on the NANUM grantees and brought together a large number of mathematicians from the developing world, active and potential sponsors, and colleagues with particular interest in supporting mathematics in developing countries. There were outstanding lectures that showed how mathematics has helped shape countries, lives, individuals and communities. These lectures were more than a loud call to IMU to keep on going in this direction.

At the GA meeting last weekend, Wandera Ogana has been elected as the new CDC president, Herb Clemens and Kesavan as the CDC secretaries, and the three new CDC members representing Latin America, Africa and Asia come from Columbia (Alf Onshuus), Cameroon (Mama Foupouagnigni) and Philippines (Polly W. Sy). Three more CDC members will be appointed in the near future.

During the last four years the following countries have joined IMU as Associate Members: Cambodia, Gabon, Madagascar, Malaysia, Moldova, Nepal, and Oman. The applications of Papua New Guinea and Senegal were approved two days ago. They will be new IMU Associate Members as of September 1, 2014. IMU has three new Full Members: Montenegro, Algeria and Ecuador, which was upgraded from Associate to Full Member.

To summarize the membership development: IMU has now seventy-one full member countries, twelve are associate members, and four international organizations are affiliate members. At ICM 2014, we count participants from 125 countries, and thus, there is more room to grow.

Papua New Guinea and other new Associate Members are not hotspots of mathematics yet, but IMU is trying hard to develop mathematics everywhere. Let us look at Korea and let me go back with you to the year 1981. How many papers did mathematicians from the Republic of Korea, the host country of this event, publish in 1981 in international journals? Have your own guess! The almost unbelievable answer is 3! Korea is now number 11 in the world ranking of mathematical publications. What a progress! Let us hope that some of the countries that now became IMU members or associate members will experience the same steep growth that, to a large extent, is based on a strong emphasis on education.

I started my presentation with saying that I am not going to tell you visions about mathe-

matics; but I have one slide that I would like to show you - and I hope it indicates something to think about. It is my firm belief that mathematics is THE scientific endeavor of this century! The reason for my conviction is that all advanced industries and all advanced sciences have meanwhile understood that mathematics is utilized in great depth and breadth for the understanding of nature, the modelling of industrial processes, the shaping of products, etc. Companies that want to stay competitive need mathematics, and if we want to save resources and make careful use of our environment, mathematical modeling, simulation and optimization are indispensable.

All this is unfortunately not so well known in the public, and that is something we have to change by intensifying our outreach activities, one reason why the Leelavati Prize was set up. It may sound strange, but also many mathematicians are not aware of the influence that mathematics has in real life in the world around us. It seems that we have to explain also within our community how important mathematics is.

Yesterday at the MENAO event, Eric A. Hanushek, an economist investigating the influence of education on economic development, reported his finding that cognitive skills are causally related to economic development and that variations in growth rates across countries can be explained by consideration of the role of cognitive skills. He emphasized the importance of mathematical education in these processes. Hanushek's observations were corroborated by Korea's former Minister KunMo Chung, who spoke about the contribution of mathematics to the development of a country. The clear statement was that the development of skills, in particular mathematical skills, is the most important aspect of development and growth. He also stressed that education at the top level does not suffice, good mathematical education on all levels is necessary. That is a message we have to bring home to the ministries in our own countries. We have to work on the whole range of education in order to go forward and grow.

My final slide is an organizational chart. Four years ago the General Assembly decided that the International Mathematical Union should have a permanent office. Since 1920 IMU has been run by volunteers only. The GA felt that some sort of professionalization would be necessary and asked for bids for establishing a permanent IMU Secretariat. Berlin won the contest and the new office with five employees was opened in the beginning of 2011. But, of course, IMU is not dominated by the Secretariat; it is still lead by the committees and commissions I have told you about and run by volunteers who contribute to IMU in their free time. Look around and see what volunteers can achieve!

This is the end of my report about what IMU has been doing. I do hope you feel somewhat encouraged thinking about contributing to IMU and its activities in the future. Please, also consider contributing to the development of mathematics in your own country, join the mathematical organizations and societies in your scientific environment, promote our science and make it stronger.

Thank you for your attention.

## Opening Ceremony of Seoul ICM 2014



2014 Award Winners with President Park Geun-hye and Minister Yanghee Choi of Republic of Korea, Ingrid Daubechies, Martin Grötschel, and Hyungju Park



Dance performance at the Ceremony



Martin Grötschel, Ingrid Daubechies, and Hyungju Park



Artur Avila



Manjul Bhargava



Martin Hairer



Maryam Mirzakhani



Subhash Khot



Stanley Osher



Phillip Griffiths



2014 Award Winners

# Closing Ceremony (21 August 2014)

## Closing Remark and Result Report

### Hyungju Park, Chairman of the Seoul ICM 2014 Organizing Committee

Ladies and gentlemen and my fellow mathematicians, I am so happy to be here in this very gracious ending of what we have been going through for nine days. We wanted this to be, instead of a formality, we wanted this Closing Ceremony to be more festive and filled with expectations for the next one. Here, you will soon hear from the next ICM Chair about their plans and we already started looking forward to that. So I will briefly report some numbers and some statistics about this ICM so that maybe you can all share what we have achieved and what we have done for the past nine days. Before we move on, during our Conference Dinner, I confessed some of our mistakes and people thought that was fun. So I will reveal some more of our mistakes today.

Early this spring, we were told that the Pope will be visiting on the day of the Opening of ICM. So we all panicked because we were hoping that the President of Korea will come to our Opening and award the medals and prizes and it's not easy to compete against the Pope. We were very worried and eventually the Vatican graciously changed their schedule. My suspicion is that some of the members of the Local Organizing Committee who are Catholics maybe made some extra efforts to change the Vatican's mind.

Another big crisis hit us several months ago also when we discovered that COEX, this convention center, was undergoing a huge renovation project. If you actually go down to the basement, where hundreds of small restaurants and cafeterias are, the whole area has been renovated. That means that poor mathematicians will be busy listening to the lectures but will be starving. So that was another crisis we had. We worked with many people and the COEX people and they actually helped us a great deal. They rented us big rooms, C1, for free. That room was given to us for free so that we can run a small food court there. That food court was offered to us free of charge and that solved the problem. But, in a sense, it actually was even better than usual because mathematicians could all come there and chat over lunches and I think it worked better that way. At this point, I would like to show my gratitude to the COEX management for being so flexible.

Of course, a TV station called EBS was very gracious and they actually put the Opening Ceremony on air, live, during the Opening. It was like a sports event being broadcasted live with commentators. So this time, our Opening Ceremony was broadcasted to whole Korea live with two math professors as commentators. It created a new job. Now mathematicians can be hired as commentators.

You might have noticed we have filmed all the plenary lectures, invited lectures and important events. They are being uploaded to Youtube and they are all available with links on our homepage. So you can actually re-watch. If you missed any interesting lectures, you can watch them through our homepage now. We didn't hire professionals for that. Twenty-some volunteers teamed up and they rented equipment. They filmed all the things, did the editing and stayed until 1 a.m. each day to finish editing and put them up on Youtube. Because our

Wi-Fi bandwidth is limited, they couldn't do it during daytime so they did it until 1 a.m. in very late nights. And I think it's really touching that these volunteers were not paid at all. We just covered their transportations. I think that's a huge act of sacrifice and dedication. So, thanks to their efforts. I think the whole math community of the world now can watch and savor the memories and interesting mathematics that was presented in this Congress.

There were other things but I would just like to say one thing, though. We have offered to invite 1,000 mathematicians from developing countries. Not all of them made it here. I wish to express my sorrow for that. It turned out many of them didn't have an international travel experience before so assumed that visas can be obtained on the spot on the same day, which is not usually the case. Whenever somebody called us asking for help in obtaining a visa to Korea, we called the Korean embassy in that country and we requested emergency actions and usually it worked. The Korean foreign ministry supported our initiatives a lot and they acted to issue the visa in an expedited manner. However still, many of the NANUM recipients didn't get visas and didn't let us know and I feel very sorry that many NANUM recipients couldn't come because of the visa problems. As far as I know, there are two mathematicians who wanted to come but could not because of political reasons. I hope the international math community would address this in the future so that this is not repeated. I will not release the identities of those two but I know as a fact that they did not obtain passports from their government because of their political beliefs.

We are very grateful for all the things. We are grateful for the good things that happened and we are grateful for your forgiveness for what we didn't do well. And I thank you for just seeing the bright side and telling us that we did a good job. I know we probably screwed up some of the things but thank you for not telling that to us now. But later, feel free to let us know so that we can pass that onto the next organizers so they don't repeat our mistakes.

Ok, so I guess I will show some numbers. I will do it very quickly. We had four Fields Medalists. Let us congratulate them again. We have more.

We had these presentations. This is scientific programs. So we had a lot. We had minimal, I think, no-shows. Every conference has some no-shows. That's not avoidable, I think. But we had very little. The number of invited lecture presenters, we had 188. We had many abstracts. You see the number is a little too much because the Congress participants, we had 4,680 regular participants plus 537 accompanying persons. So we had a total of 5,217 people who are registered. We had many media representatives here, many journalists. And there were over 20,000 high school kids and from the general public visiting us. How did we count them? Because we asked them to sign up for the Simons lecture and these big events and then the people who just came without prior signing-up, we charged them one dollar for entrance. So we can now count the money. So that's how we came up with 21,227 participants of this Congress. I thank the general public for the enthusiasm. So that's the number of participants. There were a huge number of Korean participants and there were many colleagues from USA. That's the statistics regarding the regions and that's the number of participants.

NANUM. Eventually I told you that we issued invitations but many people couldn't come and that's the end result. Those are the ones who actually brought their documents and got the reimbursements. So those were the reimbursements we issued. I hope that everybody got their reimbursement so far. By the way, there were 85 countries represented in the NANUM program.

We had, again, a lot from the general public; especially the exhibitions were very well visited. We had many booths. We had 564 staff members: 63 members of the Local Organizing Committee and we had, more than anything, 282 volunteers on site. Again, my deepest gratitude goes to the volunteers. These boys and girls really did a great job and they are still

here helping us. By the way, after this Closing Ceremony, we're throwing a party for them so they will have a party all night.

Believe it or not, there have been close to 1,500 media coverages about this Congress during this Congress. This doesn't count the coverages made before the Congress. I think this is unheard-of, unprecedented, and this will be what we will start with. Other people in other science disciplines in Korea are just envying mathematicians for having done this. So these are the articles that appeared. By the way, this LED is too bright that no pointer will work with it. I'm very sorry to the plenary speakers who gave talks here, who couldn't use their pointers because pointers simply do not work with this. So you see that the Korean public was especially thrilled to see the first female Fields Medalist together with the first woman president of Korea together with the first woman president of IMU.

I know I overused my time. I was pretty on time during the Opening so this is a payback. So thanks a lot and we will go on with others, especially the next ICM plan will be exciting. So let us hear from others.

## Awards Ceremony

### Ingrid Daubechies, President of the International Mathematical Union

The time has now come for the Awards Ceremony for the Leelavati Prize.

The Leelavati Prize was awarded for the first time during the Closing Ceremony of ICM 2010 in Hyderabad. The prize was established by the IMU and by the government of India; it is presently funded by Infosys, as a permanent IMU prize to be awarded at every ICM.

The Leelavati Prize accords high recognition for outstanding contributions to increasing public awareness of mathematics as an intellectual discipline and the crucial role it plays in diverse human endeavors.

Following the precedent set at its first awarding at ICM 2010, it was decided that the Leelavati Prize winner will be announced during the Opening Ceremony, but that the award ceremony itself is part of the closing exercises of each ICM. Adrián Paenza, the 2014 Leelavati Prize winner, gave an exciting public lecture yesterday in this hall, in which he inspired kids (who mobbed him afterwards) to not give up on math, so they would end up seeing the beauty in it, and he encouraged all mathematicians to be more involved in the teaching of mathematics in schools so that we can show them the "right door" to which to enter math.

Adrián Paenza: I would like to call you forward to hand out the prize to you. You have been awarded the prize for your decisive contributions to changing the mind of a whole country about the way it perceives mathematics in daily life and, in particular, for your books, your TV programs and your unique gift of enthusiasm and passion in communicating the beauty and joy of mathematics.

Infosys, the company funding this prize, had hoped to be able to send an officer to represent them, but in the end scheduling problems prevented this. Instead, Mr. Narayana Murthy, the founder of Infosys and the Chairman of their Board, asked me to read the following statement.

"Infosys is proud to sponsor the Leelavati Prize, which recognizes contributions in public outreach in mathematics. I would like to congratulate Adrián Paenza on winning this prize. Mathematics is often viewed as complicated by students and adults alike. Adrián has translated his love for the subject into work that addresses this issue via popular media like books and television. I'm sure this has helped remove some mysticism and phobia around mathe-

mathematics for many. We are pleased to recognize his accomplishments and enthusiasm in these fields to the Leelavati Prize. Thank you.”

### **Adrián Paenza, 2014 Leelavati Prize Winner**



It is a great honor for me to be here. But as an Argentinian mathematician, I am not unique in being honored this year. Argentinian mathematicians have been doing great work for so many years, and this year several of us are being honored. Miguel Walsh, just 26 years old, got the Ramanujan Prize; Alicia Dickenstein was elected as Vice-President of the IMU a few days ago, and now it is my turn, receiving the Leelavati Prize. South America in general can

be proud – with Arthur Avila receiving a Fields Medal at this ICM, last week. None of these achievements are isolated; seeing them as part of a whole framework makes us even happier.

Math is great. We need to grant public and free education for everybody. We also need to understand that education is a basic human right. We, mathematicians, should be more involved, as Ingrid was saying, in trying to lead our kids thru the ‘right door’.

So, thank you very much and I also want to say a couple of words in Spanish, thanking also my students and my colleagues at the University of Buenos Aires. Without them, my work would definitely have been impossible. So...

“Muchisimas gracias a todos los argentinos también. Sientan este premio como que es un premio para ustedes. Mi gratitud para todo mi país”. (Translation: “Many thanks to all the argentinians too. Feel this award as an award to all of you. My gratitude to everybody in my country”.)

### **Address of IMU President**

#### **Ingrid Daubechies**

This is the last event of the ICM. There will be many IMU activities after the ICM; let me tell you a little bit about them. In January 2015, a new Executive Committee will take over. Today we have here our new President, Shigefumi Mori, who will speak after me (this is the last time I get to address you all in my capacity of IMU President!) Helge Holden will be the new Secretary. Alicia Dickenstein and Vaughan Jones will take office as Vice Presidents, and we have also new Members-at-Large of the Executive Committee. I will still be present at Executive Committee meetings but without a vote. So I am fading away, and happy to do so.

At the General Assembly in Gyeongju, the IMU created the IMU Circle, which consists of former organizers, committee chairs, EC members of the IMU and other people who have been a great service to the IMU. Their list will soon be posted on the IMU website, which is, as you all should know, [mathunion.org](http://mathunion.org).

We also have new members for the International Commission on the History of Mathematics (ICHM) and new members will soon be added to the Committee on Electronic Information and Communication (CEIC). All this information can be found on [mathunion.org](http://mathunion.org). The new members of the Commission for Developing Countries I would like to introduce

explicitly. They are President Wandera Ogana, C. Herbert Clemens, Srinivasan Kesavan, Alf Onshuus, Mama Foupouagnigni and Polly Sy from the Philippines.

At this ICM, apart from the programs that the Programs Committee put together, there were also three extra panels that were organized by the Committee on Electronic Information and Communication: Mathematical Massive Open Online Courses (MOOCs), Future of Publishing and World Digital Mathematics Library (WDML). Blogs on the first and the third topics will be accessible via [mathunion.org](http://mathunion.org) as always. On the second topic, there did not seem to be any apparent feeling that a blog was needed. If many of you feel that the blog on publishing should be reopened, please let us know.

Just prior to the ICM, CDC organized a one-day symposium on “Mathematics in Emerging Nation: Achievements and Opportunities”, which we call MENAO for short. Lots of the material that was available, talks and other materials will be posted on the CDC page on [mathunion.org](http://mathunion.org). (At the end of this, I hope you will know, almost like a mantra, that you need to go to [mathunion.org](http://mathunion.org) to find news about the IMU!)

There was also a longer report written on IMU in the developing world in the context of this meeting, which is called “The IMU in the Developing World”. That report, since it was suggested by the organization Friends of the IMU (FIMU), has been posted on FIMU’s website, [friends-imu.org](http://friends-imu.org).

Yesterday, as you saw on the video, DonAuction culminated with a live prize drawing emceed by Cédric Villani, who did a wonderful job, as always. There were, in the end, over 400 different people who contributed, here and online. Over 400 people... And the total was over ten million Korean Won, I think that deserves your applause! All of you, you are friends of the IMU.

For the transparency of the whole drawing process, we will post everything on [donauction.org](http://donauction.org). Please give us a little time to get home and to get over sleep deprivation and then everything will be posted. If you missed the whole DonAuction initiative, don’t worry. There will be other opportunities on Friends of the IMU to donate to its efforts to raise money for the CDC. These efforts with DonAuction are a mirror at a much smaller scale, at a scale we can afford, of the fantastic gifts that I announced at the Opening Ceremony of the five Breakthrough Prize winners and of the 2014 Chern Prize Winner, who together directed 750,000 dollars towards the efforts of the CDC.

That was what I wanted to tell you about IMU after the ICM. I am now very happy and honored to introduce to you our wonderful new IMU President, Shigefumi Mori.

## **Address of IMU President-Elect**

### **Shigefumi Mori, President-Elect of the International Mathematical Union**

On behalf of the newly elected members of the Executive Committees (ECs) of IMU and its commissions, I would like to express our sincere gratitude to the Nominating Committee and the Election Committee both chaired by Professor Ragni Piene and Delegates of General Assembly (GA) of IMU.

Following a Chinese proverb that encourages one to study the past to learn new things, I would like to comment on the spectacular success of ICM 2014. The success depended on the following people.

EC members, in particular, President Ingrid Daubechies, Secretary Martin Grötchel. Committees involved in the Local Organization in Korea: Executive Organizing Committee (OC)

chaired by Prof. Hyungju Park, IMU GA OC chaired by Prof. Jeong Han Kim, and Local Program Committee chaired by Prof. JongHae Keum. I only mentioned three, but there are more committees and especially many more people behind them, without whom this congress would not have been this successful.

For the MENAO event just before ICM 2014, Korean Government launched a project called NANUM which invited 1,000 researchers from developing countries. This is quite significant, and they have set a new style of contribution.

Although not apparent from the surface, the academic content of ICM 2014 was designed by the Program Committee chaired by Professor Carlos König. Without this committee the whole congress is just impossible, and again behind it there are Panel Committees and so many people involved. Although Hyungju Park mentioned 5,000 people from overseas there are many more people involved.

ICM 2014 is also receiving generous support from many organizations, including Friends of IMU, Simons Foundation, Niels Hendrik Abel Board, and Mathematics Societies of many countries.

Very helpful volunteers gave a personal touch to ICM 2014.

Finally I should add Medalists, who volunteered to show up in social events and they were really helpful.

IMU is fortunate to be supported by so many people as above. I feel happy to be part of it, which is one of the reasons why I accepted to be the President nominee.

Now looking ahead, I would like to mention my colleagues in the newly elected IMU EC. Though I have some experience of IMU, various things have changed since then, and I would like to learn especially from the current members, and it was very fortunate that all the people elected just happened to be here at the Congress and we could meet.

The members made not only excellent academic achievements but also sincere services to the mathematics community. For instance, the Secretary elect Professor Helge Holden has a broad experience with European Mathematical Community and we already started working together and going to make a good team. I can continue talking about other members, but this is not the right moment and I just say that I am confident in my fellow EC members.

I should also mention the stable IMU Secretariat in Weierstrass Institute at Berlin. This is new to me since this did not exist when I was involved in EC more than ten years ago.

Furthermore IMU circle, as mentioned by President Daubechies, was formed this year, which consists of mathematicians who have made sustained and distinguished contributions to IMU.

These are the new people with whom we will work together and we will also have a new committee for women in mathematics. Though its name is not fixed, President Daubechies will continue to be the key person in the committee.

There are new things we have to cope with. Mathematical communities are emerging in developing countries and MENAO(Mathematics in Emerging Nations: Achievements and Opportunities) organized by CDC(Commission for Developing Countries) was very timely for this direction. Education is indispensable for mathematics in developing world, and ICMI(International Commission on Mathematical Instruction) should take part. They should work with the help of ICSU(International Council for Science), that is, under the umbrella of ICSU. There are also problem of world digital mathematical library, and the list continues.



I do not have any intention to pull IMU in any specific direction. I would like to listen to various people, which is the Japanese or Asian way. My only main concern is to contribute to the promotion of the international cooperation in mathematics. This is how I view the IMU Presidency.

## Introducing ICM 2018

### Marcelo Viana, Chair of the ICM 2018 Organizing Committee



Dear Colleagues,

On August 11, at Gyeongju, the General Assembly of the IMU unanimously approved the Brazilian bid to organize the International Congress of Mathematicians ICM 2018 in Rio de Janeiro. This will be the first time, in its more than centennial history, that the ICM will take place in the Southern Hemisphere.

We are honored by the IMU-GA's decision and thrilled by the perspective of bringing the

ICM—and all that it embodies—to Latin America. Ours is a young region of the world, where the Congress can and will be a powerful tool to disseminate Mathematics in the whole society, especially among the younger generations. Indeed we have chosen “Sowing Seeds” as the theme for the Rio de Janeiro Congress.

We are also daunted by the challenge of following on our Korean colleagues footsteps: Hyungju Park and his team did a terrific job in making this year's Congress a big success and theirs will indeed be a tough act to follow. But be assured that we will put the best of Brazilian creativity and ingenuity to the task of making ICM 2018 an equally memorable event.

Actually, preparations for the Congress are already actively under way. The website went live a few days ago (check [www.icm2018.org](http://www.icm2018.org)) and it is now possible to submit proposals for satellite events. Most specially, I invite you all to sign-up for the ICM 2018 Newsletter: just go to the website, click on Newsletter and fill-in your name, email address and country. It only takes a small fraction of a minute! And it will help us keep you current with the preparations.

Até breve no Rio de Janeiro! (See you soon in Rio de Janeiro!)

## Vote of Thanks

### Ingrid Daubechies

Every day, ICM participants told me how much they were enjoying the Congress. So many people worked very hard to make this ICM a success – and the time has come to thank them!

The Program Committee and the members of the Panels for the different disciplines carefully put together the scientific program that we all enjoyed. The prize selection committees did their thoughtful and considered work to select the Prize Winners. The Plenary and Invited Speakers, and the people who assisted them in preparing their talks, surely deserve our thanks for preparing carefully and giving us clear presentations of their field and their work. The Chairs of all the sessions helped keeping the complex schedule on track. The different Panels, organized by the ICM and the IMU, led to an interesting dialog with the audience on

the topics that were their focus. The contributed talks provided a rich and diverse collection of results. The exhibitors gave us even more mathematical food for thought.

The receptions and parties were wonderful and we are grateful to each organization that hosted one! We owe special thanks to the Mayor of Seoul, who hosted the banquet for the Congress. And the IMU is extremely grateful to the President of the Republic of Korea, who did us the great honor of attending our Opening Ceremony and awarding the Fields Medals, the Nevanlinna Prize, the Gauss Prize and the Chern Medal Award.

But the people we should thank most of all are the local organizers! The Local Organizing Committee, led by its Chair, Hyungju Park, worked incredibly hard for more than four years, in order to get everything organized into the smallest detail – and to deal brilliantly with every crisis as it came up. They were assisted by a veritable army of incredibly devoted volunteers, who we thank most wholeheartedly as well. The NANUM program was a wonderfully generous initiative and it made it possible for so many mathematicians to come to an ICM, an experience that would otherwise have remained just a dream for them... Let's thank NANUM's Chair, Dongsu Kim, and the whole Committee for its hard work.

Finally, I want to thank all of you, participants who came from far away or from nearby, and who, with your enthusiasm for mathematics, certainly contributed a lot to the success of the 2014 International Congress of Mathematicians in Seoul!

## Closing Ceremony of Seoul ICM 2014



Myung-Hwan Kim, Hyungju Park, Ingrid Daubechies, Shigefumi Mori, and Marcelo Viana



Members of the Organizing Committee



Screening of a slide show



Participants at the Ceremony

**The Work of the Winners of the Fields Medal,  
the Nevanlinna Prize, the Gauss Prize,  
and the Chern Medal**



# The Work of the Winners of the Fields Medal, the Nevanlinna Prize, the Gauss Prize, and the Chern Medal

## *Fields Medal*

### **Étienne Ghys**

The work of Artur Avila 47

### **Benedict H. Gross**

The work of Manjul Bhargava 56

### **Ofer Zeitouni**

The work of Martin Hairer 65

### **Curtis T. McMullen**

The work of Maryam Mirzakhani 73

## *Rolf Nevanlinna Prize*

### **Sanjeev Arora**

The work of Subhash Khot 81

## *Carl Friedrich Gauss Prize*

### **Ron Fedkiw, Jean-Michel Morel, Guillermo Sapiro, Chi-Wang Shu, and Wotao Yin**

The work of Stanley Osher 90

## *Chern Medal Award*

### **Mark L. Green**

The work of Phillip Griffiths 114

## **Artur Avila**

Ph.D. in Mathematics, Instituto Nacional de Matemática Pura e Aplicada (IMPA), 2001

### **Positions held**

Assistant, Collège de France, 2001–2003

Researcher, Centre National de la Recherche Scientifique (CNRS), 2003–2008

Fellow, Clay Mathematics Institute, 2006–2009

Research Director, CNRS, 2008–present

Researcher, IMPA and CNRS, 2009–present

# The work of Artur Avila

Étienne Ghys

**Abstract.** Artur Avila is awarded a Fields Medal for his profound contributions to dynamical systems theory, which have changed the face of the field, using the powerful idea of renormalization as a unifying principle.

## 1. Introduction

The citation for Avila's award states:

“Avila leads and shapes the field of dynamical systems. With his collaborators, he has made essential progress in many areas, including real and complex one-dimensional dynamics, spectral theory of the one-frequency Schrödinger operator, flat billiards and partially hyperbolic dynamics. Avila's work on real one-dimensional dynamics brought completion to the subject, with full understanding of the probabilistic point of view, accompanied by a complete renormalization theory. His work in complex dynamics led to a thorough understanding of the fractal geometry of Feigenbaum Julia sets. In the spectral theory of one-frequency difference Schrödinger operators, Avila came up with a global description of the phase transitions between discrete and absolutely continuous spectra, establishing surprising stratified analyticity of the Lyapunov exponent. In the theory of flat billiards, Avila proved several long-standing conjectures on the ergodic behavior of interval-exchange maps. He made deep advances in our understanding of the stable ergodicity of typical partially hyperbolic systems. Avila's collaborative approach is an inspiration for a new generation of mathematicians.”

Avila has published a huge number of papers, many of them solving long standing conjectures, with many collaborators. It is impossible to give an overview of his contribution in a small number of pages, even in rough outlines.

Fortunately, on the rather recent occasion of the Brin prize for Avila, two detailed papers were published, giving an excellent presentation of his work, at least in two of his main areas of research: one dimensional dynamics and the billiards dynamics [2, 3]. The interested reader is strongly encouraged to read these reviews.

I chose the option of following very closely the oral “laudation” that I presented during ICM Seoul. I had to select a very small number of results among many other possibilities.

It is intended for the general mathematician, certainly not for the expert, and its only purpose is to catch a glimpse of Avila's work.

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

## 2. The setting

If I had to give a summary of four centuries of research in dynamics, in a few sentences, I would write, following a joke by Yulij Ilyashenko, that there are three main stages in this history.

The *first stage* was initiated by Newton:

*“You are given an ordinary differential equation and your task is to find its solutions”.*

Differential calculus has been indeed remarkably successful.

The *second stage* was initiated by Poincaré at the turn of the twentieth century, when he realized that in most cases it is simply impossible to find a formula for solutions. This corresponds for instance to the birth of chaos theory.

*“You are given an ordinary differential equation and your task is to say something about its solutions.”*

If possible something useful, for instance something describing the qualitative behavior when time goes to infinity.

The *third stage* began when mathematicians realized that, in practice, physicists never know exactly the differential equation they want to solve. There are always unknown quantities, which may be small, but which do have some influence on the motion, some tiny friction for instance. One could say that this period began in the 1960's with Smale and Thom:

*“You are NOT given an ordinary differential equation and your task is to say something about its solutions.”*

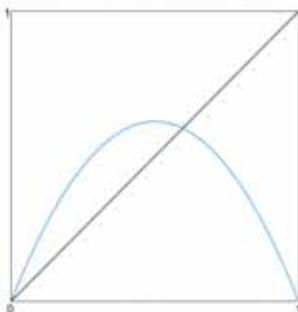
This is the field of research of Artur Avila. Most of his results turn around the question:

*“What does a typical dynamical system look like?”*

## 3. One dimensional dynamics and renormalization

Let us start with a basic example.

Consider a *unimodal map*  $f$  from an interval to itself, that is, a map having a single maximum. Assume that the second derivative is negative at the maximum.



Pick a point  $x$  in the interval, take its image by  $f$  and iterate the process. One gets the *orbit* of  $x$ , denoted  $\{f^n(x)\}$ . The main question is to describe the sequence  $f^n(x)$ . Where

does it go? Where does it accumulate? According to Smale-Thom’s message one should not try to answer this question for *every*  $f$ , but for a *typical*  $f$ .

Here is one of the very first great results of Artur, jointly with Misha Lyubich and Wellington de Melo, right after his PhD, improved a bit later in a joint work with Moreira.

*In a non-trivial real analytic family  $f_\lambda(\lambda \in \Lambda)$  of unimodal maps (where  $\Lambda$  is some finite dimensional parameter space), there is a dichotomy: for Lebesgue almost every  $\lambda$ , the map  $f_\lambda$  is either Regular or Stochastic.*

Of course, one should be more precise about the words used in this statement.

In the *regular* case, Lebesgue almost every orbit converges to some attracting cycle. After some time, the dynamics becomes essentially periodic: no chaos appears. This is the easy situation. The set of values of the parameter  $\lambda$  for which this regular case holds is typically an open and dense set in the parameter space  $\Lambda$  (but not of full Lebesgue measure).

The second case, *stochastic*, is *chaotic*. But chaos should not be understood as a negative word. It does not mean that one cannot describe the motion. There is some *absolutely continuous* measure on the interval such that for Lebesgue almost every initial condition  $x$  the sequence  $f^n(x)$  is *asymptotically distributed* according to this measure, unless it converges to a periodic cycle. So, this chaotic mode is still well understood since a single good measure describes the dynamics. The set of values of the parameter  $\lambda$  for which this happens has *positive* Lebesgue measure.

The theorem is that the union of regular and stochastic dynamics has *full Lebesgue measure* in the parameter space.

This result has a very long history and it is not possible to mention here all preliminary steps. The reader is referred to Misha Lyubich survey paper [3]. This “Regular or Stochastic dichotomy” was the first occasion confirming the general Palis conjecture on the behavior of almost all orbits for typical dissipative dynamical systems.

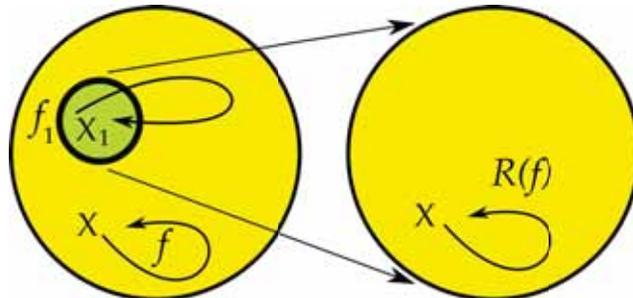
As Lyubich writes “we have reached a full probabilistic understanding of real analytic unimodal dynamics, and Artur Avila has been the key player in the final stage of the story”.

Of course, I cannot give any description of the proof of such a difficult theorem but I would like at least to explain one of the key tools. The so called *renormalization operator* has certainly not been invented by Artur but he knows better than anybody else how to use it! It quickly became his magic stick: he uses it in most of his papers. That was the topic of his plenary lecture in the previous Congress, in Hyderabad [1].

Start from a dynamical system, say a map  $f$  from a space  $X$  to itself.

Choose some small part  $X_1$  of  $X$  and assume that the orbit of every point in  $X_1$  comes back in  $X_1$ , maybe after many iterates.

Let us consider the map  $f_1$  from  $X_1$  to  $X_1$  which maps every point of  $X_1$  to its first return in  $X_1$  under the iterates of  $f$ .

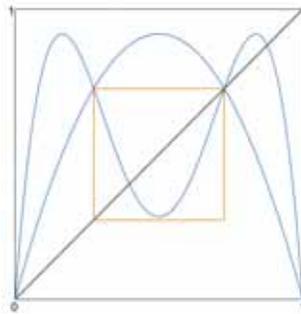


In many cases, the small  $X_1$  is somehow similar to the big  $X$  and there is a zooming out, from  $X_1$  to  $X$ , so that one can “renormalize”  $f_1$  as a map from  $X$  to  $X$ . Let us denote this new map by  $R(f)$ . Therefore, one can think of  $R$  as an operator sending a dynamical system  $f$  from  $X$  to  $X$  to some other dynamical system from  $X$  to  $X$ . This is called the *renormalization operator*. The magical fact is that there is a strong correspondence between the dynamics of  $R$ , acting on the space of maps  $f$ , and the dynamics of a typical element  $f$ .

As Adrien Douady used to say:

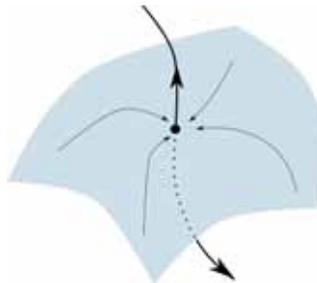
*“We first plough in the dynamical plane and then harvest in parameter plane”.*

Let us have a look at the first historical example. Consider a unimodal map  $f$  from the interval to itself. The graph of  $f$ , together with the graph of its square  $f^2$  (i.e.  $f \circ f$ ), may look like in the following picture.



In this case there is a subinterval invariant by  $f^2$ . Restricting  $f^2$  to this interval, zooming out and flipping, one gets back to the initial interval equipped with another unimodal map  $R(f)$ .

The general picture for this renormalization operator in this very special case has been a conjecture for many years. This figure illustrates the dynamics of the operator  $R$  on the infinite dimensional space of unimodal maps.



Coulet-Tresser and Feigenbaum, in the late 70's, had the intuition, based on numerical evidence, that there is a fixed point for the renormalization operator  $R$ , the so-called *Feigenbaum map*. Moreover the linearization at this fixed point has a one-dimensional expanding direction and is contracting on some hypersurface. It was a wonderful joint venture of many mathematicians to transform this intuition into a theorem. Among them, Lanford, Sullivan and McMullen. Avila and Lyubich could eventually achieve Sullivan's dream: instead of a computer assisted proof, they produced a “brain assisted proof” using some sophisticated technical preparation and then, just the standard Schwarz Lemma. A “proof from the book” as Erdős would have said.

#### 4. Billiard tables

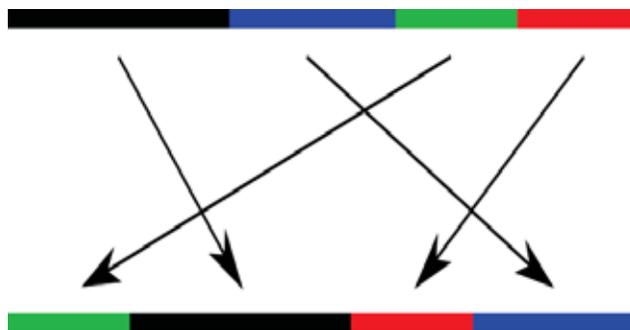
Imagine a box containing some perfect ideal gas, a huge number of bouncing molecules. For simplicity, let us make the assumption, not very realistic physically, that the gas is so dilute that the molecules don't collide between themselves.

Each individual molecule travels along straight lines in the box and bounces from time to time on the boundary. Molecules follow the orbits of a classical billiard ball game.

Let us make things simple and suppose that the box is actually 2 dimensional: a polygon in the plane. Choose a point  $x$  on the boundary of the polygon, which is a finite union of segments, and choose an initial velocity  $v$ , say of norm 1. Hit a ball there in that direction and wait until the ball bounces again on the boundary in  $x'$  and gets off in some other direction  $v'$ . This defines a dynamical system  $T$  which maps  $(x, v)$  to  $(x', v')$ . Let us make the even stronger assumption that the angles of this polygon are rational multiple of  $\pi$ . The rationality of the angles implies that the directions of the travelling ball can only take a finite number of values.

On a rectangle for instance, the velocity vector takes only four values. Therefore one can reduce the dynamics from dimension 2 to dimension 1. Now the configuration space will be a finite union of intervals. Each side of the polygon defines a finite number of intervals, one for each direction.

This kind of map is called an *interval exchange transformation*. Formally, the definition is the following. Take the unit interval  $[0, 1]$  and split it into  $k$  subintervals. Now reorganize the intervals according to some permutation. This defines a bijection from  $[0, 1]$  to itself. Don't worry about the endpoints. It is somehow like a generalized cards shuffling: you split your deck into several intervals and you permute them. Therefore, the dynamics of a rational polygonal billiard table is reduced to the dynamics of interval exchanges maps. Note that the space of interval exchange maps, with a given number  $k$  of subintervals, is parameterized by the product of the permutation group on  $k$  objects and a simplex, describing the lengths of the subintervals. In particular in this case, the space of dynamical systems under consideration is finite dimensional.



Now, let me state a theorem, due to Avila and Forni, again in the spirit of Smale-Thom. *Almost all interval exchange transformations are weakly mixing (except for trivial situations).*

I should explain the words and say at least something about the proof.

“Almost all” should be clear since the space of interval exchange maps is finite dimensional so that one has the Lebesgue measure at our disposal.

Let me define “mixing” first. Let  $f$  from  $X$  to  $X$  be a transformation preserving a probability measure  $\mu$ . One says that  $f$  is *mixing* if, for every pair  $A, B$  of measurable

subsets of  $X$  we have  $\lim_{n \rightarrow \infty} \mu(A \cap f^n(B)) = \mu(A)\mu(B)$ . This means that when times goes to infinity the dynamics somehow forgets the past: the events  $A$  and  $f^n(B)$  have a tendency to become independent. So a mixing dynamical system is a good approximation to randomness. Katok showed, however, that an interval exchange map is never mixing.

“Weak mixing” is, of course, a weakening of the concept of mixing. It simply means that  $\mu(A \cap f^n(B))$  converges to  $\mu(A)\mu(B)$  in a weaker sense: restricting  $n$  to some subset  $E$  of the integers, of density 1. Almost as good as mixing.

Avila-Forni’s theorem is a major progress in the understanding of the dynamics of billiards. The main tool to prove this theorem is again renormalization. The renormalization operator in this context acts on the space of interval exchange maps, which is a finite union of finite dimensional simplices. The important fact is that, even though each interval exchange is a rather simple dynamical system, this renormalization operator turns out to be very chaotic. This chaoticity in parameter space is the key to the understanding of a typical interval exchange map. For many more details, see the survey paper by Giovanni Forni [2].

## 5. Schrödinger operators

This is a topic in which the dynamical insight of Artur radically changed the landscape.

Imagine a 1-dimensional discrete quantum particle. Its state is described by some  $l^2$  function  $\psi$  on  $\mathbf{Z}$  with complex values. One can think that the probability that the particle is located at a point  $n$  is the square of the modulus of  $\psi(n)$ .

The time evolution of  $\psi$ , as usual, is described by the Schrödinger equation: the time derivative of  $\psi$  is  $iH\psi$  where  $H$  is the Schrödinger operator:

$$H(\psi)(n) = \psi(n+1) + \psi(n-1) + V(n)\psi(n).$$

The first two terms give a discrete version of the Laplace operator and  $V(n)$  is some bounded potential describing the environment of the particle.

Note that  $H$  is a bounded self adjoint operator on  $l^2$ . Everything depends on the *spectrum* of  $H$  and the *spectral measure*.

Let me recall that the *spectrum* is the set of energies  $E$  such that  $H - E.Id$  is not invertible. It is a compact set  $\sigma(H)$  in  $\mathbf{R}$ .

The *spectral measure* associated to some  $\psi$  is the measure  $\mu_\psi$  (supported on  $\sigma(H)$ ) such that for every continuous real valued function  $g$ , one has  $\langle \psi, g(H)\psi \rangle = \int g d\mu_\psi$ .

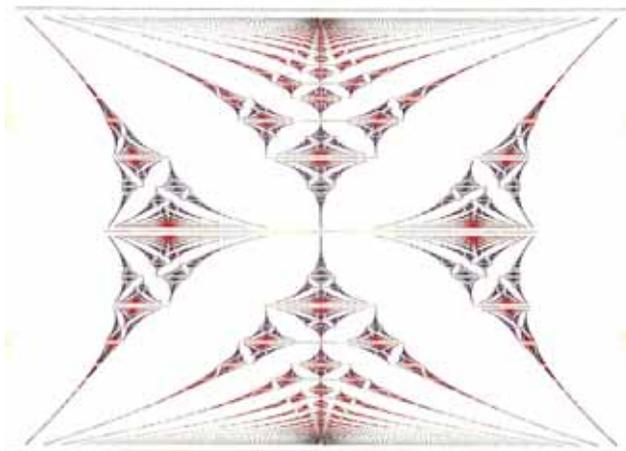
The spectral measures provide a key to the understanding of the dynamics of the quantum particle. To say things in a non precise way:

- The particle “travels freely” if  $\mu_\psi$  is absolutely continuous: the medium is conductor.
- The particle “travels a little bit” if  $\mu_\psi$  is singular continuous.
- The particle “does not travel” if  $\mu_\psi$  is pure point. The medium is insulator.

The most interesting case occurs when  $V$  is quasi-periodic. One can think for instance of a quasicrystal. The special case of  $V(n) = 2\lambda \cos(2\pi n\alpha)$  arises in this context as the simplest example. This is called the *almost Mathieu operator*.

Based on numerics, the shape of the spectrum was conjectured to be a Cantor set when  $\alpha$  is irrational. In 1981, Mark Kac offered ten Martinis for a proof of this fact. Barry Simon coined the term *Ten Martini Problem*.

The following picture is the famous *Hofstadter butterfly*.



Slicing this butterfly by a vertical line with first coordinate  $\alpha$ , one gets the spectrum for the critical case  $\lambda = 1$ . Many papers were devoted to the ten Martini problem and other conjectures in the 1980's and 1990's. It is probably fair to say that spectral theorists had exhausted their toolboxes. New ideas and approaches were needed. Artur introduced new dynamical methods in the problem and could solve the most difficult conjectures. Here is a sample of some results:

**Theorem** (Avila-Jitomirskaya 2009). *For all  $\lambda \neq 0$ , and all irrational  $\alpha$ , the spectrum  $\sigma_{\lambda,\alpha}$  is a Cantor set.*<sup>1</sup>

**Theorem** (Avila-Krikorian 2006).  $Leb(\sigma_{\lambda,\alpha}) = 4|1 - |\lambda||$ .

This was already known by Jitomirskaya and Krasovsky in the non critical case, when  $\lambda$  is not equal to 1.

**Theorem** (Avila, Damanik, 2008). *For all irrational  $\alpha$  and  $|\lambda| < 1$ , the spectrum is purely absolutely continuous.*

The key tool in the proofs of these difficult theorem is again *renormalization*.

Let me also mention, without giving any explanation that Artur created recently a global theory of one frequency Schrödinger operators, describing in detail what he calls the *stratified analyticity* of the Lyapounov exponent and the boundary of non uniform hyperbolicity.

Artur started his career by solving a number of long standing problems and conjectures but he is also an exceptional theory builder. The whole theory was developed by Artur and this required outstanding insight and exceptional technical abilities.

## 6. A gem

Let me finish by mentioning a puzzling theorem of Artur, which is somehow isolated in his work. This is not directly related to dynamics: this is a pure partial differential equations

<sup>1</sup>They could not get the ten Martinis since meanwhile Mark Kac had unfortunately passed away.

result. It is easy to state and Artur told me that almost every mathematician listening to this theorem for the first time is immediately convinced that this is very easy and that he can provide a short simple proof. But, this is not so...

Let  $f$  be a diffeomorphism of class  $C^1$  of some compact manifold of class  $C^\infty$ . It is well known, and easy to prove, that you can approximate  $f$  by  $C^\infty$  diffeomorphisms in the  $C^1$  topology.

Artur's theorem is that *if the manifold is equipped with a  $C^\infty$  volume form and if  $f$  preserves the volume, it can be approximated in the  $C^1$  topology by  $C^\infty$  diffeomorphisms which are volume preserving.*

Artur's proof starts with a triangulation and does the approximation by induction on the skeleton. It reminds me of the wonderful proofs by Gromov of his  $h$ -principles in PDE.

Avila's contributions are amazing: I convinced that this is just a beginning.

## References

- [1] Avila, Artur, *Dynamics of renormalization operators*, Proceedings of the International Congress of Mathematicians, Volume I, 154–175, Hindustan Book Agency, New Delhi, 2010.
- [2] Forni, Giovanni, *On the Brin Prize work of Artur Avila in Teichmüller dynamics and interval-exchange transformations*, J. Mod. Dyn. **6** (2012), no. 2, 139–182.
- [3] Lyubich, Mikhail, *Forty years of unimodal dynamics: on the occasion of Artur Avila winning the Brin Prize*, J. Mod. Dyn. **6** (2012), no. 2, 183–203.

UMPA, CNRS ENS Lyon, 46 Allée d'Italie, 69340 Lyon, France

E-mail: etienne.ghys@ens-lyon.fr

## **Manjul Bhargava**

A.B. in Mathematics, Harvard University, 1996

M.A. in Mathematics, Princeton University, 1998

Ph.D. in Mathematics, Princeton University, 2001

### **Positions held**

Clay Mathematics Institute Long-Term Prize Fellow, 2000–2005

Member, Institute for Advanced Study, Princeton, 2001–2002

Visiting Fellow, Harvard University, 2002–2003

Professor of Mathematics, Princeton University, 2003–present

# The work of Manjul Bhargava

Benedict H. Gross

**Abstract.** He has developed powerful new methods in the geometry of numbers and applied them to count rings of small rank and to bound the average rank of elliptic curves.

The geometry of numbers was introduced by C.F. Gauss in the *Disquisitiones Arithmeticae* [28, Art. 302], to estimate the class numbers of binary quadratic forms. It was developed by G.L. Dirichlet and R. Dedekind, in the computation of the residue of the zeta function of a number field at  $s = 1$ , and treated systematically by H. Minkowski in his book *Geometrie der Zahlen* [34]. Many leading mathematicians, such as C.L. Siegel and H. Davenport, have contributed to its modern development. Manjul Bhargava has continued in this great tradition and taken it to new heights, adding several new ideas and techniques of his own. He has applied them brilliantly to the study of orbits in integral representations, obtaining a number of ground-breaking results on the distribution of number fields and their class groups, and on rational points on elliptic and hyperelliptic curves. His work has changed the way that we now approach the subject.

The geometry of numbers typically gives results on average, not for any chosen example. For example, Gauss defined the class number  $H(D)$  as the number of orbits of the group  $\mathrm{SL}_2(\mathbb{Z})$  on the set of integral, positive-definite binary quadratic forms  $ax^2 + bxy + cy^2$  with negative discriminant  $b^2 - 4ac = -D$ . Based on extensive calculations, he was able to guess that  $H(D)$  grows at essentially the same rate as  $\sqrt{D}$ . He was not able to prove this; even today we lack an effective proof. But Gauss gave evidence that the result holds on average: as  $T \rightarrow \infty$ ,

$$\sum_{D < T} H(D) \sim \frac{\pi}{18} T^{3/2}.$$

The proof (given in Mertens [33]) has three main steps. First, every  $\mathrm{SL}_2(\mathbb{Z})$ -orbit contains a unique form whose coefficients satisfy

$$-a < b < a < c \quad \text{or} \quad 0 \leq b \leq a \leq c$$

(see [28, Art. 171-172]). Hence the sum of class numbers for  $D < T$  is the number of integral points in the corresponding region

$$A(T) = \{(x, y, z) : -x \leq y \leq x \leq z, 4xz - y^2 \leq T\}$$

of three-space (up to a negligible error as  $T$  grows). The next step is to show that volume of the region  $A(T)$  is finite, and equal to  $\frac{\pi}{18} T^{3/2}$ . A basic idea in the geometry of numbers

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

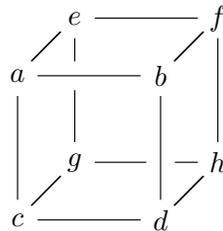
is to estimate the number of integral points in a region by its volume. If the region  $A(T)$  were bounded, simple geometric arguments would then yield the asymptotic estimate on the number of integral points. However the region has a cusp, and the final step is to determine the number of integral points which lie in the cusp, and to show that they do not change the asymptotic estimate given by the volume.

Nowadays we would add a fourth step, which is to provide an arithmetic interpretation of the orbits. In this case, the orbits of primitive forms of discriminant  $-D$  (those with  $\gcd(a, b, c) = 1$ ) correspond bijectively to the ideal classes (the quotient of the group of invertible fractional ideals by the subgroup of principal ideals) of the quadratic ring  $R = \mathbb{Z}[\frac{D+\sqrt{-D}}{2}]$ . In this bijection, Gauss' famous composition law for classes of binary quadratic forms corresponds to the group structure on the quotient. When  $-D$  is fundamental (not of the form  $-df^2$  for another discriminant  $-d$ ), every form is primitive and the orbits correspond bijectively to the elements of the ideal class group of the imaginary quadratic field  $\mathbb{Q}(\sqrt{-D})$ . Using a sieve, one can obtain an estimate for the corresponding sum of class numbers, for negative fundamental discriminants of absolute value less than  $T$ .

It is often the arithmetic interpretation of the integral orbits that is most striking in Bhargava's work. He has developed a deep understanding of the representations of arithmetic groups, which allows him to reformulate many central counting problems in number theory in terms of estimates for the number of integral orbits of bounded height. In his PhD thesis [1], Bhargava studied the representation of the group

$$G = \mathrm{SL}_2(\mathbb{Z})^3 \quad \text{on} \quad M = \mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2,$$

The symmetric square of  $M$  contains the adjoint representation of  $G$ , and Bhargava presents these three binary quadratic forms in a simple manner. Namely, he associates to each element  $m$  in  $M$  a labeling of the cube



If we slice the cube into pairs of  $2 \times 2$  matrices in three different ways

$$\begin{aligned} M_1 &= \begin{bmatrix} a & b \\ c & d \end{bmatrix}, & N_1 &= \begin{bmatrix} e & f \\ g & h \end{bmatrix} \\ M_2 &= \begin{bmatrix} a & c \\ e & g \end{bmatrix}, & N_2 &= \begin{bmatrix} b & d \\ f & h \end{bmatrix} \\ M_3 &= \begin{bmatrix} a & e \\ b & f \end{bmatrix}, & N_3 &= \begin{bmatrix} c & g \\ d & h \end{bmatrix} \end{aligned}$$

we obtain three binary quadratic forms

$$\begin{aligned} Q_1(x, y) &= -\det(M_1x - N_1y) \\ Q_2(x, y) &= -\det(M_2x - N_2y) \\ Q_3(x, y) &= -\det(M_3x - N_3y). \end{aligned}$$

The three binary quadratic forms  $Q_1, Q_2,$  and  $Q_3$  all have the same discriminant  $d$ ; this gives a quartic polynomial which generates the ring of polynomial invariants on  $M$ . The  $G$ -orbit of  $m$  is determined by the  $\mathrm{SL}_2(\mathbb{Z})$ -orbits of these binary quadratic forms, or equivalently by a triple of ideal classes  $I_1, I_2$  and  $I_3$  for the quadratic ring of discriminant  $d$  (which can now be positive, negative, or zero). Furthermore, Bhargava shows that the product of these three classes is trivial, and that every triple with trivial product arises from such an orbit. This gives a beautiful, symmetric reformulation of Gauss's law for the composition of binary quadratic forms—the composition of three binary forms is in the trivial class (given by a form which represents the integer 1) if and only if the three forms arise from a cube!

Bhargava has found arithmetic interpretations of the orbits in 14 integral representations which generalize the representation corresponding to Gauss composition [1–5]. For example, the symmetrization of  $M$  gives the action of the diagonal  $\mathrm{SL}_2(\mathbb{Z})$  on the space of binary cubic forms of the type  $ax^3 + 3bx^2y + 3cxy^2 + dy^3$ . In this case he shows that the orbits correspond to quadratic rings with an ideal class  $I$  whose cube is the trivial class. These representations (or their duals) all come from parabolic subgroups in Chevalley groups—the group which acts is a subgroup of the Levi factor and the representation occurs on the abelianization of the unipotent radical (cf. [27]). The representation  $\mathbb{Z}^2 \otimes \mathbb{Z}^2 \otimes \mathbb{Z}^2$  of  $\mathrm{SL}_2(\mathbb{Z})^3$  giving Gauss's composition law comes from a maximal parabolic subgroup in the Chevalley group of type  $D_4$ , and three of the representations on his list come from distinguished maximal parabolic subgroups in the Chevalley groups of type  $G_2, F_4,$  and  $E_8$ . These three representations are

- the action of  $\mathrm{GL}_2(\mathbb{Z})$  on  $\mathrm{Sym}^3(\mathbb{Z}^2) \otimes \det^{-1}$
- the action of  $\mathrm{SL}_3(\mathbb{Z}) \times \mathrm{GL}_2(\mathbb{Z})$  on  $\mathrm{Sym}^2(\mathbb{Z}^3) \otimes \mathbb{Z}^2$
- the action of  $\mathrm{SL}_5(\mathbb{Z}) \times \mathrm{GL}_4(\mathbb{Z})$  on  $\wedge^2(\mathbb{Z}^5) \otimes \mathbb{Z}^4$

They were studied over  $\mathbb{C}$  by Sato and Kimura [37], and over general fields by Wright and Yukie [39], as part of a program to generalize the results of Davenport and Heilbronn [23] counting cubic fields. In each case, there is a polynomial invariant  $D$ , which has degree 4, 12, and 40 respectively and generates the full ring of invariants. Over the field of complex numbers these representations have an open orbit where  $D \neq 0$  with stabilizers isomorphic to the symmetric groups  $S_3, S_4,$  and  $S_5$  respectively [29].

Bhargava shows that the integral orbits where  $D$  is non-zero correspond to cubic, quartic, and quintic rings respectively (that is, to commutative rings which are free abelian groups of rank 3, 4, and 5 over  $\mathbb{Z}$ ) and that  $D$  is equal to the discriminant of the ring. This was known for cubic rings, by work of Delone and Fadeev [24], but the quartic and quintic cases were much more difficult and involved the introduction of auxiliary resolvents. Bhargava then applies the geometry of numbers to count the number of rings of discriminant less than  $T$  as  $T \rightarrow \infty$ . Using a sieve, he converts this count to an asymptotic estimate for the number of number fields of degree 3, 4 and 5 over  $\mathbb{Q}$  with discriminant less than  $T$ . For each degree  $n \leq 5$  and signature  $(r_1, r_2)$  with  $r_1 + 2r_2 = n$ , Bhargava shows that this number grows like  $c(r_1, r_2)T$ , where  $c(r_1, r_2)$  is an explicit constant [6, 7]. As mentioned above, the case of cubic fields had been treated in the work of Davenport and Heilbronn [22, 23], but the quartic and quintic cases were much more complicated and the conjecture that the growth was proportional to  $T$  had been open for thirty years. This work led Bhargava to a guess for the constant  $c(r_1, r_2)$  for all degrees  $n$  [8]. Applying similar counting methods to some of the other representations led Bhargava to the resolution of some new cases of the Cohen–Lenstra–Martinet heuristics [19] for the ideal class groups of number fields [6].

All of the integral representations which generalize Gauss composition have the property that there is a single polynomial invariant, whose non-vanishing defines an open orbit over the complex numbers. Starting with his work with Arul Shankar on the Selmer groups of elliptic curves, Bhargava turned to the study of integral orbits in representations with a more complicated ring of invariants. For example, the action of  $\mathrm{PGL}_2(\mathbb{Z})$  on the space  $\mathrm{Sym}^4(\mathbb{Z}^2) \otimes \det^{-2}$  of binary quartic forms has a ring of invariants generated by a quadratic polynomial  $I$ , a cubic polynomial  $J$ , and a sextic polynomial  $\Delta$  with the single relation  $27\Delta = 4I^3 - J^2$ . In terms of the binary quartic form

$$F(x, y) = ax^4 + bx^3y + cx^2y^2 + dxy^3 + ey^4$$

these invariants are given by

$$I = 12ae - 3bd + c^2$$

$$J = 72ace + 9bcd - 27ad^2 - 27eb^2 - 2c^3.$$

The rational orbits where  $\Delta \neq 0$  are stable in the sense of geometric invariant theory—they are closed with finite stabilizers. The stable orbits correspond to certain curves  $X$  of genus one, defined by the equation

$$z^2 = F(x, y).$$

Furthermore, the Jacobian of  $X$  is isomorphic to the elliptic curve  $E$  whose equation is determined by the polynomial invariants  $I$  and  $J$

$$v^2 = u^3 - (I/3)u - (J/27)$$

and whose discriminant is equal to  $\Delta$ . Hence  $X$  is a principal homogeneous space for the algebraic group  $E$ . The covariants in this representation give an unramified covering  $\pi : X \rightarrow E$  of degree 4, such that  $\pi(x + e) = \pi(x) + 2e$  [40]. The fiber above the origin of  $E$  consists of the four points on the curve  $X$  where  $z = 0$ , and this gives a principal homogeneous space  $X[2]$  for the 2-torsion subgroup  $E[2]$ . The class of this homogeneous space in  $H^1(\mathbb{Q}, E[2])$  determines the rational orbit (whose stabilizer is the finite group scheme  $E[2]$ ). Moreover, in the map of principal homogeneous spaces  $H^1(\mathbb{Q}, E[2]) \rightarrow H^1(\mathbb{Q}, E)$ , the class of  $X[2]$  maps to the class of the curve  $X$ .

When the curve  $X$  has points over all completions of  $\mathbb{Q}$ , the corresponding coverings are called locally solvable. The locally solvable orbits with invariants  $(I, J)$  correspond bijectively to the classes in the 2-Selmer group  $\mathrm{Sel}_2(E)$  of the elliptic curve  $E$ . The coverings where the curve  $X$  has a rational point are called solvable and correspond to the classes in the subgroup  $E(\mathbb{Q})/2E(\mathbb{Q})$  of  $\mathrm{Sel}_2(E)$ . Birch and Swinnerton-Dyer [18] showed that the locally solvable rational orbits all have integral representatives (at least away from the prime  $p = 2$ ).

Bhargava and Shankar [9] estimate the number of integral orbits with bounded invariants  $I$  and  $J$  using the geometry of numbers. They also develop sieve methods which allow them to convert this estimate into a count of 2-Selmer elements. Their main theorem is that the average size of the 2-Selmer group of elliptic curves over  $\mathbb{Q}$  is equal to 3, in the following sense. Every elliptic curve  $E$  over  $\mathbb{Q}$  has a unique model of the form

$$y^2 = x^3 + Ax + B$$

where  $A$  and  $B$  are integers with  $4A^3 + 27B^2 \neq 0$ , which are not simultaneously divisible by  $p^4$  and  $p^6$  respectively, for any rational prime  $p$ . Define the height of  $E$  by the formula

$H(E) = \text{Max}(4|A^3|, 27B^2)$ . Then the number of elliptic curves with height  $H(E) < T$  is finite and grows at the same rate as a constant times  $T^{5/6}$ . Bhargava and Shankar prove that the limit, as  $T \rightarrow \infty$ , of the ratio

$$\sum_{H(E) < T} \# \text{Sel}_2(E) / \sum_{H(E) < T} 1$$

exists, and is equal to 3. The delicate analysis and geometry involved in their calculation is astounding. For example, the theorem that the average order is 3 arises from a volume calculation, essentially equivalent to the fact that the Tamagawa number of the group  $\text{PGL}_2$  is equal to 2. This gives the average number of non-identity elements in the Selmer group, as the identity class is the only one that appears significantly in the cusp of the fundamental domain! For a beautiful summary of this remarkable paper, see the Bourbaki talk of B. Poonen [35].

Their calculation of the average value of the order of the 2-Selmer group  $\text{Sel}_2(E)$  gives an upper bound for the average order of its subgroup  $E(\mathbb{Q})/2E(\mathbb{Q})$ , and hence on the rank of the group  $E(\mathbb{Q})$  of rational points. In this case the upper bound they obtain on the average rank is  $3/2$ . Before their work, it was not even known that the average rank was finite! But Bhargava and Shankar push much further. They also study integral orbits in the analogous representations (with two generating invariants over  $\mathbb{Q}$ ):

- the action of  $\text{PGL}_3(\mathbb{Z})$  on  $\text{Sym}^3(\mathbb{Z}^3) \otimes \det^{-1}$
- the action of  $\text{SL}_2(\mathbb{Z}) \times \text{SL}_4(\mathbb{Z})$  on  $\mathbb{Z}^2 \otimes \text{Sym}^2(\mathbb{Z}^4)$
- the action of  $\text{SL}_5(\mathbb{Z}) \times \text{SL}_5(\mathbb{Z})$  on  $\mathbb{Z}^5 \otimes \wedge^2(\mathbb{Z}^5)$ .

The stable locally soluble rational orbits in these representations correspond to elements in the 3-, 4-, and 5-Selmer groups of elliptic curves over  $\mathbb{Q}$  [21, 26]. (These representations also arise in Vinberg's theory, using the exceptional groups of type  $F_4$ ,  $E_7$ , and  $E_8$ . The action on binary quartic forms comes from the exceptional group of type  $G_2$  [29]). Bhargava and Shankar show that the average order of the 3-, 4-, and 5-Selmer groups is equal to 4, 7, and 6 respectively [10–12]. This led to the conjecture that the average order of the  $m$ -Selmer group of elliptic curves over  $\mathbb{Q}$  is equal the sum of the divisors of  $m$ . In the case when  $m = 3$ , A.J. de Jong had obtained slightly weaker results, with the field  $\mathbb{F}_q(T)$  in place of  $\mathbb{Q}$  [32].

As a corollary of their calculation of the average order of these Selmer groups, Bhargava and Shankar are able to prove that the average rank is less than 1 (we suspect it is equal to  $1/2$ ) and that at least 80% of elliptic curves over  $\mathbb{Q}$  have rank less than or equal to 1 (we suspect that 50% have rank 0 and that 50% have rank 1). Analogous results for other families of elliptic curves (for example, curves with a marked rational point other than the origin) have been obtained by Bhargava and W. Ho [17]. For the family of curves related to the congruent number problem, there were earlier results of D.R. Heath-Brown [31].

The conjecture of Birch and Swinnerton-Dyer, that the rank of  $E(\mathbb{Q})$  is equal to the order of vanishing of the  $L$ -function of  $E$  at the point  $s = 1$ , is known to be true when the order of vanishing is less than or equal to 1 [30]. Starting from this point, Bhargava, C. Skinner, and W. Zhang have recently shown, in a brilliant pair of papers [15, 16], that the conjecture of Birch and Swinnerton-Dyer holds for at least 66% of all elliptic curves over  $\mathbb{Q}$  (all of which have rank 0 or 1). This uses the results of Bhargava and Shankar on the average order of the 5-Selmer group, as well as deep work of his co-authors on the Iwasawa conjecture [38] and Kolyvagin's conjecture [41].

Bhargava's methods also extend to the study of hyperelliptic curves of genus  $g \geq 2$  over  $\mathbb{Q}$ . Such a curve has a homogeneous equation of the form

$$z^2 = F(x, y) = f_0x^{2g+2} + f_1x^{2g+1}y + \dots + f_{2g+2}y^{2g+2}$$

where  $x$  and  $y$  have degree 1,  $F(x, y)$  is a binary form of degree  $2g + 2$  with integral coefficients and non-zero discriminant, and  $z$  has degree  $g + 1$ . G. Faltings, in work which was awarded the Fields Medal in 1986, proved that there are only finitely many relatively prime integral solutions, or equivalently, rational points on the projective curve [25]. One of Bhargava's most remarkable results is that for a majority of equations of a fixed genus  $g \geq 2$ , ordered by the size of the coefficients, there are no rational points at all! This is true even if we only consider curves which have points over every completion of  $\mathbb{Q}$ , and the proportion of curves with no rational points tends to 1 very quickly as  $g$  becomes large. For example, when  $g \geq 10$  at least 99% of hyperelliptic equations of genus  $g$  have no rational solutions [13].

Bhargava has also obtained precise results on the family of hyperelliptic curves of genus  $g \geq 2$  over  $\mathbb{Q}$  with a rational Weierstrass point: the average order of the 2-Selmer group of the Jacobian is equal to 3 [14]. These curves can all be given by equations with  $f_0 = 0$ ,  $f_1 = 1$ , and  $f_2 = 0$ , so by affine equations of the form

$$z^2 = x^{2g+1} + c_2x^{2g-1} + \dots + c_{2g+1}.$$

Each curve has a unique equation of this form, provided that the coefficients  $c_k$  are all integers and there is no rational prime  $p$  with  $p^{2k}$  dividing  $c_k$  for all  $k$ . The estimate on the 2-Selmer group is obtained through a study of integral and rational orbits in the symmetric square representation of the split special orthogonal group  $\mathrm{SO}_{2g+1}$ . (This generalizes the action of  $\mathrm{SO}_3 = \mathrm{PGL}_2$  on binary quartic forms.) Since the bound  $3/2$  obtained on the average rank of the Jacobian is less than the genus, the  $p$ -adic methods introduced by Chabauty and developed by Coleman [20] can be used to effectively bound the number of rational points on most of these curves [14]. B. Poonen and M. Stoll [36] have refined this method at the prime  $p = 2$  to show that for  $g \geq 3$  most of these curves have only the one obvious rational point at infinity, and that the proportion with only one rational point tends rapidly to 1 as  $g$  becomes large. In particular, a monic polynomial of odd degree  $\geq 7$  will rarely represent a square!

Manjul Bhargava's ideas are remarkably original, yet once discovered, form a natural continuation of previous great work in the subject. His papers are written with great care, in a distinctive style, and his lectures convey the unity and the beauty of mathematics. Andrew Wiles (who was Bhargava's PhD thesis advisor) wrote the following about his impact: "The sense of a new field opening up before one's eyes with an elegance and clarity that remind one of one's first encounters with classical mathematics is unique and sometimes breathtaking."

## References

- [1] M. Bhargava, *Higher composition laws I: A new view on Gauss composition, and quadratic generalizations*, Ann. of Math. **159** (2004), no. 1, 217–250.
- [2] ———, *Higher composition laws II: On cubic analogues of Gauss composition*, Ann. of Math. **159** (2004), no. 2, 865–886.

- [3] ———, *Higher composition laws III: The parametrization of quartic rings*, Ann. of Math. **167** (2004), no. 2, 1329–1360.
- [4] ———, *Higher composition laws IV: The parametrization of quintic rings*, Ann. of Math. **159** (2008), no. 2, 53–94.
- [5] ———, *Higher composition laws and applications*, International Congress of Mathematicians. Vol II 271–295, Eur. Math. Soc., Zürich, 2006.
- [6] ———, *The density of discriminants of quartic rings and fields*, Ann. of Math. **162** (2005), no. 2, 1031–1063.
- [7] ———, *The density of discriminants of quintic rings and fields*, Ann. of Math. **172** (2010), no. 3, 1559–1591.
- [8] ———, *Mass formulae for extensions of local fields, and conjectures on the density of number field discriminants*, Int. Math. Res. Not. IMRN **17** (2007).
- [9] M. Bhargava and A. Shankar, *Binary quartic forms having bounded invariants, and the boundedness of the average rank of elliptic curves*, Ann. of Math. (to appear) arXiv:1006.1002.
- [10] ———, *Ternary cubic forms having bounded invariants, and the existence of a positive proportion of elliptic curves having rank 0*, Ann. of Math. (to appear) arXiv:1007.0052.
- [11] ———, *The average number of elements in the 4-Selmer groups of elliptic curves is 7*, arXiv:1312.7333.
- [12] ———, *The average number of elements in the 5-Selmer groups of elliptic curves is 6, and the average rank is less than 1*, arXiv:1312.7859.
- [13] M. Bhargava, *Most hyperelliptic curves over  $\mathbb{Q}$  have no rational points*, arXiv:1308.0395.
- [14] M. Bhargava and B. H. Gross, *The average size of the 2-Selmer group of the Jacobians of hyperelliptic curves having a rational Weierstrass point*, Automorphic representations and L-functions 23–91, Tata Inst. Fundam. Res. Stud. Math. **22**, Mumbai, 2013.
- [15] M. Bhargava and C. Skinner, *A positive proportion of elliptic curves over  $\mathbb{Q}$  have rank one*, arXiv:1401.0233.
- [16] M. Bhargava, C. Skinner, and W. Zhang, *A majority of elliptic curves over  $\mathbb{Q}$  satisfy the Birch and Swinnerton-Dyer conjecture*, arXiv:1407.1826.
- [17] M. Bhargava and W. Ho, *Coregular spaces and genus one curves*, arXiv:1306.4424.
- [18] B. J. Birch and H. P. F. Swinnerton-Dyer, *Notes on elliptic curves I*, J. Crelle **212** (1963), 7–25.
- [19] H. Cohen and J. Martinet, *Étude heuristique des groupes de classes des corps de nombres*, J. Crelle **404** (1990), 39–76.
- [20] R. Coleman, *Effective Chabauty*, Duke Math. J. **52** (1985), 765–770.
- [21] J. Cremona, T. Fisher, and M. Stoll, *Minimization of 2-, 3-, and 4- coverings of elliptic curves*, arXiv:1908.1741.
- [22] H. Davenport, *On the class-numbers of binary cubic forms I, II*, J. London Math. Soc. **26** (1951), 183–198.
- [23] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields II*, Proc. Roy. Soc. London Ser. A **322** (1971), 405–420.

- [24] B. N. Delone and D. K. Faddeev, *The theory of irrationalities of the third degree*. AMS Translations of Math, **10** (1964).
- [25] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983) (3): 349–366.
- [26] T. Fisher, *Invariant theory for the elliptic normal quintic I,II*, arXiv:1110.3520, arXiv:1303.2550.
- [27] W-T. Gan, B. H. Gross, and G. Savin, *Fourier coefficients of modular forms on  $G_2$* , Duke Math J. **115** (2002), 105–169.
- [28] C. F. Gauss, *Disquisitiones Arithmeticae*, Translated into English by A. Clark (1966), Yale University Press, New Haven.
- [29] B. H. Gross, *On Bhargava’s representations and Vinberg’s invariant theory*, Frontiers of mathematical sciences, International Press (2011), 317–321.
- [30] ———, *Lectures on the conjecture of Birch and Swinnerton-Dyer*, Arithmetic of  $L$ -functions, AMS PCMI Publications (2011), 169–210.
- [31] D. R. Heath-Brown, *The size of Selmer groups for the congruent number problem*, Invent. Math. **111** (1993), 171–195.
- [32] A. J. de Jong, *Counting elliptic curves over finite fields*, Moscow Math. J. **2** (2002), 281–311.
- [33] H. Mertens, *Über einiger asymptotische Gesetze der Zahlentheorie*, J. Crelle **77** (1874), 289–338.
- [34] H. Minkowski, *Geometrie der Zahlen* (1910), Teubner, Berlin.
- [35] B. Poonen, *Average rank of elliptic curves (after Manjul Bhargava and Arul Shankar)*, Sémin. Bourbaki, arXiv:1203.0809.
- [36] B. Poonen and M. Stoll, *Most odd degree hyperelliptic curves have only one rational point*, Annals of Math. (to appear) arXiv:1302.0061.
- [37] M. Sato and T. Kimura, *A classification of irreducible prehomogeneous vector spaces and their relative invariants*, Nagoya Math. J. **65** (1977), 1–155.
- [38] C. Skinner and E. Urban, *The Iwasawa main conjecture for  $GL_2$* , Invent. Math. **195** (2014), 1–277.
- [39] D. J. Wright and A. Yukié, *Prehomogenous vector spaces and field extensions*, Invent. Math. **110** (1992), 283–314.
- [40] A. Weil, *Remarques sur une mémoire d’Hermite*, Arch. Math. **5** (1954), 197–202.
- [41] W. Zhang, *Selmer groups and the indivisibility of Heegner points*, Preprint (2014).

Department of Mathematics, Harvard University, One Oxford Street, Cambridge, MA 02138, USA  
E-mail: gross@math.harvard.edu

## **Martin Hairer**

B.Sc. in Mathematics, University of Geneva, 1998

M.Sc. in Physics, University of Geneva, 1998

Ph.D. in Physics, University of Geneva, 2001

### **Positions held**

Postdoctoral Fellow, Swiss NSF, 2002–2003

Advanced Fellow, Swiss NSF, 2003–2004

Assistant Professor, University of Warwick, 2004–2006

Associate Professor, University of Warwick, 2006–2009

Associate Professor, Courant Institute, New York University, 2009

Professor, University of Warwick, 2009–present

# The work of Martin Hairer

Ofer Zeitouni

**Abstract.** Martin Hairer has been awarded the Fields medal for his groundbreaking work in the theory of Stochastic Partial Differential Equations (SPDEs). A short account of his main contributions is presented below.

**Mathematics Subject Classification (2010).** Primary 60H15; Secondary 35R60.

**Keywords.** Martin Hairer. Stochastic PDEs. Regularity structures.

## 1. Introduction

Martin Hairer has established several ground breaking results in the theory of stochastic partial differential equations (SPDEs). We focus this short account on his most striking contributions.

## 2. Background

§1 Recall that a Wiener process  $\{W_t\}_{t \geq 0}$  (also called Brownian motion) is a centered Gaussian process, built over a probability space  $(\Omega, \mathcal{F}, P)$ , with covariance  $EW_t W_s = \min(s, t)$ . Its formal time derivative, the *white noise*  $\dot{W}_t$ , is defined by its linear action on  $L^2$  test functions: for deterministic  $f \in L^2(\mathbb{R}_+)$ ,  $\int_0^t f(s) \dot{W}_s =: \int_0^t f(s) dW_s$  is the centered Gaussian process with covariance  $R(t, t') = \int_0^{\min(t, t')} f^2(s) ds$ . In case  $f \in \{f : f(t) = \int_0^t h(s) ds, h \in L^2(\mathbb{R}_+)\}$ , this notion coincides with formal integration by parts:

$$\int_0^t f(s) dW_s = W_t f(t) - \int_0^t W_s f'(s) ds.$$

In addition, because  $t \mapsto W_t$  is almost surely a continuous process, one can give a sense to the equation

$$X_t = X_0 + \int_0^t g(X_s) ds + W_t$$

as soon as  $g$  is a Lipschitz function, in such a way that the map  $W. \rightarrow X.$  is almost surely a continuous map  $C_0(\mathbb{R}_+) \rightarrow C(\mathbb{R}_+)$ .

Itô observed that the integral  $\int_0^t f(s) dW_s$  can be defined for certain *random* functions: let  $\mathcal{F}_t$  be the  $\sigma$ -algebra generated by the random variables  $\{W_s, s \leq t\}$ . A random function

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

$f : \mathbb{R}_+ \times \Omega \rightarrow \mathbb{R}$  is called *adapted* (to  $\mathcal{F}_t$ ) if  $f_t := f(t, \omega)$  is  $\mathcal{F}_t$ -measurable for all  $t$ . Let  $\mathbf{L}_2$  denote the class of adapted random functions equipped with the norm  $\|h\|_2 = (E \int_0^\infty h^2(s) ds)^{\frac{1}{2}}$ . Then  $\int_0^\infty f_s dW_s$  extends to an isometry from  $\mathbf{L}_2$  to  $L^2(\Omega, P)$ . Using this notion (the *Itô integral*), Itô [11] gave precise sense to solutions of the equation

$$X_t = X_0 + \int_0^t g_1(X_s) ds + \int_0^t g_2(X_s) dW_s \quad (2.1)$$

when both  $g_1$  and  $g_2$  are smooth Lipschitz functions; in the last display,  $W$  can be a multi-dimensional Wiener process (with each coordinate an independent Wiener process), with  $X$  also multidimensional.

§2 The Itô theory has had enormous success in both pure mathematics (through links with the theory of second order parabolic PDEs) and applied mathematics. However, the map  $W \mapsto X$  does not behave well under approximations: Wong and Zakai [16] (see also [14] for the multidimensional case) observed that if one approximates the Wiener process by piecewise linear interpolation over intervals of size  $\epsilon$ , calling the approximation  $W^\epsilon$ , then the solution of the ordinary differential equation

$$\frac{d}{dt} X_t^\epsilon = g_1(X_t^\epsilon) + g_2(X_t^\epsilon) \cdot \frac{d}{dt} W_t^\epsilon$$

does not converge (as  $\epsilon \rightarrow 0$ ) to the solution of (2.1), but rather to the solution of

$$X_t = X_0 + \int_0^t g_1(X_s) ds + \int_0^t g_2(X_s) dW_s + \frac{1}{2} \int_0^t g_2'(X_s) g_2(X_s) ds. \quad (2.2)$$

One may combine the two last terms in (2.2) to a new type of stochastic integral, called the *Stratonovich* integral. However, the nature of the correction term in (2.2) could depend on the nature of the approximation. Further, in the multidimensional case, there is no general way to extend the map  $W \rightarrow X$  to a continuous endomorphism on  $V = C_0(\mathbb{R}_+; \mathbb{R}^d)$ .

§3 The situation can be remedied for SDEs using a device introduced by T. Lyons [13], which he called rough paths analysis. The basic observation is that while the Itô map is not continuous on  $V$ , one could add more data, namely the iterated integrals of the Wiener path (or more precisely, certain antisymmetric combinations of such integrals). Those are viewed as elements of the tensor algebra  $\mathcal{V} = \bigoplus V^{\otimes k}$ . Putting an appropriate norm on  $V^{\otimes k}$  (the  $p$ -variation norm), Lyons showed that if a path together with its first  $[p]$  iterated integrals has a finite  $p$ -variation norm, then there is a unique extension to all  $V^{\otimes m}$  for  $m > [p]$ , which is of finite  $p$ -variation norm, and further the Itô map can be extended to a continuous map on  $\mathcal{V}$ -valued paths. Probability theory now enters only in giving a meaning to the first  $[p]$  iterated integrals; in the case of multidimensional SDEs, only the Wiener process and its *Lévy area*  $\int_0^t W_s^i dW_s^j - \int_0^t W_s^j dW_s^i$  are needed. Lyons expansion is thus a *universal* expansion: one is given a base of universally defined elements, and one writes the solution of SDEs in terms of those basis elements.

§4 One defines a space-time white noise  $\xi$  in a way similar to the definition for temporal white noise: for deterministic  $f \in L^2(\mathbb{R}_+ \times \mathbb{R}^d)$ , one declares the integral  $\xi(f) = \int_{\mathbb{R}_+ \times \mathbb{R}^d} f(s, x) \xi(s, x) ds dx$  to be a centered Gaussian variable of zero mean and variance  $\|f\|_2^2 = \int_{\mathbb{R}_+ \times \mathbb{R}^d} f^2(s, x) ds dx$ . One can then give an interpretation to certain linear parabolic

stochastic partial differential equation

$$\partial_t u(t, x) = \Delta u(t, x) + \xi, \quad u(0, x) = g(x) \tag{2.3}$$

in terms of the Green function  $G(s, y; t, x) = G(t - s, x - y)$  for the heat equation, by

$$u(t, x) = \int_{\mathbb{R}^d} G(t, x - y)g(y)dy + \int_0^t \int_{\mathbb{R}^d} G(t - s, x - y)\xi(s, y)dsdy. \tag{2.4}$$

The solution  $u$  is then a function only in dimension  $d = 1$  (but can be given an interpretation as a random distribution in higher dimension). One also can give sense to certain quasi-linear equations (where nonlinear terms are added to the right side of (2.3)), as long as the added terms are still well defined as function. One also can define similarly a space-only white noise (in this case, the solution to (2.3) would be a function for  $d \leq 3$ ).

§5 The biggest difference between space-time white noise and time-only white noise lies in the lack of the notion of adapted processes: without it, one cannot extend  $\xi(f)$  to a rich enough class of random integrands  $f$ . In addition, in the SPDE case, one is interested not only in adding to the right side nonlinearities of the form  $g(u)\xi$ , but also terms of the form  $(u_x)^2$  or other nonlinearities. However, the solutions to (2.3) can be checked to be only Hölder continuous (with exponent  $< 1/2$ ) in space, and terms like  $(u_x)^2$  simply do not make sense. We have now set the stage for Hairer’s theory of regularity structures.

### 3. Solving nonlinear PDEs - the KPZ example

§6 It is natural to attempt to give sense to an equation driven by rough noise by first mollifying the noise with an approximate Dirac function, solving the mollified equation, and then tuning out the mollification. More precisely, given an equation

$$\partial_t u = L(u, \xi),$$

for some nonlinear differential operator  $L$ , and a positive bump function  $\phi(x, t)$  integrating to 1, define the approximation

$$\xi^\epsilon(t, x) = \epsilon^{-d-2} \int_{\mathbb{R}_+ \times \mathbb{R}^d} \phi\left(\frac{x - y}{\epsilon}, \frac{t - s}{\epsilon^2}\right) \xi(y, s) dy ds$$

and solve the equation

$$\partial_t u^\epsilon = L(u^\epsilon, \xi^\epsilon).$$

One would like to know that  $u^\epsilon$  converges to a limit, and that the limit does not depend on the bump function  $\phi$  used. As the discussion in §2 showed, this is too much to expect - one may need to add a correction term, or in the case of SPDEs, a correction term that may blow up as  $\epsilon \rightarrow 0$ . The first truly singular example in which this program was carried out successfully was in Hairer’s construction of solution to the KPZ equation, to which we turn next.

§7 The KPZ equation (after Kardar-Parisi-Zhang) is the following SPDE:

$$\partial_t u = \Delta u + (\partial_x u)^2 + \xi. \tag{3.1}$$

The KPZ equation was introduced in order to model the evolution of (spatial) fluctuations of interfaces [12]. It is now established, both rigorously [2], [1] and experimentally [15], that the fluctuations of many physical systems converge to solutions of the (one dimensional) KPZ equation, and moreover the latter are linked to fluctuations in random matrix models. For the reasons explained in §4 and §6, the equation (3.1) does not make sense as written; until Hairer’s work, the only way to give a meaning to (3.1) was to consider the Hopf-Cole transformation  $u = \log h$ , note that  $h$  formally solves a stochastic heat equation (for which the methods of linear equations described in §4 apply), and then *define* the solution to (3.1) by taking logarithms. With the following theorem, Hairer changed all that. In the statement,  $\xi^\epsilon$  is a mollification of  $\xi$  as defined in §6.

**Theorem 3.1** ([4]). *There exists a sequence of constants  $C^\epsilon \rightarrow \infty$  depending on the bump function  $\phi$  such that the solutions to the equation*

$$\partial_t u^\epsilon = \Delta u^\epsilon - C^\epsilon + (\partial_x u^\epsilon)^2 + \xi^\epsilon, \quad u(0, x) = h_0(x) \quad (3.2)$$

converge as  $\epsilon \rightarrow 0$  (in Hölder norm) to a limit  $u$ .

The limit in Theorem 3.1 does not depend on the bump function used in the mollification, and coincides with the logarithm of the solution to the stochastic heat equation. A-priori, the solution is constructed only locally in time (i.e., up to a random explosion time), but because the solution coincides with the logarithm of the solution to the stochastic heat equation, which exists for all time, one deduces that the explosion time is infinite almost surely.

§8 Hairer’s solution replaces the universal basis employed in Lyons’ rough path theory by a basis derived from the solution to the linearized equation; the basis is thus both problem dependent and local. More explicitly, he constructs a (Polish) space  $\mathcal{X}$  consisting of choices of finitely many distributions, together with a base of the linear span of these distributions at each space-time point, with the constraint that the bases satisfy appropriate compatibility conditions and analytical bounds. He then constructs a measurable map  $\Psi$  from the probability space  $(\Omega, \mathcal{F}, P)$  (supporting the white noise) to  $\mathcal{X}$ , and a continuous map  $\mathcal{S}_R : C^\beta \times \mathcal{X} \rightarrow C(\mathbb{R}_+, C^{1/2-\beta})$  which gives  $u = \mathcal{S}_R(h_0, \Psi(\omega))$ . The map  $\mathcal{S}_R$  is constructed as fixed point of a Picard iteration, while the construction of  $\Psi$  involves a renormalization procedure, where the mollification  $\epsilon$  plays the role of small parameter.

The constant  $C_\epsilon$  in (3.2), that may blow up, is reminiscent of renormalization procedures in mathematical physics, and arises from the need to tame singularities. In the case of other equations, one may need to subtract more complicated terms in order to achieve convergence. In all cases, the solution eventually appears as the unique fixed point of the sequence of renormalizations.

Some of the results in [4] have roots in the earlier work [3].

## 4. Regularity structures

§9 Hairer’s solution for the KPZ equation was shortly followed by a general method to handle SPDEs, as long as their solution “should make sense” using renormalization. In order to do that, Hairer develops a systematic theory for the construction of base elements (that are used to locally expand the solutions), together with a transfer rule between points; crucially, only finitely many terms representing singularities need to be retained at each point. The

following, taken from [5], are representative examples of applications of the theory.

$$\begin{aligned} \partial_t h &= \Delta h + (\partial_x h)^2 + \xi, & (d = 1, \text{KPZ}) \\ \partial_t \phi &= \Delta \phi - \phi^3 + \xi, & (d = 3, \Phi_3^4) \\ \partial_t u &= \Delta u + g_{ij}(x) \partial_i u \partial_j u + f(u) \eta, & (d = 2, \text{Parabolic Anderson}) \end{aligned}$$

(In the standard parabolic Anderson model, one has  $g = 0$  and  $f(u) = u$ , and  $\eta$  is spatial white noise.) In both the KPZ and  $\Phi_3^4$  equation, the noise  $\xi$  is space-time white noise. Note that even without the nonlinear term, the solution to the  $\Phi_3^4$  equation is only a distribution, not a function.

The mollified versions of the above equations are given as follows.

$$\begin{aligned} \partial_t h^\epsilon &= \Delta h^\epsilon + (\partial_x h^\epsilon)^2 - C_\epsilon + \xi^\epsilon, & (d = 1, \text{KPZ}) \\ \partial_t \phi^\epsilon &= \Delta \phi^\epsilon - (\phi^\epsilon)^3 + C_\epsilon \phi^\epsilon + \xi^\epsilon, & (d = 3, \Phi_3^4) \\ \partial_t u^\epsilon &= \Delta u^\epsilon + g_{ij}(x) (\partial_i u^\epsilon \partial_j u^\epsilon - C_\epsilon \delta_{ij}) + f(u^\epsilon) (\eta^\epsilon - \bar{C}_\epsilon f'(u^\epsilon)), & (d = 2, \text{PA}) \end{aligned}$$

In these equations,  $C_\epsilon, \bar{C}_\epsilon$  are constants that may blow up, and depend on the mollifier employed.

**Theorem 4.1** ([5]). *There exist choices of constants  $C_\epsilon, \bar{C}_\epsilon$  such that the solutions converge, for a wide family of mollifiers, locally in time.*

Here “converges” is in a topology adapted to the regularity of the problem. Specifically, in the KPZ case, it was in Hölder norm with exponent smaller than  $1/2$  while in the  $\phi_3^4$  case, it is in the space of distribution with (negative) regularity exponent in  $(-2/3, -1/2)$ . Also, “local in time” means that there may be a finite (but a.s. strictly positive) explosion time and the solution is defined (and the mollifications converge) only up to the explosion time. The putative invariant measure for the  $\Phi_3^4$  equation, which is not shown to exist due to possible explosion, is the  $\Phi_3^4$  Euclidean quantum field theory measure.

§10 To achieve these convergence results, Hairer expresses the solutions to the SPDEs in a local basis, in such a way that the solution is determined uniquely at a point by only finitely many irregular terms. The basis elements themselves are chosen to fit the structure of the solution (often, just defined from the solution of the linearized equation), and are localized. Hairer also introduces a renormalization map that allows one to relate expansions for different values of  $\epsilon$  and eventually prove the convergence. The renormalization map is realized in terms of explicit (finite dimensional) matrices. Hairer also developed a diagrammatic calculus that allows one to keep track of the different basis elements needed in the description of the solution. Hairer coined the term *regularity structures* to emphasize the fact that the non-trivial part of the solution of the SPDE is determined by a finite dimensional space that depends on the regularity of the solution and on the nature of nonlinearities in the equation; the renormalization procedure for mollified noise is then performed in that space.

## 5. Other noteworthy results

§11 Earlier work of Hairer touched upon many aspects of the theory of SPDEs. A particularly strong line of work was jointly with J. Mattingly, where they studied the ergodicity

of two dimensional stochastic Navier-Stokes equation, when the noise enters only through finitely many modes. More explicitly, consider the stochastic Navier-Stokes equation on the two-dimensional torus

$$dw = \nu \Delta w dt + B(Kw, w)dt + QdW_t, \quad B(u, w) = -(u \cdot \nabla)w,$$

where  $K$  is a linear, divergence free operator (given in Fourier representation),  $Q$  is an operator representing coloring in space of noise, which influences only a finite number of modes. Using the Malliavin calculus and an *asymptotic* form of the strong Feller property (weak dependence on past initial conditions) which they introduce, Hairer and Mattingly [7],[8],[9] give sharp conditions for ergodicity as function of support of noise; they get also regularity of the solutions.

§12 In another direction, Hairer has come back full circle to corrections of the Wong-Zakai type for SPDEs. This is done in [6] for Burgers-type parabolic SPDEs and, more recently, in [10] using the machinery of regularity structures, for more general one dimensional SPDEs driven by space-time white noise, where correction terms that are not present in the SDE setup show up.

## 6. Summary

§13 Martin Hairer's work has established a completely new approach to problems in the theory of SPDEs that have remained largely untouchable for years. The framework he established with his theory of regularity structures has transformed the field and is paving the way to striking advance in the theory of SPDEs and in core models of mathematical physics.

## References

- [1] Gideon Amir, Ivan Corwin, and Jeremy Quastel, *Probability distribution of the free energy of the continuum directed random polymer in  $1 + 1$  dimensions*, Comm. Pure Appl. Math. **64** (2011), no. 4, 466–537.
- [2] Lorenzo Bertini and Giambattista Giacomin, *Stochastic Burgers and KPZ equations from particle systems*, Comm. Math. Phys. **183** (1997), no. 3, 571–607.
- [3] Martin Hairer, *Rough stochastic PDEs*, Comm. Pure Appl. Math. **64** (2011), no. 11, 1547–1585.
- [4] ———, *Solving the KPZ equation*, Ann. of Math. (2) **178** (2013), no. 2, 559–664.
- [5] ———, *A theory of regularity structures*, Inventiones Math. **198** (2014), no. 2, 269–504.
- [6] Martin Hairer and Jan Maas, *A spatial version of the Itô-Stratonovich correction*, Ann. Probab. **40** (2012), no. 4, 1675–1714.
- [7] Martin Hairer and Jonathan C. Mattingly, *Ergodic properties of highly degenerate 2D stochastic Navier-Stokes equations*, C. R. Math. Acad. Sci. Paris **339** (2004), no. 12, 879–882.

- [8] ———, *Ergodicity of the 2D Navier-Stokes equations with degenerate stochastic forcing*, Ann. of Math. (2) **164** (2006), no. 3, 993–1032.
- [9] ———, *Spectral gaps in Wasserstein distances and the 2D stochastic Navier-Stokes equations*, Ann. Probab. **36** (2008), no. 6, 2050–2091.
- [10] Martin Hairer and Étienne Pardoux, *A Wong-Zakai theorem for stochastic pdes*, arXiv:1409.3138.
- [11] Kiyosi Itô, *On a stochastic integral equation*, Proc. Japan Acad. **22** (1946), no. 1-4, 32–35.
- [12] Mehran Kardar, Giorgio Parisi, and Yi-Cheng Zhang, *Dynamic scaling of growing interfaces*, Physical Review Letters **56** (1986), 889–892.
- [13] Terry J. Lyons, *Differential equations driven by rough signals*, Rev. Mat. Iberoamericana **14** (1998), no. 2, 215–310.
- [14] Daniel W. Stroock and S. R. S. Varadhan, *On the support of diffusion processes with applications to the strong maximum principle*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. III: Probability theory, Univ. California Press, Berkeley, Calif., 1972, pp. 333–359.
- [15] Kazumasa A. Takeuchi, Masaki Sano, Tomohiro Sasamoto, and Herbert Spohn, *Growing interfaces uncover universal fluctuations behind scale invariance*, Scientific Reports **1** (2011), 34.
- [16] Eugene Wong and Moshe Zakai, *On the relation between ordinary and stochastic differential equations*, Internat. J. Engrg. Sci. **3** (1965), 213–229.

Faculty of Mathematics, Weizmann Institute, Rehovot, Israel and Courant Institute, New York University, USA

E-mail: ofer.zeitouni@weizmann.ac.il

## **Maryam Mirzakhani**

B.Sc. in Mathematics, Sharif University of Technology, Tehran, 1999

Ph.D. in Mathematics, Harvard University, 2004

### **Positions held**

Research Fellow, Clay Mathematics Institute, 2004

Professor, Princeton University, 2004–2008

Professor, Stanford University, 2008–present

# The work of Maryam Mirzakhani

Curtis T. McMullen

**Abstract.** Maryam Mirzakhani has been awarded the Fields Medal for her outstanding work on the dynamics and geometry of Riemann surfaces and their moduli spaces.

## 1. Introduction

Mirzakhani has established a suite of powerful new results on orbit closures and invariant measures for dynamical systems on moduli spaces. She has also given a new proof of Witten's conjecture, which emerges naturally from a counting problem for simple closed geodesics on Riemann surfaces. This note gives a brief discussion of her main results and their ramifications, including the striking parallels between homogeneous spaces and moduli spaces that they suggest.

## 2. The setting

We begin with a résumé of background material, to set the stage.

Let  $\mathcal{M}_g$  denote the moduli space of curves of genus  $g \geq 2$ . This space is both a complex variety, with  $\dim_{\mathbb{C}} \mathcal{M}_g = 3g - 3$ , and a symplectic orbifold. Its points are in bijection with the isomorphism classes of compact Riemann surfaces  $X$  of genus  $g$ .

The dimension of  $\mathcal{M}_g$  was known already to Riemann. Rigorous constructions of moduli space were given in the 1960s, by Ahlfors and Bers in the setting of complex analysis and by Mumford in the setting of algebraic geometry. Today the theory of moduli spaces is a meeting ground for mathematical disciplines ranging from arithmetic geometry to string theory.

The symplectic form  $\omega$  on  $\mathcal{M}_g$  arises from the hyperbolic metric on  $X$ . As shown by Wolpert, in the length–twist coordinates coming from a pair of pants decomposition of  $X$ , one can write

$$\omega = \sum_1^{3g-3} dl_i \wedge d\tau_i.$$

The complex structure on  $\mathcal{M}_g$  arises from the natural isomorphism

$$T_X^* \mathcal{M}_g = Q(X) = \{\text{holomorphic forms } q = q(z) dz^2 \text{ on } X\}$$

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

between the cotangent space to  $\mathcal{M}_g$  at  $X$  and the space of holomorphic quadratic differentials on  $X$ . The *Teichmüller metric* on  $\mathcal{M}_g$  also emerges from its complex structure: on the one hand, it is dual to the  $L^1$  norm

$$\|q\| = \int_X |q(z)| |dz|^2 = \text{area}(X, |q|)$$

on  $T_X^* \mathcal{M}_g$ ; on the other hand, it agrees with the intrinsic Kobayashi metric on  $\mathcal{M}_g$  (Royden).

Moduli space can be presented as the quotient  $\mathcal{M}_g = \mathcal{T}_g / \text{Mod}_g$  of Teichmüller space — its universal cover, a contractible bounded domain in  $\mathbb{C}^{3g-3}$  — by the action of the mapping–class group of a surface.

One of the challenges of working with moduli space is that it is *totally inhomogeneous*: for example, the symmetry group of  $\mathcal{T}_g$  (as a complex manifold) is simply the discrete group  $\text{Mod}_g$  (for  $g > 2$ ). One of Mirzakhani’s remarkable contributions is to show that, nevertheless, dynamics on moduli space displays many of the same rigidity properties as dynamics on homogeneous spaces (see §4).

### 3. From simple geodesics to Witten’s conjecture

We begin with Mirzakhani’s work on simple geodesics. In the 1940s, Delsarte, Huber and Selberg established the *prime number theorem* for hyperbolic surfaces, which states that the number of (oriented, primitive) closed geodesics on  $X \in \mathcal{M}_g$  with length  $\leq L$  satisfies

$$\pi(X, L) \sim \frac{e^L}{L}.$$

(The usual prime number theorem says that the number of prime integers with  $0 < \log p \leq L$  is asymptotic to  $e^L/L$ .)

The number of *simple* closed geodesics  $\sigma(X, L)$  behaves quite differently; it only has polynomial growth, and in 2004 Mirzakhani proved that

$$\sigma(X, L) \sim C_X L^{6g-6}.$$

In contrast to the prime number theorem, the right–hand side here depends on both the genus and geometry of  $X$ .

Although the statement above involves only a single Riemann surface  $X$ , Mirzakhani’s proof involves integration over moduli space and leads to a cascade of new results, including a completely unexpected proof of the Witten’s conjecture. The latter conjecture, established by Kontsevich in 1992, relates the intersection numbers on moduli space defined by

$$\langle \tau_{d_1}, \dots, \tau_{d_n} \rangle = \int_{\overline{\mathcal{M}}_{g,n}} c_1(E_1)^{d_1} \cdots c_1(E_n)^{d_n}$$

to a power series solution to the KdV hierarchy (an infinite system of differential equations satisfying the Virasoro relations). Here  $\overline{\mathcal{M}}_{g,n}$  is the Deligne–Mumford compactification of the moduli space of Riemann surfaces  $X$  with marked points  $(p_1, \dots, p_n)$ , and  $c_1(E_i)$  denotes the first Chern class of the line bundle  $E_i \rightarrow \overline{\mathcal{M}}_{g,n}$  with fibers  $T_{p_i}^* X$ .

Mirzakhani’s investigation of  $\sigma(X, L)$  also leads to formulas for the frequencies of different topological types of simple closed curves on  $X$ ; for example, a random simple curve

on a surface of genus 2 has probability  $1/7$  of cutting  $X$  into two pieces of genus 1. These frequencies are always rational numbers, and they depend only on  $g$ , not  $X$ .

At the core of these results is Mirzakhani's novel, recursive calculation of the volume of the moduli space of Riemann surfaces of genus  $g$  with  $n$  geodesic boundary components with lengths  $(L_1, \dots, L_n)$ . This volume is defined by

$$P_{g,n}(L_1, \dots, L_n) = \int_{\mathcal{M}_{g,n}(L_1, \dots, L_n)} \omega^{3g-3+n};$$

for example, one can show that  $P_{1,1}(L_1) = (1/24)(L_1^2 + 4\pi^2)$ . In general,  $P_{g,n}$  is a polynomial whose coefficients (which lie in  $\mathbb{Q}(\pi)$ ) can be related to frequencies and characteristic classes, yielding the results discussed above. Previously only the values of  $P_{g,n}(0, \dots, 0)$  were known. The proofs depend on intricate formulas for dissections of surfaces along hyperbolic geodesics; see [8], [6] and [7]. Mirzakhani has also studied the behavior of  $\mathcal{M}_g$  as  $g \rightarrow \infty$ ; see [9], [11].

#### 4. Complex geodesics in moduli space

We now turn to Mirzakhani's work on moduli spaces and dynamics. Her contributions to this area include a prime number theorem for closed geodesics in  $\mathcal{M}_g$ , counting results for orbits of  $\text{Mod}_g$  on  $\mathcal{T}_g$ , and the classification of  $\text{Mod}_g$ -invariant measures on the space of measured laminations  $\mathcal{ML}_g$ . But perhaps her most striking work — which we will present here — is a version of Ratner's theorem for moduli spaces.

**Complex geodesics.** It has been known for some time that the Teichmüller geodesic flow is ergodic (Masur, Veech), and hence almost every geodesic  $\gamma \subset \mathcal{M}_g$  is dense. It is difficult, however, to describe the behavior of *every single geodesic*  $\gamma$ ; already on a hyperbolic surface, the closure of a geodesic can be a fractal cobweb, and matters only get worse in moduli space.

Teichmüller showed that moduli space is also abundantly populated by *complex* geodesics, these being holomorphic, isometric immersions

$$F : \mathbb{H} \rightarrow \mathcal{M}_g.$$

In fact there is a complex geodesic through every  $X \in \mathcal{M}_g$  in every possible direction.

In principle, the closure of a complex geodesic might exhibit the same type of pathology as a real geodesic. But in fact, the opposite is true. In a major breakthrough, Mirzakhani and her coworkers have shown:

*The closure of any complex geodesic is an algebraic subvariety  $V = \overline{F(\mathbb{H})} \subset \mathcal{M}_g$ .*

This long sought-after rigidity theorem was known previously only for  $g = 2$ , with some restrictions on  $F$  [5]. (In the case of genus two,  $V$  can be an isometrically immersed curve, a Hilbert modular surface, or the whole space  $\mathcal{M}_2$ .)

**Dynamics over moduli space.** The proof of this rigidity theorem involves the natural action of  $\text{SL}_2(\mathbb{R})$  on the sphere bundle

$$Q_1\mathcal{M}_g \rightarrow \mathcal{M}_g,$$

consisting of pairs  $(X, q)$  with  $q \in Q(X)$  and  $\|q\| = 1$ .

To describe this action, consider a Riemann surface  $X = P/\sim$  presented as the quotient of a polygon  $P \subset \mathbb{C}$  under isometric edge identifications between pairs of parallel sides. Such identifications preserve the quadratic differential  $dz^2|_P$ , so a polygonal model for  $X$  actually determines a pair  $(X, q) \in Q\mathcal{M}_g$  with  $\|q\| = \text{area}(P)$ . Conversely, every nonzero quadratic differential  $(X, q) \in Q\mathcal{M}_g$  can be presented in this form.

Since  $\text{SL}_2(\mathbb{R})$  acts linearly on  $\mathbb{R}^2 \cong \mathbb{C}$ , given  $A \in \text{SL}_2(\mathbb{R})$  we can form a new polygon  $A(P) \subset \mathbb{C}$ , and use the corresponding edge identifications to define

$$A \cdot (X, q) = (X_A, q_A) = (A(P), dz^2)/\sim .$$

Note that  $[X_A] = [X]$  if  $A \in \text{SO}_2(\mathbb{R})$ . Thus the map  $A \mapsto X_A$  descends to give a map

$$F : \mathbb{H} \cong \text{SL}_2(\mathbb{R})/\text{SO}_2(\mathbb{R}) \rightarrow \mathcal{M}_g,$$

which is the complex geodesic *generated* by  $(X, q)$ .

The proof that  $\overline{F(\mathbb{H})} \subset \mathcal{M}_g$  is an algebraic variety involves the following three theorems, each of which a substantial work in its own right.

1. Measure classification (Eskin and Mirzakhani). *Every ergodic,  $\text{SL}_2(\mathbb{R})$ -invariant probability measure on  $Q_1X$  comes from Euclidean measure on a special complex-analytic subvariety  $A \subset Q\mathcal{M}_g$  (The variety  $A$  is linear in period coordinates).*

This is the deepest step in the proof; it uses a wide variety of techniques, including conditional measures and a random walk argument inspired by the work of Benoist and Quint [1].

2. Topological classification (Eskin, Mirzakhani and Mohammadi). *The closure of any  $\text{SL}_2(\mathbb{R})$  orbit in  $Q_1X$  is given by  $A \cap Q_1X$  for some special analytic subvariety  $A$ .*
3. Algebraic structure (Filip). *Any special analytic subvariety  $A$  is in fact an algebraic subvariety of  $Q\mathcal{M}_g$ .* Thus its projection to  $\mathcal{M}_g$ ,  $V = \overline{F(\mathbb{H})}$ , is an algebraic subvariety as well.

See [2], [3] and [4] for these developments.

**Ramifications: Beyond homogeneous spaces.** This collection of results reveals that the theory of dynamics on homogeneous spaces, developed by Margulis, Ratner and others, has a *definite resonance* in the highly inhomogeneous, but equally important, world of moduli spaces.

The setting for homogeneous dynamics is the theory of Lie groups. Given a lattice  $\Gamma$  in a Lie group  $G$ , and a Lie subgroup  $H$  of  $G$ , one can consider the action

$$H \curvearrowright G/\Gamma$$

by left multiplication, just as in the setting of moduli spaces we have considered the action

$$\text{SL}_2(\mathbb{R}) \curvearrowright Q_1\mathcal{T}_g/\text{Mod}_g .$$

One of the most powerful results in homogeneous dynamics is *Ratner's theorem*. It implies that if  $H$  is generated by unipotent elements, then every orbit closure  $\overline{Hx} \subset G/\Gamma$  is a *special submanifold* — in fact, it has the form

$$\overline{Hx} = Jx \subset G/\Gamma$$

for some Lie subgroup  $J$  with  $H \subset J \subset G$ . A similar statement holds for invariant measures. Since  $\mathrm{SL}_2(\mathbb{R})$  is generated by unipotent elements (matrices such as  $\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$  and its transpose), one might hope for a version of Ratner's theorem to hold in moduli spaces. This is what Mirzakhani's work confirms.

**Hodge theory versus geometry.** For another perspective, recall that  $\mathcal{M}_g$  embeds into the moduli space of Abelian varieties  $\mathcal{A}_g = \mathfrak{H}_g / \mathrm{Sp}_{2g}(\mathbb{Z})$ , a locally symmetric space amenable to the methods of homogeneous dynamics. But the complex geodesics in  $\mathcal{M}_g$  become inhomogeneous when mapped into  $\mathcal{A}_g$ , so they cannot be analyzed by these methods. Mirzakhani's work shows that one can work effectively and directly with  $\mathcal{M}_g$  rather than with  $\mathcal{A}_g$ , by geometric analysis on Riemann surfaces themselves.

**Ramifications: Billiards.** The  $\mathrm{SL}_2(\mathbb{R})$  action on  $Q_1\mathcal{M}_g$  is also connected with the theory of *billiards in polygons* — an elementary branch of dynamics in which difficult problems abound.

Let  $T \subset \mathbb{C}$  be a connected polygon with angles in  $\pi\mathbb{Q}$ . The behavior of billiard paths in  $T$  is closely related to the behavior of the complex geodesic generated by a quadratic differential  $(X, q)$  obtained by 'unfolding' the table  $T$ .

Indeed, the first examples of complex geodesics such that  $V = \overline{F(\mathbb{H})} \subset \mathcal{M}_g$  is an algebraic curve — i.e. the image of the complex geodesic is as small as possible — were constructed by Veech in his analysis of billiards in regular polygons. In this case the stabilizer of the corresponding quadratic differential is a lattice  $\mathrm{SL}(X, q) \subset \mathrm{SL}_2(\mathbb{R})$ , which serves as the *renormalization group* for the original billiard flow.

The work of Mirzakhani has bearing on several open conjectures in the field of billiard dynamics. For example, it provides progress on the open problem of showing that, for any table  $T$ , there is an algebraic number  $C_T$  such that the number  $N(T, L)$  of types of *primitive, periodic* billiard paths in  $T$  of length  $\leq L$  satisfies

$$N(T, L) \sim \frac{C_T L^2}{\pi \operatorname{area}(T)}.$$

Eskin and Mirzakhani have shown that an asymptotic equation of this form holds after averaging over  $L$ , and that  $C_T$  can assume only countably many values.

## 5. Dynamics of earthquakes

We conclude by discussing Mirzakhani's work on the earthquake flow, and a measurable bridge between the symplectic and holomorphic aspects of  $\mathcal{M}_g$ .

A classical construction of Fenchel and Nielsen associates to a simple closed geodesic  $\gamma \subset X \in \mathcal{M}_g$  and  $t \in \mathbb{R}$  a new Riemann surface

$$X_t = \operatorname{tw}_{t\gamma}(X) \in \mathcal{M}_g,$$

obtained by cutting  $X$  open along  $\gamma$ , twisting by length  $t$  to the right, and then regluing. The resulting *twist path* in  $\mathcal{M}_g$  is periodic; if  $\gamma$  has length  $L$ , then  $X_{t+L} = X_t$ .

On the other hand, one can also twist along *limits* of weighted simple geodesics, called *measured laminations*. As shown by Thurston, the space of measured laminations forms a

PL manifold  $\mathcal{ML}_g \cong \mathbb{R}^{6g-6}$  with a natural volume form, and the limiting twists, called *earthquakes*, are defined for all time.

Earthquakes are a natural feature of the symplectic geometry of moduli space. While they can be defined geometrically by fracturing and regluing  $X$  along the (possibly fractal) support of  $\lambda \in \mathcal{ML}_g$ , they also arise more conventionally as the *Hamilton flows* associated to the functions  $Y \mapsto \text{length}(\lambda, Y)$ .

The earthquake flow lives on the bundle  $L_1\mathcal{M}_g$  of unit length laminations over  $\mathcal{M}_g$ . Mirzakhani has shown that, with respect to the natural measure on  $L_1\mathcal{M}_g$ :

*Thurston’s earthquake flow is ergodic.*

Prior to this result, the dynamics of earthquakes seemed completely opaque. Not a single example of a dense earthquake path in  $\mathcal{M}_g$  was known; we can now assert that almost every earthquake path is dense and uniformly distributed.

**Bridging the symplectic/holomorphic divide.** The proof of ergodicity of the earthquake flow uses a remarkable bridge between the symplectic and holomorphic sides of moduli space.

In more detail, recall that the *horocycle flow* on  $Q_1\mathcal{M}_g$  is defined by the action of the 1-parameter group  $N = \{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} : t \in \mathbb{R} \} \subset \text{SL}_2(\mathbb{R})$ . Drawing on ideas from Thurston’s work on stretch maps, Mirzakhani shows there is a measure-preserving map  $\beta : L_1\mathcal{M}_g \rightarrow Q_1\mathcal{M}_g$  which transports the earthquake flow to the horocycle flow. In other words, we have a commutative diagram of the form

$$\begin{array}{ccc}
 \text{earthquake flow} \circlearrowleft L_1\mathcal{M}_g & \xrightarrow{\beta} & Q_1\mathcal{M}_g \circlearrowright \text{horocycle flow} \\
 \downarrow & & \downarrow \\
 \mathcal{M}_g & & \mathcal{M}_g
 \end{array}$$

But the horocycle flow on  $Q_1\mathcal{M}_g$  is well-known to be ergodic (this is a formal corollary of ergodicity of the geodesic flow [12, Thm. 2.4.2]), so the same is true for the earthquake flow [10]. (It is an open problem to establish Ratner-type rigidity for these flows.)

**Summary.** Mirzakhani’s research has integrated, with great originality, a broad range of mathematical disciplines — including algebraic and symplectic geometry, low-dimensional topology, and random processes. Her breakthroughs have transformed our perspective on moduli spaces, and led the way to mathematical frontiers where striking developments are still unfolding.

**References**

[1] Y. Benoist and J.-F. Quint, *Mesures stationnaires et fermés invariants des espaces homogènes*. *Annals of Math.* **174** (2011), 1111–1162.  
 [2] A. Eskin and M. Mirzakhani, *Invariant and stationary measures for the  $\text{SL}_2(\mathbb{R})$  action on moduli space*. Preprint, 2/2014.

- [3] A. Eskin, M. Mirzakhani, and A. Mohammadi, *Isolation, equidistribution, and orbit closures for the  $SL_2(\mathbb{R})$  action on moduli space*. Preprint, 6/2013.
- [4] Simion Filip, *Splitting mixed Hodge structures over affine invariant manifolds*. Preprint, 11/2013.
- [5] C. McMullen, *Dynamics of  $SL_2(\mathbb{R})$  over moduli space in genus two*. *Annals of Math.* **165** (2007), 397–456.
- [6] M. Mirzakhani, *Simple geodesics and Weil–Petersson volumes of moduli spaces of bordered Riemann surfaces*. *Inv. math.* **167** (2007), 179–222.
- [7] ———, *Weil–Petersson volumes and intersection theory on the moduli spaces of curves*. *J. Amer. Math. Soc.* **20** (2007), 1–23.
- [8] ———, *Growth of the number of simple closed geodesics on hyperbolic surfaces*. *Annals of Math.* **168** (2008), 97–125.
- [9] ———, *Growth of Weil–Petersson volumes and random hyperbolic surfaces of large genus*. *J. Differential Geom.* **94** (2013), 267–300.
- [10] ———, *Ergodic theory of the earthquake flow*. *Int. Math. Res. Not.* (2008), Art. ID rnm116, 39 pp.
- [11] M. Mirzakhani and P. Zograf, *Towards large genus asymptotics of intersection numbers on moduli spaces of curves*. Preprint, 12/2011.
- [12] R. Zimmer, *Ergodic Theory and Semisimple Groups*. Birkhäuser, 1984.

Department of Mathematics, Harvard University, One Oxford Street, Cambridge, MA 02138, USA  
E-mail: ctm@math.harvard.edu

## **Subbash Khot**

Ph.D. in Computer Science, Princeton University, 2003

### **Positions held**

Member of School of Mathematics, Institute for Advanced Study, 2003–2004

Assistant Professor, College of Computing at Georgia Institute of Technology, 2004–2007

Associate Professor, New York University, 2007–2011

Visiting Faculty, University of Chicago with the theory group, 2011–2013

Professor, New York University, 2013–present

# The work of Subhash Khot

Sanjeev Arora

**Abstract.** Subhash Khot, the winner of the 2014 Nevanlinna Prize, has brought new clarity to the study of approximation algorithms for NP-hard problems, and opened new avenues of research. Several of these concern his Unique Games Conjecture, which has led to optimal inapproximability results that exactly characterize the approximability of the problem.

The  $P \neq NP$  conjecture implies that NP-hard problems have no efficient algorithms (where efficient means the running time has to be *polynomial* in the input size). Since the class of NP-hard problems includes thousands of practical problems we still need to devise ways to handle them in practice. An *approximation algorithm* is an attractive possibility. Such an algorithm finds, *for every input*, a solution of *value* at least  $\alpha$  times the optimum. (If the problem involves minimization instead of maximization, the approximation algorithm finds a solution of *cost* at most  $\alpha$  times the optimum.) The parameter  $\alpha$  is called the *approximation ratio*. Designing such approximation algorithms has been a fertile research area in the the past few decades. This study also injects new ideas into mathematics since it calls for approximate characterizations of optimality.

## 1. Background: Approximation can be NP-hard:

Surprisingly, for a variety of problems it can be shown that achieving certain approximation ratios is also NP-hard —i.e., no easier than exact optimization. To put it another way, a good enough approximation algorithm can be used to solve the exact problem as well. Formally, this is shown by giving a *reduction* from the exact optimization (which is NP-hard) to approximation. This field took off with the proof of the PCP Theorem in the early 1990s [5, 6], which, after rapid development, resulted in a 1997 paper of Håstad [19] that proves so-called *threshold* results: these exhibit an approximation ratio  $\alpha$  for the problem such that an efficient algorithm can achieve this ratio, and yet achieving approximation ratio  $\alpha + \epsilon$  is NP-hard, where  $\epsilon = o(1)$ . The existence of threshold results is somewhat unexpected: *a priori*, one can imagine a continuous scale of difficulty, whereby achieving a ratio  $\alpha$  is polynomial-time, a ratio  $\beta$  is NP-hard for some  $\beta > \alpha$ , and ratios in  $(\alpha, \beta)$  have computational complexity intermediate between P and NP-hard. By ruling out this intermediate complexity, a threshold result gives a precise characterization of the approximability of the problem.

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

**An example:** Consider the solvability of a system of linear equations modulo 2, which is of course testable in polynomial time via Gaussian elimination. But what if the system is unsatisfiable? Can we give an approximation of some sort? A natural notion is to look for an assignment that satisfies as many equations as possible. Computing the best such assignment is NP-hard (which implies there is no maximization version of Gaussian elimination), but there is a simple  $1/2$  approximation: just assign a random value to each variable. For any constraint of the form  $x_1 + x_2 + \dots + x_k = b \pmod{2}$  the probability that the random assignment satisfies it is  $1/2$ , so linearity of expectation implies that the fraction of equations satisfied overall is  $1/2$ . Since the best assignment can at most satisfy all the equations, we conclude that the random assignment is a  $1/2$ -approximation. Håstad showed that achieving a  $1/2 + \epsilon$ -approximation is NP-hard. This is a threshold result.

Though threshold results have strong appeal, unfortunately, only a few examples came out of Håstad's techniques.

## 2. Khot's contributions

Khot made several seminal contributions to the study of inapproximability. First, in a sequence of papers he introduced new kinds of PCP Theorems tailored to proving inapproximability results for a host of open problems: Shortest Lattice Vector [22], Hypergraph Coloring [31], Agnostic Learning of Halfspaces [17], Bipartite Clique [23] etc. As part of this study he came up with a promising new direction for proving threshold results via the *Unique Games Conjecture*. Through substantial work by him and others, this has had many unexpected consequences, which we now describe.

**Unique Games Conjecture:** In 2004 Khot [21] introduced a problem called UNIQUE GAMES and conjectured that computing a good approximation to it—in the sense described below—is NP-hard. He supported this so-called *Unique Games Conjecture* (UGC) by pointing out that the problem seems resistant to usual techniques of designing approximation algorithms—specifically, linear programming (LP) and semidefinite programming (SDP), two popular techniques from convex optimization. He also showed that assuming his conjecture, he could prove new threshold results for problems that had seemed resistant to Håstad's techniques. In subsequent work, often with coauthors, he greatly extended this research program, which led to new threshold results for a variety of problems. These threshold results say that computing an  $\alpha$ -approximation to the problem (where  $\alpha$  is problem-dependent) would give a good polynomial-time approximation algorithm for the UNIQUE GAMES problem that refutes the UGC.

The original statement of the UGC was notation-heavy but an important paper of Khot, Kindler, Mossel, and O'Donnell [26] has yielded an equivalent statement that is cleaner. The conjecture posits the NP-hardness of the following problem for every positive  $\epsilon < 0.001$ : given a system of linear equations modulo a prime  $p > p(\epsilon)$ , where each equation involves only two variables, it is NP-hard to distinguish between the following two cases: (a) there is an assignment to the variables that satisfies at least  $(1 - \epsilon)$  fraction of the equations (b) Every assignment satisfies fewer than  $1/2$  the equations.

In part (b),  $1/2$  can be replaced by any other constant less than 1. Furthermore, without loss of generality the system can be assumed to consist only of equations of the type  $x_i - x_j = c_{ij} \pmod{p}$ .

As noted in our earlier example, the theory of NP-hardness rules out a robust polynomial-time analog of Gaussian elimination, i.e., an efficient algorithm that finds approximate solutions to linear systems that satisfy “almost all” or “most” equations. The UGC rules such robust analogs even for linear systems with a very simple structure.

**New threshold results and inapproximability results.** The UGC led to a raft of threshold results including for MAX CUT [26, 37], VERTEX COVER [30] and CONSTRAINT SATISFACTION PROBLEMS [38]. This confirmed Khot’s original insight that UGC should prove useful for threshold results. More unexpectedly, the UGC also implies the best inapproximability results we know of for several other well-studied problems such as GROTHENDIECK CONSTANT, SPARSEST CUT, GRAPH COLORING. Such results had proved difficult to obtain via traditional PCP Theorems. See Khot [25] and Trevisan [40] for a survey.

Such results suggest that though the unique games problem may at first sight seem a simple deviation from classical linear algebra, in fact it lies solidly in combinatorial optimization. The rich tapestry of its connections to classical problems of combinatorial optimization have convinced most experts that to make further algorithmic progress on a host of open problems, we have to first tackle (prove or disprove) the UGC.

**2.1. Unexpected Impacts of UGC.** As mentioned, the study of UGC has had many consequences that were unexpected, including by Khot himself (personal communication).

**Proof of optimality of specific approximation algorithms.** A surprising aspect of the above nonapproximability results—compared for example to results proved earlier using the PCP Theorem or Håstad’s Theorem—is that they end up proving the optimality of a *specific* algorithm (usually, LP or SDP based) in the following very strong sense: *any* problem instance where this specific algorithm doesn’t do well (i.e., for which the value of the solution is indeed  $\alpha$  factor from optimal) can be used in the reduction from UNIQUE GAMES to the problem. At first sight such an implication seems to make sense since a threshold result ought to give some insight into the optimum algorithm—after all, it implies that the optimum algorithm does no better than an  $\alpha$  approximation on some instances. But this implication is actually in the reverse direction: it turns a single a bad instance for a specific algorithm into a hardness result (implying that *no other* algorithm works) which is unprecedented in computational complexity. This phenomenon was strongly suggested in the works of Khot et al. [26] and Khot-O’Donnell [29], but got its clearest explanation in the work of Raghavendra [38], which yielded threshold results for the entire class of *constraint satisfaction problems* (CSPs). Raghavendra shows the existence of a problem-specific threshold  $\alpha$  without giving any finite algorithm to compute it. (The explicit value of this  $\alpha$  is known for a handful of problems.) He does this by showing, as hinted above, that any specific instance where the SDP approximation is worse than  $\alpha$  can be used as a *gadget* in a reduction that proves that every algorithm (not only SDP) that does better than  $\alpha + \epsilon$  would yield a polynomial-time algorithm for UNIQUE GAMES that refutes the UGC.

Let’s explain this mysterious connection in more detail for a concrete problem, MAX CUT (given a graph, find a partition of vertices into two sets so as to maximise the number of edges going between them), where the approximation threshold is known to be 0.878... This approximation ratio is achieved by an SDP-based algorithm of Goemans and Williamson [18]. We know this computes no better than a 0.878... approximation, thanks to Feige and Schechtman [16]. Their counterexample is a geometric graph: it contains vertices that correspond

to a dense set of points in  $\mathbb{R}^k$  for sufficiently large constant  $k$  and edges corresponding to vertex pairs  $u, u'$  whose corresponding vectors make an angle of about 138 degrees.

In the paper [26] this counterexample graph is turned into an actual reduction from UNIQUE GAMES to MAX CUT that shows that any efficient algorithm that computes better than a  $0.878 + \epsilon$ -approximation to MAX CUT will falsify the UGC.

**New results in analysis of boolean functions and isoperimetry.** The statements in the previous paragraph—which would have seemed miraculous and unprovable to any expert 10 years ago—rest upon some new mathematical advances. To see the need for these, one must delve into how one reduces UNIQUE GAMES to MAX CUT. If the UNIQUE GAMES instance consists of equations mod  $p$  then we produce an instance of MAX CUT, such that for each variable in the UNIQUE GAMES instance the new graph contains a copy of the boolean hypercube  $\{-1, 1\}^p$ . If the UNIQUE GAMES instance included the equation say  $x - y \equiv 11 \pmod{p}$ , then we connect in the new graph any pair  $u, u'$  where  $u$  is in the hypercube for  $x$ , and  $u'$  is a vertex in  $y$ 's hypercube, and they make an angle equal to about 138 degrees once the coordinates of  $u'$  are shifted by 11 places mod  $p$ . (This is a rough description; we are omitting details.) Thus the algebraic structure of the linear system is encoded in the edge interconnections of the new graph, and furthermore, this encoding uses a graph related to the hard instance for MAX CUT described earlier.

To prove the correctness of this reduction, one has to prove that the ability to approximate MAX CUT to a factor better than  $0.878 + \epsilon$  allows one to solve the original UNIQUE GAMES instance better than allowed by the UGC. The crux of this proof is a characterization of high-capacity cuts (i.e., those where the number of cut edges is close to optimal) in the above hypercube graph. Such a characterization requires *fourier analysis on boolean hypercube* (surveyed by Kalai and Safra [20]). The new insight of [26] was a robust characterization of the maximum cuts of the above version of the boolean hypercube, encapsulated in their *majority is stablest* conjecture. The cut corresponding to the MAJORITY function (namely, one that is 1 iff a majority of the input coordinates are 1) have capacity proportional to 0.878. Any cut with significantly higher capacity must correspond to a *junta*, i.e. largely determined by  $O(1)$  coordinates. They showed that this robust characterization, if true, implies that near-optimum cuts in the overall graph can be “decoded” to an assignment by using one of these junta coordinates for the hypercube as the value for the corresponding variable in the UNIQUE GAMES instance. Furthermore, this assignment satisfies a lot of equations in the UNIQUE GAMES instance.

The majority is stablest conjecture was essentially proved shortly thereafter by Mossel et al. [37]. In the ensuing years the above template of designing reductions has been used in other papers and has also led to more results in analysis of boolean functions and isoperimetry (see Khot’s survey [24]).

**New nonembeddability results for geometric embeddings of metric spaces.** How well does a *metric space*  $(X, d_1)$  resemble a metric space  $(Y, d_2)$ ? A natural measure is the minimum distortion  $C \geq 1$  for which there exists a map  $f: X \rightarrow Y$  and a scaling constant  $\gamma > 0$  such that

$$\gamma d_1(x, y) \leq d_2(f(x), f(y)) \leq C \gamma d_1(x, y).$$

Starting with the work of Linial, London, and Rabinovich [36], characterizing such distortion for interesting pairs of metric spaces has become an important research area due to its

algorithmic applications. An example of such rich applications was the *Goemans-Linial* conjecture, which stated that the lowest distortion for embedding *negative type* metrics into  $\ell_1$  is  $O(1)$ . If true, this conjecture would yield an approximation ratio  $O(1)$  for the SPARSEST CUT problem in graphs via SDP. At the time that conjecture was made, the best upperbound on distortion was  $O(\log n)$ , which follows from a much more general bound of Bourgain [9] for embedding any  $n$ -point metric space into  $\ell_2$ . In 2005 Arora et al. [4] made progress on the conjecture by improving the upper bound to  $O(\sqrt{\log n \log \log n})$ . This raised hopes that the Goemans-Linial conjecture may be true, but Khot and Vishnoi [32] dashed such hopes by disproving the conjecture. Note a disproof requires showing the *nonexistence* of an embedding with  $O(1)$  distortion, for which very few general techniques are known. Khot and Vishnoi's construction uses the boolean hypercube and they prove the nonexistence of a suitable embedding via fourier analysis similar to the one sketched above. This construction was inspired by the fact that the Goemans-Linial conjecture had been shown to violate UGC [10, 32], so intuitions about why UNIQUE GAMES is hard must yield a counterexample to the Goemans-Linial conjecture. Their construction is beautiful and its analysis uses fourier analysis of boolean functions analogous to the one sketched above. The techniques were extended by Khot and Naor [28] to prove nonembeddability results for other metric spaces such as edit-distance or Levenshtein metrics used in biology and other fields. More recently the works of Lee and Naor [35], and Cheeger, Kleiner, and Naor [11, 12] have greatly improved the Khot-Vishnoi result for negative type metrics.

**New flowering of higher order spectral graph theory.** Not everyone is convinced about the truth of the UGC, leading to some serious attempts to disprove it. Arora, Khot, Kolla, Steurer, Tulsiani and Vishnoi [3] showed that the UNIQUE GAMES problem is easy for random and random-like graphs. This suggested that finding hard instances of the problem is non-trivial. Then Arora, Barak and Steurer [2] showed that the UNIQUE GAMES problem has very *subexponential* algorithms, running in time  $\exp(n^\epsilon)$ . Though this doesn't disprove the UGC, it does suggest that the problem has some structure not shared by other NP-hard problems (as far as we know). This structure involves the higher eigenvalues of the *connection graph* associated with the UNIQUE—this is the graph that contains a vertex for each variable, and an edge corresponding to each pair of variables that are involved in a linear constraint. This fresh insight has since led to a host of other results in spectral graph theory, including versions of higher-order Cheeger-type inequalities for graphs (see [34] and its bibliography), and ingenious constructions of families of graphs [7] showing the near-tightness of the analysis of the algorithm of [2], which means that UGC has likely survived this attack and lives on.

**Progress on Kelvin's Problem in  $\mathfrak{R}^d$ :** In the 19th century Lord Kelvin posed the problem of determining the minimum surface area of a surface  $\Lambda$  in  $\mathfrak{R}^3$  such that  $\Lambda + Z^3$  tiles all of  $\mathfrak{R}^3$ . This problem is still open. The problem has also been studied for  $\mathfrak{R}^d$  for  $d > 3$ , where it is even unclear how the area of such a  $\Lambda$  should grow with  $d$ . Since the unit cube gives a suitable tiling, its surface area of  $2d$  is an upper bound. Furthermore, since  $\Lambda$  has volume 1 its area is at least that of the unit sphere, which is  $\Omega(\sqrt{d})$ . Inspired by a failed attempt to prove the UGC via a technique called *parallel repetition*, Kindler, O'Donnell, Rao, Wigderson [33] exhibited a suitable  $\Lambda$  of area  $O(\sqrt{d})$ , which matches the lower bound up to constant factors. This was the first major advance on Kelvin's problem in  $\mathfrak{R}^d$ .

### 3. Current status of UGC

It is only fair to end this remarkable story with some perspective and a report on our current beliefs about the UGC.

Some researchers suspect a disproof of the UGC could come out of an algorithm using the SDP hierarchies of Lasserre and Parillo, which can be seen as feasible versions of *Sum of Squares* proof systems (inspired, as the name suggests, by work on Hilbert's 17th problem). These hierarchies are known to be at least as powerful as the higher-order spectral method of [2]; see Barak and Steurer's survey [8]. A better algorithm for unique games would also imply new algorithms for the SMALL-SET EXPANSION problem on graphs [39].

Khot continues to work on proving the UGC, and has made partial progress in joint work with Moshkovitz [27].

Very likely, only one of these two directions will succeed, but it is also conceivable that they meet in the middle, and the UNIQUE GAMES problem has intermediate complexity—i.e., unsolvable in polynomial time yet not NP-hard. (The author finds this outcome most plausible.) In this case, all the above inapproximability results would be still valid and useful—since the approximation problems in question have been shown at least as hard as UNIQUE GAMES, we would conclude that they are also unsolvable in polynomial time. But whether they are NP-hard would become an open question.

### References

- [1] S. Arora, L. Babai, J. Stern, and Z. Sweedyk, *The hardness of approximate optima in lattices, codes, and systems of linear equations*, Journal of Computer and System Sciences, Volume 54, Issue 2 (1997), 317–331.
- [2] S. Arora, B. Barak, and D. Steurer, *Subexponential Algorithms for Unique Games and Related Problems*, In Proc. 51th IEEE Symposium on Foundations of Computer Science, 2010.
- [3] S. Arora, S. Khot, A. Kolla, D. Steurer, M. Tulsiani, and N. Vishnoi, *Unique Games on Expanding Constraint Graphs are Easy*, In Proc. 34th Annual ACM Symposium on Theory of Computing, 2008.
- [4] S. Arora, J. Lee, and A. Naor, *Euclidean distortion and the sparsest cut*, In Proc. 37th ACM Symposium on Theory of Computing, pages 553–562, 2005.
- [5] S. Arora, C. Lund, R. Motawani, M. Sudan, and M. Szegedy, *Proof verification and the hardness of approximation problems*, Journal of the ACM, **45**(3) (1998), 501–555.
- [6] S. Arora and S. Safra, *Probabilistic checking of proofs : A new characterization of NP*, Journal of the ACM, **45** (1) (1998), 70–122.
- [7] B. Barak, P. Gopalan, J. Håstad, R. Meka, P. Raghavendra, and D. Steurer, *Making the Long Code Shorter*, In Proc. IEEE Symposium on Foundations of Computer Science, 2012, 370–379.
- [8] B. Barak and D. Steurer, *Sum-of-Squares Proofs and the Quest toward Optimal Algorithms*, In Proc. of the International Congress of Mathematicians, 2014.
- [9] J. Bourgain, *On Lipschitz embeddings of finite metric spaces in Hilbert space*, Israel Journal of Mathematics, **52**(1985), 46–52.

- [10] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar, *On the hardness of approximating multicut and sparsest-cut*, In Proc. 20th IEEE Conference on Computational Complexity, pages 144–153, 2005.
- [11] J. Cheeger and B. Kleiner, *Differentiating maps into  $L^1$  and the geometry of BV functions*, Ann. Math., Second Series, Vol. 171, No. 2, 2010.
- [12] J. Cheeger, B. Kleiner, and A. Naor, *Compression bounds for Lipschitz maps from the Heisenberg group to  $L_1$* , Acta Mathematica, volume 207, issue 2 (2011), 291–373.
- [13] ———, *A  $(\log n)^{\Omega(1)}$  integrality gap for the sparsest cut SDP*, In Proc. 50th IEEE Symposium on Foundations of Computer Science, 2009.
- [14] N. Devanur, S. Khot, R. Saket, and N. Vishnoi, *Integrality gaps for sparsest cut and minimum linear arrangement problems*, In Proc. 38th ACM Symposium on Theory of Computing, 2006.
- [15] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy, *Interactive proofs and the hardness of approximating cliques*, Journal of the ACM **43**(2) (1996), 268–292.
- [16] U. Feige and G. Schechtman, *On the optimality of the random hyperplane rounding technique for max cut*, Random Struct. Algorithms **20**(3) (2002), 403–440.
- [17] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami, *On Agnostic Learning of Parities, Monomials, and Halfspaces*, SIAM J. Comput. **39**(2) (2009), 606–645.
- [18] M. Goemans and D. Williamson, *0.878 approximation algorithms for MAX-CUT and MAX-2SAT*, In Proc. 26th ACM Symposium on Theory of Computing (1994), 422–431.
- [19] J. Hastad, *Some optimal inapproximability results*, Journal of ACM **48** (2001), 798–859.
- [20] G. Kalai and S. Safra, *Threshold Phenomena And Influence with Some Perspective from Mathematics, Computer Science, and Economics*, Center for the Study of Rationality, Issue 398 (2005).
- [21] S. Khot, *On the power of unique 2-prover 1-round games*, In Proc. 34th ACM Symposium on Theory of Computing, 2002.
- [22] ———, *Hardness of approximating the shortest vector problem in lattices*, J. ACM **52** (5) (2005), 789–808.
- [23] ———, *Ruling Out PTAS for Graph Min-Bisection, Dense  $k$ -Subgraph, and Bipartite Clique*, SIAM J. Comput. **36**(4) (2006), 1025–1071.
- [24] ———, *Inapproximability of NP-complete problems, Discrete Fourier Analysis, and Geometry*, In Proc. of the International Congress of Mathematicians, 2010.
- [25] ———, *On the Unique Games Conjecture*, In Proc. IEEE Conference on Computational Complexity, 2010.
- [26] S. Khot, G. Kindler, E. Mossel, and R. O’Donnell, *Optimal inapproximability results for max-cut and other 2-variable CSPs?*, In Proc. 45th IEEE Symposium on Foundations of Computer Science (2004), 146–154.
- [27] S. Khot and D. Moshkovitz, *NP-Hardness of Approximately Solving Linear Equations over Reals*, SIAM J. Comput. **42**(3) (2013), 752–791.
- [28] S. Khot and A. Naor, *Nonembeddability theorems via Fourier analysis*, Mathematische Annalen, 334, number 4 (2006), 821–852. Prelim version: Proc. IEEE Foundations of

- Computer Science (2005), 101–112.
- [29] S. Khot and R. O’Donnell, *SDP gaps and UGC-hardness for Max-Cut-Gain*, Theory of Computing **5** (2009), 83–117. Prelim. version IEEE FOCS ’06.
  - [30] S. Khot and O. Regev, *Vertex cover might be hard to approximate to within  $2 - \epsilon$* , In Proc. 18th IEEE Conference on Computational Complexity, 2003.
  - [31] S. Khot and R. Saket, *Hardness of Coloring 2-Colorable 12-Uniform Hypergraphs with  $2^{(\log n)^{\Omega(1)}}$  Colors*, In Proc. IEEE Symposium on Foundations of Computer Science, 2014.
  - [32] S. Khot and N. Vishnoi, *The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into  $\ell_1$* , In Proc. 46th IEEE Symposium on Foundations of Computer Science, 2005.
  - [33] G. Kindler, A. Rao, R. O’Donnell, and A. Wigderson, *Spherical Cubes: Optimal Foams from Computational Hardness Amplification*, Communications of the ACM (55)-10 (2012), 90–97.
  - [34] T.C. Kwok, L. C. Lau, Y. T. Lee, S. Oveis Gharan, and L. Trevisan, *Improved Cheeger’s Inequality: Analysis of Spectral Partitioning Algorithms through Higher Order Spectral Gap*, In Proc. of 45th ACM Symposium on Theory of Computing, 2013.
  - [35] J. R. Lee and A. Naor,  *$l_p$  metrics on the Heisenberg group and the Goemans-Linial conjecture*, In Proc. 47th IEEE Symposium on Foundations of Computer Science (2006), 99–108.
  - [36] N. Linial, E. London, and Y. Rabinovich, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, **15**(2) (1995), 215–245.
  - [37] E. Mossel, R. O’Donnell, and K. Oleszkiewicz, *Noise stability of functions with low influences: invariance and optimality*, In Proc. 46th IEEE Symposium on Foundations of Computer Science (2005), 21–30.
  - [38] P. Raghavendra, *Optimal algorithms and inapproximability results for every CSP?*, In Proc. ACM Symposium on Theory of Computing (2008), 245–254.
  - [39] P. Raghavendra and D. Steurer, *Graph expansion and the unique games conjecture*, In Proc. 42nd ACM Symposium on Theory of Computing, 2010.
  - [40] L. Trevisan, *On Khot’s Unique Games Conjecture*, Bulletin of the AMS, **49**(1) (2012), 91–111.

Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ 08540, USA  
E-mail: arora@cs.princeton.edu

## **Stanley Osher**

B.S. in Mathematics and Physics, Brooklyn College, 1962

M.S. in Mathematics, New York University, 1964

Ph.D. in Mathematics, New York University, 1966

### **Positions held**

Assistant, Associate Mathematician, Brookhaven National Laboratory, 1966–1968

Assistant Professor, University of California, Berkeley, 1968–1970

Associate, Full Professor, State University of New York, Stony Brook, 1970–1977

Professor, University of California, Los Angeles, 1977–present

Director of Special Projects, Institute of Pure and Applied Mathematics, UCLA, 2001–present

# The work of Stanley Osher

Ron Fedkiw, Jean-Michel Morel, Guillermo Sapiro, Chi-Wang Shu, and Wotao Yin

**Abstract.** In this paper we briefly present some of Stanley Osher's contributions in the areas of high resolution shock capturing methods, level set methods, partial differential equation (PDE) based methods in computer vision and image processing, and optimization. His numerical analysis contributions, including the Engquist-Osher scheme, total variation diminishing (TVD) schemes, entropy conditions, essentially non-oscillatory (ENO) and weighted ENO (WENO) schemes and numerical schemes for Hamilton-Jacobi type equations have revolutionized the field. His level set contributions include new level set calculus, novel numerical techniques, fluids and materials modeling, variational approaches, high codimension motion analysis, geometric optics, and the computation of discontinuous solutions to Hamilton-Jacobi equations. As we will further detail in this paper, the level set method, together with his total variation contributions, have been extremely influential in computer vision, image processing, and computer graphics. On top of that, such new methods have motivated some of the most fundamental studies in the theory of PDEs in recent years, completing the picture of applied mathematics inspiring pure mathematics. On optimization, he introduced Bregman algorithms and applied them to problems in a variety of contexts such as image processing, compressive sensing, signal processing, and machine learning. Finally, we will comment on Osher's entrepreneurship and how he brought his mathematics to industry.

**Mathematics Subject Classification (2010).** Primary 65M06, 65M08, 35L67, 65D18, 65D19, 65K99

**Keywords.** Shock capturing method, level set method, computer vision, image processing, optimization

## 1. Introduction

Shock capturing numerical methods have seen revolutionary developments over the past 40 years. These are methods which deal with the numerical solutions of partial differential equations (PDEs) with discontinuous solutions. Such PDEs include nonlinear hyperbolic systems such as Euler equations of compressible gas dynamics. The problems are difficult because traditional linear numerical methods are either too diffusive, or give unphysical oscillations near the discontinuities which can lead to nonlinear instabilities. The class of high resolution numerical methods overcomes this difficulty to a large extent.

Level set methods have seen tremendously expanded applications in many areas over the past 25 years. This has been made possible by the flexibility of the level set formulation in dealing with dynamic evolutions and topological changes of curves and surfaces, and by the mathematical theory and numerical tools developed in the past 25 years in studying these methods.

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

PDEs based methods in computer vision, image processing, and computer graphics have been actively studied in the past few years. Again, the rapid development of mathematical models, solution tools such as level set methods, and high resolution numerical schemes has made PDEs based methods one of the major tools in computer vision and image processing.

A large number of inverse problems such as those arising in image processing, compressive sensing, signal processing, and machine learning are formulated as optimization problems typically with nonsmooth objective functions, dense data, and solutions that are sparse in a certain sense. The recent advances in algorithms for such problems have led to a large number of successful applications across many areas of science and engineering.

Stanley Osher has made influential contributions to all these fields. A distinctive feature of his research is that he emphasizes both fundamental problems in algorithm design and analysis, and practical considerations for the applications of the algorithms. Osher's work has been highly influential, an indication of this being the citation statistics. For example, according to the Web of Science (ISI) database, which lists papers in selected journals of high impact, the 235 papers of Osher listed there have been collectively cited 28,016 times (as of June 13, 2014, the same below). Among these, 46 papers have been cited over 100 times each. The h-index of Osher is 64. The top five highly cited papers of Osher include the paper of Osher and Sethian [123] on level set methods, cited 4,896 times; the paper of Rudin, Osher and Fatemi [132] on total variation based noise removal, cited 3,151 times; the paper of Sussman, Smereka and Osher [142] on level set approach for two-phase flows, cited 1,568 times; the paper of Shu and Osher [138] on finite difference essentially non-oscillatory (ENO) schemes and total variation diminishing (TVD) Runge-Kutta methods, cited 1,447 times, and the paper of Harten, Engquist, Osher and Chakravarthy [66] on ENO schemes, cited 1,183 times. The more recent paper by Goldstein and Osher [56] on split Bregman algorithms has already been cited 517 times. Thomson Reuters has chosen this paper as a Hot Paper in Computer Science in July 2011. Recently (in 2014) Thompson-Reuters listed the most highly cited authors of significant papers published between 2002-2012. Osher was ranked in the top 1% in both Mathematics and Computer Science.

Before ending this section, we remark that early in his career, Osher did a lot of research on the study of linear stability for finite difference and other numerical methods for hyperbolic, parabolic, and other types of PDEs, as well as well-posedness for those PDEs, especially for initial-boundary value problems. This includes for example the work in [104] which followed up on a seminal paper of Kreiss [80] and used Toeplitz matrices in an elegant way to derive what was later called the GKS condition [61], and the work in [105] where stability conditions for initial-boundary value problems for parabolic equations were obtained, generalizing the work of Varah [146]. In [93], Majda and Osher extended Kreiss' well posedness condition for initial-boundary value problems for hyperbolic equations to those with uniformly characteristic boundaries. In [92], Majda and Osher analyzed the reflection of singularities at the boundary for nongrazing reflection for hyperbolic equations. In [94], Majda and Osher showed how error propagates globally within the domain of dependence for numerical approximations to coupled hyperbolic systems. The paper [91] by Majda, McDonough and Osher was the first to recommend the use of smooth cutoff functions on the frequency domain for spectral methods to confine errors to local regions near propagating discontinuities and for stability. Sharp estimates on the region of propagation were obtained. These cutoffs are now widely used in the literature and the paper is still frequently cited, 71 times total (Web of Science), including many in recent years. Finally, in [35], Engquist, Osher and Zhong obtained wavelet based fast algorithms for linear hyperbolic and parabolic

equations, and in [32, 41, 42], Engquist, Fatemi and Osher considered numerical methods for high frequency asymptotics for geometric optics. These might be considered nonlinear, since the eikonal equation is. We shall not review in detail these early works of Osher on linear methods in the remaining part of this paper.

## 2. High resolution shock capturing methods

Shock capturing methods refer to a class of numerical methods for solving problems containing discontinuities (shocks, contact discontinuities or other discontinuities), which can automatically “capture” these discontinuities without special effort to track them. A typical situation would be the solution of hyperbolic conservation laws, such as the Euler equations for inviscid fluid flow dynamics. Almost all shock capturing schemes, including those developed by Osher and his collaborators, are of the conservation form. However, there are certain situations where a relaxation on the strict conservation would be beneficial and would not hurt the convergence to weak solutions under suitable additional assumptions. The work of Osher and Chakravarthy [111] on the “weak conservation form” for schemes on general curvilinear coordinates, and the work of Fedkiw et al. on “ghost fluid” method [45], which treats the fluid interface in a non-conservative fashion, are such examples.

**2.1. First order monotone schemes.** In the late 70s and early 80s, designing good first order monotone schemes for solving scalar conservation laws which give monotone shock transitions and can be proven to converge to the physically relevant weak solutions (e.g. Crandall and Majda [30]), with suitable generalization to systems, was an active research area. The Godunov scheme is a scheme with the least numerical dissipation among first order monotone schemes, however it is costly to evaluate for complex flux functions, and its flux is only Lipschitz continuous but not smoother. The Lax-Friedrichs scheme is easy to evaluate and very smooth but is excessively dissipative.

In [33] and [34], Engquist and Osher designed monotone schemes for the transonic potential equations and for general scalar conservation laws, which are relatively easy to evaluate, are  $C^1$  smooth, and have a small dissipation almost comparable with Godunov schemes. These Engquist-Osher schemes soon became very popular, especially for implicit type methods and steady state calculations, for which the extra smoothness of the numerical flux helped a lot. Similar schemes for Hamilton-Jacobi equations were given by Osher and Sethian [123].

Later, Osher [106] and Osher and Solomon [125] generalized these schemes to systems of conservation laws, obtaining what was later referred to as the Osher scheme in the literature. The Osher scheme for systems has a closed form formula (for Euler equations of gas dynamics and many other systems), hence no iterations are needed, unlike the Godunov scheme. It is smoother ( $C^1$ ) than the Godunov scheme and also has smaller dissipation than the simpler Lax-Friedrichs scheme. Applications of the Osher scheme to the Euler equations can be found in Chakravarthy and Osher [17].

In [121], Osher and Sanders designed a conservative procedure to handle locally varying time and space grids for first order monotone schemes, and proved convergence to entropy solutions for such schemes. These ideas have been used later by Berger and Colella on their adaptive methods, e.g. [7].

**2.2. High resolution TVD schemes.** First order monotone schemes are certainly nice in their stability and convergence to the correct entropy solutions, however they are too diffusive for most applications. One would need to use many grid points to get a reasonable resolution, which seriously restricts their usefulness for multidimensional simulations.

In the 70s and early and mid 80s, the so-called “high resolution” schemes, i.e. those schemes which are at least second order accurate and are stable when shocks appear, were developed. These started with the earlier work of, e.g., the flux-corrected transport (FCT) methods of Boris and Book [8], and the monotonic upstream-centered scheme for conservation laws (MUSCL) of van Leer [145], and moved to Harten’s TVD schemes [65]. Osher and his collaborators did extensive research on TVD schemes, and contributed significantly towards the analysis of such methods, during this period. These include the schemes developed and analyzed in [108, 109, 112], and the very high order (measured by truncation errors in smooth, monotone regions) TVD schemes in [113].

**2.3. Entropy conditions.** The entropy condition is an important feature for conservation laws. Because weak solutions are not unique, entropy conditions are needed to single out a unique, physically relevant solution. Osher and his collaborators did extensive research on designing and analyzing entropy condition satisfying numerical methods for conservation laws.

In [95], Majda and Osher proved that the traditional second order Lax-Wendroff scheme, although linearly stable, is not  $L^2$  stable when solving nonlinear conservation laws with discontinuous solutions. They then provided a recipe of adding artificial viscosities, such that the scheme maintained second order accuracy yet could be proven convergent to the entropy solution. This scheme is however oscillatory, hence not very practical in applications.

In [108], Osher provided a general framework to study systematically entropy conditions for numerical schemes. This was followed by the work of Osher and Chakravarthy [112] in the study of high resolution schemes and entropy conditions, the work of Osher [109] on generalized MUSCL schemes, the work of Osher and Tadmor [126] on entropy condition and convergence of high resolution schemes, and the work of Brenier and Osher [10] on entropy condition satisfying “maxmod” second order schemes. Entropy condition satisfying approximations for the full potential equation of transonic flow were given in [117].

**2.4. ENO and WENO schemes.** In the mid 80s it was realized that TVD schemes, despite their excellent stability and high resolution properties, have serious deficiency in that they degenerate to first order at *smooth* extrema of the solution [112]. Thus, even though TVD schemes can be designed to any order of accuracy in smooth monotone regions, see for example the schemes up to 13th order accurate in [113], practical TVD schemes are referred to as second order schemes since the global  $L^1$  errors of any TVD scheme can only be second order, even for smooth but non-monotone solutions.

In [67], Harten and Osher relaxed the TVD restriction, and replaced it by a uniformly non-oscillatory (UNO) restriction, in that the total number of numerical extrema does not increase and their amplitudes could be allowed to increase slightly. The UNO scheme in [67] is uniformly second order accurate including at smooth extrema. However, it was soon realized that the UNO restriction was still too strong and excluded schemes of higher than second order. Thus, the concept of ENO, or essentially non-oscillatory, schemes was first given by Harten, Engquist, Osher and Chakravarthy [66] in 1987. The clever idea is that of an adaptive stencil, which is chosen based on the local smoothness of the solution, measured

by the Newton divided differences of the numerical solution. Such schemes allow both the number of numerical extrema and their amplitudes to increase, however such additional oscillations are controlled on the level of truncation errors even if the solution is not smooth. ENO schemes have been extremely successful in applications, because they are simple in concept, allow arbitrary orders of accuracy, and generate sharp, monotone (to the eye) shock transitions together with high order accuracy in smooth regions of the solution including at the extrema.

The original ENO schemes in [66] are in the cell averaged form, namely they are finite volume schemes approximating an integrated version of the PDE. Finite volume schemes have the advantage of easy handling of non-uniform meshes and general geometry in multi-space dimensions, however they are extremely costly in multi-space dimensions, when the order of accuracy is higher than two. Later, Shu and Osher [138, 139] developed conservative finite difference based ENO schemes using point values of the numerical solution, which are more efficient in multi-dimensions.

Also in [138], a class of nonlinearly stable high order Runge-Kutta time discretization methods is developed. Termed TVD time discretizations, these Runge-Kutta methods have become very popular and have been used in many schemes. See, e.g. [57] for a review of such methods.

Analysis of ENO schemes was given in Harten et al. [68]. Applications of ENO schemes to two and three dimensional compressible flows, including turbulence and shear flow calculations, were given in Shu et al. [140]. Triangle based second order non-oscillatory schemes were given in Durlafsky, Engquist and Osher [31]. Non-oscillatory self-similar maximum principle satisfying high order shock capturing schemes were given in Liu and Osher [88]. Efficient characteristic projection in upwind difference schemes was given in Fedkiw, Merriman and Osher [48]. Convex ENO schemes without using field-by-field projection were given in Liu and Osher [89]. Chemically reactive flows were simulated in Ton et al. [143] and in Fedkiw, Merriman and Osher [47].

The popularity of ENO schemes is demonstrated by the citation statistics: among Osher's five most highly cited papers mentioned in the introduction, two of them are about ENO schemes, i.e. [138] and [66]. The top cited paper of Osher [123], is on level set methods but also uses second order ENO schemes for the numerical solutions and is where the construction of ENO schemes for general Hamilton-Jacobi equations began.

An improvement of ENO scheme is the WENO (weighted ENO) scheme, which was first developed by Liu, Osher and Chan [90]. WENO improves upon ENO in robustness, better smoothness of fluxes, better steady state convergence, better provable convergence properties, and more efficiency.

**2.5. Hamilton-Jacobi equations.** We will now move to the description of Osher's work in designing schemes for solving Hamilton-Jacobi equations. Further discussions on this topic will also be given in the next section on level set methods.

Viscosity solutions for Hamilton-Jacobi equations were first proposed by Crandall and Lions [28] in order to pick out the physically relevant solution. In addition, monotone first order accurate numerical methods were first proven to converge by Crandall and Lions in [29]. In [107], Osher gave explicit formulas for solutions to the Riemann problems for non-convex conservation laws and Hamilton-Jacobi equations. See also the multidimensional Riemann solver of Bardi and Osher [4]. These are important for numerical schemes such as Godunov schemes using such Riemann solvers as building blocks.

In [123], Osher and Sethian, in the context of discussing level set methods, provided a first order monotone scheme (an adaptation of the Engquist-Osher scheme [34]) and a second order ENO scheme based on the framework of [138] and [139]. In [124], Osher and Shu developed high order ENO schemes for solving Hamilton-Jacobi equations, using various building blocks including the Lax-Friedrichs flux, the local Lax-Friedrichs flux, and the Roe flux with an entropy fix. In [81], Lafon and Osher developed high order two dimensional triangle based non-oscillatory schemes for solving Hamilton-Jacobi equations.

More recently, Osher and his collaborators have studied the fast sweeping methods for efficiently solving steady state Hamilton-Jacobi equations. In [144], Tsai et al. developed fast sweeping method for Godunov-type first order schemes for a class of Hamilton-Jacobi equations, which converges very fast for computing steady state solutions. In [74], Kao, Osher and Qian developed fast sweeping method for Lax-Friedrichs type schemes. In [76], Kao, Osher and Tsai developed fast sweeping method based on Legendre transform of the numerical Hamiltonian using an explicit formula. Finally, in [75], Kao, Osher and Qian generalized the previous techniques to triangulated meshes.

**2.6. Additional topics.** Even though it does not exactly fit the title of this section, the work of Lagnado and Osher [82, 83] is worth mentioning. These papers concern solving an inverse problem to compute the volatility in the European options Black-Scholes model, and they were the first to use PDE techniques to solve this inverse problem, via gradient descent and Tychonoff regularization, allowing the volatility, which is a coefficient in a parabolic equation, to be a function of the independent variables, stock price and time. These papers have attracted a lot of attention after their publication.

Also worth mentioning is the work of Fatemi, Jerome and Osher [43] on using ENO schemes to solve the hydrodynamic models of semiconductor device simulations. This was the first work of using high order shock capturing methods in semiconductor device simulations, and has led to many further developments later in the literature.

### 3. Level set methods

Osher's most cited paper is [123], which introduced the level set method for dynamic implicit surfaces. The key idea was the Hamilton-Jacobi approach to numerical solutions of a time dependent equation for a moving implicit surface. In a series of papers that followed [123], Osher and coworkers introduced a level set calculus for the practical treatment of discretized implicit surfaces defined by time evolving partial differential equations. We summarize some of the main points below, but refer the interested reader to the review article of Osher and Fedkiw [115] and the references within. In addition, we refer the reader to the book [116] by Osher and Fedkiw.

Suppose that the surface is implicitly defined as the zero isocontour of a function  $\phi$ . Then the local sign of  $\phi$  can be used to define the inside and outside regions of the domain. Implicit functions make simple Boolean operations easy to apply. The gradient of the implicit function is perpendicular to the isocontours of  $\phi$  pointing in the direction of increasing  $\phi$ , and therefore can be used to define the normal to the interface. Similarly the mean curvature of the interface is defined as the divergence of the normal. One can readily define the characteristic function, define the Heaviside functions, compute volume integrals, define the Dirac delta function as the directional derivative of the Heaviside function in the normal direction,

compute surface integrals, etc.

A key factor for the success of level set methods is the use of high order high resolution type schemes reviewed in section 2, for the conservation laws and Hamilton-Jacobi equations. These include in particular the ENO and WENO schemes. Stanley Osher has been *the* leader in this area, and the level set method would not be what it is today without the creative and vigorous approach to the numerical analysis aspect of the method that he has shown. Implicit surface models are not new by any means, but robust and rigorous numerical methods that make them extremely powerful are new and are due primarily to Osher's ideas and leadership.

Even with these high order accurate approaches to solving the Hamilton-Jacobi equations, one can obtain surprisingly inaccurate results when the level set function solution becomes too steep or too flat, i.e. discontinuous or poorly conditioned. In [25], Chopp considered an application where certain regions of the flow had level sets piling up on each other increasing the local gradient, and other regions of the flow had level sets that separated from each other flattening out  $\phi$ . In order to reduce the numerical errors caused by both the steeping and flattening effects, [25] introduced the notion that one should reinitialize the level set function periodically throughout the calculation. In [131], Rouy and Tourin proposed a numerical method for the shape from shading problem that was later generalized into the modern day reinitialization equation of Sussman, Smereka and Osher [142]. Unfortunately, this straightforward reinitialization routine can be slow, especially if it needs to be done every time step, although [142] noted that just a few time iterations are usually needed. In order to obtain reasonable run times, [25] restricted the calculations of the interface motion and the reinitialization to a small band of points near the  $\phi = 0$  isocontour. This idea of computing solutions to Hamilton-Jacobi equations local to the interface has been studied further in the work of Adalsteinsson and Sethian [1] and Peng et al. [127].

**3.1. Fluids and materials.** Chronologically, the first attempt to use the level set method for flows involving external physics was in the area of two phase inviscid compressible flow. Mulder, Osher and Sethian [100] appended the level set equation to the standard equations for one phase compressible flow. The level set was advected using the velocity of the compressible flow field so that the zero level set of  $\phi$  corresponds to particle velocities and can be used to track an interface separating two different compressible fluids. Later, Karni [77] pointed out that such method suffered from spurious oscillations at the interface and proposed a non-conservative fix. A more robust fix was later proposed by Fedkiw et al. [45] by creating a set of fictitious ghost cells on each side of the interface, and populating these ghost cells with a specially chosen ghost fluid that implicitly captures the Rankine-Hugoniot jump conditions across the interface. This method was referred to as the ghost fluid method. Later extensions included the treatment of shocks, detonations and deflagrations [46], interfaces separating compressible flows from incompressible flows [15], and interfaces separating Eulerian discretizations of fluids from Lagrangian discretizations of solids [44]. More recently, both [103] and [54] have proposed fully conservative versions of this ghost fluid method. Moreover, the method proposed in [103] is easy to implement in multiple spatial dimensions, works for contacts, shocks, detonations and deflagrations, and has been shown to prevent the one grid cell per time step spurious wave instabilities (identified by [26]) that occur in stiff under-resolved detonation waves.

The earliest real success in the coupling of the level set method to problems involving external physics came in computing two-phase incompressible flow, in particular see Suss-

man, Smereka and Osher [142] and Chang et al. [20]. The Navier-Stokes equations were used to model the fluids on both sides of the interface. Generally, the fluids will have different densities and viscosities and the presence of surface tension forces causes the pressure to be discontinuous across the interface as well. Although these early works smeared out these discontinuous quantities across the interface, this was later remedied by Kang, Fedkiw and Liu [73] using the methods developed by Liu, Fedkiw and Kang [87]. Later, Nguyen, Fedkiw and Kang [102] extended these techniques to treat low speed flames.

A level set regularization procedure was proposed in Harabetian and Osher [63] for ill-posed problems such as vortex motion in incompressible flows. This regularization, coupled with non-oscillatory numerical methods for the resulting level set equations, provides a regularization which is topological and is automatically accomplished through the use of numerical schemes whose viscosity shrinks to zero with the grid size. There is no need for explicit filtering, even when singularities appear in the solution. The method also has the advantage of automatically allowing topological changes such as merging of surfaces.

An application of this procedure for incompressible vortex motion was given in Harabetian, Osher and Shu [64]. An Eulerian, fixed grid, approach to solve the motion of an incompressible fluid, in two and three dimensions, in which the vorticity is concentrated on a lower dimensional set, is provided. The numerical variables for the level sets are actually smooth, thus allowing for accurate numerical simulations. Numerical examples including two and three dimensional vortex sheets, two dimensional vortex dipole sheets and point vortices, are given.

Level set type analysis was also used to obtain rigorous results identifying the Wulff minimizing shape and the evolution of growing crystals moving with normal velocity defined as a given positive function of the normal direction, thus verifying a conjecture of Gross. Moreover it was also shown that the Wulff energy decreases monotonically under such an evolution to its minimum [119]. A spinoff came in [128] where it was proven that any two dimensional Wulff shape can be interpreted as the solution a corresponding Riemann problem for a scalar conservation law – jumps in the direction of the normal correspond to contact discontinuities, smoothly varying thin flat faces correspond to rarefaction curves and planar facets correspond to constant states. The work in [119] also motivated the derivation of a new class of isoperimetric inequalities for convex plane curves [59].

Molecular beam epitaxy (MBE) is a method for growing extremely thin films of material. A new continuum model for the epitaxial growth of thin films has been developed. This new island dynamics model has been designed to capture the larger length scale features. The key idea involves the level set based motion of islands of various integer levels – see for example [21, 62, 99].

**3.2. Other applications.** In [154] a variational level set approach was developed. Key ideas were the use of a single level set function for each phase, the gradient projection method of [132] to prevent overlap and / or vacuum, and the liberal use of the level set calculus as described earlier. See also [155], [122] and [18]. Typically level set methods are used to model codimension one objects, e.g. curves in  $R^2$  or surfaces in  $R^3$ . In [11], this technology was extended to treat codimension two objects, e.g. curves in  $R^3$ , using the intersection of the zero level sets of two functions. See also [24]. In [114] a level set based approach for ray tracing and for the construction of wavefronts in geometric optics was introduced. The approach automatically handles the multivalued solutions that appear and automatically resolves the wavefronts. The key idea, first introduced in [36] in a “segment projection”

(rather than a level set) approach, is to use the linear Liouville equation in twice as many independent variables and solve in this higher dimensional space via the idea introduced in [11]. See also [141] and [134]. Level set methods have been applied to a variety of other problems as well. They have been used to compute solutions to Stefan problems to study crystal growth [22, 78], to simulate water and fire for computer graphics applications [37, 50, 101], and to reconstruct three dimensional models from arbitrary unorganized data points [156, 157].

#### 4. Image processing and computer vision

The use of partial differential equations (PDE's) and curvature driven flows in image processing and computer vision has become an active research topic in the past two decades, thanks in part to the pioneering contributions of Stanley Osher on level set and total variation (TV) methods. The basic idea is to deform a given curve, surface, or image with a PDE, and obtain the desired result as the solution of this PDE. Sometimes, as in the case of color images, a system of coupled PDEs is used. The art behind this technique is in the design, analysis, and numerical implementation of these PDEs. The attributes of PDEs in image processing are discussed for example in [16, 136]. In [2] the authors prove that a few basic image processing principles naturally lead to PDEs.

When considering PDEs for image processing and numerical implementation, we are dealing with derivatives of non-smooth signals, and the right framework must be defined, connecting this with Osher's contributions in shock capturing schemes and numerical analysis in general. As introduced in [2, 3], the theory of viscosity solutions provides a framework for rigorously employing a partial differential formalism, in spite of the fact that the image may not be smooth enough to give a classical definition to the derivatives involved in the PDE. These works also showed with a very elegant axiomatic approach the importance of PDEs in image processing. This is also the framework that brings rigorousness to the level set methods developed by Osher and collaborators.

Ideas on the use of PDEs in image processing go back at least to Gabor [51] and to Jain [72]. The field took off thanks to the independent works of Koenderink [79] and Witkin [148]. These researchers rigorously introduced the notion of *scale-space*, that is, the representation of images simultaneously at multiple scales. In their work, the multi-scale image representation is obtained by Gaussian filtering. This is equivalent to deforming the original image via the classical heat equation, obtaining in this way an isotropic diffusion flow. In the late 80s, Hummel [69] noted that the heat flow is not the only parabolic PDE that can be used to create a scale-space, and indeed argued that an evolution equation which satisfies the maximum principle will define a scale-space as well. The maximum principle appears to be a natural mathematical translation of *causality*. Koenderink once again made a major contribution into the PDEs arena when he suggested to add a thresholding operation to the process of Gaussian filtering. As later suggested by Merriman, Bence and Osher [97, 98] and by Ruuth, Merriman and Osher [133], and proved by a number of groups [5, 39, 70, 71], this leads to a curvature motion geometric PDE, one of the most famous among geometric PDEs.

The approach in Merriman, Bence and Osher (MBO) [97] leads to a series of mathematical work on threshold dynamics type approximation schemes for propagating fronts, see for example Ishii, Pires and Souganidis [71]. The MBO algorithm was also extended to net-

works of interfaces in the original papers. Recently Esedoglu and Otto [38] extended it to arbitrary surface tensions by making it into a variational problem. Also, the algorithm was found to be particularly useful in segmentation of imaging data on graphs, e.g., in Merkurjev, Sunu and Bertozzi [96].

In [135], Ruuth et al. extended this approach to diffusion generated motion of curves in  $R^3$ . Solving a vector heat equation and thresholding lead to moving the curve in the direction of the normal with velocity equal to its curvature.

Perona and Malik's work [129] on anisotropic diffusion, together with the work by Rudin, Osher and Fatemi on total variation [132] and by Osher and Rudin on shock filters [120], have been among the most influential papers in the area, explicitly showing the importance of understanding non-linear PDEs theory to deal with images. They proposed to replace the linear Gaussian smoothing, equivalent to isotropic diffusion via the heat flow, by a selective non-linear diffusion that preserves edges. Their work opened a number of theoretical and practical questions that continue to occupy the PDEs image processing community, see, e.g., [3, 130].

The TV model [132] is basically the predecessor of compressed sensing, and has been one of the most influential papers in the modern era of inverse problems. The TV model is frequently used as a regularization term for inverse problems.

Many of the PDEs used in image processing and computer vision are based on moving curves and surfaces with curvature based velocities. In this area, the level set numerical method developed by Osher and Sethian [123], is very influential, and is the standard in critical applications like medical image segmentation, and implemented in the most popular packages in the area, e.g., ITK.

It should be noted again that a number of the above approaches rely quite heavily on a large number of mathematical advances in differential geometry for curve evolution [58] and in viscosity solutions theory for curvature motion (see e.g., [23, 40].) It is fascinating that Osher's work in this subject not only leads to state-of-the-art applications but has inspired some of the top mathematical minds of this century to investigate the underlying theory behind such beautiful methods.

One of the basic ideas behind this area is: the fact that images are represented in digital computers in the form of discrete objects should not limit the tools to those of discrete mathematics. It is "legal" to use tools from differential equations and differential geometry, and then deal with the computer implementation of the algorithms from the point of view of numerical analysis. Here is where Stanley Osher got his call.

The frameworks of PDEs and geometry driven diffusion have been applied to many problems in image processing and computer vision, since the seminal works mentioned above. Examples include continuous mathematical morphology, invariant shape analysis, shape from shading, segmentation, tracking, object detection, optical flow, stereo, image denoising, image sharpening, contrast enhancement, and image quantization. Today PDEs are a standard tool in these areas, something that was almost unheard of before Osher's fundamental TV, shock capturing, and level set papers. Osher's ideas have led to renewal of automatic image segmentation methods, a subject of uttermost importance in medical image analysis. In [19], multiple phases and their boundaries, represented via the level set method, evolve and interact in time, to minimize a bulk-surface energy. Combining several level set functions together, triple junctions were also represented and evolved in time. Based on these ideas, Chan and Vese presented a multi-phase level set model for image segmentation. Triple junctions and complex topologies are segmented using more than one level set function.

## 5. Computer graphics and the movies

As will be elaborated on in section 7 on industrial entrepreneurship, Stanley Osher has never been content simply sitting in his office quietly writing papers. He has been at the heart of the UCLA style of mathematics since that term was coined and has always believed that good schemes are used by people, many people, and therefore has been more of a crusader than a mathematician. This approach to the work is what fostered its proliferation in many application areas, and here we briefly discuss computer graphics and the film industry.

Over the last fifteen years, the entire field of computer graphics has evolved from a novice community of hackers in regards to fluid dynamics into a fairly savvy group of researchers that regularly collaborate with numerical analysts on topics of interest to the *Journal of Computational Physics*, *International Journal for Numerical Methods in Engineering*, *Computer Methods in Applied Mechanics and Engineering*, etc – and some of them even publish papers in these journals. The main reason for this is that people in the industry have found the UCLA style of mathematics in regards to both fluid dynamics and level set methods quite useful. In fact 15 years ago the community was quite convinced that one only needed a triangle to solve every problem. This started to change dramatically when two papers, [49] and [50], were published in 2001 in the largest computer graphics conference, SIGGRAPH. These papers outline an approach for computational fluid dynamics for smoke and interface treatment for liquids that changed how the computer graphics community viewed the computational physics community. Since then there have been too many computer graphics papers to count on fluids, interfaces and level sets. But it is worth noting that one of Osher's students received an Academy Award for using level sets and related technology to create water in many movies, including the giant whirlpool maelstrom in *Pirates of the Caribbean*. To date, Industrial Light & Magic has used Stanley Osher's technologies for 30 to 40 films and they are also being used by every single major film company including Pixar, Disney, Weta, Dreamworks, etc. Who would have thought that the Tar Monster in *Scooby Doo* would make its way from cartoons to the big screen via the level set method?

Thanks to Osher's work, drawing on surfaces has become much simpler and the applications in computer graphics are striving. Once again, the precise equations are given in [6].

## 6. Optimization and Bregman methods

Osher made significant contributions in optimization through introducing Bregman algorithms [56, 110, 118, 153] and popularizing them in a variety of contexts such as image processing (e.g., [13, 14, 55]), compressive sensing (e.g., [12, 153]), signal processing (e.g., [27, 149]), and machine learning (e.g., [52, 137]). His effort gives rise to both faster algorithms and higher quality solutions for a broad spectrum of optimization problems, leading to a large number of successful applications in many areas.

Osher's Bregman algorithms are based on the so-called Bregman divergence [9]. Given a convex function  $r$ , the Bregman divergence between two points  $x$  and  $y$  is  $D(x, y) = r(x) - (r(y) + \langle \nabla r(y), x - y \rangle)$ . In  $x$ , it is the difference between  $r$  and its linearization at  $y$ . In case  $r$  is non-differentiable,  $\nabla r(y)$  is replaced by a subgradient  $p \in \partial r(y)$ . Although it is not a distance,  $D(x, y)$  is larger if  $x$  and  $y$  are further apart on the same straight line. So, it is also called the Bregman distance.

Osher's Bregman algorithms have important applications in *inverse problems*, which traditionally arise from the study of physical phenomena, formulated via forward problems, and solved to reconstruct the information or causes from the observations, or sometimes to obtain desired effects. Most inverse problems are *ill-posed* in the sense that the observations are not sufficient to uniquely determine the solution, or small noise in the observations leads to large errors in the solution, or both. To address ill-posedness, a common approach is to minimize a regularization function  $r(x)$  so that the solution tends to have the desired property. To have tractable computation, *convex* regularization functions are typically used. For example, given an underdetermined linear system  $Ax = b$ , minimizing  $\|x\|_1$  tends to give a sparse, and sometimes the sparsest, solution to the system. In [110], Osher et al. propose to replace  $r(x)$  by the  $r$ -induced Bregman divergence and iteratively minimize  $\mu D(x, x^{k-1})$ , where  $\mu$  is a scalar and  $x^{k-1}$  is the previous solution.

This simple change has significant benefits. (i) When the observation is contaminated by random noise, this approach can yield a better solution than directly minimizing  $r(x)$ . (ii) When the observation is noise free, the iteration  $x^k \leftarrow \min_x \mu D(x, x^{k-1}) + \frac{1}{2} \|Ax - b\|_2^2$  converges quickly to an  $r$ -minimum solution to  $Ax = b$ . In particular, when  $r$  is convex and piece-wise linear (e.g.,  $\ell_1$  norm), the iteration converges finitely; in practice, it often takes as few as 5–10 iterations.

Since random noise exists in numerous inverse problems, benefit (i) is widely appreciated. For example, in MR imaging [86], it improves image quality and preserves more important features during image reconstruction. In compressive sensing, given the same amount of noise, its solution is sparser and cleaner than the  $\ell_1$  solution [110, 153]. In sparse logistic regression for feature selection, restricted to the same amount of *loss*, it uses fewer features (fewer typically means more accurate feature selection).

Although Osher et al. later find that their Bregman algorithm is equivalent to the method of multipliers (see §3.4 of [153]), also known as the augmented Lagrangian method, they discover an interesting property called *error forgetting* [152], when the algorithm is applied to sparse optimization. In a nutshell, error forgetting means that the error at each iteration, as long as below a certain threshold, does not accumulate and can even cancel each other iteratively until the solution reaches the true solution at around the machine precision. This explains why the Bregman algorithm is especially efficient on sparse optimization problems. Despite the equivalence to an existing algorithm, benefit (ii) and error forgetting of the algorithm open the way to extremely successful and wide applications of Bregman and split Bregman algorithms.

Split Bregman [56], as the name indicates, is the Bregman algorithm applied to the problem in which the objective function has a split form, namely,  $r(x) = r_1(x_1) + r_2(x_2)$ , where  $x = [x_1; x_2]$ . At each iteration, one solves two subproblems, subproblem  $i$  minimizing  $r_i(x_i)$  plus a quadratic function of  $x_i$ ,  $i = 1, 2$ . Unlike alternating minimization, this approach can handle linear constraints on  $x_1$ ,  $x_2$ , or both of them. This is a real power! Through simple transforms and additional variables, it provides simple solutions to problems with awkward combinations of objective functions and/or constraints, for instance, the joint constraints of matrix  $X$  being symmetric positive semi-definite and having nonnegative entries, the sum of  $\ell_1$  and nuclear-norm objectives, the composite objective of  $f(g(x))$  where one of  $f$  and  $g$  is nonsmooth, etc. Without splitting, it is numerically challenging to handle them, but the split Bregman subproblems are often easy, and even have closed form solutions; see, for example, [53, 60, 84, 85, 147, 150, 151]. Omitting the details on parameter selection, one can split the awkward combinations of objectives and constraints and develop

a split Bregman based algorithm in just a few hours for many optimization problems, yet also find it nearly state-of-the-art in terms of both speed and solution accuracy. Although the same algorithm dates back to work in PDE computation during the 1950's, which was later developed into an optimization algorithm in 1980's, the algorithm lost favor around 1990–2005 in nonlinear optimization. Osher's split Bregman algorithm stimulates its revival and makes unique and important contributions to solving many modern sparse optimization problems that have recently arisen *across many areas*. As mentioned in the introduction, less than five years after the publication of Osher's split Bregman paper [56] in 2009, the paper has been cited over 500 times according to the Web of Science database.

## 7. Industrial entrepreneurship: Pushing applied math to the limit

One distinctive feature of Osher's research style is that he pays a lot of attention to convert state of the art research results immediately to high tech applications, by directly involving in founding companies.

The first such effort is the company Cognitech, inc., cofounded with Rudin. This company helped introduce PDE techniques, especially TV based restoration, of image processing, to practical forensic image processing, using TV methods to clean up video images. The foundation of these techniques is the famous paper of Rudin, Osher and Fatemi [132]. The sensational success of Cognitech during the 1992 Los Angeles riot has been written up several times in popular press. It is now a very successful company, owned by Rudin, specializing in law enforcement surveillance videos, etc.

The second such effort is the company Level Set Systems, which is in business since 1999. It has developed the state-of-the-art point cloud compression package. Many industries ranging from geospatial data collection/storage to 3D medical imagery can benefit from this compression algorithm. Other projects such as hyperspectral imaging have useful applications in mining to detect desired materials, as well as security applications detecting chemical or biological agents. Graphics and exapixel camera processing are among various projects Level Set Systems has worked on. One major foundation of all these developments is the research in level set methods.

The third such effort is the company Luminescent Technologies. This company, cofounded with Eli Yablonovitch and others, is the first commercial company that introduced "Inverse Lithography Technology (ILT)" to the field of Electronic Design Automation to solve the traditional Optical Proximity Correction (OPC) problems, using latest research results in level set methods and efficient algorithms for Hamilton-Jacobi equations. With this technology, non-intuitive mask patterns, which are beyond the capabilities of traditional approaches, are generated with superior performance in printing patterns of only a few dozens of nanometers on chips. It is now widely used by most EDA companies in one way or another. It is considered as one of the most significant advancement in the EDA industry.

## 8. Concluding remarks

Stanley Osher exemplifies mathematics and provides the perfect picture of an applied mathematician. His work has all the components expected from leaders in the area, from solving

real problems to motivating deep mathematical studies. While most researchers hope to make it big once in their career, Stanley Osher made it multiple times. He has co-authored a handful of breakthrough and area opening papers and has influenced industry from movie post-production to chip manufacturers.

## References

- [1] Adalsteinsson, D. and Sethian, J., *A fast level set method for propagating interfaces*, Journal of Computational Physics **118** (1995), 269–277.
- [2] Alvarez, L., Guichard, F., Lions, P. L., and Morel, J. M., *Axioms and fundamental equations of image processing*, Archive for Rational Mechanics and Analysis **123** (1993), 199–257.
- [3] Alvarez, L., Lions, P. L., and Morel, J. M., *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM Journal on Numerical Analysis **29** (1992), 845–866.
- [4] Bardi, M. and Osher, S., *The nonconvex multi-dimensional Riemann problem for Hamilton-Jacobi equations*, SIAM Journal on Numerical Analysis **22** (1991), 344–351.
- [5] Barles, G. and Georgelin, C., *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM Journal on Numerical Analysis **32** (1995), 484–500.
- [6] Bertalmio, M., Cheng, L. T., Osher, S., and Sapiro, G., *Variational problems and partial differential equations on implicit surfaces*, Journal of Computational Physics **174** (2001), 759–780.
- [7] Berger, M. and Colella, A., *Local adaptive mesh refinement for shock hydrodynamics*, Journal of Computational Physics **82** (1989), 64–84.
- [8] Boris, J. P. and Book, D. L., *Flux corrected transport I, SHASTA, a fluid transport algorithm that works*, Journal of Computational Physics **11** (1973), 38–69.
- [9] Bregman, L. M., *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics **7** (1967), 200–217.
- [10] Brenier, Y. and Osher, S., *The discrete one-sided Lipschitz condition for convex scalar conservation laws*, SIAM Journal on Numerical Analysis **25** (1988), 8–23.
- [11] Burchard, P., Cheng, L.-T., Merriman, B., and Osher, S., *Motion of curves in three spatial dimensions using a level set approach*, Journal of Computational Physics **165** (2001), 463–502.
- [12] Cai, J. F., Osher, S., and Shen, Z., *Linearized Bregman iterations for compressed sensing*, Mathematics of Computation **78** (2009), 1515–1536.
- [13] ———, *Linearized Bregman iterations for frame-based image deblurring*, SIAM Journal on Imaging Sciences **2** (2009), 226–252.
- [14] ———, *Split Bregman methods and frame based image restoration*, Multiscale Modeling and Simulation **8** (2009), 337–369.

- [15] Caiden, R., Fedkiw, R., and Anderson, C., *A numerical method for two-phase flow consisting of separate compressible and incompressible regions*, Journal of Computational Physics **166** (2001), 1–27.
- [16] Caselles, V., Morel, J. M., Sapiro, G., and Tannenbaum, A., Editors, *Special Issue on Partial Differential Equations and Geometry-Driven Diffusion in Image Processing and Analysis*, IEEE Transaction on Image Processing **7**, March 1998.
- [17] Chakravarthy, S. and Osher, S., *Numerical experiments with the Osher upwind scheme for the Euler equations*, AIAA Journal **21** (1983), 1241–1248.
- [18] Chan, T. and Vese, L., *An active contour model without edges*, in Scale-Space Theories in Computer Vision, M. Neilsen, P. Johansen, O.F. Olson and J. Weickert (eds.), Lecture Notes in Computer Science, volume 1682, Springer-Verlag, Berlin/New York, 1999, 141–151.
- [19] Chan, T. and Vese, L., *A level set algorithm for minimizing the Mumford-Shah functional in image processing*, Proceedings of the IEEE Workshop on Variational and Level Set Methods in Computer Vision, IEEE, 2001, 161–168.
- [20] Chang, Y., Hou, T., Merriman, B., and Osher, S., *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows*, Journal of Computational Physics **124** (1996), 449–464.
- [21] Chen, S., Merriman, B., Kang, M., Caffisch, R., Ratsch, C., Guyre, M., Fedkiw, R., Anderson, C., and Osher, S., *A level set method for thin film epitaxial growth*, Journal of Computational Physics **167** (2001), 475–500.
- [22] Chen, S., Merriman, B., Osher, S., and Smereka, P., *A simple level set method for solving Stefan problems*, Journal of Computational Physics **135** (1997), 8–29.
- [23] Chen, Y. G., Giga, Y., and Goto, S., *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, Journal of Differential Geometry **33** (1991), 749–786.
- [24] Cheng, L.-T., Burchard, P., Merriman, B., Osher, S., *Motion of curves constrained on surfaces using a level set approach*, Journal of Computational Physics **175** (2002), 604–644.
- [25] Chopp, D., *Computing minimal surfaces via level set curvature flow*, Journal of Computational Physics **106** (1993), 77–91.
- [26] Colella, P., Majda, A. and Roytburd, V., *Theoretical and numerical structure for reacting shock waves*, SIAM Journal on Scientific and Statistical Computing **7** (1986), 1059–1080.
- [27] Combettes, P. L. and Pesquet, J.-C., *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer Optimization and Its Applications, 2011, 185–212.
- [28] Crandall, M. and Lions, P.-L., *Viscosity solutions of Hamilton-Jacobi equations*, Transactions of the American Mathematical Society **277** (1983), 1–42.
- [29] ———, *Two approximations of solutions of Hamilton-Jacobi equations*, Mathematics of Computation **43** (1984), 1–19.
- [30] Crandall, M. and Majda, A., *Monotone difference approximations for scalar conservation laws*, Mathematics of Computation **34** (1980), 1–21.

- [31] Durlofsky, L. J., Engquist, B., and Osher, S., *Triangle based adaptive stencils for the solution of hyperbolic conservation laws*, Journal of Computational Physics **98** (1992), 64–73.
- [32] Engquist, B., Fatemi, E., and Osher, S., *Numerical solution of the high frequency asymptotic expansion for hyperbolic equations*, Proceedings of the 10th Annual Review of Progress in Applied and Computational Electromagnetics, Monterey, California, 1994, 32–45.
- [33] Engquist, B. and Osher, S., *Stable and entropy satisfying approximations for transonic flow calculations*, Mathematics of Computation **34** (1980), 45–57.
- [34] ———, *One-sided difference approximations for nonlinear conservation laws*, Mathematics of Computation **36** (1981), 321–351.
- [35] Engquist, B., Osher, S., and Zhong, S., *Fast wavelet based algorithms for linear evolution equations*, SIAM Journal on Scientific Computing **15** (1994), 755–775.
- [36] Engquist, B., Runborg, O., and Tornberg, A.-K., *High frequency wave propagation by the segment projection method*, Journal of Computational Physics **178** (2002), 373–390.
- [37] Enright, D., Marschner, S., and Fedkiw, R., *Animation and rendering of complex water surfaces*, Siggraph 2002 Annual Conference, ACM TOG 21, 2002, 736–744.
- [38] Esedoglu, S. and Otto, F., *Threshold dynamics for networks with arbitrary surface tensions*, Communications on Pure and Applied Mathematics, to appear. DOI: 10.1002/cpa.21527
- [39] Evans, L. C., *Convergence of an algorithm for mean curvature motion*, Indiana University Mathematics Journal **42** (1993), 553–557.
- [40] Evans, L. C. and Spruck, J., *Motion of level sets by mean curvature, I*, Journal of Differential Geometry **33** (1991), 635–681.
- [41] Fatemi, E., Engquist, B., and Osher, S., *Numerical solution of the high frequency asymptotic expansion of the scalar wave equation*, Journal of Computational Physics **120** (1995), 145–155.
- [42] ———, *Finite difference methods for the nonlinear equations of perturbed geometric optics*, ACES Journal **11** (1996), 90–98.
- [43] Fatemi, E., Jerome, J., and Osher, S., *Solution of the hydrodynamic device model using high order nonoscillatory shock capturing algorithms*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **10** (1991), 232–243.
- [44] Fedkiw, R., *Coupling an Eulerian fluid calculation to a Lagrangian solid calculation with the ghost fluid method*, Journal of Computational Physics **175** (2002), 200–224.
- [45] Fedkiw, R., Aslam, T., Merriman, B., and Osher, S., *A non-oscillatory Eulerian approach to interfaces in multimaterial flows (the ghost fluid method)*, Journal of Computational Physics **152** (1999), 457–492.
- [46] Fedkiw, R., Aslam, T., and Xu, S., *The ghost fluid method for deflagration and detonation discontinuities*, Journal of Computational Physics **154** (1999), 393–427.
- [47] Fedkiw, R., Merriman, B., and Osher, S., *High accuracy numerical methods for thermally perfect gas flows with chemistry*, Journal of Computational Physics **132** (1997), 175–190.

- [48] ———, *Efficient characteristic projection in upwind difference schemes for hyperbolic systems; the complementary projection method*, *Journal of Computational Physics* **141** (1998), 22–36.
- [49] Fedkiw, R., Stam, J., and Jensen, H. W., *Visual simulation of smoke*, *Siggraph 2001 Annual Conference*, 2001, 15–22.
- [50] Foster, N. and Fedkiw, R., *Practical animation of liquids*, *Siggraph 2001 Annual Conference*, 2001, 23–30.
- [51] Gabor, D., *Information theory in electron microscopy*, *Laboratory Investigation* **14** (1965), 801–807.
- [52] Gilboa G. and Osher, S., *Nonlocal linear image regularization and supervised segmentation*, *Multiscale Modeling and Simulation* **6** (2007), 595–630.
- [53] Gilles J. and Osher, S., *Bregman implementation of meyer’s G-norm for cartoon textures decomposition*, *UCLA CAM Report 11-73*, 2011.
- [54] Glimm, J., Xia, L., Liu, Y., and Zhao, N., *Conservative front tracking and level set algorithms*, *Proceedings of the National Academy of Sciences* **98** (2001), 14198–14201.
- [55] Goldstein, T., Bresson, X., and Osher, S., *Geometric applications of the split Bregman method: Segmentation and surface reconstruction*, *Journal of Scientific Computing* **45** (2010), 272–293.
- [56] Goldstein, T. and Osher, S., *The split Bregman method for  $\ell_1$  regularized problems*, *SIAM Journal on Imaging Sciences* **2** (2009), 323–343.
- [57] Gottlieb, S., Shu, C-W., and Tadmor, E., *Strong stability preserving high order time discretization methods*, *SIAM Review* **43** (2001), 89–112.
- [58] Grayson, M., *The heat equation shrinks embedded plane curves to round points*, *Journal of Differential Geometry* **26** (1987), 285–314.
- [59] Green, M. and Osher, S., *Steiner polynomials, Wulff flows and some new isoperimetric inequalities for convex plane curves*, *Asian Journal of Mathematics* **3** (1999), 659–676.
- [60] Guo, Z. and Osher, S., *Template matching via  $\ell_1$  minimization and its application to hyperspectral data*, *UCLA CAM Report 11-78*, 2011.
- [61] Gustafsson, B., Kreiss, H.-O., and Sundstrom, A., *Stability theory of difference approximations for mixed initial boundary value problems. II*, *Mathematics of Computation* **26** (1972), 649–686.
- [62] Gyure, M., Ratsch, C., Merriman, B., Caffisch, R., Osher, S., Zinck, J., and Vvedensky, D., *Level set Methods for the simulation of epitaxial phenomena*, *Physical Review E* **58** (1998), 6927–6930.
- [63] Harabetian, E. and Osher, S., *Regularization of ill-posed problems via the level set approach*, *SIAM Journal on Applied Mathematics* **58** (1998), 1689–1706.
- [64] Harabetian, E., Osher, S., and Shu, C.-W., *An Eulerian approach for vortex motion using a level set regularization procedure*, *Journal of Computational Physics* **127** (1996), 15–26.
- [65] Harten, A., *High resolution schemes for hyperbolic conservation laws*, *Journal of Computational Physics* **49** (1983), 357–393.

- [66] Harten, A., Engquist, B., Osher, S., and Chakravarthy, S., *Uniformly high-order accurate essentially non-oscillatory schemes III*, Journal of Computational Physics **71** (1987), 231–303.
- [67] Harten, A. and Osher, S., *Uniformly high-order accurate non-oscillatory schemes, I*, SIAM Journal on Numerical Analysis **24** (1987), 279–304.
- [68] Harten, A., Osher, S., Engquist, B., and Chakravarthy, S., *Some results on uniformly high order accurate essentially non-oscillatory schemes*, Applied Numerical Mathematics **2** (1986), 347–377.
- [69] Hummel, R.A., *Representations based on zero-crossings in scale-space*, in Readings in computer vision: issues, problems, principles, and paradigms, Morgan Kaufmann Publishers Inc., San Francisco, California, 1987, 753–758.
- [70] Ishii, H., *A generalization of Bence, Merriman, and Osher algorithm for motion by mean curvature*, In Curvature Flows and Related Topics, A. Damlamian, J. Spruck, and A. Visintin (eds), Gakkôtoshô, Tokyo, 1995, 111–127.
- [71] Ishii, H., Pires, G. E., and Souganidis, P. E., *Threshold dynamics type schemes for propagating fronts*, Journal of the Mathematical Society of Japan **51** (1999), 267–308.
- [72] Jain, A. K., *Partial differential equations and finite-difference methods in image processing, Part I: Image representation*, Journal of Optimization Theory and Applications **23** (1977), 65–91.
- [73] Kang, M., Fedkiw, R., and Liu, X.-D., *A boundary condition capturing method for multiphase incompressible flow*, Journal of Scientific Computing **15** (2000), 323–360.
- [74] Kao, C.-Y., Osher, S., and Qian, J.-L., *Lax-Friedrichs sweeping scheme for static Hamilton-Jacobi equations*, Journal of Computational Physics **196** (2004), 367–391.
- [75] ———, *Legendre-transform-based fast sweeping methods for static Hamilton-Jacobi equations on triangulated meshes*, Journal of Computational Physics **227** (2008), 10209–10225.
- [76] Kao, C.-Y., Osher, S., and Tsai, Y.-H., *Fast sweeping methods for static Hamilton-Jacobi equations*, SIAM Journal on Numerical Analysis **42** (2005), 2612–2632.
- [77] Karni, S., *Multicomponent flow calculations by a consistent primitive algorithm*, Journal of Computational Physics **112** (1994), 31–43.
- [78] Kim, Y.-T., Goldenfeld, N., and Dantzig, J., *Computation of dendritic microstructures using a level set method*, Physical Review E **62** (2000), 2471–2474.
- [79] Koenderink, J.J., *The structure of images*, Biological Cybernetics **50** (1984), 363–370.
- [80] Kreiss, H.-O., *Difference approximations for the initial-boundary value problem for hyperbolic differential equations*, in Numerical Solutions of Nonlinear Differential Equations, D. Greenspan (ed), John Wiley, New York, 1966, 141–166.
- [81] Lafon, F., and Osher, S., *High order two dimensional nonoscillatory methods for solving Hamilton-Jacobi equations*, Journal of Computational Physics **123** (1996), 235–253.
- [82] Lagnado, R., and Osher, S., *Reconciling differences*, Risk Magazine, **10** (1997), 79–83.
- [83] ———, *A technique for calibrating derivative security pricing models, numerical*

- solution of an inverse problem*, Journal of Computational Finance **1** (1997), 13–25.
- [84] Lai, R. and Osher, S., *A splitting method for orthogonality constrained problems*, Journal of Scientific Computing **58** (2014), 431–449.
- [85] Langer, A., Osher, S., and Schonlieb, C. B., *Bregmanized domain decomposition for image restoration*, Journal of Scientific Computing **54** (2013), 549–576.
- [86] Liu, B., King, K., Steckner, M., Xie, J., Sheng, J., and Ying, L., *Regularized sensitivity encoding (sense) reconstruction using Bregman iterations*, Magnetic Resonance in Medicine **61** (2008), 145–152.
- [87] Liu, X.-D., Fedkiw, R., and Kang, M., *A boundary condition capturing method for Poisson’s equation on irregular domains*, Journal of Computational Physics **160** (2000), 151–178.
- [88] Liu, X.-D. and Osher, S., *Nonoscillatory high order accurate self similar maximum principle satisfying shock capturing schemes*, SIAM Journal on Numerical Analysis **33** (1996), 760–779.
- [89] ———, *Convex ENO high-order multidimensional schemes without field by field projection or staggered grids*, Journal of Computational Physics **142** (1998), 304–330.
- [90] Liu, X.-D., Osher, S., and Chan, T., *Weighted essentially non-oscillatory schemes*, Journal of Computational Physics **115** (1994), 200–212.
- [91] Majda, A., McDonough, J., and Osher, S., *The Fourier method for nonsmooth initial data*, Mathematics of Computation **32** (1978), 1041–1081.
- [92] Majda, A. and Osher, S., *Reflections of singularities at the boundary*, Communications on Pure and Applied Mathematics **28** (1975), 479–499.
- [93] ———, *Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary*, Communications on Pure and Applied Mathematics **28** (1975), 607–675.
- [94] ———, *Propagation of error into regions of smoothness for accurate difference approximations to hyperbolic equations*, Communications on Pure and Applied Mathematics **30** (1977), 671–705.
- [95] ———, *A systematic approach for correcting nonlinear instabilities: the Lax-Wendroff scheme for scalar conservation laws*, Numerische Mathematik **30** (1978), 429–452.
- [96] Murkerjev, E., Sunu, J. and Bertozzi, A. L., *Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video*, Proceedings of the International Conference in Image Processing, 2014, to appear.
- [97] Merriman, B., Bence, J., and Osher, S., *Diffusion generated motion by mean curvature*, in Computational Crystal Growers Workshop, J. E. Taylor (ed), American Mathematical Society, Providence, Rhode Island, 1992, 73–83.
- [98] ———, *Motion of multiple junctions: A level-set approach*, Journal of Computational Physics **112** (1994), 334–363.
- [99] Merriman, B., Caffisch, R., and Osher, S., *Level set methods with an application to modeling the growth of thin films*, in Free boundary value problems, theory and applications, I. Athanasopoulos, G. Makreakis and J. Rodrigues (eds.), CRC Press, Boca Raton, 1999, 51–70.

- [100] Mulder, W., Osher, S., and Sethian, J., *Computing interface motion in compressible gas dynamics*, Journal of Computational Physics **100** (1992), 209–228.
- [101] Nguyen, D., Fedkiw, R., and Jensen, H., *Physically based modeling and animation of fire*, Siggraph 2002 Annual Conference, ACM TOG 21 (2002), 721–728.
- [102] Nguyen, D., Fedkiw, R., and Kang, M., *A boundary condition capturing method for incompressible flame discontinuities*, Journal of Computational Physics **172** (2001), 71–98.
- [103] Nguyen, D., Gibou, F., and Fedkiw, R., *A fully conservative ghost fluid method & stiff detonation waves*, 12th International Detonation Symposium, San Diego, California, 2002.
- [104] Osher, S., *Systems of difference equations with general homogeneous boundary conditions*, Transactions of the American Mathematical Society **137** (1969), 177–201.
- [105] ———, *Stability of parabolic difference approximations to certain mixed initial boundary value problems*, Mathematics of Computation **26** (1972), 13–39.
- [106] ———, *Numerical solution of singular perturbation problems and one-sided difference schemes*, North Holland Math. Studies 47, 1981.
- [107] ———, *The Riemann problem for nonconvex scalar conservation laws and the Hamilton-Jacobi equations*, Proceedings of the American Mathematical Society **89** (1983), 641–646.
- [108] ———, *Riemann solvers, the entropy condition and difference approximations*, SIAM Journal on Numerical Analysis **21** (1984), 217–235.
- [109] ———, *Convergence of generalized MUSCL schemes*, SIAM Journal on Numerical Analysis **22** (1985), 947–961.
- [110] Osher, S., Burger, M., Goldfarb, D., Xu, J., and Yin, W., *An iterative regularization method for total variation-based image restoration*, Multiscale Modeling and Simulation **4** (2005), 460–489.
- [111] Osher, S. and Chakravarthy, S., *Upwind schemes and boundary conditions with applications to Euler equations in general geometries*, Journal of Computational Physics **50** (1983), 447–481.
- [112] ———, *High resolution schemes and the entropy condition*, SIAM Journal on Numerical Analysis **21** (1984), 955–984.
- [113] ———, *Very high order accurate TVD schemes*, in IMA Volumes in Mathematics and Its Applications **2** (1986), Springer-Verlag, 229–274.
- [114] Osher, S., Cheng, L.-T., Kang, M., Shim, H., and Tsai, Y.-H., *Geometric optics in a phase space and Eulerian framework*, Journal of Computational Physics **179** (2002), 622–648.
- [115] Osher, S. and Fedkiw, R., *Level set methods: an overview and some recent results*, Journal of Computational Physics **169** (2001), 463–502.
- [116] ———, *Level sets and dynamic implicit surfaces*, Springer-Verlag, New York, 2002.
- [117] Osher, S., Hafez, M., and Whitlow Jr., W., *Entropy condition satisfying approximations for the full potential equation of transonic flow*, Mathematics of Computation **44** (1985), 1–29.
- [118] Osher, S., Mao, Y., Dong, B., and Yin, W., *Fast linearized Bregman iteration for com-*

- pressive sensing and sparse denoising*, *Communications in Mathematical Sciences*, **8** (2010), 93–111.
- [119] Osher, S. and Merriman, B., *The Wulff shape as the asymptotic limit of a growing crystalline interface*, *Asian Journal of Mathematics* **1** (1997), 560–571.
- [120] Osher, S. and Rudin, R.T., *Feature-oriented image enhancement using shock filters*, *SIAM Journal on Numerical Analysis* **27** (1990), 919–940.
- [121] Osher, S. and Sanders, R., *Numerical approximations to nonlinear conservation laws with locally varying time and space grids*, *Mathematics of Computation* **41** (1983), 321–336.
- [122] Osher, S. and Santosa, F., *Level set method for optimization problems involving geometry and constraints, I. Frequencies of a two-density inhomogeneous drum*, *Journal of Computational Physics* **171** (2001), 277–288.
- [123] Osher, S. and Sethian, J., *Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations*, *Journal of Computational Physics* **79** (1988), 12–49.
- [124] Osher, S. and Shu, C.-W., *High order essentially non-oscillatory schemes for Hamilton-Jacobi equations*, *SIAM Journal on Numerical Analysis* **28** (1991), 902–921.
- [125] Osher, S. and Solomon, F., *Upwind difference schemes for hyperbolic systems of conservation laws*, *Mathematics of Computation* **38** (1982), 339–374.
- [126] Osher, S. and Tadmor, E., *On the convergence of difference approximations to scalar conservation laws*, *Mathematics of Computation* **50** (1988), 19–51.
- [127] Peng, D., Merriman, B., Osher, S., Zhao, H.-K., and Kang, M., *A PDE-based fast local level set method*, *Journal of Computational Physics* **155** (1999), 410–438.
- [128] Peng, D., Osher, S., Merriman, B. and Zhao, H.-K., *The geometry of Wulff crystal shapes and its relations with Riemann problems*, in *Nonlinear Partial Differential Equations*, G.Q. Chen and E. DeBenedetto (eds.), *Contemporary Mathematics*, volume 238, American Mathematical Society, Providence, Rhode Island, 1999, 251–303.
- [129] Perona, P. and Malik, J., *Scale-space and edge detection using anisotropic diffusion*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** (1990), 629–639.
- [130] Romeny, B., Editor, *Geometry Driven Diffusion in Computer Vision*, Kluwer, 1994.
- [131] Rouy, E. and Tourin, A., *A viscosity solutions approach to shape-from-shading*, *SIAM Journal on Numerical Analysis* **29** (1992), 867–884.
- [132] Rudin, L.I., Osher, S., and Fatemi, E., *Nonlinear total variation based noise removal algorithms*, *Physica D* **60** (1992), 259–268.
- [133] Ruuth, S., Merriman, B., and Osher, S., *Convolution generated motion as a link between cellular automata and continuum pattern dynamics*, *Journal of Computational Physics* **151** (1999), 836–861.
- [134] ———, *A fixed grid method for capturing the motion of self-intersecting interfaces and related PDEs*, *Journal of Computational Physics* **163** (2000), 1–21.
- [135] Ruuth, S., Merriman, B., Xin, J., and Osher, S., *A diffusion generated approach to the curvature motion of filaments*, *Journal of Nonlinear Science* **61** (2001), 473–494.
- [136] Sapiro, G., *Geometric Partial Differential Equations and Image Processing*, Cam-

- bridge University Press, New York, 2001.
- [137] Shi, J., Yin, W., Osher, S., and Sajda, P., *A fast hybrid algorithm for large scale  $L1$ -regularized logistic regression*, Journal of Machine Learning Research, **11** (2008), 713–741.
- [138] Shu, C.-W. and Osher, S., *Efficient implementation of essentially non-oscillatory shock capturing schemes*, Journal of Computational Physics **77** (1988), 439–471.
- [139] ———, *Efficient implementation of essentially non-oscillatory shock capturing schemes II*, Journal of Computational Physics **83** (1989), 32–78.
- [140] Shu, C.-W., Zang, T.A., Erlebacher, G., Whitaker, D., and Osher, S., *High-order ENO schemes applied to two- and three-dimensional compressible flow*, Applied Numerical Mathematics **9** (1992), 45–71.
- [141] Steinhoff, J., Fang, M., and Wang, L., *A new Eulerian method for the computation of propagating short acoustic and electromagnetic pulses*, Journal of Computational Physics **157** (2000), 683–706.
- [142] Sussman, M., Smereka, P., and Osher, S., *A level set approach for computing solutions to incompressible two-phase flow*, Journal of Computational Physics **114** (1994), 146–159.
- [143] Ton, V. T., Karagozian, A. R., Osher, S., and Engquist, B., *Numerical simulation of high speed chemically reactive flow*, Theoretical and Computational Fluid Dynamics **6** (1994), 161–179.
- [144] Tsai, Y.-H., Cheng, L.-T., Osher, S., and H.-K. Zhao, *Fast sweeping algorithms for a class of Hamilton-Jacobi equations*, SIAM Journal on Numerical Analysis **41** (2003), 673–694.
- [145] van Leer, B., *Towards the ultimate conservative difference scheme V. A second order sequel to Godunov’s method*, Journal of Computational Physics **32** (1979), 101–136.
- [146] Varah, J., *Stability of difference approximations to the mixed initial boundary value problem for parabolic systems*, SIAM Journal on Numerical Analysis **8** (1971), 598–615.
- [147] Warren, R., Osher, S., and Vanderbeek, R., *Multiple aerosol unmixing by the split Bregman algorithm*, IEEE Transactions on Geoscience and Remote Sensing, **50** (2012), 3271–3279.
- [148] Witkin, A., *Scale-space filtering*, International Joint Conference on Artificial Intelligence **2** (1983), 1019–1021.
- [149] Xu, J. and Osher, S., *Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising*, IEEE Transactions on Image Processing **16** (2007), 534–544.
- [150] Yang, Y., Ma, J., and Osher, S., *Seismic data reconstruction via matrix completion*, Inverse Problems and Imaging **7** (2013), 1379–1392.
- [151] Yang, Y., Moller, M., and Osher, S., *A dual split Bregman method for fast  $\ell1$  minimization*, Mathematics of Computation **82** (2013), 2061–2085.
- [152] Yin W. and Osher, S., *Error forgetting of Bregman iteration*, Journal of Scientific Computing, **54** (2013), 684–695.
- [153] Yin, W., Osher, S., Goldfarb, D., and Darbon, J. *Bregman iterative algorithms for*

- l1*-minimization with applications to compressed sensing, *SIAM Journal on Imaging Sciences*, **1** (2008), 143–168.
- [154] Zhao, H.-K., Chan, T., Merriman, B., and Osher, S., *A variational level set approach to multiphase motion*, *Journal of Computational Physics* **127** (1996), 179–195.
- [155] Zhao, H.-K., Merriman, B., Osher, S., and Wang, L., *Capturing the behavior of bubbles and drops using the variational level set approach*, *Journal of Computational Physics* **143** (1998), 495–518.
- [156] Zhao, H.-K., Osher, S., and Fedkiw, R., *Fast surface reconstruction using the level set method*, 1st IEEE Workshop on Variational and Level Set Methods, in conjunction with the 8th International Conference on Computer Vision (ICCV), Vancouver, Canada, 2001, 194–202.
- [157] Zhao, H.-K., Osher, S., Merriman, B., and Kang, M., *Implicit nonparametric shape reconstruction from unorganized points using a variational level set method*, *Computer Vision and Image Understanding* **80** (2000), 295–314.

Department of Computer Science, Stanford University, Stanford, CA 94305, USA

E-mail: rfedkiw@stanford.edu

Department of Mathematics, CMLA, ENS Cachan, France

E-mail: morel@cmla.ens-cachan.fr

Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

E-mail: guillermo.sapiro@duke.edu

Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

E-mail: shu@dam.brown.edu

Department of Mathematics, UCLA, Los Angeles, CA 90095, USA

E-mail: wotaoyin@math.ucla.edu

## **Phillip Griffiths**

B.S., Wake Forest University, 1959

Ph.D. in Mathematics, Princeton University, 1962

### **Positions held**

Miller Fellow, University of California, Berkeley, 1962–1964

Faculty Member, University of California, Berkeley, 1964–1967

Visiting Professor, Princeton University, 1967–1968

Professor, Princeton University, 1968–1972

Professor, Harvard University, 1972–1983

Provost and James B. Duke Professor, Duke University, 1983–1991

Director, Institute for Advanced Study, 1991–2003

Professor, Institute for Advanced Study, 2004–2009

Professor Emeritus, Institute for Advanced Study, 2009–present

# The work of Phillip Griffiths

Mark L. Green

## 1. Overview of Some Major Accomplishments

- (1) Algebraic equivalence  $\neq$  homological equivalence [37]
- (2) Geometry of period domains (w. Wilfried Schmid) [38–40, 51, 52]
- (3) Non-rationality of the cubic 3-fold (w. Herb Clemens) [18]
- (4) Holomorphic maps (w. James Carlson and MG) [15, 25, 31]
- (5) Brill-Noether Problem (w. Joseph Harris) [48]
- (6) Infinitesimal Variation of Hodge Structure (w. James Carlson, MG, Joseph Harris) [13, 14, 16, 21, 22, 35, 47]
- (7) Homotopy Theory of Kähler Manifolds (w. Pierre Deligne, John Morgan, Dennis Sullivan) [20]
- (8) Exterior Differential Systems (w. Robert Bryant, S.S. Chern, Robert Gardner, Hubert Goldschmidt) [6–8]
- (9) Conservation Laws of Exterior Differential Systems (w. Robert Bryant and Lucas Hsu) [10, 11]
- (10) Isometric Embeddings (w. Eric Berger, Robert Bryant, Deane Yang) [3, 4, 12]
- (11) Tangent Space to Algebraic Cycles (w. MG) [23, 24]
- (12) Mumford-Tate Groups (w. MG, Matt Kerr) [27]

The goal of this paper is to give a treatment of a few of these topics [(1), (3), (4), (10), (11), (12) in the list above] at a level that will be accessible to a general mathematical audience. Topics (1) and (10) will be treated in the most depth. This is not intended as a survey of Griffiths' work for experts.

Besides the work discussed here, Griffiths is known for his many students and mathematical descendants, and for his many important books [1, 2, 6, 24, 26, 27, 29, 36, 45, 46, 49, 50] both expository and research monographs, many significant expository articles [9, 17, 19, 30, 32–34, 41–44], and his extensive work in the developing world.

## 2. Griffiths' Work in Hodge Theory and Algebraic Cycles

With the development of algebraic topology by Henri Poincaré, a variety of geometric methods flourished. It was soon realized that certain classes of geometric objects were especially

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

worth studying. An outstanding example of this was the work by Solomon Lefschetz, who found a number of surprising facts about the topology of **smooth projective varieties**.

Complex projective space  $\mathbf{CP}^n$  is  $(\mathbf{C}^{n+1} - 0)/\sim$ , where

$$\lambda(z_1, \dots, z_{n+1}) \sim (z_1, \dots, z_{n+1})$$

for  $\lambda \in \mathbf{C}^*$ . A smooth projective variety is a smooth complex submanifold  $M$  of  $\mathbf{CP}^N$  for some  $N$ . By a theorem of Chow, such an  $M$  is defined by algebraic equations, i.e. by the vanishing of a collection of homogeneous polynomials in  $z_1, \dots, z_{N+1}$ . An example of the phenomena discovered by Lefschetz, is that the cohomology groups  $H^k(M, \mathbf{C})$  have even dimension when  $k$  is odd, a generalization of the fact that for a compact Riemann surface,  $H^1(M, \mathbf{C})$  has dimension  $2g$ . Once Poincaré and Élie Cartan had developed the calculus of differential forms, which are the essential tool for doing integration on manifolds, it became possible via De Rham's Theorem to express cohomology in terms of differential forms. If  $A^k(M)$  are the  $C^\infty$   $k$ -forms on a smooth real manifold  $M$ , there is a generalization of grad, curl and div, the **exterior derivative**

$$d: A^k(M) \rightarrow A^{k+1}(M)$$

with  $d^2 = 0$ . The **De Rham cohomology**

$$H_{DR}^k(M) = \frac{\text{Ker}(d: A^k(M) \rightarrow A^{k+1}(M))}{\text{Im}(d: A^{k-1}(M) \rightarrow A^k(M))}$$

then satisfies

**Theorem** (De Rham).

$$H_{DR}^k(M) \cong H^k(M, \mathbf{R}) = \frac{\{\text{closed } k\text{-forms on } M\}}{\{\text{exact } k\text{-forms on } M\}}.$$

Thus every topological cohomology class is represented by a closed form, but not uniquely. Inspired by Maxwell's equations in electromagnetics, Hodge discovered a way to choose a unique representative in a geometric manner.  $M$  must be a compact oriented Riemannian manifold, i.e. there is an analogue of the usual inner product on the tangent space to each point of  $M$ , varying smoothly. In this circumstance, there is an adjoint to the exterior derivative

$$d^*: A^k(M) \rightarrow A^{k-1}(M).$$

The **Laplacian**

$$\Delta = dd^* + d^*d: A^k(M) \rightarrow A^k(M)$$

turns out to be an elliptic operator. The **harmonic forms** are defined by

$$\mathcal{H}^k(M) = \text{Ker}(\Delta: A^k(M) \rightarrow A^k(M)).$$

Then

$$\Delta(\omega) = 0 \Leftrightarrow d\omega = 0 \text{ and } d^*\omega = 0.$$

There is thus a natural map

$$\mathcal{H}^k(M) \rightarrow H_{DR}^k(M);$$

by using elliptic operator theory, Hodge argued that this map is an isomorphism, and thus that every cohomology class is represented by a unique harmonic form.

Once we have a complex manifold, then switching to complex-valued forms and cohomology, there is a decomposition

$$A^k(M) = \bigoplus_{p+q=k} A^{p,q}(M).$$

Here, if  $z_1, \dots, z_n$  are local complex coordinates, then if  $z_j = x_j + iy_j$ , the complex-valued 1-forms have a pointwise basis  $dz_1, \dots, dz_n, d\bar{z}_1, \dots, d\bar{z}_n$ , where

$$dz_j = dx_j + idy_j, \quad d\bar{z}_j = dx_j - idy_j.$$

Now pointwise, the  $k$  forms decompose into those of **type  $p, q$** , i.e. sums of terms

$$dz_{i_1} \wedge \dots \wedge dz_{i_p} \wedge d\bar{z}_{j_1} \wedge \dots \wedge d\bar{z}_{j_q}.$$

It would then be natural to hope that the decomposition of forms by type passed to De Rham cohomology, but this does not happen for an arbitrary compact complex manifold. The manifold needs to possess a special type of Hermitian metric, a **Kähler metric**. Complex projective space has such a metric, the Fubini-Study metric, and this restricts to give one on any smooth projective variety  $M$ .

Complex manifolds decompose

$$d = \partial + \bar{\partial},$$

where

$$\partial: A^{p,q}(M) \rightarrow A^{p+1,q}(M), \quad \bar{\partial}: A^{p,q}(M) \rightarrow A^{p,q+1}(M).$$

We then have

$$H^{p,q}(M) = \frac{\text{Ker}(\bar{\partial}: A^{p,q}(M) \rightarrow A^{p,q+1}(M))}{\text{Im}(\bar{\partial}: A^{p,q-1}(M) \rightarrow A^{p,q}(M))}.$$

A Hermitian metric allows one to construct adjoints for these operators, and hence Laplacians

$$\Delta_{\partial} = \partial\partial^* + \partial^*\partial, \quad \Delta_{\bar{\partial}} = \bar{\partial}\bar{\partial}^* + \bar{\partial}^*\bar{\partial}$$

Note that

$$\Delta_{\partial}: A^{p,q}(M) \rightarrow A^{p,q}(M), \quad \Delta_{\bar{\partial}}: A^{p,q}(M) \rightarrow A^{p,q}(M).$$

The key fact for Kähler manifolds is that

$$\Delta_{\partial} = \Delta_{\bar{\partial}} = \frac{1}{2}\Delta.$$

If

$$\mathcal{H}^{p,q}(M) = \text{Ker}(\Delta_{\bar{\partial}}: A^{p,q}(M) \rightarrow A^{p,q}(M)),$$

then this identity plus elliptic operator theory for  $\Delta_{\bar{\partial}}$  gives a decomposition

$$\mathcal{H}^k(M) = \bigoplus_{p+q=k} \mathcal{H}^{p,q}(M)$$

and isomorphisms

$$H^{p,q}(M) \cong \mathcal{H}^{p,q}(M).$$

The result is the **Hodge decomposition** for a compact Kähler manifold

$$H_{DR}^k(M) \cong \bigoplus_{p+q=k} H^{p,q}(M).$$

Although the harmonic representatives depend on the choice of Kähler metric, the decomposition of the cohomology groups does not.

Because the conjugate of  $H^{p,q}(M)$  is  $H^{q,p}(M)$ , we see that

$$\dim(H^{2k+1}(M)) = 2 \sum_{p=0}^k \dim(H^{p,2k+1-p}(M))$$

and hence is even, the discovery of Lefschetz alluded to earlier.

Once one understands what happens for a single  $M$ , it is reasonable to study families. A **family of smooth manifolds** is a manifold  $\mathcal{M}$  together with a smooth map  $p: \mathcal{M} \rightarrow B$ , where  $B$  is the unit ball and  $dp$  is surjective at all points, and thus  $M_t = p^{-1}(t)$  is a smooth manifold for  $t \in B$ . Nothing interesting happens in the category of smooth manifolds, as the  $M_t$  are all diffeomorphic, but if we move to the category of complex manifolds, with  $p$  holomorphic and  $B$  a ball in  $\mathbf{C}^r$ , then the  $M_t$  can be inequivalent as complex manifolds. Such families were studied by Kodaira and Spencer, and they created an appropriate **deformation theory of complex manifolds**.

If one has a holomorphic family of compact Kähler manifolds, because the  $M_t$  are all diffeomorphic in a natural way, one can identify the  $H^k(M_t, \mathbf{C})$  and ask whether the Hodge decomposition varies analytically, i.e. whether the  $H^{p,q}(M_t)$  are holomorphically varying families of complex subspaces of what we may think of as the fixed vector space  $H^k(M_t, \mathbf{C})$ . The answer is, unfortunately, they do not.

One of Griffiths' first contributions was to change the question. He defined the **Hodge filtration**

$$F^p H^k(M, \mathbf{C}) = H^{k,0}(M) \oplus H^{k-1,1}(M) \oplus \dots \oplus H^{p,k-p}(M) = \bigoplus_{r \geq p} H^{r,k-r}(M).$$

This is a decreasing filtration

$$\begin{aligned} & H^k(X, \mathbf{C}) \\ &= F^0 H^k(M, \mathbf{C}) \supseteq F^1 H^k(M, \mathbf{C}) \supseteq \dots \supseteq F^k H^k(M, \mathbf{C}) \supseteq F^{k+1} H^k(M, \mathbf{C}) = 0. \end{aligned}$$

**Theorem** (Griffiths). *In an analytic family, the spaces  $F^p H^k(M_t, \mathbf{C})$  vary analytically.*

There was, however, a surprise. Because, for example,

$$\frac{d}{dt}(v(t) \wedge w(t)) = \frac{dv}{dt} \wedge w + v \wedge \frac{dw}{dt},$$

as we vary  $M_t$ , at most one  $dz_j$  can be changed into a linear combination involving  $d\bar{z}_i$ 's by a first derivative. The result is:

**Theorem** (Griffiths' Infinitesimal Period Relation).

$$\frac{d}{dt} F^p H^k(M_t, \mathbf{C}) \subseteq F^{p-1} H^k(M_t, \mathbf{C}).$$

The space of possible Hodge filtrations of given dimensions, satisfying various other conditions (the Hodge-Riemann relations) is called the **period domain**. The infinitesimal period relation, also known as **Griffiths transversality**, gives a natural linear space of 1-forms which much vanish on the image of the period mapping for any family. These 1-forms constitute an **exterior differential system** which in general is not integrable. One important consequence of this is that while for  $k = 1$ , all Hodge polarized structures arise from geometry, this cannot be the case for  $k \geq 2$ .

For  $M$  a smooth projective variety, an **algebraic subvariety** of  $M$  is the locus defined by a collection of homogeneous polynomials. There is a notion of dimension, and the codimension of  $Y$  is  $\dim(M) - \dim(Y)$ . An **algebraic cycle of codimension  $p$**  is a finite formal linear combination

$$Z = \sum_i n_i Z_i,$$

where  $n_i \in \mathbf{Z}$  and  $Z_i$  is a codimension  $p$  algebraic subvariety. We write

$$Z^p(M) = \left\{ \sum_i n_i Z_i \mid n_i \in \mathbf{Z}, \text{ sum is finite} \right\}$$

and

$$Z^p(M, \mathbf{Q}) = \left\{ \sum_i n_i Z_i \mid n_i \in \mathbf{Q}, \text{ sum is finite} \right\}$$

If  $\dim(M) = n$ , then an algebraic subvariety  $Y$  of codimension  $p$  has a homology class  $[Y] \in H_{2n-2p}(M, \mathbf{Z})$ . By **Poincaré Duality**, for  $M$  compact and connected,

$$H_{2n-2p}(M, \mathbf{Z}) \cong H^{2p}(M, \mathbf{Z}),$$

and the image of  $[Y]$  under this isomorphism is

$$\eta_Y \in H^{2p}(M, \mathbf{Z}),$$

the Poincaré dual class of  $Y$ . The defining formula for  $\eta_Y$  is

$$\int_Y \omega = \int_M \omega \wedge \eta_Y$$

for  $\omega$  a closed  $(2n - 2p)$ -form on  $M$ —technically, we should pull back to a desingularization of  $Y$ . Hodge noted that only forms  $\omega$  of type  $n - p, n - p$  can have non-zero integral on  $Y$ , and thus

$$\eta_Y \in H^{p,p}(M).$$

For a cycle

$$Z = \sum_i n_i Z_i,$$

set

$$\eta_Z = \sum_i n_i \eta_{Z_i}.$$

If  $i: \mathbf{Z} \rightarrow \mathbf{C}$  is the coefficient map, by Hodge,

$$\eta_Z \in H^{p,p}(M) \cap i_* H^{2p}(M, \mathbf{Z}).$$

We thus have a map

$$\eta: Z^p(M) \rightarrow H^{p,p}(M) \cap i_* H^{2p}(M, \mathbf{Z})$$

and similarly a map

$$\eta_{\mathbf{Q}}: Z^p(M, \mathbf{Q}) \rightarrow H^{p,p}(M) \cap i_* H^{2p}(M, \mathbf{Q}).$$

An element of  $H^{p,p}(M) \cap i_* H^{2p}(M, \mathbf{Z})$  is called a **Hodge class** and an element of  $H^{p,p}(M) \cap i_* H^{2p}(M, \mathbf{Q})$  a **rational Hodge class**. In either case, if the class is in the image of  $\eta$  or  $\eta_{\mathbf{Q}}$ , it is said to be represented by an algebraic cycle, and:

**Conjecture** (Hodge Conjecture).  $\eta_{\mathbf{Q}}$  is surjective, i.e. some positive multiple of any Hodge class is represented by an algebraic cycle.

There are three important notions of equivalence of algebraic cycles. If  $[Z] = 0$ , or equivalently  $\eta_Z = 0$ ,  $Z$  is said to be **homologically equivalent to 0**. The other two are algebro-geometric in character. If  $C$  is a connected compact Riemann surface, and  $\mathcal{Z} \in Z^p(M \times C)$ , we let  $Z_x = \mathcal{Z} \cdot (M \times \{x\})$  for  $x \in C$ , where  $\cdot$  denotes intersection of algebraic cycles—this has some subtleties. The cycles  $Z_x, Z_y$  for  $x, y \in C$  are said to be **algebraically equivalent**, and algebraic equivalence is the equivalence relation generated by such equivalences for all choices of  $C, x, y$ . If we restrict  $C$  to be a  $\mathbf{CP}^1$ , we get the notion of two cycles being **rationally equivalent**. These three equivalences will be denoted

$$Z \equiv_{\text{hom}} 0, \quad Z \equiv_{\text{alg}} 0, \quad Z \equiv_{\text{rat}} 0.$$

Likewise,  $Z^p(M)_{\text{hom}}$  will denote the codimension  $p$  algebraic cycles homologically equivalent to 0, etc.

The implications are

$$Z \equiv_{\text{rat}} 0 \implies Z \equiv_{\text{alg}} 0 \implies Z \equiv_{\text{hom}} 0.$$

At this point, it is useful to review what these concepts mean in the simplest case, that where  $M$  is a connected compact Riemann surface. Here

$$Z^1(M) = \left\{ \sum_{p \in M} n_p p \mid n_p \in \mathbf{Z}, \text{ sum is finite} \right\}.$$

These are the same as topological 0-chains. For  $Z = \sum_p n_p p$ ,  $\eta_Z = (\sum_p n_p) 1_M$ , where  $1_M$  is the generator of  $H^2(M, \mathbf{Z})$ . Classically, the **degree of  $Z$**  is

$$\text{deg}(Z) = \sum_p n_p$$

and elements of  $Z^1(M)$  are called **divisors**, i.e.

$$Z^1(M) = \text{Div}(M).$$

and

$$Z^1(M)_{\text{hom}} = \text{Div}_0(M) = \text{Ker}(\text{deg}: \text{Div}(M) \rightarrow \mathbf{Z}).$$

The maximum principle implies that  $M$  has no non-constant global holomorphic functions. However, there will be global meromorphic functions.

$$\mathbf{C}(M) = \{\text{global meromorphic functions } f \text{ on } M\}$$

and

$$\mathbf{C}^*(M) = \{f \in \mathbf{C}(M) \mid f \text{ not identically } 0\}.$$

Given  $f \in \mathbf{C}^*(M)$ , at any point  $p \in M$  we can define the **local degree**  $\nu_p(f)$  of  $f$  at  $p$  by, in terms of a local holomorphic coordinate  $z$  at  $p$ ,

$$f(z) = (z - p)^{\nu_p(f)} g(z),$$

where  $g(z)$  is defined locally, is holomorphic, and  $g(p) \neq 0$ . There is then a map

$$\text{div}: \mathbf{C}^*(M) \rightarrow \text{Div}(M)$$

defined by

$$f \mapsto \sum_p \nu_p(f) p.$$

It is a classical result that

$$\deg(\text{div}(f)) = 0 \quad \text{for } f \in \mathbf{C}^*(M)$$

and thus

$$\text{div}: \mathbf{C}^*(M) \rightarrow \text{Div}_0(M)$$

Unwinding the definitions,

$$Z \equiv_{\text{rat}} 0 \Leftrightarrow Z \in \text{Im}(\text{div}).$$

and also

$$Z \equiv_{\text{alg}} 0 \Leftrightarrow Z \in \text{Div}_0(M).$$

Thus in the case of connected compact Riemann surfaces,

$$\equiv_{\text{hom}} = \equiv_{\text{alg}} \neq \equiv_{\text{rat}} .$$

It was an important question, formulated by Grothendieck, whether  $\equiv_{\text{hom}} = \equiv_{\text{alg}}$  holds for cycles of codimension  $\geq 2$ . It was this problem that Griffiths solved—indeed, he showed that there are cycles  $Z \equiv_{\text{hom}} 0$  such that no positive integral multiple of  $Z$  is  $\equiv_{\text{alg}} 0$ .

To understand the background of his solution, it is helpful to explore further the classical theory of connected compact Riemann surfaces. In the case of the Riemann sphere, i.e.  $\mathbf{C}P^1$ , a surface of genus 0, we have

$$\mathbf{C}^*(\mathbf{C}P^1) = \{p(z)/q(z) \mid p(z), q(z) \text{ polynomials not identically } 0\}.$$

It follows that for  $Z \in \text{Div}(\mathbf{C}P^1)$ ,

$$Z \in \text{Im}(\text{div}) \Leftrightarrow \deg(Z) = 0 \quad \text{for genus } 0.$$

In the case  $g = 1$ , we have a complex torus

$$T = \mathbf{C}/\Lambda,$$

where

$$\Lambda = \{m + n\lambda \mid m, n \in \mathbf{Z}\},$$

where  $\lambda \in \mathbf{C} - \mathbf{R}$ . It is worth noting that different  $\lambda$ 's can give complex tori which are not biholomorphic as complex manifolds.  $T$  inherits the structure of an abelian group from  $\mathbf{C}$ , and we will denote addition by  $\oplus$ . This allows us to define a natural map

$$\oplus: \text{Div}_0(T) \rightarrow T$$

by

$$\sum_p n_p p \mapsto \oplus_p n_p p.$$

The Abel-Jacobi Theorem for a torus states:

**Theorem** (Abel-Jacobi for a torus).

- (i)  $Z \in \text{Im}(\text{div}) \Leftrightarrow \text{deg}(Z) = 0$  and  $\oplus(Z) = 0$ ;
- (ii)  $\oplus: \text{Div}_0(T) \rightarrow T$  is surjective.

In order to deal with curves of higher genus, we can reformulate the map  $\oplus$  as follows: if  $Z \in \text{Div}_0(T)$ , write  $Z = \partial\gamma$  for some 1-chain  $\gamma$ , then

$$\oplus(Z) = \int_{\gamma} dz.$$

If we change  $\gamma$  by a 1-cycle on  $T$ , we change the value of the integral by an element of  $\Lambda$ .

On a Riemann surface of genus  $g$ , the analogue of  $dz$  are global objects  $\omega$  that can be written as  $f(z)dz$  locally, where  $f$  is a holomorphic function and  $dz$  transforms as a differential when we change coordinates, i.e.  $dw = (dw/dz)dz$ . These objects are called **abelian differentials** or **holomorphic 1-forms**. It is an important classical result that the dimension of the space of holomorphic 1-forms, called the **analytic genus**, is equal to the topological genus  $g$ . If  $\omega_1, \dots, \omega_g$  is a basis for the abelian differentials, and  $Z \in \text{Div}_0(M)$  is written as  $Z = \partial\gamma$ , then define

$$AJ_M(Z) = \left( \int_{\gamma} \omega_1, \dots, \int_{\gamma} \omega_g \right),$$

modulo

$$\Lambda = \left\{ \left( \sum_{i=1}^{2g} n_i \int_{\lambda_i} \omega_1, \dots, \sum_{i=1}^{2g} n_i \int_{\lambda_i} \omega_g \right) \mid n_i \in \mathbf{Z} \right\},$$

where  $\lambda_1, \dots, \lambda_{2g}$  is a basis for  $H_1(M, \mathbf{Z})$ . We may think of  $AJ(M) \in \mathbf{C}^g/\Lambda$ , which turns out to be a torus, called the **Jacobian variety** of  $M$ , denoted  $J(M)$ . The Abel-Jacobi Theorem now states:

**Theorem** (Abel-Jacobi Theorem).

- (i)  $Z \in \text{Im}(\text{div}) \Leftrightarrow \text{deg}(Z) = 0$  and  $AJ_M(Z) = 0$ ;
- (ii)  $AJ_M: \text{Div}_0(M) \rightarrow J(M)$  is surjective.

As a consequence, for compact connected Riemann surfaces,

$$AJ_M: \frac{Z^1(M)_{\text{hom}}}{Z^1(M)_{\text{rat}}} \cong J(M).$$

In terms of Hodge theory, for a compact connected Riemann surface  $M$ ,

$$H^1(M) = H^{1,0}(M) \oplus H^{0,1}(M),$$

and

$$H^{1,0}(M) = \{\text{abelian differentials of } M\}$$

and

$$H^{0,1}(M) = \{\text{conjugates of abelian differentials of } M\}$$

An intrinsic way to write the Jacobian is

$$J(M) \cong \frac{H^{0,1}(M)}{H^1(M, \mathbf{Z})}.$$

Although all complex tori are complex manifolds, indeed Kähler manifolds, only certain ones are smooth projective varieties—these are called **abelian varieties**. It is a wonderful fact that although the construction of  $J(M)$  given here is transcendental, in fact  $J(M)$  is always an abelian variety. Perhaps inspired by this fact, André Weil constructed a generalization of the Jacobian in higher dimensions that is always an abelian variety. The story is that Griffiths was slated to give a seminar on Weil’s construction, but when it came time to prepare his talk, he discovered the journal was missing from the library. He therefore decided to reconstruct the signs in Weil’s rather intricate construction by assuming that the intermediate Jacobian varies holomorphically in families. In fact, Weil’s intermediate Jacobian does not have this property, and thus Griffiths had inadvertently defined a new intermediate Jacobian, different from Weil’s. It was eventually realized that Griffiths’ construction was the more productive one, and it proved an essential element in his proof that homological equivalence and algebraic equivalence are distinct.

Griffiths’ definition can be expressed rather simply as

$$J^p(M) = \frac{H^{2p-1}(M, \mathbf{C})}{F^p H^{2p-1}(M, \mathbf{C}) + H^{2p-1}(M, \mathbf{Z})}.$$

This agrees with the definition for Riemann surfaces, i.e.  $J^1(M) = J(M)$ . For  $p = 2$ , we have

$$J^2(M) = \frac{H^{1,2}(M) \oplus H^{0,3}(M)}{H^3(M, \mathbf{Z})}.$$

For  $p \geq 2$ ,  $J^p(M)$  is in general not an abelian variety, but it does vary holomorphically in families.

A very nice feature of Griffiths’ definition is that one can define an Abel-Jacobi map that varies holomorphically on families of cycles. If

$$Z \in Z^p(M)_{\text{hom}},$$

then write  $Z = \partial\gamma$  for a  $2n - 2p + 1$ -chain  $\gamma$ . Then for  $\omega \in F^p H^{2p-1}(M, \mathbf{C})$ ,

$$\int_{\gamma} \omega$$

is well-defined, and this gives a map

$$AJ_M: Z^p(M)_{\text{hom}} \rightarrow J^p(M).$$

Griffiths showed that  $Z \in Z^p(M)_{\text{rat}}$  implies that  $AJ_M(Z) = 0$ . This gives a well-defined map

$$AJ_M: \frac{Z^p(M)_{\text{hom}}}{Z^p(M)_{\text{rat}}} \rightarrow J^p(M).$$

For  $p \geq 2$ , it is known that this map is far from being either injective or surjective in general.

Now assume that  $Z \in Z^p(M \times C)$ , where  $C$  is a compact connected Riemann surface. We may view this as a family of algebraic cycles  $Z_t \in M \times \{t\}$ , which we think of as a family of cycles in  $M$ . If  $t_0 \in C$  is a base-point, then

$$Z_t - Z_{t_0} \in Z^p(M)_{\text{alg}} \subseteq Z^p(M)_{\text{hom}}.$$

This gives a holomorphic map

$$C \rightarrow J^p(M)$$

by

$$t \mapsto AJ_M(Z_t - Z_{t_0}).$$

We also have a map  $C \rightarrow J(C)$  by

$$t \mapsto AJ_C(t - t_0).$$

By the universal property of the Jacobian of a curve, the first map factors through a map  $J(C) \rightarrow J^p(M)$ . Now recall that  $J(C)$  is an abelian variety but  $J^p(M)$  need not be. From the above it follows that

$$AJ_M(Z_t - Z_{t_0}) \in J^p(M)_{\text{ab}},$$

where  $J^p(M)_{\text{ab}}$  denotes the maximal abelian variety contained in  $J^p(M)$ . This implies:

**Theorem** (Griffiths).  $AJ_M$  takes  $Z^p(M)_{\text{alg}}$  to  $J^p(M)_{\text{ab}}$ .

There is thus a well-defined map

$$AJ_M: \frac{Z^p(M)_{\text{hom}}}{Z^p(M)_{\text{alg}}} \rightarrow \frac{J^p(M)}{J^p(M)_{\text{ab}}}.$$

In order to show that homological and algebraic equivalence of algebraic cycles are distinct, it is enough to find an example where this map is non-zero.

The problem is that it is difficult to actually evaluate this map. However, the derivative of this map is easier to compute, and this is the strategy adopted by Griffiths.

The argument uses a construction that goes back to Poincaré and Lefschetz, and which Lefschetz used in proving the Hodge Conjecture in codimension 1—**normal functions**. Let  $M \subseteq \mathbb{C}P^N$  be a smooth projective variety. A cycle  $Z \in Z^p(M)$  is called **primitive** if, letting  $H$  be a general hyperplane,

$$[Z] \cdot [H] = 0 \text{ in } H_{2n-2p-2}(M \cap H).$$

Denote this as  $Z^p(M)_{\text{prim}}$ . If  $Z \in Z^p(M)_{\text{prim}}$ , then

$$AJ_{M \cap H}(Z \cdot H) \in J^p(M \cap H).$$

We thus get a map

$$\nu: \{\text{hyperplanes in } \mathbf{C}P^N\} \rightarrow \bigcup_H J^p(M \cap H).$$

If we pick a pencil of hyperplanes  $\{t_1H_1 + t_2H_2\} \cong \mathbf{C}P^1$ , chosen so that  $M \cap H$  has at worst mild singularities for hyperplanes in the pencil, then we may still define the intermediate Jacobians. Let  $\mathcal{J} = \bigcup_H J^p(M \cap H)$ . Restricting  $\nu$  gives a holomorphic map

$$\nu_Z: \mathbf{C}P^1 \rightarrow \mathcal{J}$$

and it is known that

$$\nu_Z = 0 \Leftrightarrow Z \equiv_{\text{hom}} 0.$$

For  $V$  a smooth 4-fold in  $\mathbf{C}P^5$  defined by a homogeneous polynomial of degree 5, Griffiths was able to show that  $Z^p(V)_{\text{prim}}/Z^p(V)_{\text{hom}} \neq 0$  and that for a general hyperplane  $H$ ,

$$J^p(V \cap H)_{\text{ab}} = 0.$$

This is done by assuming  $J^p(V \cap H)_{\text{ab}} \neq 0$  and differentiating the condition that this deforms with  $H$ , and showing by an infinitesimal argument that  $J^p(V \cap H)_{\text{ab}} = 0$  for a general  $H$ . Now for a non-zero  $Z \in Z^p(V)_{\text{prim}}/Z^p(V)_{\text{hom}}$ ,  $\nu_Z \neq 0$  and therefore for a general  $H$ ,

$$AJ_{V \cap H}(Z \cdot H) \notin J^p(V \cap H)_{\text{ab}}.$$

Thus we cannot have

$$Z \cdot H \equiv_{\text{alg}} 0,$$

but by primitivity

$$Z \cdot H \equiv_{\text{hom}} 0.$$

This proves that algebraic and homological equivalence are different.

### 3. The Work of Clemens and Griffiths on Non-Rationality of the Cubic Three-fold

For  $M$  a compact connected complex manifold, the field of meromorphic functions of  $M$  will be denoted  $\mathbf{C}(M)$ . If  $M \subseteq \mathbf{C}P^N$  is a smooth projective variety of dimension  $n$ , then

$$\mathbf{C}(M) = \left\{ \frac{P(z_0, \dots, z_N)}{Q(z_0, \dots, z_N)} \Big|_M \mid P, Q \text{ homogeneous of the same degree and } Q|_M \text{ not } \equiv 0 \right\}.$$

Note that  $\mathbf{C}(\mathbf{C}P^n) \cong \mathbf{C}(x_1, \dots, x_n)$ , where  $\mathbf{C}(x_1, \dots, x_n)$  is the field of rational functions in  $n$  variables. It is known that for  $M$  smooth projective of dimension  $n$ ,  $\mathbf{C}(M)$  is a finite algebraic extension of  $\mathbf{C}(x_1, \dots, x_n)$ .

$M$  is said to be a **rational variety** if  $\mathbf{C}(M) \cong \mathbf{C}(x_1, \dots, x_n)$ .  $M$  is said to be **unirational** if  $\mathbf{C}(M)$  is a finite degree subfield of  $\mathbf{C}(x_1, \dots, x_n)$ —this is equivalent to the geometric condition that  $M$  is the image of  $\mathbf{C}P^n$  under a generically finite rational map.

It was a celebrated problem to know whether unirational implies rational. In dimension 1, this is Luroth's Theorem, and in dimension 2, it is a theorem of Castelnuovo.

A smooth cubic threefold  $M$  is a smooth projective subvariety of  $\mathbf{C}P^4$  defined by a homogeneous polynomial of degree 3. There is a classical geometric construction that shows that such an  $M$  is unirational. Clemens and Griffiths proved that smooth cubic threefolds are not rational.

The essential tool is to study the intermediate Jacobian  $J^2(M)$ . For cubic threefolds,  $H^{3,0}(M) = 0$ , and it follows that  $J^2(M)$  is an abelian variety. In order for  $M$  to be rational, it is necessary that

$$J^2(M) \cong \bigoplus_i J(C_i)$$

for some finite collection of compact connected Riemann surfaces  $C_i$ . For cubic threefolds,  $H^{2,1}(M)$  has dimension 5, and thus so does  $J^2(M)$ . It is known that for  $g \geq 4$ , not every principally polarized abelian variety is a Jacobian or sum of Jacobians. It is thus necessary to find some geometric invariant that distinguishes which abelian varieties cannot be direct sums of Jacobians of compact Riemann surfaces.

For a principally polarized abelian variety  $A$  of dimension  $g$ , there is a **theta-divisor**  $\Theta$  and the classes

$$\eta_k = \frac{1}{(g-k)!} \eta_\Theta^{(g-k)}$$

are Hodge classes.  $A$  is said to have **level**  $k$  if

$$\eta_k = \eta_W$$

for some codimension  $g-k$  algebraic subvariety  $W$ . Note that  $W$  is a subvariety, i.e. a cycle whose coefficients are all  $\geq 0$ . Because  $\Theta$  always exists, all principally polarized abelian varieties have level  $g-1$ . For  $A = J(C)$ ,

$$\eta_1 = \eta_{\text{Im}(AJ_C)},$$

so Jacobians of curves have level 1, and hence so do principally polarized abelian varieties of the form  $\bigoplus_i J(C_i)$ . The proof proceeds by showing that for a smooth cubic threefold  $X$ ,  $J^2(X)$  does not have level 1.

It is 4 conditions for a given line  $L \subset \mathbf{C}P^4$  to be contained in a given cubic threefold  $M$ . The Grassmannian of lines in  $\mathbf{C}P^4$  has dimension 6, and there is indeed a surface of lines contained in  $M$ , known as the **Fano surface of lines**  $S$ . The geometry of  $J^2(M)$  is related to  $S$  by

$$J^2(M) \cong J^2(S).$$

One can then use the geometry of  $S$  to get information about  $J^2(M)$ . In particular,

$$J^2(S) \cong J^2(X),$$

and there is thus, after picking a base point  $s_0 \in S$ , a map

$$\psi: S \rightarrow J^2(S)$$

by

$$s \mapsto AJ_S(s - s_0).$$

Looking at  $\psi(S)$  allows one to show that  $J^2(X)$  has level 2. Results from the classical geometry of the Fano surface of lines are then invoked to show that  $J^2(X)$  does not have level 1, and hence the non-rationality of the cubic threefold.

#### 4. Work of Griffiths with James Carlson and MG on Holomorphic Maps

The classical theorem of Picard is:

**Theorem** (Picard). *A holomorphic map  $f: \mathbf{C} \rightarrow \mathbf{C}P^1 - \{p, q, r\}$  is constant for distinct points  $p, q, r$ .*

This was later generalized to:

**Theorem** (Picard). *A holomorphic map  $f: \mathbf{C} \rightarrow T - \{p\}$  for a 1-dimensional complex torus  $T$  or to a compact Riemann surface  $M$  of genus  $g \geq 2$  is constant.*

What these all have in common is that their simply-connected cover is the disc  $\Delta$ , and the map  $f$  lifts to a holomorphic map  $\tilde{f}: \mathbf{C} \rightarrow \Delta$ , which then is constant by Liouville's Theorem.

Lars Ahlfors realized that the key element in the argument is the fact that the punctured Riemann surfaces have metric of constant negative curvature.

**Theorem** (Ahlfors). *Let  $f: \Delta \rightarrow M$  be a holomorphic map to a (possibly noncompact) Riemann surface that has a metric of Gaussian curvature  $\leq -1$ . Then  $f$  is distance-decreasing, where we use the Poincaré metric of constant negative curvature on  $\Delta$ .*

A way of stating the Picard theorems more uniformly is to take

$$f: \mathbf{C} \rightarrow M - D,$$

where  $M$  is a compact Riemann surface of genus  $g$  and  $D$  is a set of  $d$  points. The condition we need is that

$$2g - 2 + d > 0,$$

which can be rephrased as  $\chi(M) + d > 0$ , where  $\chi(M)$  is the Euler characteristic.

In general, given a compact Kähler manifold  $M$  and a codimension 1 algebraic cycle (called a **divisor**)  $D$  on  $M$ , there is a standard construction of a holomorphic line bundle  $L_D \rightarrow M$  and a meromorphic section  $s_D: M \rightarrow L_D$  such that  $\text{div}(s_D) = D$ , i.e.  $D$  is the divisor of  $s_D$ , counting multiplicities. It is a consequence of Hironaka's resolution of singularities that there is no loss of generality in describing a space as  $M - D$  in assuming that  $D$  has normal crossings.

Algebraic subvarieties  $D$  of codimension 1 on  $M$  are given locally as the zero locus, with multiplicity 1, of an analytic function  $h$  on  $M$ .  $D$  is said to have **normal crossings** if it is possible at every point of  $D$  to find local coordinates  $z_1, \dots, z_n$  for  $M$  such that

$$h = z_1 z_2 \cdots z_k$$

for some  $k \leq n$ .

Given a holomorphic line bundle  $L \rightarrow M$  and if a basis of sections  $s_1, \dots, s_N$  of  $L$  have no common zero, they define a holomorphic map

$$\phi_L: M \rightarrow \mathbf{C}P^{N-1}$$

so that  $z \in M$  maps to the point with homogeneous coordinates  $(s_1(z), \dots, s_N(z))$ . Which local trivialization of  $L$  is used to consider the  $s_i(z)$  as taking values in  $\mathbf{C}$  does not matter, because of the scale factor equivalence of homogeneous coordinates. If  $\phi_L$  is an embedding,

$L$  is said to be **very ample** and if  $\phi_{L^k}$  is an embedding for some  $k$  fold tensor product of  $L$ ,  $k > 0$ ,  $L$  is said to be **ample**. On a compact connected Riemann surface,  $L_D$  is ample if and only if  $\deg(D) > 0$ ; in higher dimensions,  $L_D$  is ample if, for some choice of Hermitian metric on  $L$ ,  $c_1(L)$  is a positive form.

Another way to rephrase the classical Picard theorems is that

$$c_1(\Omega_M^1) + \eta_D$$

is a positive class, or that

$$\Omega_M^1 \otimes L_D$$

is an ample bundle.

The interesting interaction between differential geometry and algebraic geometry drew the attention of Griffiths and his students.

**Theorem** (Griffiths). *Let  $M$  be a compact Kähler manifold of dimension  $n$  such that  $\Omega_M^n$  is a very ample line bundle, i.e. its sections give an embedding of  $M$  in some  $\mathbf{C}P^N$ . Then any holomorphic map*

$$f: \mathbf{C}^n \rightarrow M$$

*is degenerate, i.e.  $df$  is identically of rank  $< n$ .*

This was generalized in joint work with James Carlson:

**Theorem** (Carlson-Griffiths). *Let  $M$  be a compact Kähler manifold of dimension  $n$  and  $D$  a divisor on  $M$  with normal crossings such that  $\Omega_M^n \otimes L_D$  is an ample line bundle on  $M$ . Then any holomorphic map*

$$f: \mathbf{C}^n \rightarrow M - D$$

*is degenerate, i.e.  $df$  is identically of rank  $< n$ .*

In fact, they prove a Nevanlinna defect relation type statement, a quantitative result on how often  $f(\mathbf{C}^n)$  meets  $D$ .

The situation for non-equidimensional maps

$$f: \mathbf{C} \rightarrow M - D$$

is more complicated. An example of Peter Kiernan showed that there are non-constant maps  $f: \mathbf{C} \rightarrow \mathbf{C}P^2 - C$ , where

$$C = \{z_0^d + z_1^d + z_2^d = 0\}$$

is the Fermat curve of degree  $d$ . After considering many examples, Griffiths and I formulated the following: A map  $f: \mathbf{C} \rightarrow M - D$  is **algebraically degenerate** in the situation where  $f(\mathbf{C})$  is contained in a lower-dimensional algebraic subvariety of  $M$ .

**Conjecture** (Green-Griffiths). *For  $M, D$  as in the Carlson-Griffiths theorem above, any holomorphic map  $f: \mathbf{C} \rightarrow M - D$  is algebraically degenerate.*

This conjecture helped to inspire some conjectures of Vojta [56] and Lang [53, 54] on rational points.

The line of thought coming out of the Ahlfors Lemma and the Griffiths and Carlson-Griffiths work led to the question of what algebro-geometric objects on a smooth projective variety  $M$  might be used to show that holomorphic maps from  $\mathbf{C}$  to  $M$  must be algebraically

degenerate. The classical Picard theorem can be redone using Griffiths' argument using holomorphic 1-forms. With the work of Bogomolov on **symmetric differentials**, i.e. sections of the symmetric product  $S^k \Omega_M^1$ , it was realized that these can also be used. Griffiths and I introduced **jet differentials**, objects which pounce on the  $k$ -jet of a holomorphic map from the disc and produce a number that is changed by  $\lambda^m$  for some  $m$  when the map is reparametrized by  $z \mapsto \lambda z$ . These are sections of a vector bundle  $\mathcal{J}_{k,m} \rightarrow M$ . It is possible to compute the Euler characteristic of this bundle and to show that it grows with the highest possible degree for a variety of general type. Going from that to getting sections requires controlling the higher cohomology groups in even degrees. For surfaces, this is not difficult, but only in recent work by Demailly has it been shown that the higher cohomology groups grow more slowly than  $H^0(\mathcal{J}_{k,m})$ .

**Theorem** (Green-Griffiths for  $n = 2$ , Demailly for all  $n$ ). *For  $X$  a smooth  $n$ -fold of general type, for  $k$  sufficiently large,  $H^0(\mathcal{J}_{k,m})$  grows as a polynomial in  $m$  of the maximal possible degree.*

### 5. Work of Griffiths with Robert Bryant, Eric Berger and Deane Yang on Isometric Embeddings

On a  $C^\infty$  manifold  $M$  of dimension  $n$  with local coordinates  $x_1, \dots, x_n$ , one denotes by  $\frac{\partial}{\partial x_i}$  the vector field whose directional derivative of any function  $f$  is  $\frac{\partial f}{\partial x_i}$ . A **Riemannian metric** on  $M$  is a smoothly varying positive definite inner product  $\langle, \rangle$  where

$$g_{ij} = \langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \rangle$$

is a positive-definite symmetric matrix  $G = (g_{ij})$ .

If  $F: M_n \rightarrow \mathbf{R}^N$  is an embedding,  $F = (f_1, \dots, f_N)$ , then the **metric induced by  $F$**  is

$$g_{ij} = \langle \frac{\partial F}{\partial x_i}, \frac{\partial F}{\partial x_j} \rangle = \sum_k \frac{\partial f^k}{\partial x_i} \frac{\partial f^k}{\partial x_j}.$$

The **isometric embedding problem** is, given  $M$  with a Riemannian metric  $G$ , does there exist an  $F: M_n \rightarrow \mathbf{R}^N$  whose induced metric is  $G$ . The problem can be **local**, if we just want to do this on each small open neighborhood of  $M$ , or **global** if we want to do it on all of  $M$  at once.

The matrix  $G$  has  $\binom{n+1}{2}$  independent entries, so if  $N = \binom{n+1}{2}$ , the isometric embedding problem becomes  $\binom{n+1}{2}$  PDE's in  $\binom{n+1}{2}$  unknown functions.

**Theorem** (Cartan-Janet). *The local isometric embedding problem can be solved in the real-analytic category for  $N = \binom{n+1}{2}$ .*

John Nash [55] solved the global isometric embedding problem, which requires large values of  $N$ . Robert Greene [28] solved the  $C^\infty$  local isometric embedding problem for  $N = \binom{n+1}{2} + n$ .

A very nice uniqueness result is:

**Theorem** (Berger-Bryant-Griffiths). *For  $N \leq \binom{n}{2}$ , the solutions to the local isometric embedding problem for  $f: M_n \rightarrow \mathbf{R}^N$  for a “general”  $G$  depend on at most a finite number of constants.*

In this paper, the focus will be on existence results for the local isometric embedding problem with  $N = \binom{n+1}{2}$ .

For  $n = 2$ , the classical result is that the local  $C^\infty$  isometric embedding problem  $F: M_2 \rightarrow \mathbf{R}^3$  is solvable if the Gaussian curvature  $K$  is nowhere 0.

**Theorem** (Bryant-Griffiths-Yang). *The local  $C^\infty$  isometric embedding problem*

$$F: M_3 \rightarrow \mathbf{R}^6$$

*is solvable if the Einstein curvature is  $\neq 0$  and is not a square.*

The solution of this problem involves some very interesting issues in partial differential equations, in particular, the case when the characteristic variety is smooth but codimension 1.

For a partial differential operator, the **symbol** is the “leading term,” i.e. the term involving the highest order derivatives. For example, for the Laplacian

$$\Delta = \sum_i \frac{\partial^2}{\partial x_i^2},$$

the symbol is

$$\sum_i \xi_i^2.$$

For the wave operator

$$\frac{\partial^2}{\partial x_0^2} - \sum_i \frac{\partial^2}{\partial x_i^2},$$

the symbol is

$$\xi_0^2 - \sum_i \xi_i^2.$$

The **characteristic variety** of the operator  $D$  is the zero locus of its symbol  $\sigma_D$ , and denoted here as  $\Xi_D$ . Because  $\sigma_D$  is homogeneous, the zero locus is a cone, and we may thus regard

$$\Xi_D \subseteq \mathbf{R}P^{n-1}$$

if  $D$  in a PDO in  $n$  variables. For the Laplacian,  $\Xi_\Delta = \emptyset$ , while for the wave operator it is non-empty. **Elliptic operators** have empty characteristic variety. The difficulty in solving a system of PDE’s is related to how complicated the characteristic variety is.

A generalization of a system of PDE’s is an **exterior differential system**. These were studied by Élie Cartan and were the subject of a monograph by Bryant, Chern, Gardner, Goldschmidt and Griffiths. Although EDS will not be used in this exposition, this is the framework used by Bryant-Griffiths-Yang.

If we choose a point  $p \in M$  such that  $F(p) = (0, \dots, 0)$  and the tangent space

$$T_0(F(M)) = \{x_\nu = 0 \mid \nu > n\}$$

then we can take  $x_1, \dots, x_n$  as local coordinates and then for  $\nu > n$ ,

$$x_\nu = f^\nu(x_1, \dots, x_n) = Q^\nu(x_1, \dots, x_n) + \text{higher order terms},$$

where  $Q^\nu$  is a homogeneous quadric. This turns out to give an intrinsic map

$$H: N_p(M) \rightarrow S^2T_p(M)^*$$

by

$$\frac{\partial}{\partial x_\nu} \mapsto Q^\nu.$$

This is a classical object, the **second fundamental form of  $M$  at  $p$** . The isometric embedding equations are

$$g_{ij} = \delta_{ij} + \sum_{\nu > n} \frac{\partial f^\nu}{\partial x_i} \frac{\partial f^\nu}{\partial x_j}.$$

Bryant-Griffiths-Yang consider the variational equation we get by starting from a given embedding and seeing what the relationship is between an infinitesimal change in the  $f^\nu$  and the  $g_{ij}$  is. Letting  $\dot{g}_{ij}$  and  $\dot{f}^\nu$  represent time derivatives, we have

$$\dot{g}_{ij} = \sum_{\nu > n} \frac{\partial \dot{f}^\nu}{\partial x_i} \frac{\partial f^\nu}{\partial x_j} + \frac{\partial f^\nu}{\partial x_j} \frac{\partial \dot{f}^\nu}{\partial x_i}.$$

This is a system of first order linear differential operators—the point of considering the variational problem is that it linearizes the isometric embedding problem. We may without loss of generality assume that at  $p$  that the position of  $F(p)$  and  $T_{F(p)}(F(M))$  do not change, and thus

$$f^\nu = Q^\nu + \text{higher order terms}$$

and

$$\dot{f}^\nu = \dot{Q}^\nu + \text{higher order terms}.$$

In order to get something that does not vanish at  $p$ , we must take the second partials of the equation for  $\dot{g}_{ij}$ . We obtain

$$\begin{aligned} \frac{\partial^2 \dot{g}_{ij}}{\partial x_k \partial x_l} &= \sum_{\nu > n} \frac{\partial^2 Q^\nu}{\partial x_j \partial x_k} \frac{\partial^2 \dot{Q}^\nu}{\partial x_i \partial x_l} + \frac{\partial^2 Q^\nu}{\partial x_i \partial x_k} \frac{\partial^2 \dot{Q}^\nu}{\partial x_j \partial x_l} + \frac{\partial^2 Q^\nu}{\partial x_i \partial x_l} \frac{\partial^2 \dot{Q}^\nu}{\partial x_j \partial x_k} + \\ &\quad \frac{\partial^2 Q^\nu}{\partial x_j \partial x_l} \frac{\partial^2 \dot{Q}^\nu}{\partial x_i \partial x_k} + \text{higher order terms} \end{aligned}$$

The left-hand side of this lies in  $S^2T_p(M)^* \otimes S^2T_p(M)^*$ . The intrinsic part of this, independent of adding quadratic terms to local coordinates  $x_1, \dots, x_n$ , is the image of

$$\wedge^2T_p(M)^* \otimes \wedge^2T_p(M)^* \rightarrow S^2T_p(M)^* \otimes S^2T_p(M)^*$$

We thus look at

$$\frac{1}{2} \left( \frac{\partial^2 \dot{g}_{ij}}{\partial x_k \partial x_l} + \frac{\partial^2 \dot{g}_{kl}}{\partial x_i \partial x_j} - \frac{\partial^2 \dot{g}_{kj}}{\partial x_i \partial x_l} - \frac{\partial^2 \dot{g}_{il}}{\partial x_k \partial x_j} \right)$$

which equals

$$\sum_{\nu > n} \frac{\partial^2 Q^\nu}{\partial x_j \partial x_k} \frac{\partial^2 \dot{Q}^\nu}{\partial x_i \partial x_l} - \frac{\partial^2 Q^\nu}{\partial x_i \partial x_k} \frac{\partial^2 \dot{Q}^\nu}{\partial x_j \partial x_l} + \frac{\partial^2 Q^\nu}{\partial x_i \partial x_l} \frac{\partial^2 \dot{Q}^\nu}{\partial x_j \partial x_k} - \frac{\partial^2 Q^\nu}{\partial x_j \partial x_l} \frac{\partial^2 \dot{Q}^\nu}{\partial x_i \partial x_k} + \text{higher order terms.}$$

The right-hand side of the equation, without the higher order terms, has a symmetry which places it in  $K$ , the space of curvature like tensors, where  $K \cong (T_p^*)^{(2,2)}$ , the representation with Young diagram a  $2 \times 2$  square.

The equation above may be expressed by saying that

$$\dot{H}_p: N_p(M) \rightarrow S^2 T_p(M)^*$$

is the unknown and  $H_p$  pairs with it to give an element of  $K$ . This pairing is the polarization of the **Gauss equations** relating the (extrinsic) second fundamental form to the (intrinsic) Riemannian curvature. The symbol of this prolonged system  $D$  of linear PDE's becomes

$$\sigma_D: N_p(M)^* \otimes S^2 T_p(M)^* \rightarrow K.$$

The **characteristic sheaf**  $\mathcal{M}_D$  of a system of linear PDE's is the cokernal of dual of  $\sigma_D$ , viewed as a map of vector bundles on  $\mathbf{C}P(T_p(M))$ . In our case,

$$\sigma_D^*: K^* \otimes \mathcal{O}_{P(T_p(M))} \rightarrow N_p(M) \otimes \mathcal{O}_{P(T_p(M))}(2).$$

Thus

$$K^* \otimes \mathcal{O}_{P(T_p(M))} \rightarrow N_p(M) \otimes \mathcal{O}_{P(T_p(M))}(2) \rightarrow \mathcal{M}_D \rightarrow 0$$

is exact. This map is just pairing with  $H_p$ . The **characteristic variety of  $D$**  is

$$\Xi_D = \text{supp}(\mathcal{M}_D).$$

We are interested in the real points of  $\Xi_D$ .

Unwinding,  $\Xi_D$  is the locus where for some  $w \in N_p(M)$ , there is a point of  $P(T_p(M))$  where  $H_p(w)$  maps to 0 in  $K$ .

If  $Q = \sum_{\nu > n} w^\nu Q^\nu$  and  $Q_{ij} = \partial^2 Q / \partial x_i \partial x_j$ , then for the point  $(1, 0, \dots, 0)$  to be in  $\Xi_D$ , the condition is that for a symmetric matrix  $(u_{ij})$ , the coefficient of  $u_{11}$  in

$$Q_{jk}u_{il} + Q_{il}u_{jk} - Q_{ik}u_{jl} - Q_{jl}u_{ik}$$

is zero for all choices of indices, i.e that

$$\delta_{1i}\delta_{1l}Q_{jk} + \delta_{1j}\delta_{1k}Q_{il} - \delta_{1j}\delta_{1l}Q_{ik} - \delta_{1i}\delta_{1k}Q_{jl} = 0$$

for all choices of indices. This forces  $Q_{jk} = 0$  unless either  $j = 1$  or  $k = 1$ , as otherwise the first term is non-zero for  $i = l = 1$  and the other three terms are all 0. But  $Q_{jk} = 0$  unless either  $j = 1$  or  $k = 1$  is equivalent to saying that  $Q = x_1 L$  for some linear factor  $L$ . Conversely, if  $Q = x_1 L$ , one may easily compute to see that the coefficient of  $u_{11}$  is 0. Thus, decoupling from the choice of the point  $(1, 0, \dots, 0)$ , we have

$$\Xi_D = \{\xi \in T_p(M)^* \mid \sum_{\nu > n} w^\nu Q^\nu = \xi \mu \text{ for some } \mu \in T_p(M)^* \text{ \& some } w = (w^\nu) \in N_p(M)\}.$$

If

$$\xi = \sum_{i=1}^n \xi_i x_i, \quad \mu = \sum_{i=1}^n \mu_i x_i,$$

then the conditions for  $\xi \in \Xi_D$  are that we can find a non-zero vector  $w = (w^\nu) \in N_p(M)$  and  $\mu = (\mu_i) \in T_p(M)^*$  such that

$$\sum_{\nu > n} w^\nu Q_{ij}^\nu = \xi_i \mu_j + \xi_j \mu_i$$

for all  $i, j$ . This is  $\binom{n+1}{2}$  equations in the  $\binom{n}{2} + n$  unknowns  $w^\nu, \mu_i$ . The matrix of coefficients is a square matrix of the form  $(A \ B)$ , where  $A$  has  $\binom{n}{2}$  columns, and involves the  $Q_{ij}^\nu$  and  $B$  has  $n$  columns and involves the  $\xi_i$ . The condition that a solution exists is that the determinant of this matrix vanishes, and this is one equation homogeneous of degree  $n$  in the  $\xi_i$ . We thus conclude:

**Theorem** (Bryant-Griffiths-Yang). *For a general isometric embedding,  $\Xi_D \subseteq \mathbf{R}P^{n-1}$  is a hypersurface of degree  $n$ .*

When  $n = 2$ , this hypersurface is a quadric in  $\mathbf{R}P^1$ , i.e. it is either 2 points, one point, or empty. These correspond to the cases where the Gaussian curvature  $K$  is negative, zero or positive. In case of negative curvature, the two points are the **asymptotic directions** for the surface, the directions where the second fundamental form vanishes.

When  $n = 3$ , one has a real cubic curve in  $\mathbf{R}P^2$ , i.e. an elliptic curve. For the same reasons that a real cubic polynomial in one variable always has a real root, such a real elliptic curve is never empty.

Returning to the original linearized equations,

$$\dot{g}_{ij} = \sum_{\nu > n} \frac{\partial f^\nu}{\partial x_i} \frac{\partial \dot{f}^\nu}{\partial x_j} + \frac{\partial f^\nu}{\partial x_j} \frac{\partial \dot{f}^\nu}{\partial x_i},$$

we may rewrite this as

$$\dot{g}_{ij} = \sum_{\nu > n} \frac{\partial}{\partial x_j} \left( \frac{\partial f^\nu}{\partial x_i} \dot{f}^\nu \right) + \frac{\partial}{\partial x_i} \left( \frac{\partial f^\nu}{\partial x_j} \dot{f}^\nu \right) - 2 \frac{\partial^2 f^\nu}{\partial x_i \partial x_j} \dot{f}^\nu.$$

Now  $\dot{g}_{ij} \in S^2 T_p(M)^*$  and the last term of the equation above is just  $Q^\nu \dot{f}^\nu \in S^2 T_p(M)^*$ . If we work in  $S^2 T_p(M)^* / \text{Im}(H_p)$ , the last term goes away and we get a system of linear differential operators  $D_0$  that goes between the vector bundles

$$T_p(M)^* \rightarrow S^2 T_p(M)^* / \text{Im}(H_p).$$

If  $H_p$  is generic, and in particular if it is of maximal rank  $\dim(N_p(M)) = \binom{n}{2}$ , then  $D_0$  is a system of linear differential operators between vector bundles of the same rank! In the language of PDE's, this says that the system is **determined** when  $N = \binom{n+1}{2}$ . A similar computation for the more general situation of an isometric embedding

$$f: M_n \rightarrow \mathbf{R}^N$$

with  $H_p$  generic gives that  $D_0$  maps  $T_p(M)^*$  to a vector bundle of the following ranks:

- (1)  $\binom{n+1}{2} - (N - n)$  if  $N \leq \binom{n+1}{2} + n$ ;
- (2) 0 if  $N \geq \binom{n+1}{2} + n$ .

In case (2), one can eliminate all of the partial derivatives, and thus obtain an algebraic problem to solve for the  $f^\nu$ . This is the situation of the local theorem of Robert Greene alluded to early in this section, and also in the work of John Nash on the global isometric embedding problem. By (1), if  $\binom{n+1}{2} + n > N > \binom{n+1}{2}$ , we get an **overdetermined system**, and if  $N < \binom{n+1}{2}$ , we get an **undetermined system**.

In the context of exterior differential systems, the process of going from  $D$  to  $D_0$  is one of **deprolongation**, reversing the basic **prolongation** construction of EDS.

The final steps in solving the local isometric embedding problem for  $f: M_3 \rightarrow \mathbf{R}^6$  involve some delicate estimates from the theory of partial differential equations which are neither elliptic nor hyperbolic, and hinge on an analysis of the geometry of the real elliptic curve  $\Xi_D$ , which for example can have one or two connected components.

For higher  $n$ , the determinantal hypersurface one gets for  $\Xi_D$  can be singular, as determinantal hypersurfaces in general are singular along the locus where they drop rank by at least 2, and this has codimension 4 in general.

## 6. Work of Griffiths with MG on Tangent Spaces to Algebraic Cycles

This work is intentionally speculative. Due to Griffiths and Wilfried Schmid in one parameter and Cattani-Kaplan-Schmid in several parameters, there is a well-developed variational theory for Hodge structures. While there is a variational theory for algebraic subvarieties, there is no comparable theory for algebraic cycles. There are formidable obstacles to being able to use such a theory to go from infinitesimal information to a geometric deformation of a cycle, and even formal deformations that are unobstructed at all levels may be obstructed geometrically.

If we have a point  $p$  on a smooth projective variety  $M$ , the tangent space of 0-dimensional submanifolds of  $M$  at  $p$  is just  $T_p(M)$ . As a subvariety, if  $p$  is non-reduced, it can have a very complicated ideal, which then can vary, and one gets the **Hilbert scheme**. As an algebraic cycle, there is a different problem—by analogy with physics, a “particle and antiparticle” can appear, e.g we can deform  $p$  as  $p(t) + q(t) - r(t)$ , where  $p(0) = q(0) = r(0) = p$ .

For codimension 1, the sheaf of divisors  $\mathcal{D}_M$  sits in an exact sequence

$$0 \rightarrow \mathcal{O}_M^* \rightarrow \mathcal{M}_M^* \rightarrow \mathcal{D}_M \rightarrow 0,$$

where  $\mathcal{O}_M^*$  and  $\mathcal{M}_M^*$  are respectively the sheaves of nowhere vanishing holomorphic and meromorphic functions on  $M$ . The tangent spaces to this sequence gives an exact sequence

$$0 \rightarrow \mathcal{O}_M \rightarrow \mathcal{M}_M \rightarrow \mathcal{PP}_M \rightarrow 0,$$

where  $\mathcal{PP}_M$  is the sheaf of **principal parts on  $M$** . A family of divisors is given locally by  $\text{div}(f(t))$ , and if

$$f(t) = f + t\dot{f} + \text{higher order terms},$$

the derivative is the principal part  $\dot{f}/f$ .

For 0-cycles on  $M$  with  $n$  the dimension of  $M$ , the expression for the tangent space to cycles is

$$T_x Z^n(M) = \lim_{i \rightarrow \infty} \text{Ext}_{\mathcal{O}_{M,x}}^n(\mathcal{O}_M/m_x^i, \Omega_{M/\mathbf{Q}}^{n-1}) \cong H_x^n(\Omega_{M/\mathbf{Q}}^{n-1}).$$

Here  $m_x$  is the ideal of local holomorphic functions on  $M$  vanishing at  $x$ , and  $\Omega_{M/\mathbf{Q}}^{n-1}$  is the sheaf of **Kähler differentials** of degree  $n - 1$  on  $M$  over  $\mathbf{Q}$ . The Kähler differentials over  $\mathbf{Q}$  are defined in the same way as their more geometric relative, the Kähler differentials over  $\mathbf{C}$ , except that for constants,

$$dc = 0$$

is only required for  $c \in \mathbf{Q}$ . Thus regarding  $r$  as a function,

$$d(\pi r^2) = 2\pi r dr + r^2 d\pi.$$

We need to be in the algebraic category, i.e. no transcendental functions. Finally,  $H_x^n(\Omega_{M/\mathbf{Q}}^{n-1})$  is the local cohomology at  $x$ .

The  $p$ 'th **Chow group** is

$$CH^p(M) = \frac{Z^p(M)}{Z^p(M)_{\text{rat}}}.$$

The Bloch-Quillen theorem describes this in terms of algebraic K-theory as

$$CH^p(M) \cong H^p(\mathcal{K}_p(\mathcal{O}_M)),$$

where cohomology is taken in the Zariski topology. Spencer Bloch gave the formal relation

$$TCH^p(M) \cong H^p(\Omega_{M/\mathbf{Q}}^{p-1}).$$

Now there is a natural map

$$H_x^p(\Omega_{M/\mathbf{Q}}^{p-1}) \rightarrow H^p(\Omega_{M/\mathbf{Q}}^{p-1})$$

which is the map, for  $p = n$ ,

$$T_x Z^n(M) \rightarrow TCH^n(M).$$

Geometrically, one must take these formal statements with a grain of salt. Not only can the first-order deformation of cycles be obstructed, but also so can the first-order deformation of rational equivalences. For example, on an  $M$  with  $H^{2,0}(M) = 0$ , we have

$$H^2(\Omega_{M/\mathbf{Q}}^1) = \ker(H^2(\Omega_{M/\mathbf{C}}^1) \rightarrow \Omega_{\mathbf{C}/\mathbf{Q}}^1 \otimes_{\mathbf{Q}} H^3(\mathcal{O}_M)).$$

However, to have a geometric global 1-parameter family of codimension 2 cycles, we need to be mapping to  $J^2(M)_{\text{ab}}$ , and this is a Hodge-theoretic obstruction to a tangent vector being tangent to a geometric family.

There has been ongoing interesting work in this direction [5, 57].

## 7. Work of Griffiths with Matt Kerr and MG on Mumford-Tate Groups and Mumford-Tate Domains

A **polarized Hodge structure of weight**  $k$  is a vector space  $V_{\mathbf{Q}}$  over  $\mathbf{Q}$  together with a pairing

$$Q: V_{\mathbf{Q}} \times V_{\mathbf{Q}} \rightarrow \mathbf{Q}$$

and a decomposition of  $V_{\mathbf{C}} = V \otimes_{\mathbf{Q}} \mathbf{C}$  as

$$V_{\mathbf{C}} = \bigoplus_{p+q=k} V^{p,q},$$

where the  $V^{p,q}$  are complex subspaces satisfying  $\bar{V}^{p,q} = V^{q,p}$ . The polarization condition is that  $Q$  is symmetric when  $k$  is even and antisymmetric when  $k$  is odd, and:

- (1) **Riemann-Hodge I:**  $Q(V^{p,q}, V^{p',q'}) = 0$  unless  $(p, q) = (q', p')$ ;
- (2) **Riemann-Hodge II:** The Hermitian form  $H(v, w) = i^{p-q}Q(v, \bar{w})$  is positive-definite on  $V^{p,q}$ .

There is a natural way to make  $\bigotimes_a V \otimes \bigotimes_b V^*$  into a Hodge structure. A **(rational) Hodge class** for  $V$  is an element, if  $k = 2p$ , of  $V_{\mathbf{Q}} \cap V^{p,p}$ . A **Hodge tensor of type**  $a, b$  for  $V$  is a rational Hodge class for  $\bigotimes_a V \otimes \bigotimes_b V^*$ . The **Hodge tensors for**  $V$  are the direct sum of the Hodge tensors for  $V$  of type  $a, b$ .

The **Mumford-Tate group of**  $V$  is

$$MT(V) = \{L \in \text{Aut}(V_{\mathbf{R}}, Q) \mid L \text{ fixes the Hodge tensors of } V\}.$$

This is an algebraic group over  $\mathbf{Q}$ .

A linear isomorphism  $L \in \text{Aut}(V_{\mathbf{R}}, Q)$  preserving  $Q$  is an **isomorphism of Hodge structures** if  $L(V^{p,q}) = V^{p,q}$  for all  $p, q$ ; let  $H_V$  denote this group. The **Mumford-Tate domain associated to**  $V$  is

$$D = MT(V)/H_V.$$

The question this work is about is: What reductive algebraic groups can be Mumford-Tate groups? Characterize the ways that such a group be realized as a Mumford-Tate group? What are the associated Hodge structures having this Mumford-Tate group? What is the geometry of Mumford-Tate domains?

An especially interesting case was that of the exceptional Lie group  $G_2$  (non-compact real form), which can be a Mumford-Tate group.

The condition to be a Mumford-Tate group is that  $G_{\mathbf{R}}$  must have a real maximal compact torus, i.e. a real torus  $T$  whose dimension is the rank of  $G$ . What makes this especially interesting is that that this is the same condition as the condition that arises in representation theory for  $G$  to have **discrete series**.

## References

- [1] Arbarello, E., Cornalba, M., Griffiths, P. A., and Harris, J., *Geometry of algebraic curves*, Vol. I. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 267. Springer-Verlag, New York, 1985.

- [2] Arbarello, Enrico; Cornalba, Maurizio; Griffiths, Phillip A., *Geometry of algebraic curves*, Volume II. With a contribution by Joseph Daniel Harris. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 268. Springer, Heidelberg, 2011.
- [3] Berger, Eric; Bryant, Robert; Griffiths, Phillip, *Some isometric embedding and rigidity results for Riemannian manifolds*, Proc. Nat. Acad. Sci. U.S.A. **78** (1981), no. 8, part 1, 4657–4660.
- [4] Berger, Eric; Bryant, Robert; Griffiths, Phillip, *The Gauss equations and rigidity of isometric embeddings*, Duke Math. J. **50** (1983), no. 3, 803–892.
- [5] Bloch, S.; Esnault, E.; Kerz, M., *Deformations of algebraic cycles in characteristic zero*, preprint.
- [6] Bryant, R. L.; Chern, S. S.; Gardner, R. B.; Goldschmidt, H. L.; Griffiths, P. A., *Exterior differential systems*, Mathematical Sciences Research Institute Publications, 18. Springer-Verlag, New York, 1991.
- [7] Bryant, Robert L.; Griffiths, Phillip A., *Characteristic cohomology of differential systems. I*, General theory. J. Amer. Math. Soc. **8** (1995), no. 3, 507–596.
- [8] Bryant, Robert L.; Griffiths, Phillip A., *Characteristic cohomology of differential systems. II*, Conservation laws for a class of parabolic equations. Duke Math. J. **78** (1995), no. 3, 531–676.
- [9] Bryant, Robert; Griffiths, Phillip; Grossman, Daniel, *Exterior differential systems and Euler-Lagrange partial differential equations*, Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 2003.
- [10] Bryant, R.; Griffiths, P.; Hsu, L., *Hyperbolic exterior differential systems and their conservation laws. I*, Selecta Math. (N.S.) **1** (1995), no. 1, 21–112.
- [11] ———, *Hyperbolic exterior differential systems and their conservation laws. II*, Selecta Math. (N.S.) **1** (1995), no. 2, 265–323.
- [12] Bryant, Robert L.; Griffiths, Phillip A.; Yang, *Deane Characteristics and existence of isometric embeddings*, Duke Math. J. **50** (1983), no. 4, 893–994.
- [13] Carlson, James; Green, Mark; Griffiths, Phillip, *Variations of Hodge structure considered as an exterior differential system: old and new results*, SIGMA Symmetry Integrability Geom. Methods Appl. **5** (2009)
- [14] Carlson, James; Green, Mark; Griffiths, Phillip; Harris, Joe, *Infinitesimal variations of Hodge structure. I*, Compositio Math. **50** (1983), no. 2-3, 109–205.
- [15] Carlson, James; Griffiths, Phillip, *A defect relation for equidimensional holomorphic mappings between algebraic varieties*, Ann. of Math. (2) **95** (1972), 557–584.
- [16] Carlson, James A.; Griffiths, Phillip A., *Infinitesimal variations of Hodge structure and the global Torelli problem*, Journées de Géométrie Algébrique d'Angers, Juillet 1979/Algebraic Geometry, Angers, 1979, pp. 51–76, Sijthoff & Noordhoff, Alphen aan den Rijn–Germantown, Md., 1980.
- [17] Carlson, James; Griffiths, Phillip, *What is ... a period domain?*, Notices Amer. Math. Soc. **55** (2008), no. 11, 1418–1419.
- [18] Clemens, C. Herbert; Griffiths, Phillip A., *The intermediate Jacobian of the cubic threefold*, Ann. of Math. (2) **95** (1972), 281–356.

- [19] Cornalba, Maurizio; Griffiths, Phillip A., *Some transcendental aspects of algebraic geometry*, Algebraic geometry (Proc. Sympos. Pure Math., Vol. 29, Humboldt State Univ., Arcata, Calif., 1974), pp. 3D110. Amer. Math. Soc., Providence, R.I., 1975.
- [20] Deligne, Pierre; Griffiths, Phillip; Morgan, John; Sullivan, Dennis, *Real homotopy theory of Kähler manifolds*, Invent. Math. **29** (1975), no. 3, 245–274.
- [21] Green, Mark; Griffiths, Phillip, *Algebraic cycles and singularities of normal functions*, Algebraic cycles and motives. Vol. 1, 206–263, London Math. Soc. Lecture Note Ser. **343**, Cambridge Univ. Press, Cambridge, 2007.
- [22] ———, *Algebraic cycles and singularities of normal functions. II*, Inspired by S. S. Chern, 179–268, Nankai Tracts Math. **11**, World Sci. Publ., Hackensack, NJ, 2006.
- [23] ———, *Formal deformation of Chow groups*, The legacy of Niels Henrik Abel, 467–509, Springer, Berlin, 2004.
- [24] ———, *On the tangent space to the space of algebraic cycles on a smooth algebraic variety*, Annals of Mathematics Studies, **157**, Princeton University Press, Princeton, NJ, 2005
- [25] ———, *Two applications of algebraic geometry to entire holomorphic mappings*, The Chern Symposium 1979 (Proc. Internat. Sympos., Berkeley, Calif., 1979), pp. 41–74, Springer, New York-Berlin, 1980.
- [26] Green, Mark; Griffiths, Phillip; Kerr, Matt, *Hodge theory, complex geometry, and representation theory*, CBMS Regional Conference Series in Mathematics, 118. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 2013. iv+308 pp.
- [27] ———, *Mumford-Tate groups and domains*, Their geometry and arithmetic. Annals of Mathematics Studies, **183**, Princeton University Press, Princeton, NJ, 2012.
- [28] Greene, Robert E., *Isometric embeddings of Riemannian and pseudo-Riemannian manifolds*, Memoirs of the American Mathematical Society, No. 97 American Mathematical Society, Providence, R.I. 1970.
- [29] Griffiths, Phillip A. , *An introduction to the theory of special divisors on algebraic curves*, CBMS Regional Conference Series in Mathematics **44**, American Mathematical Society, Providence, R.I., 1980.
- [30] Griffiths, Phillip, *Hodge theory and geometry*, Bull. London Math. Soc. **36** (2004), no. 6, 721–757.
- [31] Griffiths, Phillip A., *Holomorphic mapping into canonical algebraic varieties*, Ann. of Math. (2) **93** (1971), 439–458.
- [32] ———, *Holomorphic mappings: Survey of some results and discussion of open problems*, Bull. Amer. Math. Soc. **78** (1972), 374–382.
- [33] ———, *Complex analysis and algebraic geometry*, Bull. Amer. Math. Soc. (N.S.) **1** (1979), no. 4, 595–626.
- [34] ———, *Differential geometry and complex analysis*, Differential geometry (Proc. Sympos. Pure Math., Vol. XXVII, Part 2, Stanford Univ., Stanford, Calif., 1973), pp. 43–64. Amer. Math. Soc., Providence, R. I., 1975.
- [35] ———, *Infinitesimal variations of Hodge structure. III*, Determinantal varieties and the infinitesimal invariant of normal functions. Compositio Math. **50** (1983), no. 2–3,

- 267–324.
- [36] ———, *Introduction to algebraic curves*, Translated from the Chinese by Kuniko Weltin. Translations of Mathematical Monographs, 76. American Mathematical Society, Providence, RI, 1989.
- [37] ———, *On the periods of certain rational integrals. I, II*. Ann. of Math. (2) **90** (1969), 460–495; *ibid.* (2) **90** (1969), 496–541.
- [38] ———, *Periods of integrals on algebraic manifolds. I*, Construction and properties of the modular varieties. Amer. J. Math. **90** (1968), 568–626.
- [39] ———, *Periods of integrals on algebraic manifolds. II*, Local study of the period mapping. Amer. J. Math. **90** (1968), 805–865.
- [40] ———, *Periods of integrals on algebraic manifolds. III*, Some global differential-geometric properties of the period mapping. Inst. Hautes Études Sci. Publ. Math. No. 38 (1970), 125–180.
- [41] ———, *Periods of integrals on algebraic manifolds: Summary of main results and discussion of open problems*, Bull. Amer. Math. Soc. **76** (1970), 228–296.
- [42] ———, *Poincaré and algebraic geometry*, Bull. Amer. Math. Soc. (N.S.) **6** (1982), no. 2, 147–159.
- [43] ———, *S. S. Chern: always changing, always the same*, Chern—great geometer of the twentieth century, 117–120, Int. Press, Hong Kong, 1992.
- [44] Griffiths, Phillip, *The legacy of Abel in algebraic geometry*, The legacy of Niels Henrik Abel, 179–205, Springer, Berlin, 2004.
- [45] ———, *Topics in algebraic and analytic geometry*, Written and revised by John Adams. Mathematical Notes, No. 13. Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1974.
- [46] Griffiths, Phillip, ed., *Topics in transcendental algebraic geometry*, (Princeton, N.J., 1981/1982), Ann. of Math. Stud. **106**, Princeton Univ. Press, Princeton, N.J., 1984.
- [47] Griffiths, Phillip; Harris, Joseph, *Infinitesimal variations of Hodge structure. II*, An infinitesimal invariant of Hodge classes. Compositio Math. **50** (1983), no. 2–3, 207–265.
- [48] ———, *On the variety of special linear systems on a general algebraic curve*, Duke Math. J. **47** (1980), no. 1, 233–272.
- [49] ———, *Principles of algebraic geometry*, Reprint of the 1978 original. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1994.
- [50] Griffiths, Phillip; Morgan, John, *Rational homotopy theory and differential forms*, Second edition. Progress in Mathematics **16**, Springer, New York, 2013.
- [51] Griffiths, Phillip; Schmid, Wilfried, *Locally homogeneous complex manifolds*, Acta Math. **123** (1969), 253–302.
- [52] ———, *Recent developments in Hodge theory: a discussion of techniques and results*, Discrete subgroups of Lie groups and applications to moduli (Internat. Colloq., Bombay, 1973), pp. 31–127. Oxford Univ. Press, Bombay, 1975.
- [53] Lang, Serge, *Hyperbolic and Diophantine analysis*, Bull. Amer. Math. Soc. (N.S.) **14** (1986), no. 2, 159–205.

- [54] ———, *Introduction to complex hyperbolic spaces*, Springer-Verlag, New York, 1987.
- [55] Nash, John, *C1 isometric imbeddings*, Ann. of Math. (2) **60** (1954), 383–396.
- [56] Vojta, Paul, *Diophantine approximations and value distribution theory*, Lecture Notes in Mathematics, **1239**, Springer-Verlag, Berlin, 1987.
- [57] Yang, S., *Higher algebraic K-theory and tangent spaces to Chow groups*, preprint.

Department of Mathematics, University of California, Los Angeles, CA 90095, USA

E-mail: mlgucla@gmail.com



# Plenary Lectures



# Virtual properties of 3-manifolds

*Dedicated to the memory of Bill Thurston*

Ian Agol

**Abstract.** We will discuss the proof of Waldhausen’s conjecture that compact aspherical 3-manifolds are virtually Haken, as well as Thurston’s conjecture that hyperbolic 3-manifolds are virtually fibered. The proofs depend on major developments in 3-manifold topology of the past decades, including Perelman’s resolution of the geometrization conjecture, results of Kahn and Markovic on the existence of immersed surfaces in hyperbolic 3-manifolds, and Gabai’s sutured manifold theory. In fact, we prove a more general theorem in geometric group theory concerning hyperbolic groups acting on CAT(0) cube complexes, concepts introduced by Gromov. We resolve a conjecture of Dani Wise about these groups, making use of the theory that Wise developed with collaborators including Bergeron, Haglund, Hsu, and Sageev as well as the theory of relatively hyperbolic Dehn filling developed by Groves-Manning and Osin.

**Mathematics Subject Classification (2010).** Primary 57M.

**Keywords.** hyperbolic, 3-manifold.

## 1. Introduction

In Thurston’s 1982 Bulletin of the AMS paper *Three Dimensional Manifolds, Kleinian groups, and hyperbolic geometry* [81], he asked 24 questions which have guided the last 30 years of research in the field. Four of the questions have to do with “virtual” properties of 3-manifolds:

- Question 15 (paraphrased): Are Kleinian groups LERF? [53, Problem 3.76 (Hass)]
- Question 16: “Does every aspherical 3-manifold have a finite-sheeted cover which is Haken?” This question originated in a 1968 paper of Waldhausen. [52, Problem 3.2] <sup>1</sup>
- Question 17: “Does every aspherical 3-manifold have a finite-sheeted cover with positive first Betti number?” [53, Problem 3.50 (Mess)]
- Question 18: “Does every hyperbolic 3-manifold have a finite-sheeted cover which fibers over the circle? This dubious-sounding question seems to have a definite chance for a positive answer.” [53, Problem 3.51 (Thurston)]

The goal of this talk is to explain these problems, and how they reduce to a conjecture of Wise in geometric group theory.

Note that there are now several expository works on the topics considered here [10, 11, 13, 20, 29].

---

<sup>1</sup>Proceedings of the International Congress of Mathematicians, Seoul, 2014

## 2. 3-manifold topology

Haken introduced the notion of a Haken manifold as a way to understand certain 3-manifolds via an inductive procedure by cutting along surfaces [43].

**Definition 2.1.** A closed essential surface  $f : \Sigma^2 \rightarrow M^3$  is a surface with either

- $\chi(\Sigma) \leq 0$  and  $f_{\#} : \pi_1(\Sigma) \hookrightarrow \pi_1(M)$  is injective or
- $\Sigma \cong S^2$ , and  $[f] \neq 0 \in \pi_2(M)$  (in other words,  $f$  is not homotopically trivial).

If  $M$  is a manifold, then  $M$  is termed *aspherical* if its universal cover  $\tilde{M}$  is contractible. For example, this holds if  $\tilde{M} \cong \mathbb{R}^n$ . In three dimensions,  $M$  is closed and aspherical if and only if  $\tilde{M} \cong \mathbb{R}^3$ , or equivalently  $\pi_2(M) = \pi_3(M) = 0$  (this is a non-trivial consequence of the geometrization conjecture). By the sphere theorem of Papakyriopoulos [69], equivalently  $|\pi_1(M)| = \infty$  and  $M$  is irreducible.

If  $M$  is aspherical and contains an embedded essential surface, then  $M$  is called *Haken*.

For example if  $M$  is aspherical, and  $\text{rank}(H_1(M; \mathbb{Q})) = b_1(M) > 0$ , then  $M$  is Haken. This follows from the loop theorem.

A 3-manifold  $M$  **fibers over the circle** if there is a map  $\eta : M \rightarrow S^1$  such that each point preimage  $\eta^{-1}(x)$  is a surface called a **fiber**.

If  $M$  is closed and 3-dimensional and fibers over  $S^1$ , then the fiber is a genus  $g$  surface  $F_g$ , and  $M$  is obtained as the mapping torus of a homeomorphism  $f : F_g \rightarrow F_g$ ,

$$M \cong T_f = \frac{F_g \times [0, 1]}{\{(x, 0) \sim (f(x), 1)\}}.$$

A fibered 3-manifold  $M$  has positive first betti number, and the fiber surface is essential. Therefore  $M$  is aspherical and Haken if  $g > 0$ .

A motivating question in 20th century 3-manifold topology:

*Given an immersed essential surface in a 3-manifold, does there exist an embedded essential surface of the same type?*

This has been an important question because embedded essential surfaces are easier to work with than immersed surfaces in general. For example, the theory of normal surfaces allows certain questions about embedded essential surfaces in 3-manifolds to be made algorithmic.

Examples include when  $\chi(\Sigma) \geq 0$ :

- Dehn’s Lemma [25] [69, Papakyriopoulos 1957]: If an embedded loop in  $\partial M$  is homotopically trivial, then it bounds an embedded disk.
- The Loop Theorem [69]: Similar statement for an immersed loop in  $\partial M$ .
- The Sphere Theorem [69, Papakyriopoulos 1957] [75, Stallings 1969]: If  $\pi_2(M) \neq 0$  (i.e., there’s an immersed essential sphere in  $M$ ), then there exists an embedded essential sphere in  $M$ .

---

<sup>1</sup> “Of those irreducible manifolds, known to me, which have infinite fundamental group and are not sufficiently large, some (and possibly all) have a finite cover which is sufficiently large.” [84] Waldhausen may only have been referring to small Seifert-fibered space examples that he was aware of, but the general question has been attributed to him.

- The annulus and torus theorems [49, Jaco-Shalen] and [50, Johannson]:  
In a Haken manifold, if there is an immersed essential annulus or torus, then there is an embedded one.
- The Seifert fibered space theorem (Scott, Tukia, Casson-Jungreis, Gabai): If the center  $Z(\pi_1(M)) \neq 0$  and  $M$  is aspherical, then  $M$  is Seifert-fibered.

As was known to Waldhausen, there is an infinite class of aspherical Seifert-fibered spaces which are non-Haken, so one cannot hope to extend the torus theorem to non-Haken 3-manifolds. For example, one may consider the unit tangent bundle to a turnover orbifold of euler characteristic  $< 0$ . However, these are easily shown to be virtually Haken, since they have a finite-sheeted cover homeomorphic to the unit tangent bundle of a surface. Thus, one may ask the question:

*Given an immersed essential surface in a 3-manifold, does there exist a finite-sheeted cover with an embedded essential surface of the same type?*

These classic theorems of 3-manifold topology are now superseded by the Geometrization Theorem (Question 1 from Thurston's list [81] [53, Problem 3.45 (Thurston)]). The geometrization theorem states that an irreducible 3-manifold  $M$  admits a (possibly non-orientable) embedded essential surface  $\Sigma \hookrightarrow M$  which is unique up to isotopy, such that  $\chi(\Sigma) = 0$  and each component of  $M - \Sigma$  admits a complete locally homogeneous Riemannian metric of finite volume. There are eight possible model geometries for these metrics.

This question was formulated by William Thurston at Princeton in the 1970s, and was proved by him for Haken 3-manifolds [82, 83], and conjectured to hold in general. A proof of the conjecture was given by Grigori Perelman in 2003 using Ricci flow [70], finishing a program of Hamilton who introduced the Ricci flow in the 1980s [45].

The most interesting and least understood homogeneous geometry is hyperbolic geometry.

Consider a chunk of glass sitting on a table, so that the speed of light  $n$  is proportional to the height above the table (Figure 2.1). Then light will follow a geodesic path in the glass which is a semicircle or line perpendicular to the tabletop.

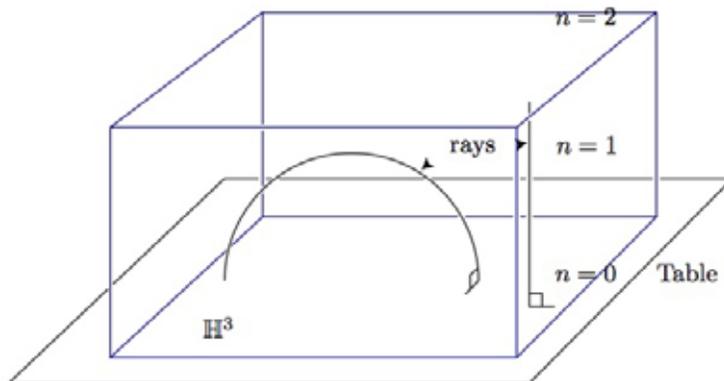


Figure 2.1. A physical model for hyperbolic space

This gives a physical model for the upper half space model of hyperbolic space.

Manifolds modeled on this geometry are *hyperbolic 3-manifolds* if they admit a complete Riemannian metric of constant curvature  $-1$ , with fundamental group a *Kleinian group* (if it is finitely generated). Classic examples of hyperbolic 3-manifolds are the **Seifert-Weber dodecahedral space**, the **figure eight knot complement**, and the **Whitehead link complement**. Given a cusped hyperbolic 3-manifold (finite-volume non-compact), Thurston showed that one may deform the hyperbolic metric to obtain hyperbolic metrics on Dehn fillings [80, Theorem 5.8.2]. A Dehn filling is obtained from a manifold with torus boundary by identifying the boundary with the boundary of a solid torus (Figure 2.2). The homeomorphism type of the Dehn filling is determined by the slope of the meridian of the torus, which may be regarded as a rational number  $\in \mathbb{P}\mathbb{Q}^1$ .

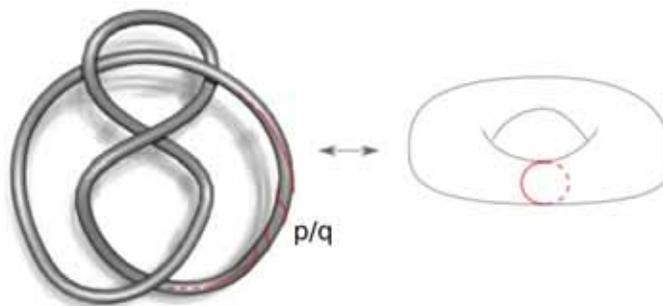


Figure 2.2. Dehn filling on the figure 8 knot complement

Thurston proved that all but finitely many slopes  $\in \mathbb{P}\mathbb{Q}^1$  give Dehn fillings on a hyperbolic 3-manifold which are hyperbolic.

An aspherical 3-manifold  $M$  whose geometric decomposition does not contain a hyperbolic piece is called a **graph manifold**. If  $M$  is not geometric, then all of the geometric pieces of the JSJ decomposition are modeled on the geometry  $\mathbb{H}^2 \times \mathbb{R}$ .

### 3. Virtual properties of 3-manifolds

We say that a property of a space holds *virtually* if it holds for a finite-sheeted cover, or a property holds virtually for a group if it holds for a finite-index subgroup.

- Recall that a compact aspherical 3-manifold  $M$  is **Haken** if it contains an embedded  $\pi_1$ -injective surface (e.g. a knot complement). The Seifert-Weber space is non-Haken [19, Burton-Rubinstein-Tillmann], as well as hyperbolic surgeries on the figure 8 knot complement [80, Corollary 4.11].
- A 3-manifold  $M$  is **virtually Haken** if there is a finite-sheeted manifold cover  $\tilde{M} \rightarrow M$  such that  $\tilde{M}$  is Haken, e.g. hyperbolic surgeries on the figure 8 knot complement are virtually Haken [27, Dunfield-Thurston].
- Waldhausen conjectured that every aspherical 3-manifold  $M$  is virtually Haken (the *virtual Haken conjecture*, Question 16).
- A fortiori, does  $M$  have a finite-sheeted cover  $\tilde{M} \rightarrow M$  with  $b_1(\tilde{M}) > 0$  (Question 17)? Recall that  $b_1(M) = \text{rank}(H_1(M; \mathbb{Q}))$ .

**Remark:** Since closed 3-manifold fundamental groups have balanced presentations, it is unlikely that a generic 3-manifold  $M$  has  $b_1(M) > 0$ , which clarifies the difficulty of this question.

- $M$  is **virtually fibered** if there exists a finite-sheeted cover  $\tilde{M} \rightarrow M$  such that  $\tilde{M}$  fibers.
- If  $M$  fibers, then  $b_1(M) > 0$ , so this is stronger than asking for virtual positive betti number.
- There have previously been several classes of hyperbolic 3-manifolds shown to virtually fiber, including 2-bridge links (Walsh), some Montesinos links (Agol, Boyer, Zhang, Guo) and certain alternating links (Aitchison-Rubinstein) as well as many examples of hyperbolic manifolds (Bergeron, Chesebro-DeBlois-Wilton, Gabai, Leininger, Reid, Wise, Aitchison-Rubinstein). Some of these constructions have the advantage that they give explicit descriptions or prescriptions for finding a cover that fibers.
- Thurston asked whether every hyperbolic 3-manifold is virtually fibered (Question 18)?

If  $M$  is a finite volume hyperbolic 3-manifold, and  $f : F_g \rightarrow M$  is an essential immersion of a surface of genus  $g > 0$ , then there is a dichotomy for the geometric structure of the surface discovered by Thurston, and proven by Bonahon in general [15].

Either  $f$  is

- **geometrically finite** or
- **geometrically infinite**.

The first case includes **quasifuchsian** surfaces. A quasifuchsian surface group is discrete and preserves a circle in  $\hat{\mathbb{C}}$  such that the convex hull of this curve has finite covolume under the group action. More generally, a geometrically finite group preserves a convex subset of hyperbolic space whose quotient by the group has finite (non-zero) volume.

In the geometrically infinite case, the surface is **virtually the fiber** of a fibering of a finite-sheeted cover of  $M$ .

The **Tameness theorem** [1, Agol], [21, Calegari-Gabai] plus the **covering theorem** of [22, Canary] implies a similar dichotomy for finitely generated subgroups of  $\pi_1(M)$ :

either a subgroup is geometrically finite, or it corresponds to a virtual fiber. The limit set of a fiber of a fibration is  $\partial_\infty \mathbb{H}^3 = \hat{\mathbb{C}}$ , but may be regarded as a sphere-filling curve [24, Cannon-Thurston]. Analogous to the loop, sphere, annulus and torus theorems, one may ask:

*Given an essential map of a surface  $f : \Sigma \rightarrow M$  with  $\chi(\Sigma) < 0$ , is there an essential embedding  $\Sigma \hookrightarrow M$ ?*

The answer to this question is no since there are examples of non-Haken 3-manifolds such as the figure 8 knot hyperbolic fillings which have virtual positive betti number, and therefore contain an immersed essential surface, but no embedded essential surface.

With further hypotheses on the surface, the answer to this question can be a qualified yes.

Gabai proved that if  $f : \Sigma \looparrowright M$  is an immersed oriented surface with  $\chi(\Sigma) \leq 0$ , and  $f_*([\Sigma]) \neq 0 \in H_2(M)$ , then there is an embedded essential surface  $\Sigma' \hookrightarrow M$  such that  $[\Sigma'] = f_*[\Sigma] \in H_2(M)$ , and  $\chi(\Sigma') \geq \chi(\Sigma)$  [30–32].

Gabai’s proof makes use of an inductive method called **sutured manifold hierarchies** to construct a foliation of the manifold with an embedded compact leaf, and obtain the desired lower bound on Euler characteristic by analyzing the Euler class of the foliation.

**Theorem 3.1** ([51, Kahn-Markovic] [53, Problem 3.75 (Waldhausen)]). *Hyperbolic 3-manifolds contain immersed quasi-fuchsian surfaces which are arbitrarily close to being totally geodesic.*

The limit sets of these surfaces in  $\partial_\infty \mathbb{H}^3$  can be made arbitrarily close to being a round circle.

There has been much previous work on this problem, by Bart, Cooper, Lackenby, Li, Long, Masters, and Zhang.

### 4. 3-manifold fundamental group properties

**Definition 4.1.** A group  $G$  is **residually finite (RF)** if for every  $1 \neq g \in G$ , there exists a finite group  $K$  and a homomorphism  $\phi : G \rightarrow K$  such that  $\phi(g) \neq 1 \in K$ .

Alternatively,

$$\{1\} = \bigcap_{[G:H] < \infty} H. \tag{4.1}$$

Examples of residually finite groups include

- finitely generated linear groups [61, Malcev];
- 3-manifold groups [46, Hempel] + Geometrization [70]; and
- mapping class groups of surfaces [35, Grossman].

**Definition 4.2.** A subgroup  $L < G$  is *separable* if for all  $g \in G - L$ , there exists  $\phi : G \rightarrow K$  finite such that  $\phi(g) \notin \phi(L)$ .

Alternatively,

$$L = \bigcap_{L \leq H \leq G, [G:H] < \infty} H \tag{4.2}$$

*Residual finiteness* of  $G$  is equivalent to  $1 < G$  is separable.

**Definition 4.3.** A subgroup  $L < G$  is *weakly separable* if for all  $g \in G - L$ , there exists  $\phi : G \rightarrow K$  such that  $\phi(L)$  is finite and  $\phi(g) \notin \phi(L)$  ( $K$  need not be finite).

**Example 4.4.**

- If  $L < G$  is finite, then  $L$  is (trivially) weakly separable in  $G$ .
- Let  $H \triangleleft G$  be a normal subgroup of  $G$ , then  $H$  is weakly separable in  $G$ . In fact, we may use the quotient  $\varphi : G \rightarrow G/H$  to weakly separate all elements of  $G - H$  from  $H$ .

**Definition 4.5.** A group  $G$  is **Locally Extended Residually Finite (LERF)** if *finitely generated subgroups of  $G$  are separable* (*local* means finitely generated).

Previously well-known examples of LERF groups include

- $\mathbb{Z}^n$ ;
- free groups [44, Hall] and surface groups [74, Scott];
- certain doubles of compression body groups [33, Gitik];
- Bianchi groups  $\mathrm{PSL}(2, \mathbb{Z}[\sqrt{-d}])$  [6, Agol-Long-Reid] and certain other arithmetic subgroups of  $\mathrm{PSL}(2, \mathbb{C})$  such as the fundamental group of the Seifert-Weber dodecahedral space;
- 3-dimensional hyperbolic reflection groups [41, Haglund-Wise].

There are examples of 3-manifold groups which are not LERF which are *graph manifold* groups [18, Burns-Karrass-Solitar].

Thurston's question 15 is whether Kleinian groups are LERF? LERF allows one to lift  $\pi_1$ -injective immersions to embeddings in finite-sheeted covers [74, Scott] (Figure 4.1).

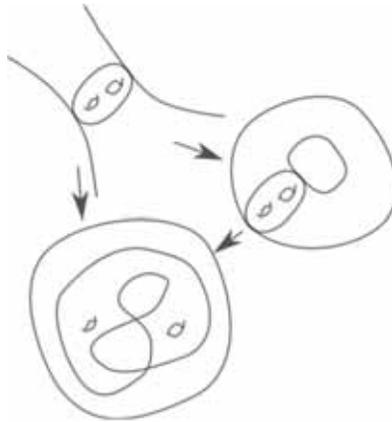


Figure 4.1. A surface immersed in a 3-manifold with separable fundamental group lifts to an embedding in a finite-sheeted cover

In fact, Matsumoto showed that there are certain graph manifolds which contain surfaces which do not lift to an embedding in any finite-sheeted covering space [64, Matsumoto]. These examples highlight the importance of hyperbolicity with respect to subgroup separability.

**4.1. Virtual fibering.** Thurston's virtual fibering question was stated for hyperbolic 3-manifolds, and does not hold for general 3-manifolds.

**Theorem 4.6** (Przytycki-Wise 2012). *If  $M$  is an aspherical closed 3-manifold which is not a graph manifold, then  $M$  is virtually fibered.*

Svetlov characterized virtually fibered graph manifolds (e.g. unit tangent bundles to closed hyperbolic surfaces are not virtually fibered), but the criterion is technical to state [79, Svetlov].

**Definition 4.7.** A group  $G$  is **Residually Finite Rationally Solvable** or **RFRS** if there is a sequence of subgroups  $G = G_0 > G_1 > G_2 > \dots$  such that  $\bigcap_i G_i = \{1\}$ ,  $[G : G_i] < \infty$  and  $G_{i+1} = \ker\{G_i \rightarrow \mathbb{Z}^{k_i} \rightarrow (\mathbb{Z}/n_i)^{k_i}\}$  for sequences  $n_i, k_i \in \mathbb{N}$ .

**Remark 4.8.** We may assume that  $G_i \triangleleft G$ , in which case  $G/G_i$  is a finite solvable group. Thus, the RFRS condition is a strong form of residual finite solvability.

We remark that if  $G$  is RFRS, then any subgroup  $H < G$  is as well.

Examples of RFRS groups are free groups, surface groups,  $\mathbb{Z}^n$  and free products of RFRS groups.

For a 3-manifold  $M$  with RFRS fundamental group, the condition is equivalent to there existing a **cofinal tower** of finite-index covers

$$M \leftarrow M_1 \leftarrow M_2 \leftarrow \dots$$

such that  $M_{i+1}$  is obtained from  $M_i$  by taking a finite-sheeted cyclic cover dual to an embedded non-separating surface in  $M_i$ . Equivalently,  $\pi_1(M_{i+1}) = \ker\{\pi_1(M_i) \rightarrow \mathbb{Z} \rightarrow \mathbb{Z}/k\mathbb{Z}\}$ .

This condition implies that  $M$  has virtual infinite  $b_1$ , unless  $\pi_1(M)$  is virtually abelian.

**Theorem 4.9** ([2, Agol]). *If  $M$  is aspherical and  $\pi_1(M)$  is RFRS, then  $M$  virtually fibers.*

The proof makes use of **sutured manifold theory**, the inductive technique mentioned before for studying foliations of 3-manifolds introduced by Gabai. For a self-contained proof, see a preprint of [29, Friedl-Kitayama].

**Theorem 4.10** ([86, Corollary 14.3, Theorem 14.29 Wise]). *Haken hyperbolic 3-manifolds are virtually fibered.*

The theorem includes non-compact hyperbolic 3-manifolds with finite volume unconditionally.

## 5. Geometric group theory

Let  $G$  be a finitely generated group, with generators  $G = \langle g_1, \dots, g_n \rangle$ . The *Cayley graph* of  $G$  with respect to the generating set  $\{g_1, \dots, g_n\}$  is a graph  $\Gamma = \Gamma(G, \{g_1, \dots, g_n\})$  with vertex set  $V(\Gamma) = G$ , and edge set  $E(\Gamma) = \{(g, g \cdot g_i) | g \in G, 1 \leq i \leq n\}$ . So the degree of each vertex  $g$  is  $2n$ .

We may regard  $\Gamma$  as a metric space, by letting edges of  $\Gamma$  have length 1, and taking the path metric. So the distance  $d(1, g)$  between vertices  $1, g \in V(\Gamma)$  is the smallest  $k$  such that  $g = g_{i_1}^{\pm 1} \dots g_{i_k}^{\pm 1}$ . Then clearly  $d(h, h \cdot g) = d(1, g)$ , since the metric is invariant under the left group action of  $G$  on  $\Gamma(G, \{g_1, \dots, g_n\})$ .

The Cayley graph  $\Gamma(F_2, \{a, b\})$  of the two generator free group  $F_2 = \langle a, b \rangle$  with respect to the free generating set  $\{a, b\}$  is an infinite 4-valent tree, with oriented edges labeled  $a, b$ . Geometric group theory is the study of properties of groups from the geometric properties of the Cayley graph. This notion has some origins in the work of [26, Dehn 1911] on the word problem for surface groups, but was introduced by [66, Milnor 1968] who studied the growth of balls in Cayley graphs of groups as a function of the radius, and [23, Cannon 1984] who studied the Cayley graphs of hyperbolic manifolds.

If  $G$  acts properly and cocompactly on a metric space  $X$  (for example,  $X = \tilde{M}$  the universal cover, where  $M$  is a compact Riemannian manifold, and  $G = \pi_1(M)$ ), then some geometric properties of  $X$  are reflected in the geometric properties of the Cayley graph  $\Gamma(G, \{g_1, \dots, g_n\})$ . So we may study properties of a group  $G$  by studying the geometric properties of  $X$ .

For example, Milnor observed that if the volumes of balls of radius  $r$  in  $X$  grow exponentially with  $r$ , then the same will hold for the balls in  $\Gamma$ , with volume replaced by the number of vertices. Exponential growth of balls holds for universal covers of compact Riemannian manifolds with negative curvature.

Cannon '84 realized that Cayley graphs of hyperbolic manifolds have a nice recursive combinatorial structure for the balls of radius  $r$ . This notion was then extended and codified by [34, Gromov 1987] in the notion of a *hyperbolic group*.

A (Gromov-)hyperbolic geodesic metric space  $X$  may be defined by Rips' "slim triangle" condition: for points  $A, B$  in the metric space, let  $[A, B] \subset X$  be a geodesic connecting  $A$  and  $B$ . Then  $X$  is called a  $\delta$ -hyperbolic metric space if for any three points  $A, B, C \in X$ ,

$$[B, C] \subset \mathcal{N}_\delta([A, B] \cup [A, C]).$$

For example, hyperbolic space  $\mathbb{H}^n$  is  $\log(1 + \sqrt{2})$ -hyperbolic and a tree is 0-hyperbolic.

If  $\Gamma(G, \{g_1, \dots, g_n\})$  is a  $\delta$ -hyperbolic metric space for some  $\delta$ , then  $G$  is called a (*Gromov*)-*hyperbolic group* (sometimes also called  $\delta$ -hyperbolic, word-hyperbolic, or just hyperbolic group).

Gromov proved many properties of these groups, such as there exists a compactification  $\Gamma(G, \{g_1, \dots, g_n\}) \cup \partial_\infty(G)$ , so that  $\partial_\infty(G)$  is independent of the generating set and  $\Gamma$ .

**Definition 5.1.** Let  $X$  be a geodesic metric space, and  $Y \subset X$ . Then  $Y$  is  $R$ -quasiconvex in  $X$  if for every  $y_1, y_2 \in Y$ , the geodesic  $[y_1, y_2] \subset X$  lies in an  $R$ -neighborhood of  $Y$ ,  $[y_1, y_2] \subset \mathcal{N}_R(Y)$ .

For example,  $X$  is  $\delta$ -hyperbolic if  $[a, b] \cup [b, c]$  is  $\delta$ -quasiconvex for every  $a, b, c \in X$ .

Let  $G$  be a hyperbolic group, with Cayley graph  $\Gamma$ . A subgroup  $H < G$  may be regarded as a subspace  $H \subset G = V(\Gamma) \subset \Gamma$ . Then we say  $H$  is *quasiconvex* if it is  $R$ -quasiconvex in  $\Gamma$  for some  $R$ . It follows from quasigeodesic stability that  $H$  will be quasiconvex in the Cayley graph with respect to any (finite) generating set of  $G$ .

Motivating examples of hyperbolic groups are Kleinian groups without  $\mathbb{Z}^2$  subgroups (e.g. fundamental groups of closed hyperbolic manifolds and convex cocompact Kleinian groups), and more generally fundamental groups of closed negatively curved manifolds. Motivating examples of quasi-convex subgroups are quasi-fuchsian surface groups (such as the fundamental groups of the essential Kahn-Markovic surfaces) in closed hyperbolic 3-manifold groups, and cyclic subgroups of arbitrary hyperbolic groups.

**Theorem 5.2** ([4, 62, Agol, Groves, Manning, Martinez-Pedrosa 2008]). *If hyperbolic groups are RF, then Kleinian groups are LERF*

So it may be possible to show that hyperbolic 3-manifold groups are LERF by showing that Gromov-hyperbolic groups are RF

**Caveat:** This approach seems quite unlikely to work, since many experts believe that there are non-RF Gromov-hyperbolic groups.

## 6. Cube complexes

A topological space  $X$  is **locally CAT(0) cubed** if  $X$  is a cube complex such that putting the standard Euclidean metric on each cube gives a locally CAT(0) metric (a form of non-positive curvature). Gromov [34] showed that this metric condition is equivalent to a purely

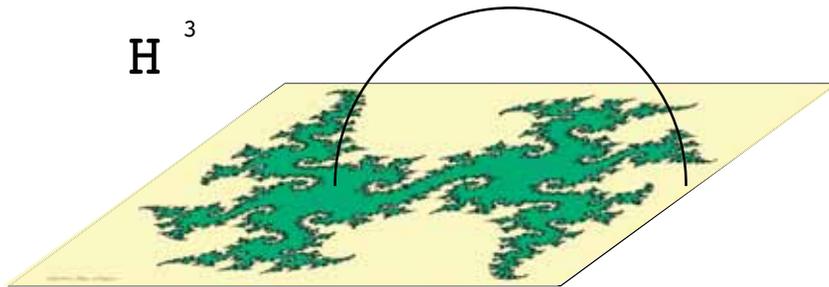


Figure 6.1. For the endpoints of each geodesic, there is a quasifuchsian limit set which separates the endpoints

combinatorial condition on the links of vertices of  $X$ , they are **flag**. A flag simplicial complex has the property that its simplices are determined by the 1-skeleton: if one sees a  $k + 1$  complete subgraph in the 1-skeleton, then there is a  $k$ -simplex spanning the subgraph. If  $X$  is locally CAT(0) and simply-connected, then it is globally CAT(0).

In a locally CAT(0) cube complex, there are canonical maps of codimension-one locally geodesic subcomplexes  $W \looparrowright X$  called **hyperplanes**, which are obtained by taking the union of midplanes in each cube. The components of the hyperplane complex correspond to equivalence classes of an equivalence relation on edges of the complex generated by edges lying on opposite sides of a square.

A locally CAT(0) square complex has the property that the link of each vertex is a graph of girth  $\geq 4$  (there are no triangles). In this picture of a square complex, the link of each vertex is a 5-cycle, so it is a CAT(0) square complex.

A topological space  $Y$  is **cubulated** if it is *homotopy equivalent* to a compact locally CAT(0) cube complex  $X \simeq Y$  (equivalently,  $Y$  is aspherical and  $\pi_1(X) \cong \pi_1(Y)$ ). We also say in this case that  $\pi_1(Y)$  is cubulated. We are interested in 3-manifolds which are cubulated.

**Remark 6.1.** If  $Y = M^3$ , and  $X \simeq Y$  is a CAT(0) cubing, then  $\dim X$  may be  $> 3$ . Tao Li has shown that there are hyperbolic 3-manifolds  $Y$  such that there is no **homeomorphic** CAT(0) cubing  $X \cong Y$  [56].

A theorem of [73, Sageev 1995] associates a cocompact action of  $\pi_1(M)$  on a (globally) CAT(0) cube complex if  $M$  contains an immersed essential surface. Sageev's construction gives a cube complex in which each immersed essential surface in a 3-manifold corresponds to an immersed hyperplane.

For example, Sageev's construction applied to a fiber surface gives an action factoring through the  $\mathbb{Z}$  action on  $\mathbb{R}$ , with quotient  $S^1$ . In the case of a geometrically infinite surface in a hyperbolic 3-manifold, Sageev's construction gives rise to a crystallographic group action.

**Theorem 6.2** ([12, Bergeron-Wise 2012]). *Closed hyperbolic 3-manifolds are cubulated.*

Bergeron-Wise give a condition for cubulation. If every geodesic in  $\mathbb{H}^3$  has the property that its endpoints in  $\partial_\infty \mathbb{H}^3$  are separated by the limit set of a quasifuchsian surface, then one may use finitely many surfaces so that Sageev's construction will give a proper cocompact action on a CAT(0) cube complex (Figure 6.1).

The surfaces produced by Kahn-Markovic have limit sets which are close to any given

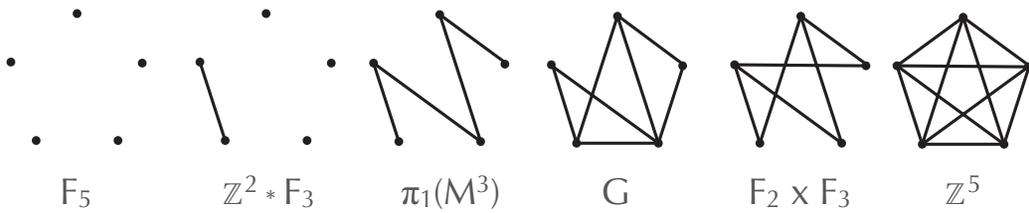


Figure 6.2. Some graphs with their associated RAAGs

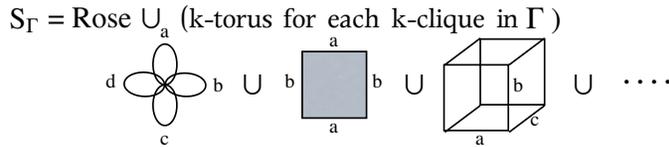


Figure 6.3. Defining the Salvetti complex  $S_\Gamma$

circle, so can separate any pair of points in  $\partial_\infty \mathbb{H}^3$ . Thus closed hyperbolic 3-manifolds are cubulated.

There were many known examples of cubulated hyperbolic 3-manifolds before this theorem, e.g. alternating link complements [7, Aitchison-Rubinstein]. Other examples come from duals to tessellations by right-angled polyhedra.

### 6.1. Right angled Artin groups.

**Definition 6.3.** Let  $\Gamma$  be a simplicial graph. The **right-angled Artin group**  $A_\Gamma$  (**RAAG**) defined by  $\Gamma$  has a generator for each vertex  $v \in V(\Gamma)$ , and relators  $vw = wv$  if  $(v, w) \in E(\Gamma)$  is an edge of  $\Gamma$ .

The **Salvetti complex**  $S_\Gamma$  associated to  $A_\Gamma$  is a  $K(A_\Gamma, 1)$  which is a locally CAT(0) cube complex, defined by taking a wedge of loops (rose), one for each generator, and attaching a  $k$ -torus for each complete subgraph ( $k$ -clique) of  $\Gamma$  (Figure 6.3). The 2-skeleton by construction gives a presentation  $\pi_1(S_\Gamma) \cong A_\Gamma$ .

The Salvetti complex has the property that the links of the vertices are flag simplicial complexes, and therefore these complexes are locally CAT(0).

Examples include

- The free group associated to the trivial graph  $\Gamma$  with no edges, for which  $S_\Gamma$  is a wedge of loops
- The  $n$ -torus associated to the complete graph on  $n$  vertices  $K_n$ , for which  $S_{K_n} \cong T^n$  the  $n$ -torus
- The complement of a chain of 4 links (Figure 6.4).

**6.2. Special cube complexes.** Special cube complexes are defined in terms of properties of their hyperplanes. Hyperplanes are embedded and 2-sided. Moreover, there are no self-osculating or inter-osculating hyperplanes (Figure 6.5). The midplane of a cube is dual to the edges of the cube it crosses. Thus, we may regard a hyperplane as an equivalence class of (oriented) edges generated by the equivalence relation of two edges lying on opposite sides of

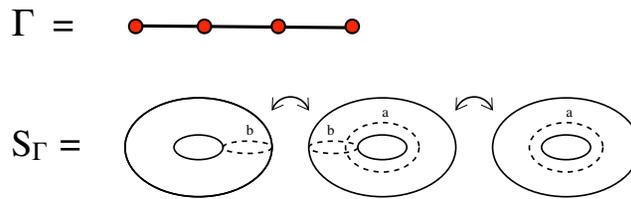


Figure 6.4. The complement of a chain of 4 links

a square. If the orientation of edges dual to a hyperplane is preserved in an equivalence class, then the hyperplane is said to be 2-sided. If no adjacent edges of a square are in the same equivalence class, then the hyperplane is embedded. If equivalent edges share a common vertex, which is the end or beginning of both edges, then we say that the hyperplane osculates (Figure 6.5 (a)). If two hyperplanes osculate at one vertex, and cross at another vertex, then we say that the hyperplanes interosculate (Figure 6.5 (b)). These are the forbidden configurations in a special cube complex.

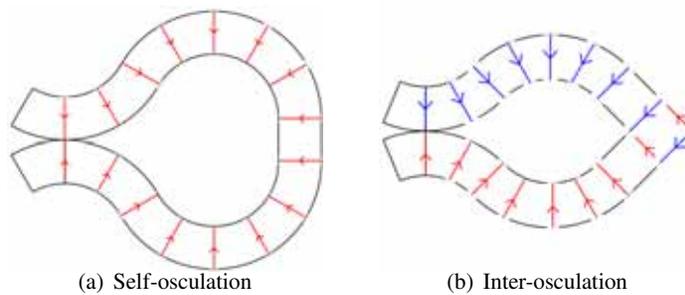


Figure 6.5. Configurations forbidden in a special cube complex

The motivating examples of special cube complexes are Salvetti complexes of RAAGs. Here's an example of a special cube complex  $X$  homeomorphic to a surface. The hyperplanes consist of six curves colored blue and red in Figure 6.6.

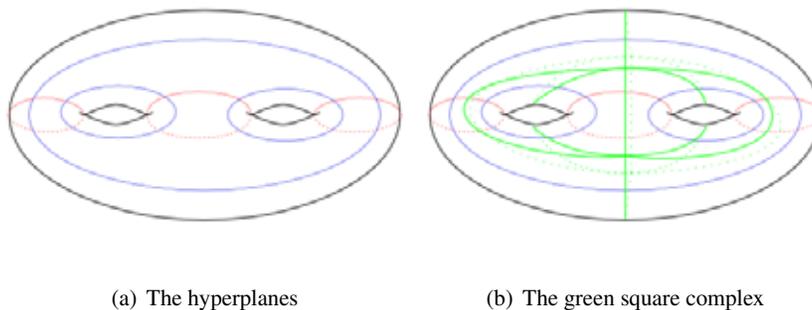
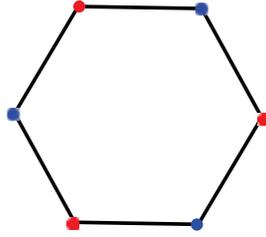


Figure 6.6. A special cube complex  $X$  homeomorphic to a surface

Figure 6.7. The crossing graph of  $X$ 

The *crossing graph*  $\Gamma(X)$  of a cube complex  $X$  has vertices corresponding to the hyperplanes of  $X$ , and two vertices of  $\Gamma(X)$  are connected by an edge if and only if the corresponding hyperplanes of  $X$  cross (Figure 6.7).

**Theorem 6.4** ([39, Haglund-Wise 2007]). *If  $X$  is a special cube complex with hyperbolic fundamental group  $\pi_1 X$  (in the sense of Gromov), then  $\pi_1 X$  embeds in a RAAG  $A_{\Gamma(X)}$  and quasi-convex subgroups of  $\pi_1 X$  are separable.*

For the proof, take the crossing graph  $\Gamma(X)$  associated to  $X$  and form the RAAG  $A_{\Gamma(X)}$ . There is a natural map from  $X$  to the Salvetti complex  $S_{\Gamma(X)}$  sending every edge dual to a hyperplane to the corresponding edge in the Salvetti complex, and extending over the higher skeleta. This map is a locally isometric immersion when  $X$  is special, and therefore  $\pi_1(X) \leq A_{\Gamma(X)}$ .

For example, applying this construction to  $S_\Gamma$  just recovers the identity isometry  $S_\Gamma \rightarrow S_\Gamma$  and the isomorphism  $\pi_1(S_\Gamma) \cong A_\Gamma$ !

The notion of a virtual retract was defined independently by [60, Long-Reid] and [38, Haglund]:

**Definition 6.5.** A subgroup  $L < G$  is a virtual retract if there exists  $G' < G$  a finite-index subgroup such that  $L < G'$  and a retract  $r : G' \rightarrow L$ , meaning  $r|_L = Id$ .

E.g. if  $\Lambda \subset \Gamma$  is a subgraph spanned by vertices, there is a natural retract  $A_\Gamma \rightarrow A_\Lambda$ , setting generators of  $A_\Gamma$  corresponding to vertices not in  $\Lambda$  to 1.

**Claim:** If  $G$  is residually finite, and  $L$  is a virtual retract of  $G$ , then  $L$  is separable in  $G$ .

Haglund proved that quasi-convex subgroups of RAAGs are virtual retracts [38]. This is proved geometrically using “canonical completions” and “canonical retracts”.

**Theorem 6.6** ([2, Agol 2008]). *If  $M^3$  is special cubulated, then  $M$  is virtually fibered.*

Since  $M$  is special cubulated,  $M \simeq X$ , where  $X$  is a CAT(0) compact special cube complex. Thus  $\pi_1(M) = \pi_1(X) < A_{\Gamma(X)}$ , a Right-Angled Artin Group. The RAAGs have the RFRS property, so it passes to  $\pi_1(M)$  and implies that  $M$  is virtually fibered.

We resolved a conjecture of Wise which implies Thurston’s questions.

**Theorem 6.7** ([86, Conjecture 19.5 Wise] [3, Agol 2012]). *Locally CAT(0) cube complexes with hyperbolic fundamental group are virtually special.*

The importance of hyperbolicity in the hypotheses of this theorem is made apparent by the following remarkable theorem:

**Theorem 6.8** ([17, Burger-Mozes 1997]). *There are simple groups which are the fundamental group of a locally CAT(0) square complex whose universal cover is a product of finite-valence regular trees.*

**Corollary 6.9.** *Let  $M$  be a closed hyperbolic 3-manifold. Then  $\pi_1 M$  is LERF, large, and  $M$  virtually fibers.*

A group  $G$  is large if there is a finite-index subgroup  $G' < G$  which surjects a free group on 2 generators.

This resolves positively Thurston’s questions 15-18. The next sections will discuss the background needed in the proof of Theorem 6.7. We remark that the proof of Theorem 6.7 makes use of ideas introduced in the context of 3-manifold topology, including hierarchies and relatively hyperbolic Dehn filling. However, to prove the theorem, it is *essential* that one work in the category of hyperbolic groups, rather than specialize to hyperbolic 3-manifold groups which are of interest for Thurston’s questions.

**6.3. Amalgamated products and HNN extensions.** Given groups  $A, B, C$ , and injections  $\varphi_1 : C \hookrightarrow A, \varphi_2 : C \hookrightarrow B$ , we may form the amalgamated product  $G = A *_C B$ , which has a (relative) presentation  $\langle A, B | \varphi_1(c) = \varphi_2(c), c \in C \rangle$ . By combinatorial group theory,  $A, B, C \hookrightarrow G$  inject.

Similarly, suppose we have two subgroups  $B, C < A$ , such that there is an isomorphism  $\varphi : B \rightarrow C$ . Then the HNN extension  $G = A *_\varphi$  has the presentation  $\langle A, t | tct^{-1} = \varphi(c), c \in B \rangle$ .

For example, all RAAGs are HNN extensions (more generally, any group  $G$  with a surjection  $G \rightarrow \mathbb{Z}$ ). For any vertex  $v$  of a graph  $\Gamma$  defining a RAAG, one has an associated HNN decomposition, where  $A$  is the RAAG  $A_{\Gamma - v}$ , where  $\Gamma - v$  is the subgraph obtained by deleting all edges adjacent to  $v$ . The subgroup defined by  $link(v)$  is both  $B$  and  $C$  in this case, where  $\varphi = Id$ , since the generator corresponding to  $v$  in  $A_\Gamma$  commutes with all the elements in  $A_{link(v)}$ . This HNN decomposition may be realized geometrically by splitting the Salvetti complex  $S_\Gamma$  along the hyperplane dual to the generator corresponding to  $v$ .

For example, applying this to the complete graph RAAG  $A_{K_n}$ , one obtains the HNN extension  $\mathbb{Z}^n = \mathbb{Z}^{n-1} *_Id$ . Another example is a 3-manifold fibered over  $S^1$ ,  $A = \pi_1(F_g)$ ,  $B = C = A$ , and  $\varphi : B \rightarrow C$  is an isomorphism induced by a homeomorphism  $f : F_g \rightarrow F_g$ . Then  $\pi_1(T_f) = A *_\varphi$ .

**6.4. Quasiconvex hierarchies.** The notion of a hierarchy originated in the study of 1-relator groups (the Magnus hierarchy), and in the study of Haken 3-manifolds (a Haken hierarchy).

**Definition 6.10.** The class of groups  $QVH$  (standing for “Quasiconvex Virtual Hierarchy”) are defined inductively by

- (1)  $1 \in QVH$
- (2) If  $G = A *_C B$  or  $G = A *_\varphi$ , with  $A, B, C \in QVH$  and quasiconvex in  $G$ , then  $G \in QVH$ .
- (3) Let  $H < G$  with  $[G : H] < \infty$ . If  $H \in QVH$  then  $G \in QVH$  (in particular with (1), any finite group  $K \in V\mathcal{H}$ ).

The class of groups  $MQH$  is defined similarly, but we require that  $C$  is malnormal in  $G$  in (2) as well.

It is not hard to show that if  $M$  is a hyperbolic 3-manifold, then  $\pi_1(M) \in QVH$  if and only if  $M$  is virtually Haken with a finite-sheeted cover containing an embedded quasifuchsian surface.

Special cube complexes with hyperbolic fundamental group are also in  $\mathcal{QVH}$ , with hierarchy induced by cutting along hyperplanes.

If we have a closed hyperbolic 3-manifold  $M$  fibering over  $S^1$  with fiber  $\Sigma$ , then  $\pi_1(\Sigma)$  is not quasi convex in  $\pi_1(M)$ , so  $\pi_1(M)$  is not necessarily contained in  $\mathcal{QVH}$ .

**Theorem 6.11** ([86, Wise 2011]). *Let  $G \in \mathcal{QVH}$ . Then  $G$  is virtually special. That is, there is a  $CAT(0)$  cube complex  $X$  so that  $G$  acts properly cocompactly on  $X$ , and a finite-index subgroup  $G' < G$  such that  $X/G'$  is a special cube complex.*

Wise showed that one-relator groups with torsion are in  $\mathcal{QVH}$ . This resolved a conjecture of [9, Baumslag 1967].

**6.5. Relatively hyperbolic Dehn filling.** Recall that the figure eight knot complement has a complete hyperbolic metric of finite volume. However, the figure eight knot group  $G$  is not a hyperbolic group, since it contains the peripheral subgroup  $\mathbb{Z}^2 = P < G$  coming from the  $\pi_1$ -injective torus that is the boundary of a tubular neighborhood of the knot.

However, I mentioned that Thurston proved that all but finitely many Dehn fillings on the figure 8 knot complement are closed hyperbolic 3-manifolds. In fact, the core of the solid torus of the Dehn filling is a closed geodesic in the hyperbolic structure on the Dehn filling.

Let  $G_{p/q}$  be the fundamental group of  $p/q$  Dehn filling on the figure eight knot complement. Then  $P \cap \ker\{G \rightarrow G_{p/q}\} = \langle (p, q) \rangle$ . In fact,  $\ker\{G \rightarrow G_{p/q}\}$  is freely generated by conjugates of the subgroup  $\langle (p, q) \rangle$ .

The group  $G$  is not hyperbolic, but it is relatively hyperbolic. Roughly, this means that if we take the coset graph of the subgroup  $P$ , then this graph is  $\delta$ -hyperbolic. This notion was suggested by Gromov, and developed by Bowditch [16] and Farb [28]. There's an extra condition needed called "bounded coset penetration".

Alternatively, Groves and Manning showed that if one attaches "combinatorial horoballs" to the cosets of the peripheral group  $P$ , then the resulting space is  $\delta$ -hyperbolic if and only if  $G$  is relatively hyperbolic to  $P$  [36].

For example, if  $F$  is a free group, and  $h \in F$  is a primitive element, then  $F$  is hyperbolic relative to  $\langle h \rangle$ .

For a relatively hyperbolic group, such as the figure eight knot complement, there is an analogue of Thurston's hyperbolic Dehn filling theorem.

**Theorem 6.12** ([36, Groves-Manning][68, Osin]). *Let  $G$  be a group which is hyperbolic relative to the subgroup  $P$ . Then there is a finite set of elements  $S \subset P - \{1\}$  so that if  $P' \triangleleft P$  is finite-index with  $S \cap P' = \emptyset$ , then the quotient  $G / \langle\langle P' \rangle\rangle$  is a hyperbolic group. Moreover,  $P \cap \langle\langle P' \rangle\rangle = P'$ .*

For example, if  $G$  is a hyperbolic group, and  $h \in G$  is a primitive element, then  $G$  is hyperbolic relative to  $\langle h \rangle$ . Then for all sufficiently large  $n$ ,  $G / \langle\langle h^n \rangle\rangle$  will also be a hyperbolic group. This result is due to Gromov (or rather small-cancellation theory), but the relatively hyperbolic Dehn filling theorem vastly generalizes this result.

**6.6. MSQT.** We state a special case of the Malnormal Special Quotient Theorem (MSQT):

**Theorem 6.13** ([86, Wise 2011]). *Let  $G$  be a virtually special hyperbolic group, and let  $h \in G$ . Then there exists  $N$  such that  $G / \langle\langle h^n \rangle\rangle$  is virtually special hyperbolic for all  $N | n$ .*

**Remark:** The hyperbolicity of  $G / \langle\langle h^n \rangle\rangle$  for  $n$  large may be proved using relatively hyperbolic Dehn filling.

The general statement of the malnormal special quotient theorem is a bit more technical to state. First we need a definition. A collection of subgroups  $\{H_1, \dots, H_m\} < G$  form an almost malnormal collection provided that for any element  $g \in G$  with  $|H_i \cap gH_jg^{-1}| = \infty$ , we must have  $i = j$  and  $g \in H_i$ . We state a strengthened version of the MSQT:

**Theorem 6.14** ([86, Theorem 12.3, Malnormal Special Quotient Theorem (MSQT)], [5]). *Let  $G$  be hyperbolic, virtually special, and  $\mathcal{H} = \{H_1, \dots, H_M\} < G$  a almost malnormal collection of quasi convex subgroups. Then there exists  $\tilde{H}_i < H_i$  such that for any  $H'_i < \tilde{H}_i$ , such  $H'_i < H_i$  and  $H_i/H'_i$  is virtually special hyperbolic, the quotient group  $\overline{G} = G / \langle\langle H'_1, \dots, H'_m \rangle\rangle$  is virtually special hyperbolic.*

**Remarks on the proof.** The original version of Wise assumes that  $[H_i : H'_i] < \infty$ . The hypothesis implies that  $(G, \mathcal{H})$  is relatively hyperbolic. Using hyperbolic Dehn filling results of Groves-Manning and Osin, one may conclude that  $\overline{G}$  is hyperbolic whenever  $H_i/\tilde{H}_i$  avoids a finite set of elements by Theorem 6.12. The difficult thing is showing that the quotient is cubulated and virtually special.

What Wise actually proves is that there is a finite-index normal subgroup  $G' < G$  which has an induced peripheral structure  $(G', \mathcal{H}')$ , so that  $\mathcal{H}'$  contains representatives of each  $G'$  conjugacy class of  $H_i \cap G'$ . Moreover, he shows that a hyperbolic Dehn filling on  $(G', \mathcal{H}')$  admits a malnormal quasiconvex hierarchy, so is in  $\mathcal{MQH}$ . Then he applies his joint work with Haglund [42] and Hsu [47] to conclude that groups with a malnormal quasiconvex hierarchy are virtually special. One may then choose the Dehn filling of  $G'$  to be induced from a Dehn filling of  $G$ , and thus the Dehn filling of  $G$  will be virtually special. The main difference in the new proof of this theorem in [5] is that we first form a malnormal hierarchy of  $G'$  which terminates in copies of  $\mathcal{H}'$ . This gives a malnormal hierarchy for any Dehn filling of  $G'$ , giving the same conclusion.

The MSQT is the key result that Wise uses to prove that groups in  $\mathcal{QVH}$  are virtually special (Theorem 6.11).

**6.7. Weak separability of subgroups.** The starting point for applying Wise's results to prove his conjecture is the following result proved in the appendix to the paper:

**Theorem 6.15** ([3, Agol-Groves-Manning, Appendix]). *Let  $G$  be a hyperbolic group, and  $H < G$  a quasi-convex virtually special subgroup. Then  $H$  is weakly separable in  $G$ .*

The proof of this result is an inductive argument using relatively hyperbolic Dehn filling. It is a direct generalization of the previously mentioned result (Theorem 5.2) that if hyperbolic groups are residually finite, then quasiconvex subgroups are separable. The proof is by induction on **height** of quasiconvex subgroups, which measures how many conjugates of a subgroup intersect in an infinite group. So finite groups have height zero, almost malnormal groups have height 1. One uses relative hyperbolic Dehn filling to reduce the height, and eventually find a quotient in which the image of the subgroup is finite. Note that the same induction on height is used by Wise in the proof of Theorem 6.11 in order to reduce the case of  $\mathcal{QVH}$  to the  $\mathcal{MQH}$  case.

Hyperbolicity is used in a crucial way in the proof of this theorem, making it inapplicable for example to the examples of Burger-Mozes (Theorem 6.8).

## 7. Outline of the proof of Wise's conjecture

The proof of Wise's conjecture (Theorem 6.7) is by induction on dimension. Let  $X$  be a compact locally CAT(0) cube complex with  $G = \pi_1 X$  hyperbolic. Let  $W \looparrowright X$  be the immersed hyperplane complex. Then the maximal dimension of cubes in  $W$  is one less than those in  $X$ , so by induction we may assume that  $W$  is virtually special. Then we may apply weak separability to conclude under these hypotheses that

**Theorem 7.1.** *There exists  $G'' \triangleleft G$ ,  $G/G'' \cong \mathcal{G}$ ,  $\tilde{X}/G'' \cong \mathcal{X}$  such that  $\mathcal{X}$  has 2-sided embedded acylindrical compact hyperplanes.*

The acylindrical hypothesis is equivalent to the condition that the fundamental groups of the hyperplanes of  $\mathcal{X}$  are malnormal in  $G''$ . If  $\mathcal{X} \rightarrow X$  were a finite-sheeted cover, then we would be done, since we would have proved that  $\pi_1(X)$  is in  $\mathcal{QVH}$ . However, the proof of the theorem produces an infinite-sheeted regular cover.

**Definition 7.2** (Crossing Graph). Let  $\Gamma(\mathcal{X})$  be a graph with vertex set  $V(\Gamma(\mathcal{X})) = \mathcal{W}$  the hyperplanes of  $\mathcal{X}$ , and edges  $(W_1, W_2) \in E(\Gamma(\mathcal{X}))$  if  $W_1 \cap W_2 \neq \emptyset$  or if there is an essential cylinder going between  $W_1$  and  $W_2$ .

**Definition 7.3** (Coloring space). Let  $[n] = \{1, \dots, n\}$ . Let

$$C_n(\Gamma) = \{c : V(\Gamma) \rightarrow [n] \mid c(W_1) \neq c(W_2), \forall (W_1, W_2) \in E(\Gamma)\}$$

denote the space of  $n$ -colorings of the graph  $\Gamma$ .

We regard  $C_n(\Gamma)$  as a closed subspace of the Cantor set  $[n]^{V(\Gamma)}$ . If  $\deg(\Gamma) \leq k$ , then  $C_{k+1}(\Gamma) \neq \emptyset$ .

A coloring  $c \in C_n(\Gamma(\mathcal{X}))$  gives rise to a hierarchy of  $\mathcal{X}$ : cut along the hyperplanes colored 1, then the hyperplanes colored 2, ..., and finally the hyperplanes colored  $n$ . What is left at the ends are stars of the vertices of  $\mathcal{X}$ , with residues of the colorings remaining on the boundary facets. We call these colored polyhedra.

The idea of the proof is to "reverse-engineer" a hierarchy of a finite-sheeted cover of  $X$ , which is modeled on the hierarchy coming from a coloring of  $\mathcal{X}$ . We want to find a finite collection of colored polyhedra which is balanced, so that the number of colorings of a face is the same for the two polyhedra containing the face.

Then we may glue together polyhedra inductively, in order to reverse-engineer a hierarchy of a finite-sheeted cover, which is therefore virtually special by Wise (Figure 7.1).

**7.1. Colorings of graphs.** I'll discuss a lemma which is used in the proof of Wise's conjecture.

Let  $\Gamma$  be a graph of bounded valence  $\leq k$ , and let  $\mathcal{G}$  be a group acting cocompactly on  $\Gamma$ .

Let  $C_n(\Gamma)$  be the space of all colorings of  $\Gamma$ . Then  $C_n(\Gamma)$  is a compact topological space, considered as a closed subspace of the Cantor set  $[n]^\Gamma$ .

**Lemma 7.4.** *There exists a probability measure  $\mu$  on  $C_{k+1}(\Gamma)$  which is  $\mathcal{G}$ -invariant.*

The proof of this lemma proceeds by coloring the vertices  $V(\Gamma)$  randomly with  $n$ -colors,  $n \geq k + 1$ . The probability that two endpoints of an edge  $e \in E(\Gamma)$  have the same color is  $1/n$ . One can produce an  $(n - 1)$ -coloring of the vertices, by sending each vertex colored  $n$  to the smallest color unused by its neighbors. By induction then, one produces a measure on

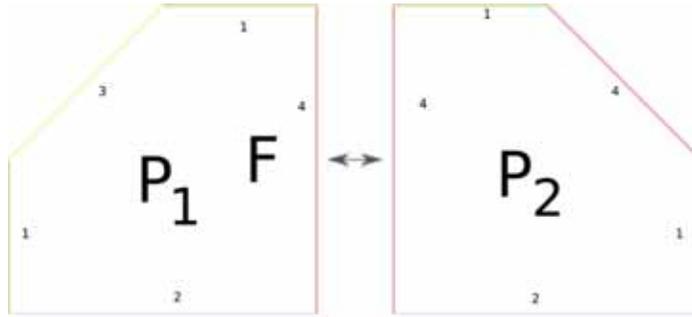


Figure 7.1. Gluing polyhedra at the 4th stage of the hierarchy should preserve the lower stages

$(k + 1)$ -colorings of  $V(\Gamma)$  which have probability of coloring the endpoints of  $e$  the same color as  $\leq 1/n$ . Taking a weak-\* limit of these measures, one obtains a  $\mathcal{G}$ -invariant measure  $\mu$  on  $V(\Gamma)^{[k+1]}$  which is supported on the colorings of  $\Gamma$ .

**7.2. Colorings and hierarchies.** The probability measure is just an artifice to construct a solution to the *gluing equations*. We want to reverse engineer a hierarchy of a finite-sheeted cover. We have a finite (non-compact) hierarchy associated to the cover  $\mathcal{X}$ . The probability measure allows us to extract some finiteness associated to this hierarchy.

**7.3. Polyhedra and facets.** Let  $\mathcal{P}$  denote the stars of vertices of  $\mathcal{X}$ , which we will call *polyhedra*. Let  $\mathcal{F}$  denote the facets of  $\mathcal{X}$ , which are dual to each edge of  $\mathcal{X}$ , and are the facets of the polyhedra  $\mathcal{P}$ . Each facet  $F \in \mathcal{F}$  will be contained uniquely in two polyhedra  $P, Q \in \mathcal{P}$ ,  $P \cap Q = F$ . There are 4 polygons in the example in Figure 6.6 up to the action of  $\mathcal{G}$  (we won't draw  $P'$  and  $Q'$  which are duplicates of  $P$  and  $Q$ ). As a concrete example, we take a covering space of  $X$  which kills the red curves, and kills the third power of the blue curves, giving a cover looking like Figure 7.2. Note that only half of the cover is drawn; the other half is obtained by doubling along the blue curves to get an infinite surface without boundary.

**7.4. Supercoloring.** Each polyhedron and facet of  $\mathcal{X}$  will correspond uniquely to one of  $X$  via the covering  $\mathcal{X} \rightarrow X$ .

We refine the  $k + 1$ -coloring of the hyperplanes  $\mathcal{W}$  by the coloring of a neighborhood of size  $j$  in  $\Gamma(\mathcal{X})$ , where  $j$  is the color of a vertex, to get *supercolored* hyperplanes. The facets  $F \in \mathcal{F}$  get supercolored by their corresponding hyperplanes, and polyhedra will be supercolored by their facets.

**7.5. Polyhedral gluing equations.** The variables for the gluing equations will be super colored polyhedra, and the gluing equations will say that for a given super colored facet  $F$ , the super colorings of  $P$  which induce the same super coloring of  $F$  must equal the super colorings of  $Q$  which induce the super coloring of  $F$ . We require that the variables are  $\mathcal{G}$ -invariant, in which case they are determined by finitely many variables corresponding to the polyhedra of  $X$  (or  $\mathcal{G}$ -orbits of super colored polyhedra of  $\mathcal{X}$ ).

The  $\mathcal{G}$ -invariant measure  $\mu$  gives a solution to the gluing equations with non-negative weights. Then we can get an integral solution to the gluing equations with non-negative

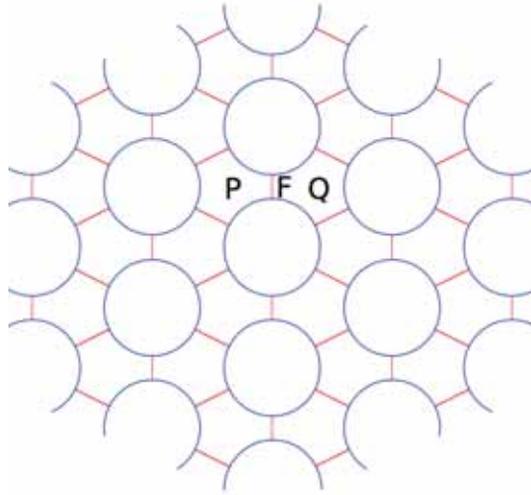


Figure 7.2. The cover  $\mathcal{X}$  of the cube complex in Figure 6.6 and polyhedra and facet

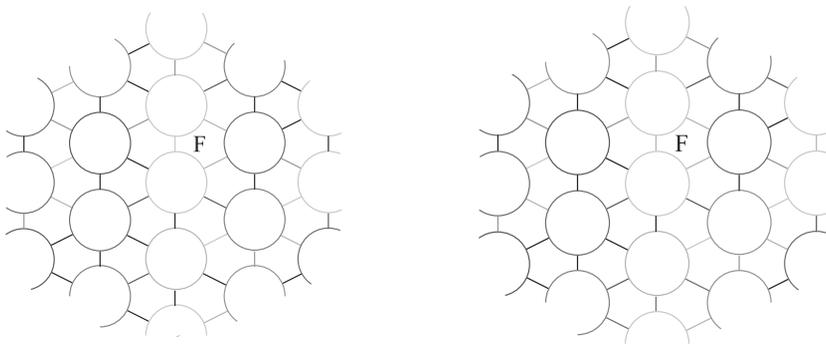


Figure 7.3. The face  $F$  has different supercolorings, even though the facet is colored the same in both colorings

weights, since the equations are linear with integral coefficients. We take the integral solution to the polyhedral gluing equations, and use them to glue up a finite-sheeted cover of  $X$ , which is “modeled” on the hierarchies associated to colorings of  $\mathcal{X}$ .

**7.6. Gluing up the hierarchy.** We construct a sequence of (usually disconnected) finite cube complexes  $\mathcal{V}_j$ ,  $k + 1 \geq j \geq 0$ , with boundary pattern  $\{\partial_1(\mathcal{V}_j), \dots, \partial_j(\mathcal{V}_j)\}$  determined by the unpaired faces colored  $j$ . The final stage  $\mathcal{V}_0$  will be a finite-sheeted cover of  $X$ . The first stage  $\mathcal{V}_{k+1}$  is obtained by taking a number of copies of each supercolored polyhedron determined by the integral solution to the gluing equations. In our example,  $k = 6$ , so the first stage is  $\mathcal{V}_7$  (Figure 7.4). If we glued the faces of the polyhedra  $\mathcal{V}_{k+1}$  together preserving colors, then we would obtain a finite-sheeted branched cover of  $X$ . So we have to be careful at each stage that the gluing extends to an unbranched covering space.

The next stage of the hierarchy  $\mathcal{V}_k$  is obtained from  $\mathcal{V}_{k+1}$  by gluing the faces labeled  $k + 1$  in pairs along matching supercolored faces (in our example,  $k + 1 = 7$  is represented

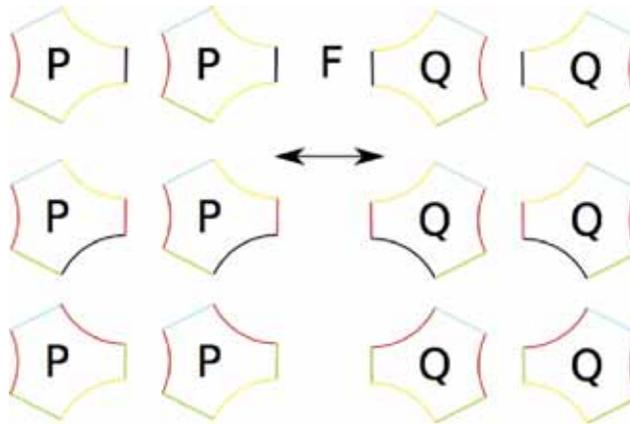


Figure 7.4. Collection of supercolored polyhedra determined by the solution to the gluing equations, giving  $\mathcal{V}_7$

by black, obtaining  $\mathcal{V}_6$ , Figure 7.5).

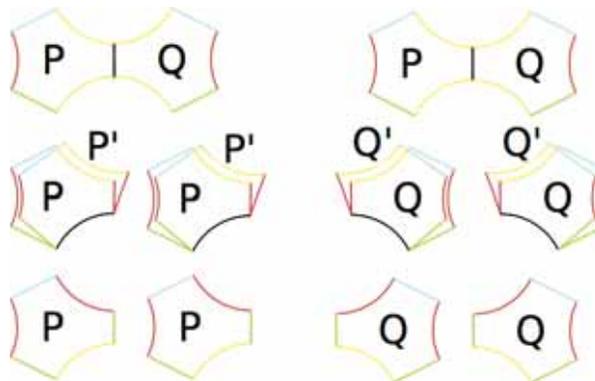


Figure 7.5. Gluing  $\mathcal{V}_7$  to get  $\mathcal{V}_6$

We glue  $\mathcal{V}_5$  from a cover  $\tilde{\mathcal{V}}_6$  of  $\mathcal{V}_6$  by gluing the boundary pattern  $\partial_6 \mathcal{V}_6$  (which in our example is colored yellow Figure 7.6):

The supercoloring guarantees that the two sides of  $\partial_6 \mathcal{V}_6$  have consistently supercolored hyperplanes, and therefore is a finite-sheeted cover of the hyperplane in a representative coloring of  $\mathcal{X}$  (Figure 7.6 (a)). The MSQT allows us to pass to a finite-sheeted cover  $\tilde{\mathcal{V}}_6$  in which both sides of  $\partial_6 \tilde{\mathcal{V}}_6$  match by an isometry (Figure 7.6 (b)).

We obtain  $\mathcal{V}_i$  from  $\mathcal{V}_{i+1}$  by finding a covering space  $\tilde{\mathcal{V}}_{i+1} \rightarrow \mathcal{V}_{i+1}$  in which the boundary pattern  $\partial_{i+1} \tilde{\mathcal{V}}_{i+1}$  may be matched up in pairs which reverse the coorientations and preserve super colorings. Constructing the cover  $\tilde{\mathcal{V}}_{i+1}$  requires another application of Wise’s MSQT.

The cube complex  $\mathcal{V}_0$  will have no boundary pattern, and thus will give a finite-sheeted covering space  $\mathcal{V}_0 \rightarrow X$  and which has by construction has embedded acylindrical hyperplanes, and therefore a malnormal hierarchy.

One more application of Wise’s theorem ( $\mathcal{MQH} \implies$  virtually special) gives a cover  $\tilde{\mathcal{V}}_0 \rightarrow X$  which is special.

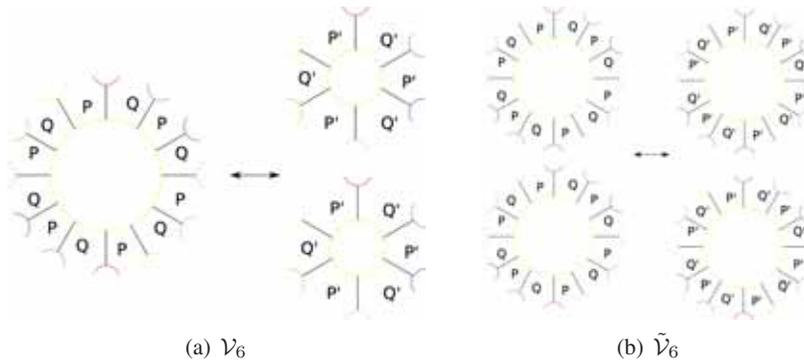


Figure 7.6. Taking a cover  $\tilde{\mathcal{V}}_6$  of  $\mathcal{V}_6$  to be able to glue to get  $\mathcal{V}_5$

### 8. 3-manifold applications

**8.1. Non-positive curvature.** We state a result that combines the statements of theorems of Liu and Przytycki-Wise:

**Theorem 8.1** ([58, Theorem 1.1] [71, Corollary 1.4]). *Let  $M$  be an aspherical compact 3-manifold. The following are equivalent:*

- (1)  $M$  admits a complete metric of non-positive curvature
- (2)  $M$  is virtually homotopic to a special cube complex
- (3)  $\pi_1(M)$  virtually embeds in a right-angled Artin group
- (4)  $\pi_1(M)$  is virtually RFRS

In particular, such manifolds are virtually fibered.

The manifolds which do not admit a metric of non-positive curvature are graph manifolds, and have been characterized by Svetlov in terms of the BKN equations [78].

A corollary of this result is that if  $M$  admits a non-positively curved metric, then  $\pi_1(M)$  is linear (in fact, embeds in  $GL(n, \mathbb{Z})$ ). It is still unresolved whether graph manifold groups are linear. It would be remarkable if there are examples of fibered graph manifolds with non-linear fundamental group, since it would imply the existence of non-linear mapping class groups.

**8.2. Virtual torsion.** Let  $K$  be a finitely generated abelian group.

**Theorem 8.2** ([76, Sun 2013]). *Given  $M$  a closed hyperbolic 3-manifold, there is a finite-sheeted cover  $\tilde{M} \rightarrow M$  such that  $H_1(\tilde{M}) \cong K \oplus L$ .*

For each summand  $\mathbb{Z}/N\mathbb{Z}$  of  $K$ , Sun constructs an immersed complex  $C_N \rightarrow M$  which has a surface with one boundary component which wraps  $N$  times around a loop, and such that  $\pi_1(C_N) \rightarrow \pi_1(M)$  is an injection, in fact with quasiconvex image. Take such a complex for each cyclic summand of  $K$ , and immerse a wedge of these complexes in  $M$  to get a quasiconvex immersion of a complex  $C \rightarrow M$  such that  $\pi_1(C) < \pi_1(M)$  is quasiconvex. We also have by construction  $H_1(C; \mathbb{Z}) \cong K$ . By the virtual retract property, there is a cover  $\tilde{M} \rightarrow M$  such that there is a retract  $r : \pi_1(\tilde{M}) \rightarrow \pi_1(C)$ . Then we have a retract

$r_* : H_1(\tilde{M}) \rightarrow H_1(C)$ . Therefore, we have  $H_1(\tilde{M}) \cong H_1(C) \oplus \ker(r_*) = K \oplus L$ , for  $L = \ker(r_*)$ .

**Theorem 8.3** ([77, Sun 2014]). *Let  $M$  be a closed hyperbolic 3-manifold. For any closed manifold  $N$ , there is a finite cover  $\tilde{M} \rightarrow M$  such that there is a degree 2 map  $\rho : \tilde{M} \rightarrow N$ .*

Since the map  $H^3(N; \mathbb{Z}/p\mathbb{Z}) \rightarrow H^3(\tilde{M}; \mathbb{Z}/\mathbb{Z})$  is an isomorphism for  $p$  odd, there is an embedding of cohomology rings  $H^*(N; \mathbb{Z}/p\mathbb{Z}) \rightarrow H^*(\tilde{M}; \mathbb{Z}/p\mathbb{Z})$ . Thus, not only can one achieve arbitrary torsion in covers of a hyperbolic 3-manifold, but one can also embed any cohomology ring of a 3-manifold, at least with odd order coefficients (one may also use rational coefficients).

**8.3. Heegaard gradient.** For  $M$  a closed 3-manifold, the **Heegaard genus**  $g(M)$  is the minimal genus of a surface  $\Sigma_g \subset M$  such that  $\Sigma_g$  bounds handlebodies to each side ( $\Sigma_g$  is a **Heegaard surface**). The **Heegaard gradient** of  $M$  is

$$\nabla g(M) = \inf_{\tilde{M} \rightarrow M \text{ finite}} \frac{2g(\tilde{M}) - 2}{[\pi_1 M : \pi_1 \tilde{M}]}$$

This notion was introduced by Lackenby to probe the virtual Haken conjecture [54].

If  $M$  is fibered, then it is easy to see that  $\nabla g(M) = 0$ .

**Conjecture 8.4** (Lackenby [54]). *Let  $M$  be a closed hyperbolic 3-manifold.  $M$  is virtually fibered if and only if  $\nabla g(M) = 0$ .*

This conjecture now follows (essentially trivially) from the virtual fibering conjecture, and therefore hyperbolic 3-manifolds have  $\nabla g(M) = 0$ . Note that Ichihara has shown that Seifert-fibered 3-manifolds with infinite fundamental group have zero Heegaard gradient, even though some of them are not virtually fibered. It remains to compute the Heegaard gradients of graph manifolds which are not virtually fibered.

**Theorem 8.5.** *A closed orientable 3-manifold  $M$  has  $\nabla g(M) \leq 0$  if and only if it is prime or  $M \cong \mathbb{RP}^3 \# \mathbb{RP}^3$ .*

*Proof.* First, note that if  $M$  is not prime or  $\mathbb{RP}^3 \# \mathbb{RP}^3$ , then  $\nabla g(M) > 0$ . The sphere decomposition of  $M$  gives a graph-of-groups decomposition of  $\pi_1 M$  with trivial edge groups. After passing to a finite-sheeted cover, one may assume that the vertices of this graph all have degree  $\geq 3$ . Then the corank of  $\pi_1 M$  is  $> 1$  (the **corank** is the maximal rank free group surjected by  $\pi_1(M)$ ). As one passes to further finite-sheeted covers, the corank grows at least linearly with the index, and therefore the corank gradient is  $> 0$ , a fortiori the rank gradient and Heegaard gradient.

If  $|\pi_1(M)| < \infty$ , then  $\nabla g(M) < 0$  by the Poincaré conjecture, and if  $|\pi_1(M)| \cong \mathbb{Z}$  or  $\mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$ , then  $\nabla g(M) = 0$ .

Now, suppose  $M$  is aspherical. If  $M$  has non-zero Gromov norm, then  $M$  virtually fibers, and therefore  $\nabla g(M) = 0$  [71]. If  $M$  has zero Gromov norm, then  $M$  is a graph manifold.

If  $M$  is Seifert-fibered (with infinite fundamental group), then this was proved by Ichihara [48]. It is easy to check that this holds for graph manifolds with non-trivial JSJ decomposition as well. There is a finite-sheeted cover in which each Seifert piece is homeomorphic to  $\Sigma \times S^1$ , for some surface with boundary  $\Sigma$ . The Heegaard genus of each piece

is  $b_1(\Sigma) + 1$ . By passing to a further cover, we may assume that the JSJ decomposition is bipartite, so that  $M = M_1 \cup_T M_2$ , where  $M_i \cong \Sigma_i \times S^1$  ( $\Sigma_i$  may be disconnected), and  $T = \partial M_1 = \partial M_2$  is the union of JSJ tori. Each  $M_i$  has a Heegaard splitting of genus  $b_1(M_i)$ , in  $M_i = H_i \cup C_i$ , where  $H_i$  is a union of handlebodies of genus  $b_1(M_i)$ , and  $T = \partial M_i \subset \partial C_i$ . We may construct a Morse function on  $M$ , which has  $T$  as a level set, and induces a perfect Morse function  $M_i$ , which is standard on  $H_i$  and  $C_i$  (although the restriction to  $M_2$  will have the indices flipped). The index 0 critical points lie in  $H_1$ , so that there are  $b_0(M_1)$  index 0 critical points, and similarly the index 3 critical points lie in  $H_2$ , so there are  $b_0(M_2)$  of them. There are  $b_1(H_1) = b_1(M_1)$  critical points of index one in  $H_1$ , and there are  $b_2(M_2) = b_1(M_2) - b_0(M_2)$  critical points of index one in  $C_2$ . Similarly, there are  $b_1(H_2) = b_1(M_2)$  critical points of index two in  $H_2$ , and  $b_2(M_1) = b_1(M_1) - b_0(M_1)$  critical points of index two in  $C_1$ . Now, detelescope the Morse function on  $M$  to get a Heegaard splitting of  $M$ , then cancel all but one of the index 0 critical points with index 1 critical points. This gives a Morse function with  $b_1(H_1) + b_1(M_2) - b_0(M_2) - b_0(M_1) = b_2(M_1) + b_2(M_2) = b_1(\Sigma_1) + b_1(\Sigma_2)$  index one critical points.

Now we observe that by passing to covering spaces  $\tilde{M} \rightarrow M$ , we may make the ratio  $(b_1(\tilde{\Sigma}_1) + b_1(\tilde{\Sigma}_2))/[\tilde{M} : M]$  arbitrarily close to zero, by unwrapping the  $S^1$  direction of  $M_1$  and  $M_2$  an arbitrarily large amount. This shows that the Heegaard gradient is zero.  $\square$

**Definition 8.6.** Let  $G$  be a group, then  $d(G)$  is the minimal number of generators needed to generate  $G$  (if  $G$  is not finitely generated, set  $d(G) = \infty$ ).

Now, suppose  $G$  is a residually finite group. Define

$$\nabla d(G) = \inf_{\tilde{G} < G, [G:\tilde{G}] < \infty} \frac{d(\tilde{G}) - 1}{[G : \tilde{G}]}.$$

If  $(G_n)$  is a chain of subgroups  $G_{n+1} < G_n < G$  with  $[G : G_n] < \infty$ , define

$$\nabla d(G, (G_n)) = \lim_{n \rightarrow \infty} \frac{d(G_n) - 1}{[G : G_n]}.$$

Clearly  $\nabla d(G) \leq \nabla d(G, (G_n))$ .

Let  $M$  be a closed aspherical 3-manifold with  $\pi_1(M) = G$ . Then  $2d(G) \leq g(M)$ , and therefore  $\nabla d(G) = \frac{1}{2} \nabla g(M) = 0$ . It is known that there are manifolds  $M$  with  $d(G) < \frac{1}{2}g(M)$  [14, 57]. If  $M$  is hyperbolic, and  $G_n < G$  are congruence subgroups, then it is shown that  $\nabla g(M, G_n) > 0$  [54, 59]. As observed by Abert-Nikolov, the fixed price conjecture of Gaboriau would imply that  $\nabla d(\pi_1 M, G_n) = 0$  for any cofinal chain  $(G_n)$ .

**Question 8.7.** For  $M$  a closed 3-manifold with  $\pi_1 M = G$ , what is

$$\inf_{[M:N] < \infty} d(\pi_1 N)/g(N)?$$

For further properties of 3-manifold groups, we refer to the comprehensive survey [8, Section 6].

## 9. Cubulated groups

**Theorem 9.1** ([63, Markovic 2012]). *Let  $\Gamma$  be a word-hyperbolic group, such that  $\partial_\infty \Gamma \cong S^2$  (and  $\Gamma$  acts effectively on  $\partial_\infty \Gamma$ ). Suppose moreover that  $\Gamma$  is cubulated. Then  $\Gamma$  is*

isomorphic to a Kleinian group. In particular, if  $\Gamma$  is torsion-free, then  $\Gamma = \pi_1(M)$  for some closed hyperbolic 3-manifold.

This gives a possible approach to Cannon’s conjecture, which is that  $\Gamma$  is a Kleinian group. One could try to carry out the technique of Kahn-Markovic to try to find quasiconvex subgroups with limit sets circles in  $\partial_\infty \Gamma$  which satisfy Bergeron-Wise’s condition, that one can separate any pair of points in  $\partial_\infty \Gamma$  by a circle limit set of a surface subgroup.

We remark that there are many other classes of cubulated hyperbolic groups to which Theorem 6.7 applies:  $C'(\frac{1}{6})$  groups (Wise), random groups at density  $< \frac{1}{6}$  (Ollivier-Wise), certain ascending HNN extensions of free groups (Button, Hagen-Wise), and isometry groups of certain polygonal complexes (Desgroseilliers-Haglund, Futer-Thomas).

## 10. Group theoretic applications

We point out a minor observation regarding Haglund-Wise’s theorem [39]:

**Theorem 10.1.** *Let  $G$  act properly cocompactly and virtually specially on a cube complex  $X$ . Then  $G$  embeds in a finite extension of a RAAG.*

The point is that there is a normal subgroup  $G' \triangleleft G$  such that  $X/G'$  is special. The embedding  $X/G' \rightarrow S_{\Gamma(X/G')}$  is functorial, in that combinatorial automorphisms of  $X/G'$  extend to  $S_{\Gamma(X/G')}$ . Thus,  $G$  embeds in an extension of  $A_{\Gamma(X/G')}$  by  $G/G'$ . Thus, all hyperbolic 3-manifold groups embed in finite extensions of RAAGs. This observation may have importance, for example, in understanding the representations of 3-manifold groups, by examining the representations of finite extensions of RAAGs.

We point out another consequence of virtual specialness. Given a RF group  $G$ , let  $\hat{G}$  denote its profinite completion.

**Definition 10.2.** A group  $G$  is good if for every finite  $\hat{G}$ -module  $M$ , there is an isomorphism  $H^*(\hat{G}, M) \cong H^*(G, M)$ .

**Theorem 10.3.** *Let  $G$  be a virtually compact special group. Then  $G$  is good.*

*Proof.* This follows by induction from [37, Proposition 3.6]. If  $G$  is virtually compact special, then it has a finite-index subgroup which admits a quasi-convex hierarchy. Then  $G = A *_C B$ , where  $A, B, C$  are virtually compact special. By induction,  $A, B, C$  are good groups. Also, by [40], the groups  $A, B, C$  are virtual retracts, and therefore are efficient. So by [37, Proposition 3.6], we conclude that  $G$  is good.  $\square$

**Remark:** We cannot apply directly [37, Proposition 3.9], since we don’t know that  $G$  is subgroup separable, only that quasiconvex subgroups are separable.

In general, cubulated hyperbolic groups are not LERF, for example by Rips’ construction [72, 85]. However, quasiconvex subgroups are separable. So it is natural to ask for which hyperbolic groups are finitely generated subgroups quasiconvex? This is a strong form of coherence.

**Theorem 10.4.** *Negatively curve square complex groups are LERF.*

A negatively curved square complex has vertex links graphs of girth  $\geq 5$ , so that it admits a CAT(-1) metric making each square a hyperbolic square with angles  $2\pi/5$ . This follows from a result of McCammond-Wise that negatively curved square complexes are locally convex [65].

## 11. Open questions

1. (Long-Reid) Can two Kleinian groups which are non-isomorphic have the same profinite completion?

**Remark:** This is equivalent to the question, given two hyperbolic 3-manifold groups, do they have the same collection of finite quotients?

2. Are compact 3-manifold fundamental groups linear?

**Remark:** The only aspherical case left is graph manifolds which don't admit a non-positively curved metric by Theorem 8.1.

3. Is there an algorithm to detect if a compact cube complex is virtually special?
4. Find a bound on the index of a cover of an aspherical 3-manifold which is Haken. The bound should be some computable function of some complexity of the 3-manifold, such as the minimal number of tetrahedra of a triangulation. In principle, there is an algorithm which will find a Haken cover. The most practical approach is likely to enumerate homomorphisms  $\rho : \pi_1(M) \rightarrow K$ ,  $K$  a finite group, and compute  $H_1(\pi_1(M); \mathbb{Q}[K])$ , which is the homology of the covering space corresponding to  $\ker(\rho)$  [27].
5. Let  $M$  be a 3-manifold with  $\text{rank}(H_1(M; \mathbb{F}_p)) \geq 4$ . Does  $M$  admit a regular  $p$ -cover  $\tilde{M}$  with  $b_1(\tilde{M}) > 0$ ? If this were true, it might yield a more practical approach to finding Haken covers [55].
6. For any two hyperbolic 3-manifolds  $M_1, M_2$ , are there fibered covers  $M'_i \rightarrow M_i$  such that there is a non-zero degree map  $M'_1 \rightarrow M'_2$  which preserves the fibering?
7. Do closed hyperbolic 3-manifolds contain immersed quasi-fuchsian surfaces of odd Euler characteristic?
8. [67, Niblo-Wise] Which 3-manifold groups are LERF? No Seifert-Seifert gluings in JSJ?
9. Consider a hyperbolic group  $G$  which acts properly on a cube complex with finitely many orbits of hyperplanes, but not necessarily cocompactly. Is  $G$  virtually special?
10. Which knot groups are RFRS?
11. Are braid groups  $B_n$  RFRS? **Remark:** Mapping class groups are not virtually RFRS in general (cf. [58, Liu]).
12. Does a finite volume hyperbolic 3-manifold  $M$  admit a cover which fibers over  $S^1$  with orientable foliation of the pseudo-Anosov map?
13. For  $M$  a finite-volume hyperbolic 3-manifold,  $\Gamma = \pi_1(M)$ , does  $\text{rank}(\Gamma) = \text{rank}(\hat{\Gamma}) = \max\{\text{rank}(\Gamma/N) \mid N \triangleleft \Gamma, [\Gamma : N] < \infty\}$ ? Note that  $M$  has a finite-sheeted cover  $\tilde{M} \rightarrow M$  which has this property, in fact such that

$$\text{rank}(\pi_1(\tilde{M})) = \text{rank}H_1(\tilde{M}; \mathbb{Z}/2\mathbb{Z}),$$

since a fibered manifold always has a finite-sheeted cover with this property.

14. Is there a strengthening of the Malnormal Special Quotient Theorem? Let  $G$  be hyperbolic and cubulated and  $(G, P)$  be relatively hyperbolic. Are there finitely many

elements that we may exclude in  $P$  so that any Dehn filling which avoids these elements is cubulated? This would be a strengthening of the MSQT (and there is an obvious generalization to multiple peripheral subgroups).

15. Which cocompact lattices in hyperbolic buildings are cubulated?

**Acknowledgements.** Agol is supported by DMS-1105738 and the Simons Foundation.

## References

- [1] Ian Agol, *Tameness of hyperbolic 3-manifolds*, preprint, May 2004, arXiv:math/0405568
- [2] ———, *Criteria for virtual fibering*, J. Topol. **1** (2008), no. 2, 269–284, arXiv:math/0707.4522.
- [3] ———, *The virtual haken conjecture*, Documenta Mathematica **18** (2013), 1045–1087, arXiv:1204.2810, appendix by Agol, Groves, Manning.
- [4] Ian Agol, Daniel Groves, and Jason Fox Manning, *Residual finiteness, qcerf and fillings of hyperbolic groups*, Geometry and Topology **13** (2009), 1043–1073, arXiv:0802.0709.
- [5] Ian Agol, Daniel Groves, and Jason Fox Manning, *An alternate proof of Wise’s malnormal special quotient theorem*, 2014.
- [6] Ian Agol, Darren D. Long, and Alan W. Reid, *The Bianchi groups are separable on geometrically finite subgroups*, Ann. of Math. (2) **153** (2001), no. 3, 599–621, math.GT/9811114.
- [7] I. R. Aitchison and J. H. Rubinstein, *An introduction to polyhedral metrics of nonpositive curvature on 3-manifolds*, Geometry of low-dimensional manifolds, 2 (Durham, 1989), Cambridge Univ. Press, Cambridge, 1990, pp. 127–161.
- [8] Matthias Aschenbrenner, Stefan Friedl, and Henry Wilton, *3-manifold groups*, preprint, 107 pages, 2012, arXiv:1205.0202.
- [9] Gilbert Baumslag, *Residually finite one-relator groups*, Bull. Amer. Math. Soc. **73** (1967), 618–620.
- [10] Nicolas Bergeron, *La conjecture des sous-groupes de surfaces (d’après Jeremy Kahn et Vladimir Marković)*, Astérisque (2013), no. 352, Exp. No. 1055, x, 429–458, Séminaire Bourbaki. Vol. 2011/2012. Exposés 1043–1058.
- [11] Nicolas Bergeron, *Toute variété de dimension 3 compacte et asphérique est virtuellement haken*, January 2014.
- [12] Nicolas Bergeron and Daniel T. Wise, *A boundary criterion for cubulation*, Amer. J. Math. **134** (2012), no. 3, 843–859.
- [13] Mladen Bestvina, *Geometric group theory and 3-manifolds hand in hand: the fulfillment of Thurston’s vision*, Bull. Amer. Math. Soc. (N.S.) **51** (2014), no. 1, 53–70.
- [14] M. Boileau and H. Zieschang, *Heegaard genus of closed orientable Seifert 3-manifolds*, Invent. Math. **76** (1984), no. 3, 455–468.
- [15] Francis Bonahon, *Bouts des variétés hyperboliques de dimension 3*, Ann. of Math. (2)

- 124** (1986), no. 1, 71–158.
- [16] B. H. Bowditch, *Relatively hyperbolic groups*, Internat. J. Algebra Comput. **22** (2012), no. 3, 1250016, 66.
- [17] Marc Burger and Shahar Mozes, *Lattices in product of trees*, Inst. Hautes Études Sci. Publ. Math. (2000), no. 92, 151–194 (2001).
- [18] R. G. Burns, A. Karrass, and D. Solitar, *A note on groups with separable finitely generated subgroups*, Bull. Austral. Math. Soc. **36** (1987), no. 1, 153–160.
- [19] Benjamin A. Burton, J. Hyam Rubinstein, and Stephan Tillmann, *The Weber-Seifert dodecahedral space is non-Haken*, Trans. Amer. Math. Soc. **364** (2012), no. 2, 911–932.
- [20] Danny Calegari, *Notes on agol's virtual haken theorem*, notes, June 2013.
- [21] Danny Calegari and David Gabai, *Shrinkwrapping and the taming of hyperbolic 3-manifolds*, J. Amer. Math. Soc. **19** (2006), no. 2, 385–446 (electronic), arXiv:math/0407161v3.
- [22] Richard D. Canary, *Covering theorems for hyperbolic 3-manifolds*, Low-dimensional topology (Knoxville, TN, 1992), Conf. Proc. Lecture Notes Geom. Topology, III, Internat. Press, Cambridge, MA, 1994, pp. 21–30.
- [23] James W. Cannon, *The combinatorial structure of cocompact discrete hyperbolic groups*, Geom. Dedicata **16** (1984), no. 2, 123–148.
- [24] James W. Cannon and William P. Thurston, *Group invariant Peano curves*, Geom. Topol. **11** (2007), 1315–1355.
- [25] M. Dehn, *Über die Topologie des dreidimensionalen Raumes*, Math. Ann. **69** (1910), no. 1, 137–168.
- [26] ———, *Über unendliche diskontinuierliche Gruppen*, Math. Ann. **71** (1911), no. 1, 116–144.
- [27] Nathan M. Dunfield and William P. Thurston, *The virtual Haken conjecture: experiments and examples*, Geom. Topol. **7** (2003), 399–441 (electronic).
- [28] B. Farb, *Relatively hyperbolic groups*, Geom. Funct. Anal. **8** (1998), no. 5, 810–840.
- [29] Stefan Friedl and Takahiro Kitayama, *The virtual fibering theorem for 3-manifolds*, October 2012, arXiv:1210.4799.
- [30] David Gabai, *Foliations and the topology of 3-manifolds*, J. Differential Geom. **18** (1983), no. 3, 445–503.
- [31] ———, *Foliations and the topology of 3-manifolds. II*, J. Differential Geom. **26** (1987), no. 3, 461–478.
- [32] ———, *Foliations and the topology of 3-manifolds. III*, J. Differential Geom. **26** (1987), no. 3, 479–536.
- [33] Rita Gitik, *Doubles of groups and hyperbolic LERF 3-manifolds*, Ann. of Math. (2) **150** (1999), no. 3, 775–806.
- [34] Mikhael Gromov, *Word hyperbolic groups*, Essays in Group Theory (S. M. Gersten, ed.), Mathematical Sciences Research Institute Publications, vol. 8, Springer-Verlag, New York, 1987, [http://link.springer.com/chapter/10.1007/978-1-4613-9586-7\\_3](http://link.springer.com/chapter/10.1007/978-1-4613-9586-7_3), pp. 75–264.

- [35] Edna K. Grossman, *On the residual finiteness of certain mapping class groups*, J. London Math. Soc. (2) **9** (1974/75), 160–164.
- [36] Daniel Groves and Jason Fox Manning, *Dehn filling in relatively hyperbolic groups*, Israel Journal of Mathematics **168** (2008), 317–429, <http://dx.doi.org/10.1007/s11856-008-1070-6>.
- [37] 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.
- [38] Frédéric Haglund, *Finite index subgroups of graph products*, Geom. Dedicata **135** (2008), 167–209, <http://dx.doi.org/10.1007/s10711-008-9270-0>.
- [39] Frederic Haglund and Daniel Wise, *Special cube complexes*, Geom. Funct. Anal. (2007), 1–69.
- [40] Frédéric Haglund and Daniel T. Wise, *Special cube complexes*, Geom. Funct. Anal. **17** (2008), no. 5, 1551–1620, <http://dx.doi.org/10.1007/s00039-007-0629-4>.
- [41] ———, *Coxeter groups are virtually special*, Adv. Math. **224** (2010), no. 5, 1890–1903.
- [42] ———, *A combination theorem for special cube complexes*, Ann. of Math. (2) **176** (2012), no. 3, 1427–1482, <http://dx.doi.org/10.4007/annals.2012.176.3.2>.
- [43] Wolfgang Haken, *Über das Homöomorphieproblem der 3-Mannigfaltigkeiten. I*, Math. Z. **80** (1962), 89–120.
- [44] Marshall Hall, Jr., *Subgroups of finite index in free groups*, Canadian J. Math. **1** (1949), 187–190.
- [45] Richard S. Hamilton, *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17** (1982), no. 2, 255–306.
- [46] John Hempel, *Residual finiteness for 3-manifolds*, Combinatorial group theory and topology (Alta, Utah, 1984), Ann. of Math. Stud., vol. 111, Princeton Univ. Press, Princeton, NJ, 1987, pp. 379–396.
- [47] Tim Hsu and Daniel T. Wise, *Cubulating malnormal amalgams*, <http://comet.lehman.cuny.edu/behrstock/cbms/program.html>, preprint.
- [48] Kazuhiro Ichihara, *Heegaard gradient of Seifert fibered 3-manifolds*, Bull. London Math. Soc. **36** (2004), no. 4, 537–546.
- [49] William H. Jaco and Peter B. Shalen, *Seifert fibered spaces in 3-manifolds*, Mem. Amer. Math. Soc. **21** (1979), no. 220, viii+192.
- [50] Klaus Johannson, *Homotopy equivalences of 3-manifolds with boundaries*, Lecture Notes in Mathematics, vol. 761, Springer, Berlin, 1979.
- [51] Jeremy Kahn and Vladimir Markovic, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, Ann. of Math. (2) **175** (2012), no. 3, 1127–1190.
- [52] Rob Kirby, *Problems in low dimensional manifold theory*, Algebraic and geometric topology (Proc. Sympos. Pure Math., Stanford Univ., Stanford, Calif., 1976), Part 2, Proc. Sympos. Pure Math., XXXII, Amer. Math. Soc., Providence, R.I., 1978, pp. 273–312.
- [53] Robion Kirby, *Problems in low-dimensional topology*, Geometric topology (Athens, GA, 1993) (Rob Kirby, ed.), AMS/IP Stud. Adv. Math., vol. 2, Amer. Math. Soc., Providence, RI, 1997, pp. 35–473.

- [54] Marc Lackenby, *Heegaard splittings, the virtually Haken conjecture and property  $(\tau)$* , Invent. Math. **164** (2006), no. 2, 317–359.
- [55] ———, *Finite covering spaces of 3-manifolds*, Proceedings of the International Congress of Mathematicians. Volume II, Hindustan Book Agency, New Delhi, 2010, pp. 1042–1070.
- [56] Tao Li, *Boundary curves of surfaces with the 4-plane property*, Geom. Topol. **6** (2002), 609–647 (electronic).
- [57] ———, *Rank and genus of 3-manifolds*, J. Amer. Math. Soc. **26** (2013), no. 3, 777–829.
- [58] Yi Liu, *Virtual cubulation of nonpositively curved graph manifolds*, preprint, 30 pages, 2011, arXiv:1110.1940.
- [59] D. D. Long, A. Lubotzky, and A. W. Reid, *Heegaard genus and property  $\tau$  for hyperbolic 3-manifolds*, J. Topol. **1** (2008), no. 1, 152–158.
- [60] D. D. Long and A. W. Reid, *Subgroup separability and virtual retractions of groups*, Topology **47** (2008), no. 3, 137–159.
- [61] A. I. Mal'cev., *On homomorphisms onto finite groups.*, American Mathematical Society Translations (2) **119** (1983), 67–79, <http://www.ams.org/bookstore?fn20&arg1trans2series&ikeyTRANS2-119>.
- [62] Jason Fox Manning and Eduardo Martínez-Pedroza, *Separation of relatively quasiconvex subgroups*, Pacific J. Math. **244** (2010), no. 2, 309–334.
- [63] Vladimir Markovic, *Criterion for Cannon's conjecture*, Geom. Funct. Anal. **23** (2013), no. 3, 1035–1061.
- [64] Saburo Matsumoto, *Non-separable surfaces in cubed manifolds*, Proc. Amer. Math. Soc. **125** (1997), no. 11, 3439–3446.
- [65] Jonathan P. McCammond and Daniel T. Wise, *Locally quasiconvex small-cancellation groups*, Trans. Amer. Math. Soc. **360** (2008), no. 1, 237–271 (electronic).
- [66] J. Milnor, *A note on curvature and fundamental group*, J. Differential Geometry **2** (1968), 1–7.
- [67] Graham A. Niblo and Daniel T. Wise, *Subgroup separability, knot groups and graph manifolds*, Proc. Amer. Math. Soc. **129** (2001), no. 3, 685–693.
- [68] Denis V. Osin, *Peripheral fillings of relatively hyperbolic groups*, Invent. Math. **167** (2007), no. 2, 295–326.
- [69] Christos D. Papakyriakopoulos, *On Dehn's lemma and the asphericity of knots*, Ann. of Math. (2) **66** (1957), 1–26.
- [70] Grisha Perelman, *Ricci flow with surgery on three-manifolds*, <http://front.math.ucdavis.edu/math.DG/0303109>, arXiv:math.DG/0303109.
- [71] Piotr Przytycki and Daniel Wise, *Mixed 3-manifolds are virtually special*, 2012.
- [72] E. Rips, *Subgroups of small cancellation groups*, Bull. London Math. Soc. **14** (1982), no. 1, 45–47.
- [73] Michah Sageev, *Ends of group pairs and non-positively curved cube complexes*, Proc. London Math. Soc. (3) **71** (1995), no. 3, 585–617.
- [74] Peter Scott, *Subgroups of surface groups are almost geometric*, Journal of the London

- Mathematical Society. Second Series **17** (1978), no. 3, 555–565.
- [75] John Stallings, *Group theory and three-dimensional manifolds*, Yale University Press, New Haven, Conn.-London, 1971, A James K. Whittemore Lecture in Mathematics given at Yale University, 1969, Yale Mathematical Monographs, 4.
- [76] Hongbin Sun, *Virtual homological torsion of closed hyperbolic 3-manifolds*, September 2013, arXiv:1309.1511.
- [77] ———, *Virtual domination of 3-manifolds*, January 2014, arXiv:1401.7049.
- [78] P. V. Svetlov, *Metrics of nonpositive curvature on infinite graph-manifolds*, Algebra i Analiz **13** (2001), no. 4, 174–195.
- [79] ———, *Graph-manifolds of nonpositive curvature are virtually fibered over a circle*, Algebra i Analiz **14** (2002), no. 5, 188–201.
- [80] William P. Thurston, <http://library.msri.org/books/gt3m/> *The geometry and topology of 3-manifolds*, Lecture notes from Princeton University, 1978–80, <http://library.msri.org/books/gt3m/>.
- [81] William P. Thurston, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. (N.S.) **6** (1982), no. 3, 357–381.
- [82] William P. Thurston, *Hyperbolic structures on 3-manifolds i: Deformation of acylindrical manifolds*, Annals of Mathematics **124** (1986), 203–246.
- [83] ———, *Hyperbolic structures on 3-manifolds, ii: Surface groups and 3-manifolds which fiber over the circle*, preprint, 1998, arXiv:math/9801045.
- [84] Friedhelm Waldhausen, *On irreducible 3-manifolds which are sufficiently large*, Ann. of Math. (2) **87** (1968), 56–88.
- [85] D. T. Wise, *Cubulating small cancellation groups*, Geom. Funct. Anal. **14** (2004), no. 1, 150–214.
- [86] Daniel Wise, *The structure of groups with a quasiconvex hierarchy*, preprint, 2011.

University of California, Berkeley, 970 Evans Hall #3840, Berkeley, CA, 94720-3840

E-mail: [ianagol@math.berkeley.edu](mailto:ianagol@math.berkeley.edu)

# *L*-functions and automorphic representations

James Arthur

**Abstract.** Our goal is to formulate a theorem that is part of a recent classification of automorphic representations of orthogonal and symplectic groups. To place it in perspective, we devote much of the paper to a historical introduction to the Langlands program. In our attempt to make the article accessible to a general mathematical audience, we have centred it around the theory of *L*-functions, and its implicit foundation, Langlands' principle of functoriality.

**Mathematics Subject Classification (2010).** Primary 11F66, 11F70, 11R37, 11F57, 22F55.

**Keywords.** *L*-functions, automorphic representations, functoriality, classical groups, discrete spectrum.

## 1. Preface

Suppose that  $f(x)$  is a monic polynomial of degree  $n$  with integral coefficients. For any prime number  $p$ , we can then write  $f(x)$  as a product

$$f(x) \equiv f_1(x) \cdots f_r(x) \pmod{p}$$

of irreducible factors modulo  $p$ . It is customary to leave aside the finite set  $S$  of primes for which these factors are not all distinct. For those that remain, we consider the mapping

$$p \longrightarrow \Pi_p = \{n_1, \dots, n_r\}, \quad n_i = \deg(f_i(x)),$$

from primes  $p \notin S$  to partitions  $\Pi_p$  of  $n$ . Here are two basic questions:

- (I) Is there some independent way to characterize the preimage

$$\mathcal{P}(\Pi) = \{p \notin S : \Pi_p = \Pi\}$$

of any partition  $\Pi$  of  $n$ ?

- (II) What is the density of  $\mathcal{P}(\Pi)$  in the set of all primes, or for that matter, the set of all positive integers.

Suppose for example that  $f(x) = x^2 + 1$ . Then  $S$  consists of the single prime 2, while

$$\mathcal{P}(1, 1) = \{p : p \equiv 1 \pmod{4}\}$$

and

$$\mathcal{P}(2) = \{p : p \equiv 3 \pmod{4}\}.$$

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

This well known supplement of the law of quadratic reciprocity gives us a striking answer to (I). As for (II),  $\mathcal{P}(1, 1)$  and  $\mathcal{P}(2)$  are each known to have density  $\frac{1}{2}$  in the set of all primes. Combined with the prime number theorem, this gives an asymptotic formula

$$\lim_{x \rightarrow \infty} |\mathcal{P}(\Pi, x)| / \left( \frac{x}{\log x} \right) = 1/2, \quad \Pi = \{1, 1\}, \{2\},$$

for the set

$$\mathcal{P}(\Pi, x) = \{p \in \mathcal{P}(\Pi) : p \leq x\}$$

of primes  $p \in \mathcal{P}(\Pi)$  with  $p \leq x$ . If the generalized Riemann hypothesis can be proved, there will be much sharper asymptotic estimates.

In general, the two questions have significance that goes well beyond their obvious initial interest. The first could perhaps be regarded as the fundamental problem in algebraic number theory. The second has similar standing in the area of analytic number theory. Both questions are central to the Langlands program.

In this article we combine an introduction to the theory of automorphic forms with a brief description of a recent development in the area. I would like to make the discussion as comprehensible as I can to a general mathematical audience. The theory of automorphic forms is often seen as impenetrable. Although the situation may be changing, the aims and techniques of the subject are still some distance from the common “mathematical canon”. At the suggestion of Bill Casselman, I have tried to present the subject from the perspective of the theory of  $L$ -functions. These are concrete, appealing objects, whose behaviour reflects the fundamental questions in the subject. I will use them to illustrate the basic tenets of the Langlands program. As we shall see in §4,  $L$ -functions are particularly relevant to the principle of functoriality, which can be regarded as a foundation of the Langlands program.

The new development is a classification [4] of automorphic representations of classical groups  $G$ , specifically orthogonal and symplectic groups, in terms of those of general linear groups  $GL(N)$ . It was established by a multifaceted comparison of trace formulas. These are the trace formula for  $G$  [1] and its stabilization [2], which is now unconditional thanks to the proof of the fundamental lemma [33], and the twisted trace formula for  $GL(N)$  [22] and its stabilization, which is still under construction [31, 41, 42]. We refer the reader to the surveys [3, 5, 6], each written from a different perspective, for a detailed description of the classification. We shall be content here to formulate a consequence of the classification in terms of our two themes,  $L$ -functions and the principle of functoriality.

Upon reflection after its completion, I observe that the article is not typical of plenary reports for an ICM. It represents a broader, and perhaps denser, introduction than is customary. I hope that the nonspecialist for whom the article is intended will find the details comprehensive enough without being overwhelming.

## 2. Classical introduction

Recall that a Dirichlet series is an infinite series

$$\sum_{n=1}^{\infty} \frac{a_n}{n^s}, \tag{2.1}$$

for a complex variable  $s$  and complex coefficients  $a_n$ . If the coefficients have moderate growth, the series converges when  $s$  lies in some right half plane in  $\mathbb{C}$ . If they are bounded,

for example, the series converges absolutely when the real part  $\operatorname{Re}(s)$  of  $s$  is greater than 1. More generally, if the coefficients satisfy a bound

$$|a_n| \leq cn^k,$$

for positive real numbers  $c$  and  $k$ , the series converges absolutely in the right half plane  $\operatorname{Re}(s) > k + 1$ . Since the convergence is uniform in any smaller right half plane, the sum of the series is an analytic function of  $s$  on the open set  $\operatorname{Re}(s) > k + 1$ .

The most famous example is the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \quad (2.2)$$

Since the coefficients are all equal to 1, this converges as an analytic function of  $s$  for  $\operatorname{Re}(s) > 1$ . Euler had studied the series earlier for real values of  $s$ . He discovered a remarkable formula

$$\zeta(s) = \prod_p \left( \frac{1}{1 - p^{-s}} \right) \quad (2.3)$$

for  $\zeta(s)$  as a product over all prime numbers  $p$ , which he proved using only the fundamental theorem of arithmetic and the formula for the sum of a geometric series. With the later theory of complex analysis, Riemann was able to extend the domain. He showed that the function has analytic continuation to a meromorphic function of  $s$  in the entire complex plane, which satisfies a functional equation

$$\zeta(s) = 2^s \pi^{s-1} \sin(\pi s/2) \Gamma(1-s) \zeta(1-s) \quad (2.4)$$

in terms of its values at points  $s$  and  $1-s$ . He observed further that if

$$L_{\mathbb{R}}(s) = \pi^{-s/2} \Gamma(s/2),$$

where  $\Gamma(\cdot)$  is the gamma function (here and in (2.4)), the product

$$L(s) = L_{\mathbb{R}}(s) \zeta(s) = L_{\mathbb{R}}(s) \cdot \prod_p \left( \frac{1}{1 - p^{-s}} \right) \quad (2.5)$$

satisfies the symmetric functional equation

$$L(s) = L(1-s). \quad (2.6)$$

The functions  $\zeta(s)$  and  $L(s)$  are both analytic in the complex plane, except for a simple pole at  $s = 1$ . Riemann conjectured that the only zeros of  $L(s)$  lie on the vertical line  $\operatorname{Re}(s) = 1/2$ . This is the famous Riemann hypothesis, regarded by many as the most important unsolved problem in mathematics. Its interest stems from the fact that the zeros  $\{\rho = 1/2 + it\}$  of  $L(s)$  on this line are in some sense dual to prime numbers, or more accurately, to logarithms  $\{\gamma = \log p^n\}$  of prime powers. We can think of the former as a set of spectral data and the latter as a set of geometric data, which are related to each other by a Fourier transform. The Riemann hypothesis implies a very sharp asymptotic estimate for the number

$$\pi(x) = |\mathcal{P}(x)| = |\{p \leq x\}|$$

of primes less than or equal to  $x$ . One particularly explicit form [34] of the estimate is

$$|\pi(x) - li(x)| \leq \frac{1}{8\pi} \sqrt{x} \log x, \quad x \geq 2658, \tag{2.7}$$

for the principle value integral

$$li(x) = \int_0^x \frac{1}{\log t} dt.$$

Because the function  $li(x)$  is easy to approximate for large values of  $x$ , and is asymptotic to a function  $\frac{x}{\log x}$  that strongly dominates the error term, this is a striking estimate indeed.

For example, if  $x = 10^{100}$ , one sees that  $\pi(x)$  and  $li(x)$  are positive integers with 97 digits each, the first 47 of which match!

Dirichlet later generalized Riemann’s construction to series of the form

$$L^N(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s}, \tag{2.8}$$

where  $\chi(n)$  is what later became known as a Dirichlet character. We recall that  $\chi$  is a complex valued function on  $\mathbb{N}$  such that

$$\chi(nm) = \chi(n)\chi(m),$$

$$\chi(n + N) = \chi(n),$$

and

$$\chi(n) = 0, \quad \text{if } \gcd(n, N) > 1,$$

where  $N$  is a positive integer called the *modulus* of  $\chi$ . One says that  $\chi$  is *primitive* if its nonzero values are not given by restriction of a Dirichlet character with modulus a proper divisor of  $N$  (in which case  $N$  is called the proper divisor of  $\chi$ ). This means that  $x$  generates the cyclic group of characters on the multiplicative group  $(\mathbb{Z}/\mathbb{Z}N)^*$ . The series (2.8) behaves very much like the Riemann zeta function. It converges if  $\text{Re}(s) > 1$ . It has an Euler product

$$L^N(s, \chi) = \prod_p \left( \frac{1}{1 - \chi(p)p^{-s}} \right). \tag{2.9}$$

It also has analytic continuation to a meromorphic function on the complex plane. However, its functional equation is a little more interesting.

To state it, we can suppose without loss of generality that  $\chi$  is primitive. We account for the gamma function and powers of  $\pi$  in the analogue of (2.4) by setting

$$L_{\mathbb{R}}(s, \chi) = (\pi)^{-(s+a)/2} \Gamma((s + a)/2),$$

where

$$a = a(\chi) = \begin{cases} 0, & \text{if } \chi(-1) = 1 \\ 1, & \text{if } \chi(-1) = -1. \end{cases}$$

The product

$$L(s, \chi) = L_{\mathbb{R}}(s, \chi)L^N(s, \chi) = \Gamma_{\mathbb{R}}(s, \chi) \cdot \prod_p \left( \frac{1}{1 - \chi(p)p^{-s}} \right) \tag{2.10}$$

then satisfies the functional equation

$$L(s, \chi) = \varepsilon(s, \chi)L(1 - s, \bar{\chi}), \tag{2.11}$$

where  $\bar{\chi}$  is the complex conjugate of  $\chi$ , and

$$\varepsilon(s, \chi) = N^{-1/2-\varepsilon}\varepsilon(\chi), \tag{2.12}$$

for

$$\varepsilon(\chi) = \varepsilon(1/2, \chi) = i^{-a}N^{-1/2}\left(\sum_{n=1}^{N-1} \chi(n)e^{2\pi in/N}\right), \quad i = \sqrt{-1}.$$

The expression in the brackets is a Gauss sum. It is the analogue for a finite field (or ring) of the classical gamma function for the field  $\mathbb{R}$ . Its generalizations are an important part of the functional equations of nonabelian *L*-functions.

Dirichlet introduced his *L*-series to study the prime numbers in an arithmetic progression. Suppose that  $\chi$  is primitive and nontrivial. Dirichlet showed that the *L*-function (2.8) (or its normalization (2.10)) is actually an entire function of  $s \in \mathbb{C}$ , and in addition, that its value  $L(1, \chi)$  at 1 is nonzero. He then used this to show that for any integer  $a$  with  $\gcd(a, N) = 1$ , the number

$$\pi(a, N) = |\mathcal{P}(a, N)| = |\{p \equiv a \pmod{N}\}|$$

of primes  $p$  in the arithmetic sequence

$$a, a + N, a + 2N, \dots$$

is infinite. The generalized Riemann hypothesis is the assertion that the only zeros of the entire function  $L(s, \chi)$  again lie on the line  $\operatorname{Re}(s) = \frac{1}{2}$ . It implies an analogue

$$|\pi(x, a, N) - \phi(N)^{-1}li(x)| < C\sqrt{x} \log x, \quad x \geq 2,$$

of the asymptotic estimate (2.7), for the number

$$\pi(x, a, N) = |\{p \in \mathcal{P}(a, N) : p \leq x\}|$$

of primes in the arithmetic sequence less than or equal to  $x$ . The familiar Euler function

$$\phi(N) = |\{a : 1 \leq a \leq N, \gcd(a, N) = 1\}|$$

equals the number of such arithmetic progressions.

These remarks illustrate the power of *L*-functions. They are directed at some of the deepest analytic questions on the distribution of prime numbers. The *L*-functions we have described are just the beginning. They are the simplest among an enormous but unified collection of *L*-functions, which have the potential to resolve fundamental arithmetic questions about prime numbers, as well as refinements of the analytic questions we have looked at.

### 3. Artin $L$ -functions and class field theory

There seems not to be complete agreement on what kind of Dirichlet series (2.1) should be called an  $L$ -function. To qualify, it should certainly converge to an analytic function in some right half plane  $\operatorname{Re}(s) > k + 1$ . It should also have an Euler product

$$\prod_p (1 + c_{p,1}p^{-s} + c_{p,2}p^{-2s} + \cdots), \quad \operatorname{Re}(s) > k + 1, \quad (3.1)$$

for a family of coefficients  $C = \{c_{p,n}\}$ . Finally, it ought to have (or at least be expected to have) analytic continuation and functional equation. We shall take these conditions as our working definition.

The Euler product often arises naturally as an incomplete product, taken over the primes outside a finite set  $S$ . However, the expected functional equation is generally best stated for the completed product, in which one adds factors  $L_p(s, C)$  for the primes  $p \in S$ , as well as a factor  $L_{\mathbb{R}}(s, C)$  for the “archimedean prime”  $\mathbb{R}$  as above. If we also write  $L_p(s, C)$  for the factors with  $p \notin S$ , we then have the given incomplete  $L$ -function

$$L^S(s, C) = \prod_{p \notin S} L_p(s, C),$$

and its better behaved completion

$$L(s, C) = L_{\mathbb{R}}(s, C) \left( \prod_{p \in S} L_p(s, C) \right) L^S(s, C). \quad (3.2)$$

Treating Dirichlet  $L$ -functions as a guide, we would be looking for a functional equation

$$L(s, C) = \varepsilon(s, C) L(1 - s, C^\vee), \quad (3.3)$$

where  $C^\vee$  is some “dual” family of coefficients attached naturally to  $C$ , and

$$\varepsilon(s, C) = b_C^{(1/2-s)} \varepsilon\left(\frac{1}{2}, C\right),$$

for a positive real number  $b_C$ , and a complex number  $\varepsilon(\frac{1}{2}, C)$  that is independent of  $s$ .

The higher  $L$ -functions that will be our topic are of two kinds, automorphic and arithmetic. The former are primarily analytic objects, while the latter are algebraic. The Riemann zeta function is the common ancestor of them all. It is a mainstay of analytic number theory. However, it can also be regarded as the “trivial” case of the arithmetic  $L$ -functions we shall describe in this section. Dirichlet  $L$ -functions  $L(s, \chi)$  are automorphic. As in the case of the zeta function, it is the application of analysis (real, complex and harmonic) to  $L(s, \chi)$  that leads to its analytic continuation and functional equation, and to the location of any poles.

An important family of arithmetic  $L$ -functions was introduced by Emile Artin. These are attached to  $N$ -dimensional representations

$$r : \Gamma_{E/\mathbb{Q}} \longrightarrow GL(N, \mathbb{C})$$

of the Galois group  $\Gamma_{E/\mathbb{Q}} = \operatorname{Gal}(E/\mathbb{Q})$  of a finite Galois extension  $E$  of  $\mathbb{Q}$ . We shall describe them here from the perspective of the questions (I) and (II) raised in the preface.

For any  $E$ , we have the finite set  $S = S_E$  of prime numbers  $p$  that ramify in  $E$ , and for any unramified  $p \notin S_E$ , a canonical conjugacy class  $\text{Frob}_p$  (the *Frobenius* class) in  $\Gamma_{E/\mathbb{Q}}$ . This is one of the first constructions encountered in algebraic number theory. To be concrete, we can take  $E$  to be the splitting field of a monic, integral polynomial  $f(x)$  of degree  $n$ , as in the preface. For any prime number  $p$ , we then have the corresponding factorization of  $f(x)$  into irreducible factors  $f_i(x)$  modulo  $p$ , with degrees  $n_i$ . This embeds  $S_E$  in the finite set  $S = S_f$  of primes  $p$  for which these factors are distinct. The choice of  $f(x)$  also identifies  $\Gamma_{E/\mathbb{Q}}$  with a conjugacy class of subgroups of the symmetric group  $S_n$ . For any unramified prime  $p \notin S_E$ ,  $\text{Frob}_p$  is then mapped to the conjugacy class in  $S_n$  defined by the partition  $\Pi_p = \{n_1, \dots, n_r\}$  of  $n$ . In particular, the set

$$\text{Spl}(E/\mathbb{Q}) = \{p \notin S : \text{Frob}_p = 1\}$$

of prime numbers that *split completely* in  $E$ , is the set of  $p$  such that  $f(x)$  breaks into linear factors modulo  $p$ . It is known [38, p. 165] that  $\text{Spl}(E/\mathbb{Q})$  characterizes  $E$ . In other words, the mapping

$$E \longrightarrow \text{Spl}(E/\mathbb{Q}),$$

from Galois extensions of  $\mathbb{Q}$  to subsets of prime numbers, is *injective*. A variant of the question (I) would be to characterize its image. This would amount to a classification of Galois extensions  $E$  of  $\mathbb{Q}$ .

For any  $p \notin S$ , the image  $r(\text{Frob}_p)$  under the representation  $r$  of  $\Gamma_{E/\mathbb{Q}}$  gives a semisimple conjugacy class in the complex general linear group  $GL(N, \mathbb{C})$ . Artin defined its local *L*-factor

$$L_p(s, r) = \det(1 - r(\text{Frob}_p)p^{-s})^{-1}, \quad p \notin S, \tag{3.4}$$

in terms of the associated characteristic polynomial. It clearly has an expansion in terms of powers of  $p^{-s}$ , and therefore has the general form of the factor of  $p$  in (3.1). Artin then defined an incomplete *L*-function as the Euler product

$$L^S(s, r) = \prod_{p \notin S} L_p(s, r), \tag{3.5}$$

which he conjectured had analytic continuation with functional equation of the general form (3.3). The question (II) will be reflected in the analytic properties of the resulting function of  $s$ . However, the analytic continuation and functional equation of (3.5), let alone the relevant analogue of the Riemann hypothesis, is a more serious proposition. Since  $L^S(s, r)$  is a fundamentally arithmetic object, it cannot be studied in terms of the kind of analysis that Dirichlet applied to the *L*-functions  $L(s, \chi)$ . Artin treated it indirectly.

Suppose that the Galois group  $\text{Gal}(E/\mathbb{Q})$  is abelian, and that  $r$  is irreducible, and therefore one-dimensional. The classes  $r(\text{Frob}_p)$  are then just nonzero complex numbers. The (incomplete) Artin *L*-function becomes a product

$$L^S(s, r) = \prod_{p \notin S} \frac{1}{1 - r(\text{Frob}_p)p^{-s}}$$

that resembles the Euler product (2.9) of a Dirichlet *L*-function. Indeed, if  $p$  divides the modulus of  $\chi$  (written  $p|N$ ),  $\chi(p)$  vanishes, and the corresponding product (2.9) can then be written

$$L^S(s, \chi) = \prod_{p \notin S} \frac{1}{1 - \chi(p)p^{-s}},$$

where

$$S = \{p : p|N\}.$$

This formal similarity between the products  $L^S(s, r)$  and  $L^S(s, \chi)$  turns out in fact to be an identity. More precisely, for any one-dimensional Galois representation  $r$ , there is a Dirichlet character  $\chi$  such that the function  $L^S(s, r)$  equals  $L^S(s, \chi)$ . The new  $L$ -function therefore has the analytic behaviour of the Dirichlet  $L$ -function  $L^S(s, \chi)$ . In particular, it has analytic continuation, and its completed  $L$ -function (3.2) (with  $r$  in place of  $C$ ) satisfies the functional equation (3.3) (with  $r^\vee = \bar{r}$  in place of  $C^\vee$ ).

The assertion that for every  $r$  there is a  $\chi$  is a rather deep classical result, known as the Kronecker-Weber theorem. It finesses the question of the behaviour of abelian Artin  $L$ -functions by imposing limits on the set of abelian extensions of  $\mathbb{Q}$ . From an elementary calculation in algebraic number theory, one obtains the converse theorem that for every  $\chi$  there is an  $r$ . More precisely, if  $\chi$  has modulus  $N$ , there is a one-dimensional representation  $r$  of the abelian Galois group of the cyclotomic Galois extension  $\mathbb{Q}(e^{2\pi i/N})$  of  $\mathbb{Q}$  such that  $\chi(p)$  equals  $r(\text{Frob}_p)$ , for any  $p$  that does not divide  $N$ . The Kronecker-Weber theorem can then be interpreted as the assertion that any finite abelian extension of  $\mathbb{Q}$  is contained in the cyclotomic extension of  $N$ th roots of 1, for some  $N$ .

The Kronecker-Weber theorem predated Artin by many years. However, Artin was working over an arbitrary number field  $F$ , a finite field extension of  $\mathbb{Q}$ , rather than  $\mathbb{Q}$  itself. The definitions we have made so far are easily extended from  $\mathbb{Q}$  to  $F$ . The ring  $\mathfrak{o} = \mathfrak{o}_F$  of algebraic integers of  $F$  does not generally have unique factorization, so one must replace prime numbers  $p$  in  $\mathbb{Z}$  with prime ideals  $\mathfrak{p}$  in  $\mathfrak{o}$ , and integers  $n$  in  $\mathbb{Z}$  with general (integral) ideals  $\mathfrak{a}$  in  $\mathfrak{o}$ . Any ideal then has a norm

$$N\mathfrak{a} = |\mathfrak{o}/\mathfrak{a}| = (N\mathfrak{p}_1)^{a_1} \cdots (N\mathfrak{p}_r)^{a_r} = |\mathfrak{o}/\mathfrak{p}_1|^{a_1} \cdots |\mathfrak{o}/\mathfrak{p}_r|^{a_r}, \tag{3.6}$$

where

$$\mathfrak{a} = \mathfrak{p}_1^{a_1} \cdots \mathfrak{p}_r^{a_r}$$

is its unique factorization into prime ideals. The definition of a Dirichlet series (2.1) for  $\mathbb{Q}$  can then be extended to  $F$  by replacing the sum over  $n$  by a sum over  $\mathfrak{a}$ , and the corresponding variable  $n^{-s}$  by  $(N\mathfrak{a})^{-s}$ . The same goes for an Euler product (3.1). A Dirichlet series for  $F$  does reduce to a Dirichlet series for  $\mathbb{Q}$ , since the norm of a prime ideal  $\mathfrak{p}$  is a power of a prime number  $p$ . However, a Dirichlet or Artin  $L$ -function over  $F$  represents something different, even though it can be regarded as a Dirichlet series (2.1) with Euler product (3.1).

Artin worked from the beginning over  $F$ . He defined the  $L$ -function  $L^S(s, r)$  for any  $N$ -dimensional representation

$$r : \Gamma_{E/F} \longrightarrow GL(N, \mathbb{C}) \tag{3.7}$$

of the Galois group of a finite Galois extension  $E/F$ . Algebraic number theory again tells us that for any  $\mathfrak{p}$  outside the finite set  $S = S_E$  of prime ideals for  $F$  that ramify in  $E$ , there is a canonical conjugacy class  $\text{Frob}_\mathfrak{p} = \text{Frob}_{E/F, \mathfrak{p}}$  in  $\Gamma_{E/F}$ . The definition (3.4) therefore does extend to  $F$  if we replace  $p^{-s}$  by  $N\mathfrak{p}^{-s}$ . Artin also introduced factors for the ramified primes  $\mathfrak{p} \in S$ , and for the archimedean “primes”  $v$  for  $F$  (now a finite set  $S_\infty$  rather than just the one completion  $\mathbb{R}$  of  $\mathbb{Q}$ ). He conjectured that the resulting product  $L(s, r)$  had functional equation (3.3), with an  $\varepsilon$ -factor  $\varepsilon(s, r)$  he formulated in terms of  $r$ .

It was in this context that Artin obtained the analytic continuation and functional equation for the abelian  $L$ -functions  $L(s, r)$ . Dirichlet characters  $\chi$  can still be defined by a variant

of the prescription for  $\mathbb{Q}$  above, and the same analysis that works for  $\mathbb{Q}$  gives the analytic continuation and functional equation of a general Dirichlet *L*-function  $L(s, \chi)$ . With this in mind, Artin established an *F*-analogue of the Kronecker-Weber theorem, known now as the Artin reciprocity law. It again asserts that for any one-dimensional Galois representation  $r$  over  $F$ , there is a Dirichlet character  $\chi$  over  $F$  such that  $\chi(\mathfrak{p})$  equals  $r(\text{Frob}_{\mathfrak{p}})$ , for every unramified prime ideal  $\mathfrak{p}$  for  $F$ . This leads to the identity  $L(s, r) = L(s, \chi)$  of *L*-functions, and therefore the desired analytic properties of the arithmetic *L*-function  $L(s, r)$ .

The Artin reciprocity law is one of the central assertions of class field theory. Unlike the general constructions above, it becomes much deeper in the passage from  $\mathbb{Q}$  to  $F$ , even though the assertion remains similar. As in the case  $F = \mathbb{Q}$ , the general law asserts that the abelian field extensions over  $F$  are limited to those attached to Dirichlet characters  $\chi$ . These may then be classified by the “reciprocity law”

$$r(\text{Frob}_{\mathfrak{p}}) = \chi(\mathfrak{p}), \quad \mathfrak{p} \notin S, \tag{3.8}$$

according again to the remark on [38, p. 165].

For completeness, we note that Artin *L*-functions are but the simplest in the general family of arithmetic *L*-functions, called motivic *L*-functions. A  $\mathbb{Q}$ -motive  $M$  over  $F$  (which we will not try to define!) also comes with a finite dimensional representation  $r_{M,\ell}$  of the absolute Galois group  $\Gamma_F = \Gamma_{\overline{F}/F}$ . In this general case, however, it takes values in an  $\ell$ -adic general linear group  $GL(N, \mathbb{Q}_{\ell})$ , for a variable prime number  $\ell \notin S$ . It therefore gives rise to a (continuous) homomorphism

$$r_M = \bigotimes_{\ell} r_{M,\ell} : \Gamma_F \longrightarrow \prod_{\ell} \widetilde{GL}(N, \mathbb{Q}_{\ell}) \tag{3.9}$$

from  $\Gamma_F$  to a large, totally disconnected group. It therefore represents a much larger quotient of  $\Gamma_F$  than does a complex valued representation (3.7). The motive should also come with a finite set  $S$  of prime ideals in  $F$  such that  $r_{M,\ell}$  is unramified in any prime  $\mathfrak{p} \notin S \cup S_{\ell}$ , where  $S_{\ell}$  is the set of primes in  $F$  that divide the prime  $\ell$  of  $\mathbb{Q}$ . The family  $\{r_{M,\ell}\}$  of  $\ell$ -adic representations is conjectured to be *compatible*, in the sense that the image  $r_{M,\ell}(\text{Frob}_{\mathfrak{p}})$  of the associated Frobenius class in  $GL(N, \mathbb{Q}_{\ell})$  is the image of a semisimple conjugacy class  $r_M(\text{Frob}_{\mathfrak{p}})$  in  $GL(N, \mathbb{Q})$  that is independent of  $\ell$ . The incomplete *L*-function of  $M$  is then given by an Euler product

$$L^S(s, M) = \prod_{\mathfrak{p} \notin S} \det(1 - r_M(\text{Frob}_{\mathfrak{p}})(N\mathfrak{p})^{-s})^{-1}, \tag{3.10}$$

which converges in some right half plane, and which is again expected to have analytic continuation with functional equation. Notice that any complex representation (3.7) of  $\Gamma_{E/F}$  that is defined over  $\mathbb{Q}$  gives a compatible family (3.9) of  $\ell$ -adic representations. It represents a  $\mathbb{Q}$ -motive over  $F$  of dimension 0.

#### 4. Automorphic *L*-functions

Dirichlet *L*-functions are the analytic counterparts of abelian Artin *L*-functions. Class field theory, the culmination of many decades of effort by number theorists past, represents a classification of the finite abelian field extensions of any number field. It tells us that any abelian

Artin  $L$ -function is a Dirichlet  $L$ -function. An  $L$ -function of the former sort therefore inherits all of the rich properties that can be made available for the latter through analysis. What are the analytic counterparts of nonabelian Artin  $L$ -functions? They are the automorphic  $L$ -functions introduced by Robert Langlands [23] in 1970.

Automorphic representations are the nonabelian generalizations of Dirichlet characters, and their abelian generalizations introduced later by Erich Hecke. They are defined for any connected, reductive algebraic group  $G$  over the number field  $F$ . Algebraic groups represent a conceptual hurdle for many, but a reader is invited to take  $G$  to be the general linear group  $GL(N)$  of invertible  $(N \times N)$ -matrices. For any ring  $A$  (abelian, with 1),  $G(A)$  is then equal to the group  $GL(N, A)$  of  $(N \times N)$ -matrices with entries in  $A$  and determinant equal to a unit in  $A$ . We want to take  $A$  to be the ring  $A = \mathbb{A}_F$  of adeles of  $F$ . This is a locally compact topological ring, in which  $F$  embeds as a discrete subring. It is often a second hurdle, but avoiding it would make matters considerably more complicated. The idea is really quite simple and natural.

By definition, the group of adelic points of  $G$  is a restricted direct product

$$G(\mathbb{A}) = \prod_v^{\sim} G(F_v), \quad (4.1)$$

taken over the valuations  $v$  on  $F$ . For any  $v$ ,  $F_v$  is the locally compact field obtained by completing  $F$  with respect to  $v$ . It is modeled on the standard case of the completion  $F_v = \mathbb{R}$  of  $F = \mathbb{Q}$  with respect to the usual absolute value  $|\cdot|_v = |\cdot|$ . We recall that the complementary valuations for  $F = \mathbb{Q}$  are the nonnegative functions

$$|u|_p = \begin{cases} p^{-r}, & \text{if } u = (a/b)p^r, \text{ for } a, b, r \in \mathbb{Z}, (a, p) = (b, p) = 1, \\ 0, & \text{if } u = 0, \end{cases}$$

on  $\mathbb{Q}$ , parametrized by prime numbers  $p$ . In general, the restricted direct product is the subgroup of elements

$$x = \prod_v x_v, \quad x_v \in G(F_v),$$

in the direct product such that for almost all valuations  $v$ ,  $x_v$  lies in the maximal compact subgroup  $G(\mathfrak{o}_v)$  of points in  $G(F_v)$  with values in the compact subring

$$\mathfrak{o}_v = \{u_v \in F_v : |u_v|_v \leq 1\}$$

of integers in  $F_v$ . It becomes a locally compact group under the appropriate direct limit topology. The group  $G(F)$  embeds in  $G(F_v)$  (as a dense subgroup). The diagonal embedding of  $G(F)$  into  $G(\mathbb{A})$  then exists (because an element in  $G(F)$  is integral at almost all valuations  $v$ ), and is easily seen to have discrete image.

Since  $G(F)$  is discrete in  $G(\mathbb{A})$ , the quotient  $G(F) \backslash G(\mathbb{A})$  is a reasonable object. It comes with a right invariant measure, which is determined up to a positive multiplicative constant. One can therefore form the associated space  $L^2(G(F) \backslash G(\mathbb{A}))$  of square-integrable functions. It is a Hilbert space, equipped with the unitary representation

$$(R(y)\phi)(x) = \phi(xy), \quad x, y \in G(\mathbb{A}), \phi \in L^2(G(F) \backslash G(\mathbb{A})),$$

of  $G(\mathbb{A})$  by right translation. One could describe an *automorphic representation* of  $G$  to be an irreducible representation of  $G(\mathbb{A})$  that occurs in the spectral decomposition of  $R$ .

This description is actually more of an informal characterization than a definition. It is also more restrictive than the formal definition in [8] and [26]. We shall take the broader definition, without recalling its two equivalent formulations established in [26]. We will then call automorphic representations that satisfy the narrower spectral condition above *globally tempered*.

Suppose for example that  $G$  is the abelian algebraic group  $GL(1)$ . Then  $G(\mathbb{A})$  is the multiplicative group  $\mathbb{A}^*$  of elements  $x$  in  $\mathbb{A}$  whose components  $x_v \in F_v$  are all nonzero and of valuation 1 for almost all  $v$ . This is the group of ideles, introduced earlier by Chevalley. A (globally tempered) automorphic representation of  $G = GL(1)$  is a character  $\chi$  on the idele class group  $F^* \backslash \mathbb{A}^*$ , or in other words, a continuous  $F^*$ -invariant homomorphism from  $\mathbb{A}^*$  to the group  $U(1)$  of complex numbers of absolute value 1. It is the generalization of a Dirichlet character introduced by Hecke, which he called a Grossencharakter, and which is now generally referred to as a Hecke character. Hecke worked in the classical context of ideals  $\mathfrak{a}$ , but his constructions are a little easier to formulate now in the language of ideles. It is an early illustration of the convenience of the language of adeles. In this regard, a Dirichlet character is just a Hecke character of finite order.

One of the remarkable discoveries of Langlands has been the fundamental role played by a certain dual group of  $G$ . The dual group is a complex connected reductive group  $\widehat{G}$ , whose Coxeter-Dynkin diagram is the dual of the diagram of  $G$ . It comes with an action of the absolute Galois group  $\Gamma_F = \Gamma_{\overline{F}/F}$ , a compact totally disconnected group, which factors through the finite quotient  $\Gamma_{E/F}$  of  $\Gamma_F$  attached to some finite Galois extension  $E$  of  $F$ . Langlands built this action into the dual group as the semidirect product

$${}^L G_E = \widehat{G} \rtimes \Gamma_{E/F},$$

or more canonically

$${}^L G = \widehat{G} \rtimes \Gamma_F,$$

that is now known as the  $L$ -group. If  $G$  equals  $GL(N)$  for example,  $\widehat{G}$  is just the complex general linear group  $GL(N, \mathbb{C})$ . In this case, the action of  $\Gamma_F$  on  $\widehat{G}$  is trivial, so we can take  $E = F$ . For the case that  $G$  is orthogonal or symplectic, the families that will be our ultimate interest, we refer the reader to the beginning of §5.

Automorphic  $L$ -functions  $L(s, \pi, r)$  were defined by Langlands for any  $G$ . They depend on an automorphic representation  $\pi$  of  $G$  and a finite dimensional representation

$$r : {}^L G \longrightarrow GL(N, \mathbb{C}) \tag{4.2}$$

of  ${}^L G$ . It is understood that  $r$  is continuous on  $\Gamma_F$  and analytic on  $\widehat{G}$ , and in particular, that it factors through a finite quotient  $\Gamma_{E/F}$  of  $\Gamma_F$ . Its image is therefore a complex group with finitely many connected components. We will review the definition in the rest of this section.

We first recall [13] that any  $\pi$  can be written as a restricted tensor product

$$\pi = \bigotimes_v \widetilde{\pi}_v, \quad \pi_v \in \Pi(G_v), \tag{4.3}$$

where  $\Pi(G_v)$  is the set of irreducible representations of the locally compact completion  $G(F_v)$  of  $G(F)$ . The interest is not so much in the individual constituents  $\pi_v$  of  $\pi$ , as in the relations they need to satisfy among themselves in order that the product be automorphic. Much of the data that characterize these representations is quite explicit. For example, it is

a consequence of what it means for (4.3) to be a restricted direct product that the irreducible representations  $\pi_v$  of  $G(F_v)$  will be unramified<sup>1</sup> for almost all  $v$ . A well known integral transform, introduced into  $p$ -adic harmonic analysis by Satake, leads to a canonical mapping<sup>2</sup>

$$\pi_v \longrightarrow c(\pi_v)$$

from the set of unramified representations  $\pi_v$  of  $G(F_v)$  to the set of semisimple conjugacy classes in  ${}^L G$ . The automorphic representation  $\pi$  thus gives rise to a family

$$c^S(\pi) = \{c_v(\pi) = c(\pi_v) : v \notin S\}$$

of semisimple classes in  ${}^L G$ .

If we are given  $r$  as well as  $\pi$ , we obtain a family

$$\{r(c_v(\pi)) : v \notin S\}$$

of semisimple conjugacy classes in  $GL(N, \mathbb{C})$ . The incomplete automorphic  $L$ -function of  $\pi$  and  $r$  is then defined in terms of the characteristic polynomials of these classes. It equals the product

$$L^S(s, \pi, r) = \prod_{v \notin S} L(s, \pi_v, r_v), \tag{4.4}$$

where

$$L(s, \pi_v, r_v) = \det(1 - r(c_v(\pi))q_v^{-s})^{-1}, \tag{4.5}$$

and

$$q_v = p_v^{f_v} = |\mathfrak{o}_v/\mathfrak{p}_v|$$

is the order of the residue class field of  $F_v$ . The product is easily seen to converge for  $s$  in some right half plane, and is clearly a Dirichlet series (2.1) with Euler product (3.1). The definition can be compared with that of the incomplete Artin  $L$ -function (3.4) and (3.5), or rather its generalization from  $\mathbb{Q}$  to  $F$ . The analogy is clear, even though the earlier definition was in terms of ideals rather than the formulation here in terms of valuations.

Langlands introduced automorphic  $L$ -functions in [23]. He conjectured that they have analytic continuation, with a very precise functional equation

$$L(s, \pi, r) = \varepsilon(s, \pi, r) L(1 - s, \pi, r^\vee), \tag{4.6}$$

where

$$L(s, \pi, r) = L_S(s, \pi, r) L^S(s, \pi, r) = \prod_v L(s, \pi_v, r_v) \tag{4.7}$$

is a completed  $L$ -function obtained by appending a finite product

$$L_S(s, \pi, r) = \prod_{v \in S} L(s, \pi_v, r_v)$$

<sup>1</sup>This means that  $F_v$  is nonarchimedean, that  $G_v = G \times_F F_v$  is quasisplit and split over an unramified extension of  $F_v$ , and that the restriction of  $\pi_v$  to a hyperspecial maximal compact subgroup  $K_v$  of  $G(F_v)$  contains the trivial 1-dimensional representation.

<sup>2</sup>The mapping becomes a bijection if one takes the restricted form  ${}^L G_E$  of the  $L$ -group, and then takes its range to be the set of semisimple conjugacy classes in  ${}^L G_E$  whose image in  $\Gamma_{E/F}$  equals the Frobenius class  $\text{Frob}_v = \text{Frob}_{E/F, v}$ , if  $v$  is unramified in  $E$ . This basic condition on the Satake transform was observed by Langlands in [23].

of suitable factors at the ramified valuations  $v \in S$  (including the archimedean valuations  $v \in S_\infty$  of  $F$ ), and

$$\varepsilon(s, \pi, r) = \prod_{v \in S} \varepsilon(s, \pi_v, r_v, \psi_v) \tag{4.8}$$

is a finite product of local monomials in  $q_v^{-s}$ . The local  $\varepsilon$ -factors on the right would depend on the local components  $\psi_v$  of a fixed, nontrivial additive character  $\psi$  on the group  $F \backslash \mathbb{A}$ , while the global product on the left hand side of (4.8) would be independent of  $\psi$ . Langlands did not define the ramified local  $L$ - and  $\varepsilon$ -factors in [23]. Nevertheless, his introduction of the general automorphic  $L$ -function in [23], with its proposed functional equation (4.6), was an enormous step beyond the abelian automorphic  $L$ -functions of Hecke.<sup>3</sup> It depends above all on the  $L$ -group  ${}^L G$  he introduced at the same time.

We have now described two fundamental families of  $L$ -functions. They are the arithmetic  $L$ -functions of the last section and the automorphic  $L$ -functions of this section. Langlands later conjectured that the former family is a subset of the latter. In other words, for any motive  $M$  (which for present purposes we can take to be an  $N$ -dimensional representation of a finite Galois group  $\Gamma_{E/F}$ ), there should be a pair  $(\pi, r)$  such that the completed  $L$ -functions satisfy

$$L(s, M) = L(s, \pi, r). \tag{4.9}$$

In particular, the analytic continuation and functional equation for  $L(s, M)$  would follow from the same properties for  $L(s, \pi, r)$ . This would be a striking and far reaching generalization of the method used by Artin to establish the analytic continuation and functional equation of abelian Artin  $L$ -functions.

There is actually a theory that applies directly to nonabelian Artin  $L$ -functions. Richard Brauer established a general property of the representations of a finite group (the Brauer induction theorem), which he used to express any nonabelian Artin  $L$ -function  $L(s, r)$  as a quotient of finite products of abelian Artin  $L$ -functions  $L(s, r_1)$  (over finite extensions  $F_1$  of  $F$ ). Combined with the results of Artin described in §2, this shows that  $L(s, r)$  has analytic continuation, with a functional equation of the desired sort. However, it does not give much control over the analytic behaviour of  $L(s, r)$ . In particular, it gives no information on a fundamental conjecture of Artin, which asserts that an irreducible, nonabelian Artin  $L$ -function is entire.

Brauer's theorem has, however, led to important results on local arithmetic  $L$ - and  $\varepsilon$ -factors. These apply more generally to the variant of the Galois group that Weil introduced as a consequence of class field theory. Like the absolute Galois group  $\Gamma_F$ , the Weil group  $W_F$  is defined if  $F$  is a local or a global field. It is a locally compact group, equipped with a continuous homomorphism  $W_F \rightarrow \Gamma_F$  with dense image and connected kernel, whose maximal abelian quotient is given by

$$W_F^{ab} = W_F / W_F^c \cong \begin{cases} F^*, & \text{if } F \text{ is local,} \\ F^* \backslash \mathbb{A}^*, & \text{if } F \text{ is global.} \end{cases} \tag{4.10}$$

If  $F$  is the global field we are discussing here,  $W_F$  comes with a conjugacy class of embed-

---

<sup>3</sup>Hecke also introduced some nonabelian  $L$ -functions for the group  $G = GL(2)$  and the standard two dimensional representation  $r$ .

dings

$$\begin{array}{ccc}
 W_{F_v} & \longrightarrow & \Gamma_{F_v} \\
 \downarrow & & \downarrow \\
 W_F & \longrightarrow & \Gamma_F
 \end{array}$$

for any  $v$ , that are compatible with the abelianization (4.10). (See [39, §1].) These properties imply that the Artin reciprocity law described in §2 extends to a canonical isomorphism from the group of 1-dimensional representations of  $W_F$  to the group of (1-dimensional) automorphic representations of  $G = GL(1)$ . Moreover, the Brauer induction theorem extends to  $N$ -dimensional representations  $r$  of  $W_F$ . It leads to a global  $L$ -function

$$L(s, r) = \prod_v L(s, r_v) = \prod_{v \in S} L(s, r_v) \prod_{v \notin S} \det(1 - r(\text{Frob}_v)q_v^{-s})^{-1} \tag{4.11}$$

that has analytic continuation and functional equation

$$L(s, r) = \varepsilon(s, r) L(s, r^\vee), \tag{4.12}$$

for a global  $\varepsilon$ -factor

$$\varepsilon(s, r) = \prod_{v \in S} \varepsilon(s, r_v, \psi_v). \tag{4.13}$$

As in the special case of Artin  $L$ -functions, we obtain little control over the analytic behaviour of the global  $L$ -functions  $L(s, r)$  in the full complex domain. The global interest in these results is therefore limited. However, Deligne used them to establish important local results [39, §2]. He showed that the local  $L$ -functions  $L(s, r_v)$  in (4.11) and  $\varepsilon$ -factors  $\varepsilon(s, r_v, \psi_v)$  in (4.13) have a canonical local definition. In particular, they can be constructed independently of the global representation  $r$ .

The global  $L$ -functions  $L(s, r)$  attached to representations of  $W_F$  are not all motivic, unlike Artin  $L$ -functions. We cannot therefore really regard them as arithmetic. On the other hand, they are not automorphic, since they are not generally defined in terms of automorphic representations. Perhaps they should be regarded as objects that lie between the two classes. In any case,  $L(s, r)$  should still be equal to an automorphic  $L$ -function. This is again among the original conjectures of Langlands in [23]. Deligne’s constructions then become important for the local classification of representations, and in particular, for comparison with the local  $L$ - and  $\varepsilon$ -factors in (4.7) and (4.8).

### 5. The principle of functoriality

Langlands’ conjectural functional equation (4.6) for a general automorphic  $L$ -function is very deep, and far from known. However, among the cases that are known, there is one that deserves special mention. It is the standard automorphic  $L$ -function, in which  $G$  equals  $GL(N)$ , and  $r$  equals the standard  $N$ -dimensional representation  $St$  of  ${}^L G_F = GL(N, \mathbb{C})$ .

Abelian Hecke  $L$ -functions are given by the further special case that  $G = GL(1)$ . Hecke established their analytic continuation and functional equation, using the classical language of ideals. Tate later simplified Hecke’s proof by introducing the ring of adèles  $\mathbb{A}$ . In his

famous thesis [37], he established the results through the interplay of multiplicative harmonic analysis on the idele class group

$$F^* \backslash \mathbb{A}^* = GL(1, F) \backslash GL(1, \mathbb{A})$$

with additive harmonic analysis on the group  $\mathbb{A}$ .

Following Langlands' paper [23], Godement and Jacquet [15] extended the method of Tate to  $GL(N)$ , with the additive group of adelic matrices  $M_{N \times N}(\mathbb{A})$  in place of  $\mathbb{A}$ . It follows from their results and later refinements [16] that the standard (completed) *L*-function

$$L(s, \pi) = L(s, \pi, St) \tag{5.1}$$

for any automorphic representation  $\pi$  of  $GL(N)$  is well defined, and has analytic continuation with functional equation

$$L(s, \pi) = \varepsilon(s, \pi) L(1 - s, \pi^\vee).$$

This method also gives further information about the analytic behaviour of standard *L*-functions. For example, if the automorphic representation  $\pi$  is cuspidal,  $L(s, \pi)$  is an entire function of  $s$  unless  $N = 1$  and  $\pi(x) = |x|^\lambda$  for some  $\lambda \in \mathbb{C}$ , in which case  $L(s, \pi)$  is analytic apart from a simple pole at  $s = 1 - \lambda$ .

For any  $G$ , we write  $\Pi_{\text{aut}}(G)$  for the set of automorphic representations of  $G$  (in the broad sense of [26] we have agreed upon). We then write

$$\mathcal{C}_{\text{aut}}(G) = \{c(\pi) : \pi \in \Pi_{\text{aut}}(G)\} \tag{5.2}$$

for the set of families  $c^S(\pi)$  of automorphic conjugacy classes, taken up to the equivalence relation defined by  $c^S \sim c_1^S$  if  $c_v = c_{1,v}$  for almost all  $v$ . We emphasize that these are concrete objects. They represent the fundamental data encompassed in the seemingly abstract notion of an automorphic representation. As we have noted, the arithmetic significance of these data is not so much in the value of any one class  $c_v(\pi)$  as in the relationships among the classes as  $v$  varies.

In his original paper [23], Langlands made a profound conjecture that later became known as the principle of functoriality. We shall state it in the restricted form that applies to the concrete families  $\mathcal{C}_{\text{aut}}(G)$ .

**Principle of Functoriality** (Langlands). *Suppose that  $G$  and  $G'$  are quasisplit<sup>4</sup> groups over the number field  $F$ . Suppose also that*

$$\rho : {}^L G' \longrightarrow {}^L G$$

*is an  $L$ -homomorphism (that is, a continuous, analytic homomorphism that commutes with the two projections onto  $\Gamma_F$ ) between their  $L$ -groups. Then if  $c' = \{c'_v\}$  lies in  $\mathcal{C}_{\text{aut}}(G')$ , the family*

$$c = \rho(c') = \{\rho(c'_v)\}$$

*lies in  $\mathcal{C}_{\text{aut}}(G)$ . In other words, if  $c' = c(\pi')$  for some  $\pi' \in \Pi_{\text{aut}}(G')$ , then  $c = c(\pi)$  for some  $\pi \in \Pi_{\text{aut}}(G)$ .*

---

<sup>4</sup>See the brief description of this property at the beginning of the next section.

The principle of functoriality is the central problem in the theory of automorphic forms. It asserts that the internal relations in an automorphic family  $c' = c(\pi')$  for  $G'$ , whatever they might be, are reflected in the internal relations in some automorphic family  $c = c(\pi)$  for  $G$ . The principle of functoriality has been established in a significant number of cases. But as challenging as these have been, they pale in comparison with the cases that have not been established.

In the same paper [23], Langlands pointed out some fundamental applications of functoriality. The first concerns the automorphic  $L$ -functions he had just introduced.

Suppose that  $G', G, \rho, c' = c(\pi')$  and  $c = c(\pi)$  are as in the assertion of functoriality. If  $r$  is an  $N$ -dimensional representation of  ${}^L G$ , the composition  $r \circ \rho$  is an  $N$ -dimensional representation of  ${}^L G'$ . We then obtain an identity

$$L^S(s, \pi', r \circ \rho) = L^S(s, \pi, r) \tag{5.3}$$

of incomplete automorphic  $L$ -functions from the definitions, and of course, the principle of functoriality. This relation might seem almost routine at first glance, certainly not the sweeping observation it actually is. But consider the special case with  $G, GL(N), c = c(\pi)$  and  $c_N = c(\pi_N)$  in place of  $G', G, c'$  and  $c$  respectively. Then  $\rho$  can be identified with an  $N$ -dimensional representation  $r$  of  ${}^L G$ , and (5.3) specializes to an identity

$$L^S(s, \pi, r) = L^S(s, \pi_N, St) = L^S(s, \pi_N).$$

The general incomplete automorphic  $L$ -function on left thus equals a standard incomplete  $L$ -function, the function for which we already have analytic continuation and functional equation. If we set<sup>5</sup>

$$L_S(s, \pi, r) = L_S(s, \pi_N)$$

and

$$\varepsilon(s, \pi, r) = \varepsilon(s, \pi_N),$$

for the supplementary terms, the completed  $L$ -function satisfies

$$L(s, \pi, r) = L_S(s, \pi, r)L^S(s, \pi, r) = L_S(s, \pi_N)L^S(s, \pi_N) = L(s, \pi_N), \tag{5.4}$$

and the general functional equation (4.6) then follows from its analogue for standard  $L$ -functions.

A second immediate application of functoriality pointed out by Langlands in [23] is to nonabelian class field theory. It concerns the seemingly trivial case of functoriality with  $G' = \{1\}$ . Despite its apparent simplicity, however, this case comes with answers to the two general questions (I) and (II) from the preface.

If  $G'$  equals  $\{1\}$ , the dual group  $\widehat{G}'$  also equals  $\{1\}$ , but the  $L$ -group  ${}^L G'$  is still the absolute Galois group  $\Gamma_F$ . We again take the second group  $G$  to be  $GL(N)$ . An  $L$ -homomorphism from  ${}^L G'$  to  ${}^L G$  will be continuous (by definition) on its totally disconnected domain  $\Gamma_F$ . It can therefore be identified with an  $N$ -dimensional representation

$$r : \Gamma_{E/F} \longrightarrow GL(N, \mathbb{C}) \tag{5.5}$$

---

<sup>5</sup>In order that the two left hand sides here depend only on  $\pi$ , we assume implicitly that  $\pi_N$  is *isobaric*, in the sense of [27, §2]. It is then the *unique* automorphic representation of  $GL(N)$  with the given eigenfamily  $c(\pi_N) = c(\pi)$ , according to [18]. Notice that we do not obtain a local construction for the factors in these supplementary terms, unlike in their analogues (4.11), (4.13) for representations of Weil groups. This requires a stronger (and more complex) assertion of functoriality as in [23], and is predicated on a local classification of representations, such as that obtained for special orthogonal and symplectic groups in [4].

of the Galois group of some finite Galois extension  $E$  of  $F$ . The only automorphic representation of  $G'$  is the trivial representation 1. However, the associated automorphic family  $c(1)$  is still interesting. It is represented by the set

$$c^S(1) = \{c_v(1) = \text{Frob}_v : v \notin S\}$$

of Frobenius conjugacy classes in  $\Gamma_{E/F}$  of primes  $v$  of  $F$  that are unramified in  $E$ , according to footnote 2 from the last section. The principle of functoriality in this case asserts that its  $r$ -image  $r(c(1))$  is automorphic for  $GL(N)$ . In other words, there is an automorphic representation  $\pi$  of  $GL(N)$  such that

$$c_v(\pi) = r(c_v(1)) = r(\text{Frob}_v), \quad v \notin S.$$

This can be regarded as a general answer to the question (I). Since it includes an analytic characterization of the set  $\text{Spl}(E/F)$  of primes of  $F$  that split completely in  $E$ , it also amounts to a classification theory for general Galois extensions of  $F$ , the long standing dream of earlier number theorists. The arithmetic data  $\{\text{Frob}_v\}$  that characterize finite Galois extensions  $E$  of  $F$  ([38, p. 165]), and that are conveniently packaged by the characters of continuous, finite dimensional representations of  $\Gamma_F$ , can be represented by the analytic data  $\{c_v(\pi)\}$  of automorphic representations of general linear groups.

Langlands' formulation of nonabelian class field theory has implications for  $L$ -functions. It follows from the definitions (3.4), (3.5), (4.5) and (4.4) that for  $G' = \{1\}$ , the automorphic  $L$ -function  $L(s, 1, r)$  on the left hand side of (5.4) is equal to the completed Artin  $L$ -function  $L(s, r)$ . It therefore equals an automorphic  $L$ -function  $L(s, \pi_N) = L(s, \pi)$  for  $GL(N)$ . We should note here that the general principle of functoriality implicitly includes some common spectral properties of the two automorphic representations  $\pi'$  and  $\pi$ . In particular, if the representation  $r$  in (5.4) is irreducible, the automorphic representation  $\pi_N$  of  $GL(N)$  should be cuspidal. If  $N \geq 2$ , this implies that the automorphic  $L$ -function  $L(s, \pi_N)$  is entire, as we noted at the beginning of the section. On the other hand, the Artin conjecture mentioned near the end of §3 asserts that the irreducible  $L$ -function  $L(s, r)$  is entire. The principle of functoriality, in the case  $G' = \{1\}$  and  $G = GL(N)$ , therefore implies this well known conjecture of almost one hundred years. By relating  $L(s, r)$  to a standard automorphic  $L$ -function  $L(s, \pi_N)$ , we would obtain what could be considered an answer to the question (II). For since we now have an understanding of the poles of  $L(s, \pi_N)$ , and can perhaps hope someday to have also an understanding of its zeros, we would have the means to estimate the distribution of the classes  $\{\text{Frob}_v\}$ . Notice that this is a beautiful generalization of the indirect method in Section 2 used by Artin to study abelian  $L$ -functions. In both cases, an analytic problem for arithmetic  $L$ -functions is solved by showing that these functions are also automorphic  $L$ -functions, for which the analytic behaviour is better understood.

In addition to the two striking consequences of functoriality in [23] we have just described, Langlands proposed two further applications. One is to the generalized Ramanujan conjecture, the other to a generalization of the conjecture of Sato-Tate. Both have implications for  $L$ -functions. For the sake of completeness, we shall say a word on each.

The generalized Ramanujan conjecture can be formulated for any  $G$ . It asserts that the local constituents  $\pi_v$  of certain natural automorphic representations  $\pi$  in the discrete spectrum of  $G$  are (locally) tempered, in the extension to the local groups  $G(F_v)$  by Harish-Chandra of the definition from classical Fourier analysis. If  $G = GL(N)$ , for example, it is the unitary cuspidal automorphic representations to which the conjecture applies. The connection

with  $L$ -functions, suggested by Langlands and reinforced by later local harmonic analysis, is that the local components  $\pi_v$  of  $\pi$  will be tempered if and only if the  $L$ -functions  $L(s, \pi, r)$  are analytic in the right half plane  $\operatorname{Re}(s) > 1$ . Given this property, Langlands deduced the generalized Ramanujan conjecture from the principle of functoriality, and the fact that a representation  $\pi$  in the discrete spectrum is automatically unitary [23, p. 43–49]. For more recent progress for the group  $G = GL(2)$ , we refer the reader to [20] and [19].

A generalized Sato-Tate conjecture is only hinted at in [23], and then just in the last paragraph on p. 49. It would apply to any automorphic representation  $\pi$  of  $G$  that satisfies the generalized Ramanujan conjecture. We assume the principle of functoriality. The Ramanujan condition is then valid. It implies that for any unramified place  $v \notin S$ , the conjugacy class  $c_v(\pi)$  in  ${}^L G$  intersects a fixed maximal compact subgroup  ${}^L K = \widehat{K} \rtimes \Gamma_F$  in  ${}^L G = \widehat{G} \rtimes \Gamma_F$ , and can therefore be identified with a conjugacy class in  ${}^L K$ . The problem is to determine the distribution of these classes as  $v$  varies. If the family  $c(\pi)$  is a proper functorial image of a family  $c(\pi')$  for some group  $G'$ , one could determine the distribution of  $c(\pi)$  from that of  $c(\pi')$ . One can therefore assume that  $\pi$  is *primitive*, in the sense that it is not a proper functorial image from some  $G'$ . We would then expect<sup>6</sup> that

$$-\operatorname{ord}_{s=1}(L(s, \pi, r)) = [r : 1]$$

for any finite dimensional representation  $\rho$  of  ${}^L G$ . That is, the order of the pole of  $L(s, \pi, r)$  at  $s = 1$  equals the multiplicity of the trivial representation of  ${}^L G$  in  $r$ . It would then follow from the Wiener-Ikehara Tauberian theorem that the distribution of the classes  $\{c_v(\pi)\}$  in  ${}^L K$  is given by the Haar measure on  ${}^L K$ . In concrete terms, one would be able to express the distribution of classes

$$\{c_v(\pi) \cap {}^L T : v \notin S\}$$

in a maximal torus  ${}^L T = \widehat{T} \rtimes \Gamma_F$  in  ${}^L K$  in terms of the explicit density function on  ${}^L T$  obtained from the Haar measure that occurs in the Weyl integration formula. (See [35, §2, Appendix].)

I have attempted to introduce the subject through Langlands' original paper [23] without discussing subsequent developments. The most famous of these is undoubtedly Wiles' work on the Shimura-Taniyama-Weil conjecture [43], which he used to prove Fermat's Last Theorem. An important foundation for Wiles' work was the Langlands-Tunnell theorem that two-dimensional representations  $r$  of solvable Galois groups  $\Gamma_{E/F}$  satisfy Artin's conjecture. This followed from base change for  $GL(2)$  [28], established by Langlands as an early application of the trace formula. (See [1, §26], for example.) I mention also that R. Taylor has established the classical Sato-Tate conjecture, which applies to the group  $G = GL(2)$ , by using base change for  $GL(N)$  [7] and other means to extract what is needed from the unproven principle of functoriality. (See [40] and the references there.)

## 6. Orthogonal and symplectic groups

The monograph [4] contains a classification of automorphic representations of quasisplit orthogonal and symplectic groups over the number field  $F$ . The groups of interest are attached

---

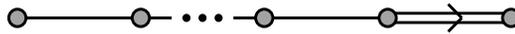
<sup>6</sup>This is actually a little stronger than the principle of functoriality, of which it represents a converse of sorts. However, Langlands' recent ideas [29] for attacking the principle of functoriality, speculative as they may be, would treat this question as well.

to the four infinite families of complex simple Lie algebras. These in turn are represented by the following four infinite families of Coxeter-Dynkin diagrams, for which I am indebted to W. Casselman.

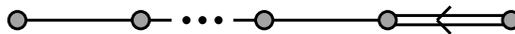
Type  $A_n$



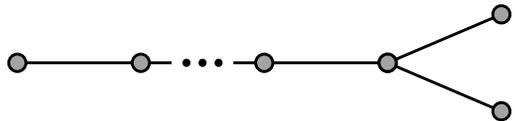
Type  $B_n$



Type  $C_n$



Type  $D_n$



For corresponding complex groups, we could take the special linear groups  $SL(n + 1, \mathbb{C})$ , the odd orthogonal groups  $SO(2n + 1, \mathbb{C})$ , the symplectic groups  $Sp(2n, \mathbb{C})$  and the even orthogonal groups  $SO(2n, \mathbb{C})$ . The family  $A_n$  is the starting point for the classification. Since the representation theory is simplest for general linear groups [18], [30], we take the reductive groups  $GL(N, \mathbb{C})$ ,  $N = n + 1$ , as the complex representatives for this family.

We want to take these groups over the number field  $F$ , which is not algebraically closed. But according to a fundamental theorem of Chevalley, any one of these complex groups corresponds naturally to a canonical group over  $F$ . It is the *split* group attached to the given diagram (and centre). Our interest is actually in *quasisplit* groups. These are obtained by twisting the Galois action on any given split group by a supplementary Galois action on the diagram. The symmetry group of a diagram is the group of bijections of the set of vertices that preserve all edges and directions. It equals  $\mathbb{Z}/2\mathbb{Z}$  in type  $A_n$ , is trivial in types  $B_n$  and  $C_n$ , and equals<sup>7</sup>  $\mathbb{Z}/2\mathbb{Z}$  in type  $D_n$ . A quasisplit group is determined by a homomorphism from the Galois group  $\Gamma_F$  to the symmetry group of the diagram. Following [5], we will not treat nonsplit, quasisplit groups of type  $A_n$ . These are unitary groups, for which we refer the reader to [32]. Since a quasisplit group of type  $B_n$  or  $C_n$  is split, we have only then to consider type  $D_n$ . In this case, a quasisplit group is determined by a quotient of  $\Gamma_F$  of order 1 or 2, or in other words, a Galois extension  $E/F$  of degree 1 or 2.

From now on,  $G$  will stand exclusively for one of our quasisplit groups of type  $B_n$ ,  $C_n$  or  $D_n$ . Its construction above relies on the identification of the supplementary Galois action on the diagram with the Galois action on the underlying split group (determined by a fixed splitting). The transfer of this action to the dual group  $\widehat{G}$  is what is used to define

<sup>7</sup>If  $n = 4$ , this group is actually isomorphic to  $S_3$ , but we agree to consider only the standard symmetries that interchange the two right hand vertices in the diagram.

the semidirect product  ${}^L G = \widehat{G} \rtimes \Gamma_F$ . (See [21, §1].) We list the three families of objects  $(G, \widehat{G}, {}^L G_E)$  explicitly, where  $E/F$  is the minimal Galois extension through which the action of  $\Gamma_F$  factors.

**Type  $\mathbf{B}_n$ :**  $G = SO(2n + 1)$  is split,  $\widehat{G} = Sp(2n, \mathbb{C}) = {}^L G_E, E = F$ .

**Type  $\mathbf{C}_n$ :**  $G = Sp(2n)$  is split,  $\widehat{G} = SO(2n + 1, \mathbb{C}) = {}^L G_E, E = F$ .

**Type  $\mathbf{D}_n$ :**  $G = SO(2n)$  is quasisplit,  $\widehat{G} = SO(2n, \mathbb{C}), {}^L G_E = SO(2n, \mathbb{C}) \rtimes \Gamma_{E/F}, \deg(E/F) \in \{1, 2\}$ .

The other family corresponds to diagrams of type  $\mathbf{A}_n$ . We are taking the underlying group in this case to be the split group  $GL(N)$ , with dual group  $GL(N, \mathbb{C})$ , and minimal  $L$ -group  ${}^L(GL(N))_E$  also equal to  $GL(N, \mathbb{C})$ , for  $N = n + 1$  and  $E = F$ .

The monograph [4] is devoted to a classification of automorphic representations of any of our groups  $G$  in terms of those of general linear groups. In the rest of this section, we shall discuss how it relates to functoriality and  $L$ -functions, the central themes of this article. The classification is based on two general cases of the principle of functoriality. We shall describe them each in turn, following the remarks at the end of §1 of [5].

**Cases of Functoriality:** 1. This case arises from the natural embedding of a complex classical group into a complex general linear group. According to our understanding,  $G$  is any one of our quasisplit classical groups of type  $\mathbf{B}_n, \mathbf{C}_n$  or  $\mathbf{D}_n$  over  $F$ . There is then a canonical embedding of the dual group  $\widehat{G}$  into a general linear group  $GL(N, \mathbb{C})$ , for  $N$  equal to  $2n, 2n + 1$  and  $2n$  respectively. If  $G$  is split over  $F$ , this extends trivially to a canonical  $L$ -embedding

$${}^L G = \widehat{G} \rtimes \Gamma_F \longrightarrow {}^L(GL(N)) = GL(N, \mathbb{C}) \times \Gamma_F$$

of the full  $L$ -group of  $G$  to that of  $GL(N)$ . In the special case of type  $\mathbf{C}_n$ , we also obtain a nonstandard  $L$ -embedding of  ${}^L G$  into  ${}^L(GL(N))$  for any quadratic extension  $E/F$ , by mapping the quotient  $\Gamma_{E/F} = \Gamma_F/\Gamma_E$  isomorphically into the central subgroup  $\{\pm 1\}$  of the image of  $O(2n + 1, \mathbb{C})$  in  $GL(N, \mathbb{C})$ . If  $G$  is not split over  $F$ , it is of type  $\mathbf{D}_n$ . The associated quadratic quotient  $\Gamma_{E/F}$  then acts on  $\widehat{G} = SO(2n, \mathbb{C})$  through the nonidentity connected component of the complex group  $O(2n, \mathbb{C})$ . This leads again to a canonical  $L$ -embedding of  $L$ -groups

$${}^L G = \widehat{G} \rtimes \Gamma_F \longrightarrow {}^L(GL(N)) = GL(N, \mathbb{C}) \times \Gamma_F.$$

2. In the second general case,  $G$  is as in the first. This time, however, we take a product

$$G' = G'_1 \times G'_2$$

of smaller such groups. We require that the dual group

$$\widehat{G}' = \widehat{G}'_1 \times \widehat{G}'_2$$

come with a natural embedding into  $\widehat{G}$ . This means that

$$\widehat{G}' = Sp(2m, \mathbb{C}) \times Sp(2n - 2m, \mathbb{C}) \subset Sp(2n, \mathbb{C}) = \widehat{G},$$

$$\widehat{G}' = SO(2m, \mathbb{C}) \times SO(2n + 1 - 2m, \mathbb{C}) \subset SO(2n + 1, \mathbb{C}) = \widehat{G},$$

and

$$\widehat{G}' = SO(2m, \mathbb{C}) \times SO(2n - 2m, \mathbb{C}) \subset SO(2n, \mathbb{C}) = \widehat{G},$$

for integers  $0 \leq m \leq n$ , when  $G$  is of type  $\mathbf{B}_n$ ,  $\mathbf{C}_n$  and  $\mathbf{D}_n$  respectively. If  $G$  is of type  $\mathbf{B}_n$ ,  $G'$  is split, and the  $L$ -embedding of  ${}^L G'$  into  ${}^L G$  extends trivially to an  $L$ -embedding of  ${}^L G'$  into  ${}^L G$ . If  $G$  is of type  $\mathbf{C}_n$ ,  $G$  and  $G'_2$  are split, but  $G'_1$  can be a quasisplit group defined by an extension  $E_1$  of  $F$  of degree 1 or 2. In this case, we obtain an  $L$ -embedding

$${}^L G' = (\widehat{G}'_1 \times \widehat{G}'_2) \rtimes \Gamma_F \longrightarrow {}^L G = \widehat{G} \times \Gamma_F$$

from the nonstandard embedding of the second factor  ${}^L G'_2$  attached to the quadratic extension  $E_1/F$ . Finally, if  $G$  is of type  $\mathbf{D}_n$ , it is the quasisplit group defined by an extension  $E = F(\sqrt{d})$  of degree 1 or 2. We can then take  $G'_1$  and  $G'_2$  to be quasisplit groups of types  $\mathbf{D}_m$  and  $\mathbf{D}_{n-m}$  defined by any extensions  $E_1 = F(\sqrt{d_1})$  and  $E_2 = F(\sqrt{d_2})$  such that  $d_1 d_2$  equals  $d$ . It is then easy to see that there is a natural  $L$ -embedding of  $L$ -groups

$${}^L G' = (\widehat{G}'_1 \times \widehat{G}'_2) \rtimes \Gamma_F \longrightarrow {}^L G = \widehat{G} \rtimes \Gamma_F.$$

We thus obtain two basic cases of the principle of functoriality by taking the  $L$ -homomorphism  $\rho$  to be any one of the  $L$ -embeddings we have just described. The first is at the heart of the classification of representations of  $G$  (both local and global) in terms of those of  $GL(N)$ . The second provides the foundation for an understanding of the precise functorial correspondence from  $G$  to  $GL(N)$ .

**Theorem 6.1.** *The principle of functoriality stated in §4 is valid if  $\rho$  is any one of the  $L$ -embeddings in the two general cases described above.*

This theorem is a consequence of the classification of representations of  $G$  in [4]. It has a significant application to Rankin-Selberg products. These are the automorphic  $L$ -functions whose arithmetic analogues correspond to tensor products of finite dimensional representations of  $\Gamma_F$  (or  $W_F$ ).

We first review the standard theory of Rankin-Selberg products for general linear groups. In this case, the underlying quasisplit group is a product  $GL(N_1) \times GL(N_2)$ , while the underlying representation  $r = r_N$  of its  $L$ -group is given by the standard representation

$$g = g_{N_1} \times g_{N_2} : X \longrightarrow g_{N_1} \times {}^t g_{N_2}, \quad g_{N_i} \in GL(N_i, \mathbb{C}),$$

of the dual group  $GL(N_1, \mathbb{C}) \times GL(N_2, \mathbb{C})$  on the  $N = N_1 N_2$ -dimensional vector space of complex  $(N_1 \times N_2)$ -matrices  $X$ . For any automorphic representation  $\pi_N = \pi_{N_1} \otimes \pi_{N_2}$  of this group, we can form the incomplete  $L$ -function

$$L^S(s, \pi_{N_1} \times \pi_{N_2}) = L^S(s, \pi_N, r_N)$$

of (4.4). In this case, it is known how to define the local  $L$ -functions

$$L(s, \pi_{N_1, v} \times \pi_{N_2, v}) = L(s, \pi_{N, v}, r_{N, v}) \tag{6.1}$$

and  $\varepsilon$ -factors

$$\varepsilon(s, \pi_{N_1, v} \times \pi_{N_2, v}, \psi_v) = \varepsilon(s, \pi_{N, v}, r_{N, v}, \psi_v) \tag{6.2}$$

in a purely local manner for all valuations  $v$ , in such a way that the completed  $L$ -function

$$L(s, \pi_{N_1} \times \pi_{N_2}) = L_S(s, \pi_{N_1} \times \pi_{N_2}) L^S(s, \pi_{N_1} \times \pi_{N_2}) = L_S(s, \pi_N, r_N) L^S(s, \pi_N, r_N)$$

has analytic continuation and functional equation

$$L(s, \pi_{N_1} \times \pi_{N_2}) = \varepsilon(s, \pi_{N_1} \times \pi_{N_2}) L(1 - s, \pi_{N_1}^\vee \times \pi_{N_2}^\vee),$$

for the associated global  $\varepsilon$ -factor

$$\varepsilon(s, \pi_{N_1} \times \pi_{N_2}) = \varepsilon(s, \pi_N, r_N).$$

The general principle of functoriality applies to the mapping  $r_N$  from  $GL(N_1, \mathbb{C}) \times GL(N_2, \mathbb{C})$  to  $GL(N, \mathbb{C})$ . However, it is far from known in this case. On the other hand, and in contrast to Langlands' first application of functoriality described in §4, the analytic continuation and functional equation of Rankin-Selberg products has been established directly. There have been two different approaches to the theory, both of which lead to the same results. The original method [10, 17, 18], [30, Appendix] combines certain integrals with the Poisson summation formula, in a way that is reminiscent of Tate's thesis [37] (which applies to the special case that  $(N_1, N_2) = (N, 1)$ ). The other approach, known as the Langlands-Shahidi method [11, 24, 36], combines Whittaker models and intertwining operators with the analytic continuation and functional equations for Eisenstein series established by Langlands in his study [25] of continuous automorphic spectra. It is capable of considerably broader application.

Our application of Theorem 6.1 is to Rankin-Selberg products for classical groups, specifically a product  $G_1 \times G_2$  of any two groups from our general family of quasisplit special orthogonal and symplectic groups. From the standard  $L$ -embeddings

$$\rho_i : {}^L G_i \longrightarrow {}^L (GL(N_i)), \quad i = 1, 2,$$

of Case 1 above, we obtain a homomorphism

$$\rho_1 \times \rho_2 : {}^L (G_1 \times G_2) \longrightarrow GL(N_1, \mathbb{C}) \times GL(N_2, \mathbb{C}).$$

If  $\pi = \pi_1 \otimes \pi_2$  is an automorphic representation of  $G_1 \times G_2$ , and  $r$  is the composition  $r_N \circ (\rho_1 \times \rho_2)$ , we can form the partial  $L$ -function

$$L^S(s, \pi_1 \times \pi_2) = L^S(s, \pi, r)$$

for the group  $G_1 \times G_2$ . We apply Theorem 6.1 to the two  $L$ -embeddings  $\rho_i$ . It attaches to the two automorphic representations  $\pi_i \in \Pi_{\text{aut}}(G_i)$  two (self-dual, isobaric) automorphic representations  $\pi_{N_i} \in \Pi_{\text{aut}}(N_i)$  for the general linear groups  $GL(N_i)$ , such that

$$L^S(s, \pi_1 \times \pi_2) = L^S(s, \pi_{N_1} \times \pi_{N_2}).$$

In other words, the partial  $L$ -function for  $G_1 \times G_2$  on the left equals its analogue for  $GL(N_1) \times GL(N_2)$  on the right. It follows from the theory we have just described for general linear groups that we can define the supplementary  $L$ -factor

$$L_S(s, \pi_1 \times \pi_2) = L_S(s, \pi_{N_1} \times \pi_{N_2}) \tag{6.3}$$

and the global  $\varepsilon$ -factor

$$\varepsilon(s, \pi_1 \times \pi_2) = \varepsilon(s, \pi_{N_1} \times \pi_{N_2}) \tag{6.4}$$

for  $G_1 \times G_2$  so that the completed *L*-function

$$L(s, \pi_1 \times \pi_2) = L_S(s, \pi_1 \times \pi_2)L^S(s, \pi_1 \times \pi_2) \tag{6.5}$$

has analytic continuation, with the functional equation

$$L(s, \pi_1 \times \pi_2) = \varepsilon(s, \pi_1 \times \pi_2) L(1 - s, \pi_1 \times \pi_2). \tag{6.6}$$

Our discussion for  $G_1 \times G_2$  does not to this point include a local theory. That is, it does not give a local construction of the factors implicit in the left hand sides of (6.3) and (6.4). This is in contrast to the theory for  $GL(N)$ , which not only gives a local construction for the factors (6.1) and (6.2) for the right hand side, but also relates them (according to the local Langlands correspondence for  $GL(N)$ ) to their arithmetic analogues in (4.11) and (4.13), for representations  $r_{N_1,v} \otimes r_{N_2,v}$  of the local Weil groups  $W_{F_v}$ . The stronger results for  $G_1 \times G_2$  follow, at least for representations  $\pi_i$  that are globally tempered, from the local and global classifications in [4].

In summary, Theorem 5.1 establishes two cases of functoriality for quasisplit orthogonal and symplectic groups. As a corollary of the first case, we also obtain the analytic continuation of the corresponding Rankin-Selberg *L*-functions (6.5), with functional equation (6.6). This last result is very much in the spirit of our earlier discussion. Like Artin’s proof of analytic continuation and functional equation for the abelian *L*-functions that bear his name, and Langlands’ reduction of the analytic properties of general Artin *L*-functions and general automorphic *L*-functions to the principle of functoriality, the approach is indirect. Rather than deal with the unknown *L*-functions directly, we establish classification theorems that limit their scope. That is, contrary perhaps to what might have been expected, the unknown *L*-functions are in fact included among a class of *L*-functions whose analytic behaviour is understood.

## 7. Remarks on the classification

We have not described the classification [4] that might have been the natural topic of this article, having chosen instead to focus on its simpler implications for our historical introduction to the Langlands program. The classification is given by Theorems 1.5.1, 1.5.2 and 1.5.3 of [4]. It is also summarized from different points of view in the three surveys [3, 5, 6]. Partial results were established earlier for generic representations in [9, 14], by quite different methods. It is now possible to see where the generic representations of these papers fit into the general classification [4, Proposition 8.3.2].

We conclude with a few very general remarks on the structure of the classification. The first of the two cases of functoriality in Theorem 6.1 gives a canonical mapping

$$\mathcal{C}_{\text{aut}}(G) \longrightarrow \mathcal{C}_{\text{aut}}(N) = \mathcal{C}_{\text{aut}}(GL(N)), \tag{7.1}$$

from automorphic eigenfamilies for our classical group  $G$  to automorphic eigenfamilies for  $GL(N)$ . The methods of [4] are designed for the representations that occur in the spectral decomposition of  $L^2(G(F)\backslash G(\mathbb{A}))$ , namely the subset  $\Pi(G) \subset \Pi_{\text{aut}}(G)$  of automorphic representations we are calling globally tempered. The version of Theorem 6.1 that arises<sup>8</sup>

---

<sup>8</sup>It is not stated explicitly in [4]. In fact, the analogue of the second case of Theorem 6.1 is not quite true for  $\mathcal{C}(G)$ , thanks to an interesting pathology discovered by Cogdell and Piatetski-Shapiro. (See [5, §3] and [6, §8].)

most directly from [4] actually applies to automorphic eigenfamilies that are globally tempered, namely the image  $\mathcal{C}(G) \subset \mathcal{C}_{\text{aut}}(G)$  of  $\Pi(G)$  under the mapping  $\pi \rightarrow c(\pi)$ . The restriction of (7.1) can be seen from [4] to give a canonical mapping

$$\mathcal{C}(G) \longrightarrow \mathcal{C}(N) \tag{7.2}$$

from  $\mathcal{C}(G)$  to the image  $\mathcal{C}(N) \subset \mathcal{C}_{\text{aut}}(N)$  of the set  $\Pi(N)$  of globally tempered automorphic representations of  $GL(N)$ . We shall comment briefly on the general steps needed to obtain a classification<sup>9</sup> of  $\Pi(G)$  from (7.2).

The general linear group  $GL(N)$  has the remarkable property that the mapping  $\pi_N \rightarrow c(\pi_N)$  from  $\Pi(N)$  to  $\mathcal{C}(N)$  is a bijection. This follows from fundamental theorems of Jacquet-Shalika [18] and Mœglin-Waldspurger [30]. (See [4, §1.3] and [6, §4].) The composition

$$\Pi(G) \longrightarrow \mathcal{C}(G) \longrightarrow \mathcal{C}(N) \xleftarrow{\sim} \Pi(N)$$

then gives a mapping  $\pi \rightarrow \pi_N$  from  $\Pi(G)$  to  $\Pi(N)$ . Langlands’ theory of Eisenstein series [25] constructs the automorphic spectrum of any group in terms of automorphic discrete spectra. For our group  $G$ , it is therefore enough to classify the subset  $\Pi_2(G)$  of representations in  $\Pi(G)$  that occur in the discrete spectrum, the subspace  $L^2_{\text{disc}}(G(F)\backslash G(\mathbb{A}))$  of  $L^2(G(F)\backslash G(\mathbb{A}))$  that decomposes under right translation by  $G(\mathbb{A})$  into a direct sum of irreducible representations. To classify automorphic representations of  $G$  in terms of those of  $GL(N)$ , we would need to give an explicit description of the restricted mapping

$$\pi \longrightarrow \pi_N, \quad \pi \in \Pi_2(G), \tag{7.3}$$

from  $\Pi_2(G)$  to  $\Pi(N)$ . Specifically, we would need to characterize the image and the kernel of this mapping.

To characterize the image of (7.3), it is necessary to analyze the (globally tempered) automorphic representations of  $GL(N)$  that are self-dual. This is not difficult to do, using the general structure of the set of self-dual,  $N$ -dimensional representations of an arbitrary Galois group  $\Gamma_{E/F}$  for guidance, and the classification in [30] of the automorphic, relatively discrete spectrum of  $GL(N)$  [4, §1.2, 1.4]. The problem, it then turns out, is to establish two necessary and sufficient conditions for a self-dual *cuspidal* automorphic representation  $\pi_N$  of  $GL(N)$  to lie in the image of (7.3). One is a familiar condition [9, 14] on the existence of a pole at  $s = 1$  of a certain automorphic  $L$ -function of  $\pi_N$ . The other is a more technical condition in harmonic analysis, which is harder to state, but which is at the centre of the argument. The two conditions are among the last things to be established in the classification, but they lead in the end to a clear description of the image of (7.3).

The fibres of (7.3) are often large. They occur in packets

$$\Pi_\psi, \quad \psi \in \Psi_2(G),$$

parametrized by a family  $\Psi_2(G)$  of objects  $\psi$  that is in canonical bijection with the image of (7.3), the subset of  $\Pi(N)$  we have just discussed. These global “parameters” have localizations  $\psi_v$  at valuations  $v$ , which are parameters in the more familiar sense. They belong to the set  $\Psi(G_v)$  of local  $L$ -homomorphisms<sup>10</sup>

$$\psi_v : L_{F_v} \times SU(2) \longrightarrow {}^L G_v,$$

---

<sup>9</sup>In principle, one can obtain a classification of the larger set  $\Pi_{\text{aut}}(G)$  from that of  $\Pi(G)$  in [4] and the criterion for automorphy in [26]. (See [6, §8].)

<sup>10</sup>In the domain,  $L_{F_v}$  is the local Langlands group. It is defined as the local Weil group  $W_{F_v}$  if  $v$  is archimedean, and the product of  $W_{F_v}$  with a separate copy of  $SU(2)$  if  $v$  is nonarchimedean.

taken up to  $\widehat{G}$ -conjugacy in the local *L*-group  ${}^L G_v$ , such that the image of  $\psi_v$  in  $\widehat{G}$  is relatively compact. A significant part of the global classification in [4] is purely local. To every local parameter  $\psi_v \in \Psi(G_v)$ , one has to attach a canonical, *finite* set  $\Pi_{\psi_v} \subset \Pi_{\text{unit}}(G_v)$  of irreducible unitary representations of  $G(F_v)$ . A global packet  $\Pi_{\psi}$  is then defined as the set of restricted tensor products

$$\Pi_{\psi} = \left\{ \pi = \bigotimes_v^{\sim} \pi_v : \pi_v \in \Pi_{\psi_v} \right\} \tag{7.4}$$

of representations from the corresponding local packets. The subset  $\Phi_{\text{bdd}}(G_v)$  of parameters  $\phi_v \in \Psi(G_v)$  that are constant on the factor  $SU(2)$  are known as (bounded) Langlands parameters. A prerequisite for the study of the general packets  $\Pi_{\psi_v}$  in Chapter 7 of [4] is the proof in Chapter 6 of [4] of the local Langlands correspondence for  $G_v$ . This asserts<sup>11</sup> that the set  $\Pi_{\text{temp}}(G_v)$  of (locally) tempered, irreducible representations of  $G(F_v)$  is a disjoint union over  $\phi_v \in \Phi_{\text{bdd}}(G_v)$  of the local Langlands packets  $\Pi_{\phi_v}$ .

The global classification thus depends on a local description of representations in order to define the global packets (7.4). The main global result of [5] is Theorem 1.5.2. It gives a multiplicity formula for any irreducible representation of  $G(\mathbb{A})$  in the discrete spectrum. More precisely, the theorem asserts that any representation in  $\Pi_2(G)$  lies in a unique global packet  $\Pi_{\psi}$ . For any  $\pi \in \Pi_{\psi}$ , it then gives an explicit multiplicity formula for  $\pi$  in  $L^2_{\text{disc}}(G(F)\backslash G(\mathbb{A}))$  in terms of invariants attached to its local constituents. Its proof is a multifaceted induction, which includes most of the other results in [4], and takes up much of the volume.

**Acknowledgements.** Supported in part by NSERC Discovery Grant A3483.

## References

- [1] J. Arthur, *An introduction to the trace formula*, in *Harmonic Analysis, the Trace Formula, and Shimura Varieties*, Clay Mathematics Proceedings, vol. 4, 2005, pp. 1–263.
- [2] —, *A stable trace formula III. Proof of the main theorems*, *Annals of Math.* **158** (2003), 769–873.
- [3] —, *The Endoscopic Classification of Representations*, in *Automorphic Representations and L-functions*, Tata Institute of Fundamental Research, 2013, pp. 1–22.
- [4] —, *The Endoscopic Classification of Representations: Orthogonal and Symplectic Groups*, Colloquium Publications, 61, 2013, American Mathematical Society.
- [5] —, *Classifying automorphic representations*, in *Current Developments in Mathematics*, Volume 2012, International Press of Boston, pp. 1–58.
- [6] —, *Eigenfamilies, characters and multiplicities*, preprint, to appear in *Actes de la conférence Laumon*.

---

<sup>11</sup>In the case that  $G = SO(2n)$  is of type  $\mathbf{D}_n$ , we actually prove something weaker. We classify only the set  $\widetilde{\Pi}_{\text{temp}}(G_v)$  of orbits in  $\Pi_{\text{temp}}(G_v)$  under the quotient  $\text{Out}_N(G_v) = SO(2n, F_v)\backslash O(2n, F_v)$ , a group of order 2. The local packets  $\widetilde{\Pi}_{\phi_v}$  and  $\widetilde{\Pi}_{\psi_v}$  are likewise  $\text{Out}_N(G_v)$ -orbits of irreducible representations, and the global packets  $\widetilde{\Pi}_{\psi}$  are formal tensor products of such objects.

- [7] J. Arthur and L. Clozel, *Simple Algebras, Base Change, and the Advanced Theory of the Trace Formula*, Ann. of Math. Studies 120, Princeton Univ. Press, Princeton, N.J., 1989.
- [8] A. Borel and H. Jacquet, *Automorphic forms and automorphic representations*, in Automorphic Forms, Representations and  $L$ -functions, Proc. Symp. Pure Math. vol. 33, Part 1, Amer. Math. Soc., 1979, pp. 189–202.
- [9] J. Cogdell, H. Kim, I. Piatetski-Shapiro, and F. Shahidi, *Functoriality for the classical groups*, Publ. Math. Inst. Hautes Études Sci. **99** (2004), 163–233.
- [10] J. Cogdell and I. Piatetski-Shapiro, *Remarks on Rankin-Selberg convolutions*, in *Contributions to Automorphic Forms, Geometry and Number Theory*, Johns Hopkins University Press, 2004, pp. 255–278.
- [11] J. Cogdell, I. Piatetski-Shapiro and F. Shahidi, *Functoriality for quasisplit classical groups*, in *On Certain  $L$ -functions*, Clay Mathematics Proceedings, vol. 13, 2011, pp. 117–140.
- [12] P. Deligne, *Les constantes des équations fonctionnelles des fonctions  $L$* , in *Modular Forms of One Variable II*, Lecture Notes in Math. 349, Springer, New York, 1973, pp. 501–597.
- [13] D. Flath, *Decomposition of representations into tensor products*, in *Automorphic Forms, Representations and  $L$ -functions*, Proc. Sympos. Pure Math. vol. 33, Part 1, Amer. Math. Soc., 1979, pp. 179–184.
- [14] D. Ginzburg, S. Rallis, and D. Soudry, *The descent map from automorphic representations of  $GL(n)$  to classical groups*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011.
- [15] R. Godement and H. Jacquet, *Zeta functions of simple algebras*, Lecture Notes in Math. 260, Springer, New York, 1972.
- [16] H. Jacquet, *Principle  $L$ -functions in Automorphic Forms, Representations and  $L$ -functions*, Proc. Sympos. Pure Math. vol. 33, Part 2, Amer. Math. Soc., 1979, 63–86.
- [17] H. Jacquet, I. Piatetski-Shapiro, and J. Shalika, *Rankin-Selberg convolutions*, Amer. J. Math. **105** (1983), 367–464.
- [18] H. Jacquet and J. Shalika, *On Euler products and the classification of automorphic representations II*, Amer. J. Math. **103** (1981), 777–815.
- [19] H. Kim, *Functoriality for the exterior square of  $GL_4$  and the symmetric fourth of  $GL_2$* , with appendix 1 by D. Ramakrishnan and appendix 2 by Kim and P. Sarnak, J. Amer. Math. Soc. **16** (2003), 139–183.
- [20] H. Kim and F. Shahidi, *Functorial products for  $GL_2 \times GL_3$  and the symmetric cube for  $GL_2$* , Ann. of Math. (2) **155** (2002), 837–893.
- [21] R. Kottwitz, *Stable trace formula: cuspidal tempered terms*, Duke Math. J. **51** (1984), 611–650.
- [22] J.-P. Labesse, and J.-L. Waldspurger, *La formule des traces tordues d’après le Friday Morning Seminar*, CRM Monograph Series **31**, American Mathematical Society, 2013.
- [23] R. Langlands, *Problems in the theory of automorphic forms*, in *Lectures in Modern Analysis and Applications*, Lecture Notes in Math. 170, Springer, New York, 1970, pp. 18–61.

- [24] —, *Euler Products*, Yale University Press, 1971.
- [25] —, *On the Functional Equations Satisfied by Eisenstein Series*, Lecture Notes in Math. 544, Springer, New York, 1976.
- [26] —, *On the notion of an automorphic representation. A supplement to the preceding paper*, in *Automorphic Forms, Representations and L-functions*, Proc. Sympos. Pure Math. vol. 33, Part 1, Amer. Math. Soc., 1979, pp. 203–208.
- [27] —, *Automorphic representations, Shimura varieties, and motives, Ein Märchen*, in *Automorphic Forms, Representations and L-functions*, Proc. Sympos. Pure Math. vol. 33, Part 2, Amer. Math. Soc., 1979, pp. 205–246.
- [28] —, *Base Change for  $GL(2)$* , Ann. of Math. Studies 96, Princeton Univ. Press, Princeton, N.J., 1980.
- [29] —, *A prologue to “Functoriality and reciprocity”, part I*, Pacific J. Math., Special issue devoted to the memory of Jonathan Rogawski, **260** (2012), 583–663.
- [30] C. Moeglin and J.-L. Waldspurger, *Le spectre résiduel de  $GL(n)$* , Ann. Scient. Éc. Norm. Sup. 4<sup>e</sup> série **22** (1989), 605–674.
- [31] —, *La partie géométrique de la formule des traces tordue*, preprint.
- [32] C.P. Mok, *Endoscopic classification of representations of quasisplit unitary groups*, to appear in Memoirs of the American Mathematical Society.
- [33] B.C. Ngô, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. Inst. Hautes Études Sci. **111** (2010), 1–269.
- [34] L. Schoenfeld, *Sharper bounds for Chebyshev functions  $\theta(x)$  and  $\Psi(x)$  II*, Mathematics of Computation **30** (1976), 337–360.
- [35] J.-P. Serre, *Abelian  $\ell$ -adic Representations and Elliptic Curves*, Benjamin, New York, 1968.
- [36] F. Shahidi, *On the Ramanujan conjecture and finiteness of poles for certain L-functions*, Annals of Math. **127** (1988), 547–584.
- [37] J. Tate, *Fourier Analysis in Number Fields and Hecke’s Zeta Functions*, in Algebraic Number Theory, Thompson, Washington, D.C., 1967, pp. 305–347.
- [38] —, *Global Class Field Theory*, in Algebraic Number Theory, Tompson, Washington, D.C., 1967, pp. 163–203.
- [39] —, *Number theoretic background*, in *Automorphic Forms, Representations and L-functions*, Proc. Sympos. Pure Math. 33, Part 2, Amer. Math. Soc., 1979, pp. 3–26.
- [40] R. Taylor, *Automorphy for some  $\ell$ -adic lifts of automorphic mod  $\ell$  Galois representations. II*, Pub. Math. Inst. Hautes Études Sci. **108** (2008), 183–239.
- [41] J.-L. Waldspurger, *Préparation à la stabilisation de la formule des traces tordue I: endoscopie tordue sur un corps local*, preprint.
- [42] —, *Préparation à la stabilisation de la formule des traces tordue II: intégrales orbitales et endoscopie sur un corps archimédien*, preprint.
- [43] A. Wiles, *Modular elliptic curves and Fermat’s Last Theorem*, Annals of Math. **142** (1995), 443–551.



# Integrable probability

Alexei Borodin

**Abstract.** The goal of the lecture is to survey the emerging field of integrable probability which aims at identifying and analyzing exactly solvable probabilistic models. The models and results are often easy to describe, yet difficult to find, and they carry essential information about broad universality classes of stochastic processes. The methods of analysis are largely algebraic, and they are deeply rooted in representation theory.

**Mathematics Subject Classification (2010).** 60K35, 82B23, 82C41.

**Keywords.** Integrability, random growth.

## Introduction

This talk is about probabilistic systems that can be analyzed by essentially algebraic methods.

The historically first example of such a system goes back to De Moivre (1738) and Laplace (1812) who considered the problem of finding the asymptotic distribution of the sum of i. i. d. random variables for Bernoulli trials, when the pre-limit distribution is explicit, and took the limit of the resulting expression. While this computation may look like a simple exercise when viewed from the heights of modern probability, in its time it likely served the role of a key stepping stone — first rigorous proofs of central limit theorems appeared only in the beginning of the XXth century.

As an example of a similar modern development, we are currently in a “De Moivre-Laplace stage” for a certain class of stochastic systems which is often referred to as the *KPZ universality class*, after an influential work of Kardar-Parisi-Zhang in mid-80’s. We will be interested in the case of one and two space dimensions.

While the class and some of its members have been identified by physicists, the first examples of convincing (actually, rigorous) analysis were provided by mathematicians, who were also able to identify the distributions that play the role of the Gaussian law. For one space dimension, they are often referred to as the *Tracy-Widom type distributions* as they had previously appeared in Tracy-Widom’s work on spectra of large random matrices.

The reason for mathematicians’ success was that there is an unusually extensive amount of algebra and combinatorics required to gain access to suitable pre-limit formulas that admit large time limit transitions. The “exactly solvable” or *integrable* members of the class should be viewed as projections of much more powerful objects whose origins lie in representation theory. In a way, this is similar to integrable systems that can also be viewed

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

as projections of representation theoretic objects; and this is one reason we use the words *integrable probability* to describe the phenomenon.

What goes below is a write-up of an hour-long talk; detailed expositions can be found in [9, 15, 20, 23].

## 1.

Suppose that one is building a tower out of unit blocks. Blocks are falling from the sky, as shown on Figure 1.1, and the tower slowly grows. If one introduces randomness by declaring the times between arrivals of blocks to be independent identically distributed (i.i.d.) random variables, then one obtains the simplest 1d random growth model. The kind of question we would like to answer is what the height  $h(t)$  of tower at time  $t$  is?

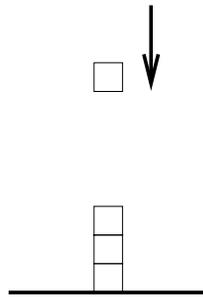


Figure 1.1. Building a tower from standard blocks.

The classical central limit theorem (see e.g. [6, Chapter 5] or [45, Chapter 4]) provides the answer:

$$h(t) \approx c_1^{-1}t + \xi c_2 c_1^{-\frac{3}{2}} t^{\frac{1}{2}},$$

where  $c_1$  and  $c_2$  are the mean and standard deviation of the times between arrivals of the blocks, respectively, and  $\xi$  is a standard normal random variable  $N(0, 1)$ .

## 2.

Let us try to generalize by introducing one space dimension.

If blocks fall independently in different columns, then one obtains a 2d growth model, as shown on Figure 2.1. When there are no interactions between blocks and the blocks are aligned (cf. the left panel of Figure 2.1), the columns grow independently and fluctuations remain of order  $t^{1/2}$  — this is called “random deposition”.

But what happens if we allow column interaction by, for example, letting blocks travel finitely many (say, no more than 1) units left or right to locate the lowest possible landing position, cf. middle panel of Figure 2.1. Such random growth processes are commonly called “random deposition with relaxation”. Or if we make blocks sticky so that they attach to the sides of the boxes in adjacent columns, as shown on the right panel of Figure 2.1 — this is known as “ballistic deposition”?

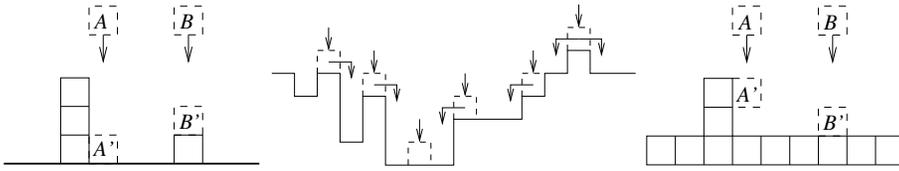


Figure 2.1. Random deposition, random deposition with relaxation, and ballistic deposition.

Computer simulations (see e.g. [5]) show that the height fluctuations in the second and the third model are of order  $t^{1/4}$  and  $t^{1/3}$ , respectively, and it is visible to a naked eye that the roughness of the interface is different, cf. Figure 2.2.

We are pretty far from fully understanding the observed phenomenon; for example, proving the  $1/3$  fluctuation exponent for ballistic deposition is way beyond currently available techniques. However, we can do something else.

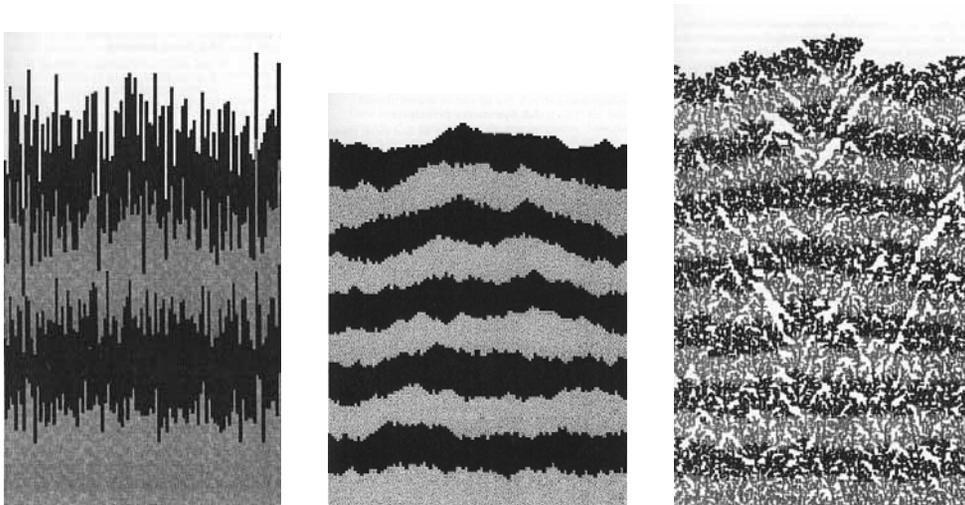


Figure 2.2. Computer simulations of the three models of random growth (from [5]).

Physicists know fairly well how to identify broad classes of random growth models that should have exactly the same asymptotic behavior of interface fluctuations — those are known as ‘universality classes’. In 0 space dimensions, the classical central limit can be viewed as an illustration of such behavior: The universality class consists of all growth models with i. i. d. intervals between block arrivals, and the Gaussian distribution describes the universal fluctuations. Although for now we can only wish for something similar in  $\geq 1$  space dimensions, we are capable of finding exactly solvable, or *integrable* models in certain universality classes and analyzing them to a great level of detail, thus identifying the (conjecturally) universal fluctuations.

## 3.

One (simple) example of an integrable model in 0 space dimensions is the case when the times between block arrivals are geometrically distributed random variables. Then the tower height  $h(t)$  at any time moment is distributed as a sum of independent Bernoulli random variables; its distribution is given by binomial coefficients, and the application of Stirling's formula proves the convergence of rescaled  $h(t)$  to the standard Gaussian. This is the celebrated De Moivre–Laplace theorem.

Let us now describe an integrable model of random growth in one space dimension. Consider the interface given by a broken line with segments of same length and of slopes  $\pm 1$ , as shown on Figure 3.1, left panel. Its growth consists of adding a new unit box at each local minimum, independently after an exponentially distributed waiting time.

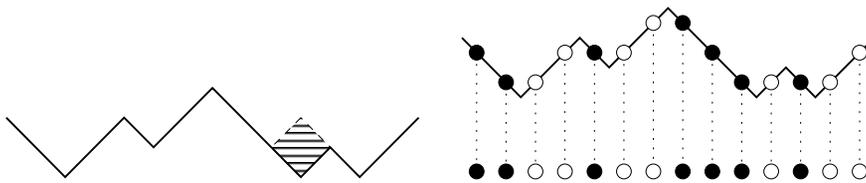


Figure 3.1. Broken line with slopes  $\pm 1$  with local minimum where a box can be added, and correspondence with particle configurations on  $\mathbb{Z}$

There is also an appealing equivalent formulation of this growth model. Project the interface to a straight line at 45 degree angle to the segments, and place “particles” at images of unit segments of slope  $-1$  and “holes” at images of segments of slope  $+1$ , see Figure 3.1, right panel. Then the random growth process is equivalent to the following particle update rule (as seen from examining the picture): Each particle jumps to the right by one unit independently of the others after an exponential waiting time (in other words, each particle jumps with probability  $dt$  in each very small time interval  $[t, t + dt]$ ) except for the exclusion constraint: Jumps to the already occupied spots are prohibited. One can view this as a simplified model of a one-lane highway with particles representing cars. This model is widely known under the name of Totally Asymmetric Simple Exclusion Process (or TASEP, for brevity), cf. [52, 53, 66].

At large times, in the first order approximation (law of large numbers type behavior also referred to as *hydrodynamic limit*, obtained when time and space coordinates are scaled in the same way) TASEP's interface evolves deterministically according to the (first order, nonlinear) inviscid Burgers equation

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}(\rho(1 - \rho)),$$

where  $\rho = \rho(x, t) \in [0, 1]$  is the local density of TASEP's particles. This is a nontrivial statement that has been proved in a fairly large generality, see e.g. the introduction of [38] for a brief survey. In particular, the shocks in the solution of this equation correspond to the traffic jams in the system of cars-particles on a one-lane highway.

Proceeding to fluctuations around the global hydrodynamic interface profile, here is the very first result that rigorously proved the existence of the  $1/3$  exponent (it was conjectured by physicists more than two decades earlier [37, 40, 46]).

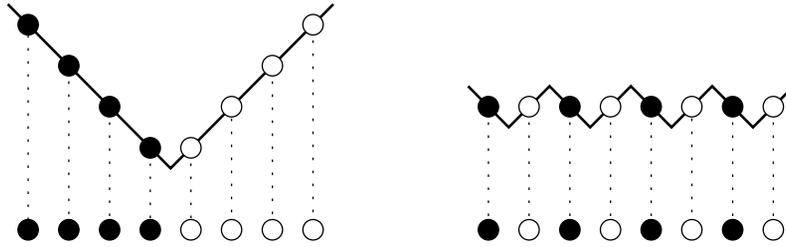


Figure 3.2. Wedge and flat initial conditions: broken lines and corresponding particle configurations.

**Theorem** ([42]). *Suppose that at time 0 the interface  $h(x; t)$  is a wedge  $h(x, 0) = |x|$  as shown on Figure 3.2, left panel. Then for every  $x \in (-1, 1)$*

$$\lim_{t \rightarrow \infty} \mathbb{P} \left( \frac{h(tx, t) - c_1(x)t}{c_2(x)t^{1/3}} \geq -s \right) = F_2(s),$$

where  $c_1(x), c_2(x)$  are certain explicit functions of  $x$ .

Here  $F_2(s)$  is a distribution that originated in *random matrix theory*, known under the name of the GUE Tracy-Widom distribution. It is the limiting distribution, as the size of the matrix tends to infinity, of the properly centered and scaled largest eigenvalue in the Gaussian Unitary Ensemble of random matrices (which is the probability measure with density proportional to  $\exp(-\text{Trace}(X^2))$  on Hermitian matrices), see [68].

At first glance, one might expect that because of universality, the fluctuations of the TASEP interface should be described by  $F_2$  for any initial condition. However, one proves the following

**Theorem** ([17, 61]). *Suppose that at time 0 the interface  $h(x; t)$  is flat as shown on Figure 3.2 (right panel). Then for every  $x \in \mathbb{R}$*

$$\lim_{t \rightarrow \infty} \mathbb{P} \left( \frac{h(x, t) - c_3t}{c_4t^{1/3}} \geq -s \right) = F_1(s),$$

where  $c_3, c_4$  are certain explicit positive constants.

Similarly to  $F_2(s)$ ,  $F_1(s)$  from the right-hand side is the GOE Tracy-Widom distribution that arises as a scaling limit of the largest eigenvalue in Gaussian Orthogonal Ensemble of real symmetric matrices distributed according to  $\text{const} \cdot \exp(-\text{Trace}(X^2))dX$ , see [69].<sup>1</sup>

#### 4.

This caveat, however, appears to be only a minor correction to the universality principle: The two theorems above conjecturally provide the distributions of the fluctuations for a whole universality class of random growth models in (1+1) dimensions which is usually referred to

<sup>1</sup>The indices “2” and “1” stand for the dimension of the base field over the reals; one can naturally define distributions  $F_\beta$  for  $\beta > 0$  that at  $\beta = 1, 2, 4$  correspond to the largest eigenvalues of Gaussian Orthogonal/Unitary/Symplectic Ensembles, see [60].

as the Kardar-Parisi-Zhang (KPZ) universality class. It is just that the asymptotic behavior becomes more delicate than in (0+1) dimensional case, namely, while for deterministic initial conditions scaling by  $t^{1/3}$  is always the same, the resulting distribution may also depend on the “subclass” of the model. Conjecturally, for deterministic initial conditions, the only two generic subclasses are the ones we have seen. They are distinguished by whether the global interface profile is locally curved or flat near the observation location.

The KPZ universality class was suggested in [46], and its members can be described as random growth models that have three key features:

- Locality of growth — distant parts of the interface evolve independently;
- Smoothing mechanism (a.k.a. relaxation) — deep holes in the interface tend to fill up, and no fractal structures appear;
- Lateral growth — the interface grows in the normal direction to its global profile (this could be relaxed to simply claiming that the vertical speed of growth depends on the slope of the global profile).

While the above description may seem vague, it works remarkably well. As an example, out of the three models of Figure 2.1 above, all three have local growth, the last two have a smoothing mechanism, and only ballistic deposition enjoys lateral growth <sup>2</sup>. A recent survey of the KPZ universality class in (1+1) dimensions can be found in [27].

## 5.

As was mentioned above, TASEP is one integrable model in the KPZ class. One way to describe its solvability is to say that its study can be reduced to that of a determinantal point process (see e.g. [7] and references therein for details on such processes). The techniques of the determinantal point processes can be utilized to solve a few other random growth models in the KPZ class. Those include (citation lists are non-exhaustive, see also references therein and survey [35])

- Discrete time TASEPs [15, 16, 18, 44];
- Discrete and continuous time PushASEPs [14, 15];
- Directed last passage percolation with geometric/Bernoulli/exponential edge weights [2, 42, 43];
- Polynuclear growth processes [3, 4, 18, 41, 59, 62].

The exact conjectures on fluctuation behavior of generic models in the KPZ class were derived from asymptotic analysis of these determinantly solvable cases.

In recent years, there has been a substantial progress in analyzing a class of integrable yet not determinantal KPZ models. Those include (again, citations lists are non-exhaustive)

- Partially Asymmetric Simple Exclusion Process (PASEP or ASEP) [13, 70];

---

<sup>2</sup>If a growth model is local, has relaxation, but its vertical speed of growth is *independent* of the global slope, then it is said to belong to the Edwards-Wilkinson (EW) universality class [34], which is much simpler than the KPZ class (in particular, its asymptotics can be described via Gaussian processes). This is the universality class to which random deposition with relaxation belongs.

- KPZ equation, or stochastic heat equation, or continuous Brownian polymer [1, 11, 25, 32, 63];
- $q$ -TASEPs [9, 10, 28, 31, 36];
- Semi-discrete Brownian polymer [9, 11, 55];
- Fully discrete log-Gamma polymer [12, 30].

As one can see, we have a fairly good control over integrable representatives of the  $(1+1)$ d KPZ universality class, and their list keeps growing. However, proving any statement about large time fluctuations for a *generic*  $(1+1)$ d KPZ model is still well beyond our reach.

## 6.

The  $(1+1)$ d KPZ fluctuation behavior that was obtained through integrable models has recently found remarkable experimental confirmations, see e.g. [67, 74].



Figure 6.1. Views of a stepped surface from three different angles.

## 7.

Let us now proceed to the case of two space dimensions. To be concrete, we will be interested in random stepped surfaces built from standard  $1 \times 1 \times 1$  cubes without holes and overhangs, as pictured on Figure 6.

Comparing to the interface interpretation of TASEP, see Figure 3.1, it is natural to consider the random growth process where standard cubes are being added at any allowed position independently with an exponential waiting time. Alas, apart from numerical simulations, nothing is known about large time behavior of such a process, even conjecturally, even on the law of large numbers (hydrodynamic) level, let alone the fluctuations.

On the other hand, uniform measures on stepped surfaces subjected to certain polygonal boundary conditions and constrained by the volume underneath are well known to form, in the limit of infinite volume, beautiful deterministic limit shapes given by explicit algebraic curves, cf. Figure 7.1, [24, 26, 51, 56, 57] and references therein.

It is therefore natural to ask whether it would be possible to find examples of integrable random growth models in  $(2+1)$  dimensions and, in particular, can those be used to grow such algebraic limit shapes.

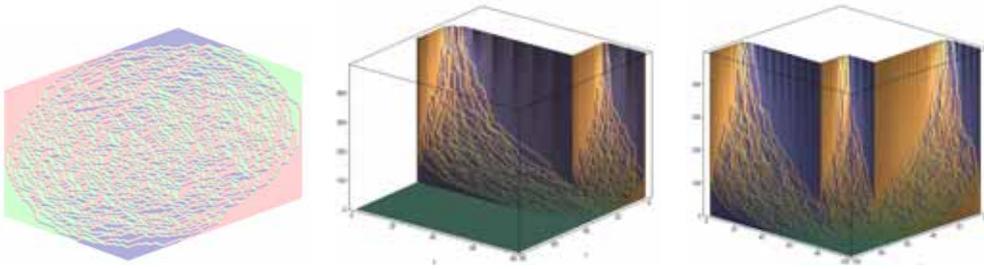


Figure 7.1. Algebraic limit shapes.

8.

Let us present an example of an integrable growth model from [15]. It is not quite as simple as randomly placing standard cubes into all possible positions randomly, but it is not too complicated either.

As for the TASEP, there are two descriptions of the model, as a growing interface and as a particle system. We start with the latter.

The state space of the particle system is a triangular array of interlacing variables

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

As initial condition, we consider a densely packed one: At time  $t = 0$  we have  $x_k^m(0) = k - m$  for all  $k, m$ , see Figure 8.1, left panel. The particle locations are pictured using the axes that are at  $\pi/3$  angle, the horizontal axes measures the particle positions  $x_k^*$ , and the (sloped) vertical axis measures the upper index. The particle locations are marked with small filled circles, the right-most ones in each row representing  $x_m^m = 0, m = 1, 2, \dots$

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

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

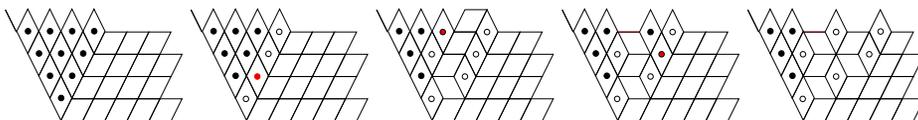


Figure 8.1. An integrable (2+1)-dimensional model of random growth.

Let us illustrate the dynamics using Figure 8.1, which shows possible few first particle jumps with four rows of particles ( $n = 4$  in the above notation). At times close to 0, only the right-most particle in each row can jump (that is, the other ones are blocked by lower/heavier

neighbors); the freedom to jump is denoted by empty circles on the second panel of Figure 8.1. The red circle denotes the free particle (it's  $x_2^2$ ) whose clock rang first. If we simply moved this particle to the right we would have violated the interlacing condition; instead, we additionally move higher/lighter particles in a minimal way that preserves the interlacing. This results in  $x_3^3$  and  $x_4^4$  also moving the right by one, and we end up with the particle configuration pictures on the third panel. Now more particles are free (empty circles), and we again assume that it is the red one ( $x_3^4$ ) that goes first. This move does not violate interlacing, and we land at the fourth panel. There  $x_3^3$  jumps and pushes  $x_4^4$ , etc.

Observe that  $\mathcal{S}^{(n_1)} \subset \mathcal{S}^{(n_2)}$  for  $n_1 \leq n_2$ , and the definition of the evolution implies that the process on  $n_1$  rows is a marginal of that on  $n_2$  rows. Thus, we can think of the stochastic evolution on the space of infinite point configurations  $\{x_k^m\}_{k=1, \dots, m, m \geq 1}$ .

The images of Figure 8.1 offer two three-dimensional interpretations — that of unit boxes being added and the other one with unit boxes being removed. Focusing on the first one, it is easily seen that the evolution of our particle system is equivalent to the following random growth recipe for the stepped surface: Each directed column of the form pictured in Figure 8.2 that can be added to the surface without creating holes or overhangs, is being added independently of the others with exponential waiting time of rate 1. For this reason we call this model the *directed column deposition model*.

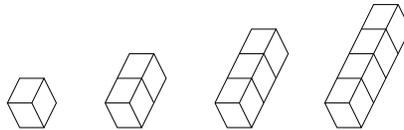


Figure 8.2. Directed sticks that are being added to the stepped surface

## 9.

Let us note a few properties of the interacting particle systems that we just described.

- The set of left-most particles  $\{x_1^m\}_{m \geq 1}$  evolves independently of the rest of the system, and its evolution is nothing but TASEP, with the initial condition  $x_1^m(0) = -m + 1$  that is commonly referred to as the *step initial condition*.
- The set of right-most particles  $\{x_m^m\}_{m \geq 1}$  also evolves independently and forms “Push-TASEP” or “long range TASEP”: If one views  $\{x_m^m + m\}_{m \geq 1}$  as particle locations in  $\mathbb{Z}$ , then when the  $x_k^k$ -clock rings, the particle  $x_k^k + k$  jumps to its right and pushes by one unit the (possibly empty) block of particles sitting next to it. If one disregards the particle labeling, one can think of particles as independently jumping to the next free site on their right with unit rate.
- For our densely packed initial condition, the evolution of each row  $\{x_k^m\}_{k=1, \dots, m}$ ,  $m = 1, 2, \dots$ , is also a Markov chain. It can be defined as Doob’s  $h$ -transform for  $m$  independent rate one Poisson processes with the harmonic function  $h$  equal to the Vandermonde determinant. In diffusive (large time) limit, it yields the Dyson Brownian Motion on spectra of the Gaussian Unitary Ensemble of random matrices, and we thus see an immediate connection between TASEPs and random matrices.

## 10.

The directed column deposition model is a representative of a fairly broad class of integrable growth models in (2+1) dimensions, cf. [8, 15, 19, 21]. In particular, such models can be used to grow the algebraic limit shapes mentioned above. As an example, Figure 10.1 shows one variant of how this may happen via growing the support of the random interface.

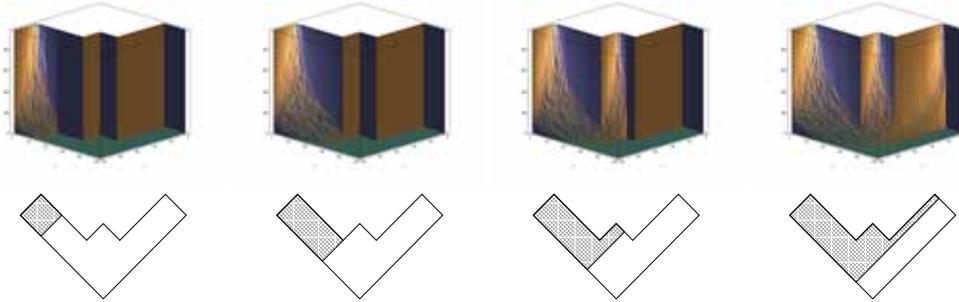


Figure 10.1. Growth of a random stepped surface by its support.

## 11.

The integrability of the models allows one to obtain very fine asymptotic results as the time of the growth tends to infinity, see Figure 11.1 a picture of the directed column deposition model at a large time. Let us describe the types of results that are achievable; each of them has been verified in at least one model, and they are expected to hold very broadly (we refer to Appendix B in the journal version of [15] for a brief survey and references).

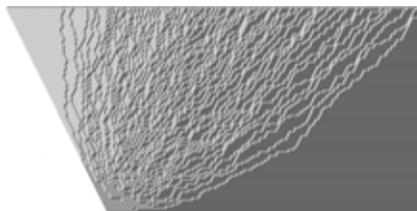


Figure 11.1. Simulation of the directed column deposition model.

- In hydrodynamic limit regime, when space and time are scaled in the same way, a deterministic limit shape forms, and it evolves according to a first order PDE of the form  $h_t(x, t) = f(x, \nabla_x h)$ , where the interface is represented as the plot of a function  $h : \mathbb{R}^2 \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ , and the function  $f$  in the right-hand side depends on the model (compare to the inviscid Burgers equation for TASEP).
- The (random) boundaries of the disordered regions asymptotically behave as interfaces from the (1+1)d KPZ universality class (which should not be surprising as we already

mentioned that the left and right boundaries of the disordered region on Figure 11.1 are indeed TASEP and PushTASEP interfaces).

- The models belong to the (2+1) dimensional Anisotropic KPZ universality class. In a little known dichotomy, the KPZ class in (2+1) dimensions splits into two subclasses, isotropic and anisotropic ones. The split is related to the (mathematically ill posed) KPZ equation  $h_t(x, t) = \Delta_x h + Q(\nabla_x h) + \{space-time\ white\ noise\}$ , where  $Q$  is a quadratic form, and this quadratic form may have either signature  $(1, 1)$  or  $(-1, -1)$  in the isotropic case, or signature  $(\pm 1, \mp 1)$  in the anisotropic one. Many natural models of random growth in (2+1) dimensions are isotropic (like the one with random placement of single unit cubes), and thus our results cannot be extended to them via the universality principle.
- The one-point fluctuations around the limit shape in the bulk of the random surface are Gaussian, with variance growing as  $\log t$  for large time  $t$ . This was predicted on the basis of (formal) one-loop expansion in renormalization group analysis of the Anisotropic KPZ equation in [73]. This claim was first verified with the appearance of the integrable models.
- The multi-point fluctuations around the limit shape are described by a remarkable object known as the *Gaussian (massless) Free Field* (GFF, for short). Let us explain this claim in more detail.

12.

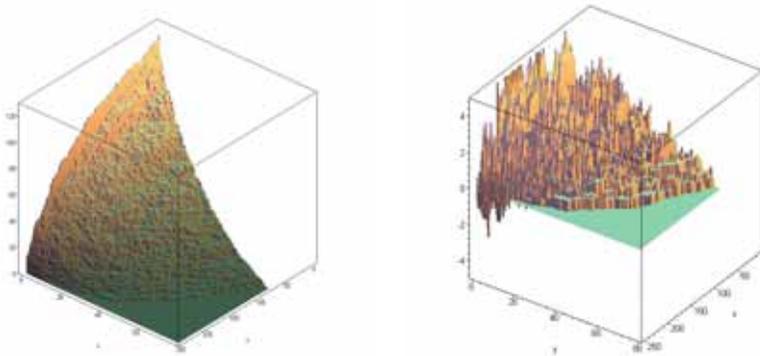


Figure 12.1. Directed column deposition model and its fluctuations.

It is instructive to look on Figure 12.1 whose right panel depicts fluctuations of the interface around the limit shape in the directed column deposition model (the random stepped surface is pictured on the left panel). The spikes have average size of  $\sqrt{\log(t)}$ , and the overall plot of the fluctuations seems too rough to represent any meaningful function. This is indeed so, and the GFF that describes the fluctuations in the large time limit is a *generalized random function*.

More exactly, there exists a bijective map

$$\Omega : \{\text{limit shape}\} \rightarrow \mathbb{H} = \{z \in \mathbb{C} : \Im z > 0\},$$

where by “limit shape” above we mean the smooth surface that approximates the disordered region of the interface, such that the fluctuations near a given point of the limit shape are given by

$$\text{GFF}(\Omega) = \sum_k \xi_k \frac{\phi_k(\Omega)}{\sqrt{\lambda_k}}, \quad \Omega \in \mathbb{H},$$

where  $\phi_k$ 's are the eigenfunctions of  $-\Delta$  on  $\mathbb{H}$  with Dirichlet (zero) boundary conditions,  $\lambda_k$ 's are the corresponding eigenvalues, and  $\{\xi_k\}$  is a collection of standard i. i. d. Gaussians.

For each  $\Omega \in \mathbb{H}$ , the above series is almost surely divergent, but if one pairs the right-hand side with a test function, the resulting series converges and defines a Gaussian random variable. A mathematical introduction to the GFF can be found in [64], [33, Section 4], [39, Section 2] and references therein.

### 13.

The GFF can be (formally) verified to constitute a time-stationary solution to the (formal) (2+1)d Anisotropic KPZ equation [65], thus one could anticipate the appearance of the GFF in the models of the corresponding universality class. However, this appearance is rather nontrivial: The bijective map  $\Omega$  is crucial for the GFF to show up, and yet its presence is in no way captured by the Anisotropic KPZ equation.<sup>3</sup>

The analysis of integrable models thus predicts that for any random growth model in (2+1)d Anisotropic KPZ class, *under a suitable identification of the limit shape with a region in the complex plane*, the fluctuations will be described by the GFF. Such a prediction would not have been possible without integrable examples.

### 14.

We have so far used mostly probabilistic language to describe our problems and results. However, the key feature of the subject of integrable probability is that the analysis is largely performed by algebraic methods.

The hierarchy of integrable models that includes most of those mentioned above, shadows that of multivariate special functions that originate from representation theory and integrable systems, as characters/zonal spherical functions for Lie groups/symmetric spaces over real/complex/finite/p-adic fields, and as eigenfunctions for integrable quantum many body systems.

*Representation theoretic tools are essential in our approach.* One example that is especially important is Casimir operators of Lie theory and their generalizations.

Figure 14.1 represents the hierarchy, with the top box corresponding to the so-called *Macdonald processes* [9] that are defined in terms of celebrated (multivariate symmetric) Macdonald polynomials. The Macdonald polynomials form a two-parameter deformation of the Schur polynomials that are well-known as irreducible characters of  $U(N)$  and  $GL(N, \mathbb{C})$ .

It is easy to deform (= add parameters to) a random growth model viewed as a proba-

---

<sup>3</sup>The map  $\Omega$  is a close relative of a similar map discovered and broadly utilized in the context of dimer models [49–51].

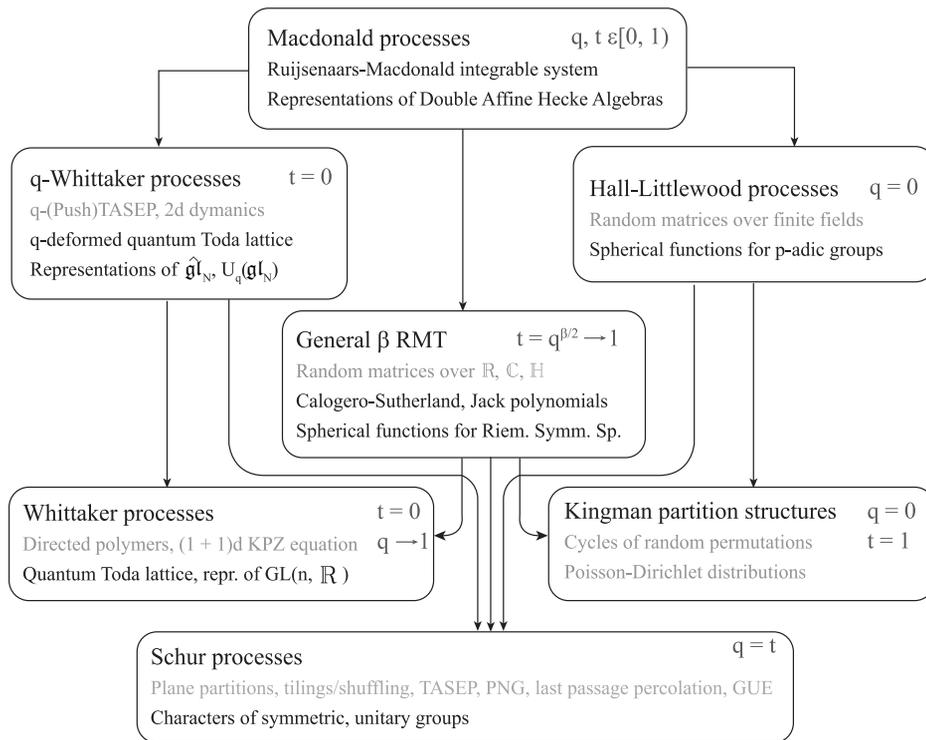


Figure 14.1. Macdonald processes and their degenerations.

bilistic object. However, most such deformations destroy integrability. The reason is that algebraic structures (in contrast to probabilistic ones) are usually very rigid. Thus, finding meaningful deformations of algebraic structures is very nontrivial.

Historically, first two different one-parameter deformations of the Schur polynomials were suggested: around 1960 by algebraists Ph. Hall and D. E. Littlewood,<sup>4</sup> and around 1970 by a statistician H. Jack. The Hall-Littlewood polynomials naturally arose in finite group theory and were later shown to be indispensable in representation theory of  $GL(n)$  over finite and  $p$ -adic fields. The Jack polynomials extrapolated the so-called zonal spherical functions arising in harmonic analysis on Riemannian symmetric spaces from three distinguished parameter values that correspond to spaces over real and complex numbers and quaternions. They are also known as eigenfunctions of the trigonometric Calogero-Sutherland integrable system.

In mid-1980's, in a remarkable development I. Macdonald [54] united the two deformations into a two-parameter deformation known as Macdonald polynomials. The two parameters are traditionally denoted as  $q$  and  $t$ . The Hall-Littlewood polynomials arise when  $q = 0$ , and the Jack polynomials correspond to the limit regime  $t = q^\theta \rightarrow 1$ , where  $\theta > 0$ . Schur polynomials correspond to  $q = t$ . Other significant limits are Schur's  $Q$ -functions (for  $q = 0, t = -1$ ); monomial symmetric functions ( $q = 0, t = 1$ ); and  $q$ -Whittaker functions arising for  $t = 0$ .

Each box of the degeneration scheme of Figure 14.1 corresponds to certain values of  $q$

<sup>4</sup>This is not the most famous mathematician with this last name, that would be J. E. Littlewood.

and  $t$ . There is a natural way to associate probability distributions to each of the boxes as well, and many of those are interesting in their own right (some of them are listed inside the scheme). Rich structural (representation theoretic!) properties of the special functions involved deliver tools that are essentially sufficient for analyzing various asymptotics of the arising probabilistic objects. This is the backbone of all the models discussed above.

It is worth noting that the bridge between representation theory and probability benefits the former as well. The corresponding part of representation theory often referred to as *asymptotic representation theory* was founded by Vershik and Kerov in late 1970's, see e.g. [22, 47, 48, 71, 72] and references therein. Its latest developments lead to new constructions of infinite-dimensional Markov processes, see [58] and references therein.

Unfortunately, a further discussion of the integrable probability and its representation theoretic origins is beyond the goals of the present talk, and we refer an interested reader to detailed expositions of different parts of this exciting and rapidly developing domain in [9, 15, 20, 23, 29].

## References

- [1] G. Amir, I. Corwin, and J. Quastel. *Probability distribution of the free energy of the continuum directed random polymer in  $1 + 1$  dimensions*. *Comm. Pure Appl. Math.* **64** (2011), 466–537, arXiv:1003.0443.
- [2] 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, arXiv:math/9810105.
- [3] J. Baik and E. M. Rains, *Symmetrized random permutations*. In: *Random matrix models and their applications*, 1–19, *Math. Sci. Res. Inst. Publ.* **40**, (2001), Cambridge Univ. Press, Cambridge, arXiv:math/9910019.
- [4] ———, *Limiting distributions for a polynuclear growth model with external sources*. *J. Stat. Phys.* **100** (2000), no. 3–4, 523–541, arXiv:math/0003130.
- [5] A.-L. Barabasi and H. E. Stanley, *Fractal Concepts in Surface Growth*, Cambridge University Press, 1995.
- [6] P. Billingsley, *Probability and Measure*, 3rd Edition, Wiley Series in Probability and Mathematical Statistics, 1995.
- [7] A. Borodin, *Determinantal point processes*. In: *Oxford Handbook of Random Matrix Theory*, G. Akemann, J. Baik, P. Di Francesco (editors), Oxford University Press, 2011, arXiv:0911.1153.
- [8] ———, *Schur dynamics of the Schur processes*. *Adv. Math.* **228** (2011), 2268–2291, arXiv:1001.3442.
- [9] A. Borodin and I. Corwin, *Macdonald Processes*, *Prob. Theory and Related Fields* **158** (2014), no. 1-2, 225–400, arXiv:1111.4408.
- [10] ———, *Discrete time  $q$ -TASEPs*. *Intern. Math. Research Notices* **05** (2013), doi:10.1093/imrn/rnt206, arXiv:1305.2972.
- [11] A. Borodin, I. Corwin, and P. Ferrari, *Free energy fluctuations for directed poly-*

- mers in random media in 1+1 dimension.* To appear in *Comm. Pure Appl. Math.*, arXiv:1204.1024.
- [12] A. Borodin, I. Corwin, and D. Remenik, *Log-Gamma polymer free energy fluctuations via a Fredholm determinant identity.* *Comm. Math. Phys.* **324** (2013), no. 1, 215–232, arXiv:1206.4573.
- [13] A. Borodin, I. Corwin, and T. Sasamoto, *From duality to determinants for  $q$ -TASEP and ASEP,* To appear in *Ann. Prob.*, arXiv:1207.5035.
- [14] A. Borodin and P. Ferrari, *Large time asymptotics of growth models on space-like paths I: PushASEP,* *Electr. J. Prob.* **13** (2008), 1380–1418, arXiv:0707.2813.
- [15] A. Borodin and P. Ferrari, *Anisotropic growth of random surfaces in  $2 + 1$  dimensions.* *Comm. Math. Phys.* **325** (2014), no. 2, 603–684, arXiv:0804.3035.
- [16] A. Borodin, P. L. Ferrari, and M. Prähofer, *Fluctuations in the Discrete TASEP with Periodic Initial Configurations and the Airy<sub>1</sub> Process.* *Int. Math. Res. Papers* (2007) 2007, doi: 10.1093/imrp/rpm002, arXiv:math-ph/0611071.
- [17] A. Borodin, P. L. Ferrari, M. Prähofer, and T. Sasamoto, *Fluctuation properties of the TASEP with periodic initial configuration.* *Jour. Stat. Phys.* **129** (2007), no. 5-6, 1055–1080, arXiv:math-ph/0608056.
- [18] A. Borodin, P. L. Ferrari, and T. Sasamoto, *Large Time Asymptotics of Growth Models on Space-like Paths II: PNG and Parallel TASEP.* *Comm. Math. Phys.* **283** (2008), no. 2, 417–449, arXiv:0707.4207.
- [19] A. Borodin and V. Gorin, *Shuffling algorithm for boxed plane partitions.* *Adv. Math.* **220** (2009), no. 6, 1739–1770, arXiv:0804.3071.
- [20] A. Borodin and V. Gorin, *Lectures on integrable probability.* Preprint, ArXiv:1212.3351.
- [21] A. Borodin, V. Gorin, and E. Rains,  *$q$ -Distributions on boxed plane partitions.* *Selecta Mathematica, New Series*, **16** (2010), no. 4, 731–789, arXiv:0905.0679.
- [22] A. Borodin and G. Olshanski, *Harmonic analysis on the infinite-dimensional unitary group and determinantal point processes.* *Annals of Mathematics* vol. **161** (2005), no. 3, 1319–1422, arXiv:math/0109194.
- [23] A. Borodin and L. Petrov, *Integrable probability: From representation theory to Macdonald processes,* *Probab. Surveys* **11** (2014), 1–58, arXiv:1310.8007.
- [24] C. Boutillier, S. Mkrtchyan and N. Reshetikhin, P. Tingley, *Random skew plane partitions with a piecewise periodic back wall,* *Annales Henri Poincaré* **13** (2012), no. 2, 271–296, arXiv:0912.3968.
- [25] P. Calabrese, P. Le Doussal and A. Rosso, *Free-energy distribution of the directed polymer at high temperature.* *Euro. Phys. Lett.* **90** (2010), 20002, arXiv:1002.4560.
- [26] H. Cohn, M. Larsen and J. Propp, *The Shape of a Typical Boxed Plane Partition.* *New York J. Math.* **4** (1998), 137–166, arXiv:math/9801059.
- [27] I. Corwin, *The Kardar-Parisi-Zhang equation and universality class.* *Random Matrices: Theory and Applications* **1** (2012), no. 1, arXiv:1106.1596.
- [28] ———, *The  $(q, \mu, \nu)$ -Boson process and  $(q, \mu, \nu)$ -TASEP.* Preprint, 2014, ArXiv:1401.3321.
- [29] ———, *Macdonald processes, quantum integrable systems and the KPZ class,* Pro-

- ceedings of the ICM-2014.
- [30] I. Corwin, N. O’Connell, T. Seppäläinen, and N. Zygouras, *Tropical Combinatorics and Whittaker functions*, To appear in Duke Math. Jour., arXiv:1110.3489.
  - [31] I. Corwin and L. Petrov, *The  $q$ -PushASEP: A New Integrable Model for Traffic in  $1+1$  Dimension*, Preprint, 2013, arXiv:1308.3124.
  - [32] V. Dotsenko, *Bethe ansatz derivation of the Tracy-Widom distribution for one dimensional directed polymers*, Euro. Phys. Lett. **90** (2010), 20003, arXiv:1003.4899.
  - [33] J. Dubedat, *SLE and the free field: Partition functions and couplings*, J. Amer. Math. Soc. **22** (2009), no. 4, 995–1054, arXiv:0712.3018.
  - [34] S. Edwards and D. Wilkinson, *The surface statistics of a granular aggregate*, Proc. R. Soc. Lond. A **381** (1982), 17–31.
  - [35] P. L. Ferrari and H. Spohn, *Random growth models*. In: *The Oxford Handbook of Random Matrix Theory*, G. Akemann, J. Baik, P. Di Francesco (editors), Oxford University Press, 2011, arXiv:1003.0881.
  - [36] P. L. Ferrari and B. Veto, *Tracy-Widom asymptotics for  $q$ -TASEP*, Preprint, 2013, arXiv:1310.2515.
  - [37] D. Forster, D. Nelson, and M. Stephen, *Large-distance and long-time properties of a randomly stirred fluid*, Phys. Rev. A. **16** (1977), 732–749.
  - [38] N. Georgiou and R. Kumar, T. Seppäläinen, *TASEP with Discontinuous Jump Rates*. ALEA Lat. Am. J. Probab. Math. Stat. **7** (2010), 293–318, arXiv:1003.3218.
  - [39] X. Hu, J. Miller and Y. Peres, *Thick points of the Gaussian free field*. Ann. Prob. **38** (2010), no. 2, 896–926, arXiv:0902.3842.
  - [40] D. A. Huse and C. L. Henley, *Pinning and roughening of domain walls in Ising systems due to random impurities*. Phys. Rev. Lett. **54** (1985), 2708–2711.
  - [41] T. Imamura and T. Sasamoto, *Fluctuations of the one-dimensional polynuclear growth model with external sources*, Nuclear Phys. B **699** (2004), no. 3, 503–544, arXiv:math-ph/0406001.
  - [42] K. Johansson, *Shape Fluctuations and Random Matrices*, Comm. Math. Phys. **209** (2000), 437–476, arXiv:math/9903134.
  - [43] ———, *Discrete orthogonal polynomial ensembles and the Plancherel measure*, Ann. of Math. (2) **153** (2001), no. 2, 259–296, arXiv:math/9906120.
  - [44] ———, *Discrete polynuclear growth and determinantal processes*, Comm. Math. Phys. **242** (2003), no. 1-2, 277–329, arXiv:math/0206208.
  - [45] O. Kallenberg, *Foundations of Modern Probability*, Second Edition, Springer, 2002.
  - [46] K. Kardar, G. Parisi and Y. Z. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett. **56** (1986), 889–892.
  - [47] S. Kerov, *Asymptotic Representation Theory of the Symmetric Group and its Applications in Analysis*, Amer. Math. Soc., Providence, RI, 2003.
  - [48] S. Kerov, G. Olshanski and A. Vershik, *Harmonic analysis on the infinite symmetric group*, Inv. Math. **158** (2004), no. 3, 551–642, arXiv:math/0312270.
  - [49] R. Kenyon, *Height fluctuations in the honeycomb dimer model*, Comm. Math. Phys. **281** (2008), no. 3, 675–709, arXiv:math-ph/0405052.

- [50] R. Kenyon, *Lectures on dimers. IAS/Park City Mathematical Series*, vol. 16: Statistical Mechanics, AMS, 2009, arXiv:0910.3129.
- [51] R. Kenyon and A. Okounkov, *Limit shapes and the complex Burgers equation*, Acta Math. **199** (2007), no. 2, 263–302, arXiv:math-ph/0507007.
- [52] T. Liggett, *Interacting Particle Systems*, Springer-Verlag, New York, 1985.
- [53] ———, *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*, Grundlehren der mathematischen Wissenschaften, volume **324**, Springer, 1999.
- [54] I. G. Macdonald, *A new class of symmetric functions*, Publ. I.R.M.A., Strasbourg, Actes 20-e Seminaire Lotharingen, (1988), 131–171.
- [55] N. O’Connell, *Directed polymers and the quantum Toda lattice*, Ann. Prob. **40** (2012), 437–458, arXiv:0910.0069.
- [56] A. Okounkov and N. Reshetikhin, *Correlation functions of Schur process with application to local geometry of a random 3-dimensional Young diagram*, J. Amer. Math. Soc. **16** (2003), 581–603, arXiv:math.CO/0107056.
- [57] A. Okounkov and N. Reshetikhin, *Random skew plane partitions and the Pearcey process*, Comm. Math. Phys. **269** (2007), no. 3 (2007), 571–609, arXiv:math.CO/0503508.
- [58] G. Olshanski, *The Gelfand-Tsetlin graph and Markov processes*, Proceedings of the ICM-2014.
- [59] M. Prähofer and H. Spohn, *Scale Invariance of the PNG Droplet and the Airy Process*, J. Stat. Phys. **108** (2002), no. 5–6, 1071–1106, arXiv:math/0105240.
- [60] J. Ramirez, B. Rider, B. Virag and Beta ensembles, *stochastic Airy spectrum, and a diffusion*, J. Amer. Math. Soc. **24** (2011), no. 4, 919–944, arXiv:math/0607331.
- [61] T. Sasamoto, *Spatial correlations of the 1D KPZ surface on a flat substrate*, J. Phys. A **38** (2005), 549–556, arXiv:cond-mat/0504417.
- [62] T. Sasamoto and T. Imamura, *Fluctuations of the one-dimensional polynuclear growth model in half-space*, J. Statist. Phys. **115** (2004), no. 3–4, 749–803, arXiv:cond-mat/0307011.
- [63] T. Sasamoto, H. Spohn, *One-dimensional KPZ equation: an exact solution and its universality*, Phys. Rev. Lett. **104** (2010), 230602, arXiv:1002.1883.
- [64] S. Sheffield, *Gaussian free fields for mathematicians*, Prob. Theory Related Fields **139** (2007), no. 3-4, 521–541, arXiv:math/0312099.
- [65] R. A. da Silveira and M. Kardar, *Nonlinear stochastic equations with calculable steady states*, Phys. Rev. E **68** (2003), no. 4, 046108, arXiv:cond-mat/0302003.
- [66] F. Spitzer, *Interaction of Markov processes*, Adv. Math. **5** (1970), no. 2, 246–290.
- [67] K. A. Takeuchi and M. Sano, *Universal Fluctuations of Growing Interfaces: Evidence in Turbulent Liquid Crystals*, Phys. Rev. Lett. **104** (2010), no. 23, 230601, arXiv:1001.5121.
- [68] C. A. Tracy and H. Widom, *Level-spacing distributions and the Airy kernel*, Comm. Math. Phys. **159** (1994), no. 1, 151–174, arXiv:hep-th/9210074.
- [69] ———, *On orthogonal and symplectic matrix ensembles*, Comm. Math. Phys. **177** (1996), 727–754, arXiv:solv-int/9509007.
- [70] ———, *Formulas and Asymptotics for the Asymmetric Simple Exclusion Pro-*

- cess, *Mathematical Physics, Analysis and Geometry* **14** (2011), no. 3, 211–235, arXiv:1101.2682.
- [71] A. M. Vershik and S. V. Kerov, *Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables*. *Soviet Math. Dokl.* **18** (1977), 527–531.
- [72] ———, *Asymptotic theory of characters of the symmetric group*, *Funct. Anal. Appl.* **15**, no. 4 (1981), 246–255.
- [73] D. E. Wolf, *Kinetic roughening of vicinal surfaces*, *Phys. Rev. Lett.* **67** (1991), 1783–1786.
- [74] P. J. Yunker, M. A. Lohr, T. Still, A. Borodin, D. J. Durian, and A. G. Yodh, *Effects of Particle Shape on Growth Dynamics at Edges of Evaporating Drops of Colloidal Suspensions*, *Phys. Rev. Lett.* **110** (2013), 035501, arXiv:1209.4137.

Department of Mathematics, Massachusetts Institute of Technology, USA; Institute for Information Transmission Problems of Russian Academy of Sciences, Russia

E-mail: borodin@math.mit.edu

# The Great Beauty of VEMs

Franco Brezzi

**Abstract.** In this paper I review the main features of the (newborn) Virtual Element Method, and of its application to the approximation of boundary value problems for Partial Differential Equations of particular relevance for applications. I will mostly concentrate on the definition of the Virtual Element spaces, that, roughly, consist of (vector valued) functions that are solution of (systems of) partial differential equations in each subdomain of a decomposition of the computational domain into polygons or polyhedra of quite general shape. Then I will give some hint on the use of these spaces for the discretization of some classical toy-problems like Heat conduction, Darcy flows, and Magnetostatic problems.

**Mathematics Subject Classification (2010).** Primary 65Nxx; Secondary 65N30.

**Keywords.** Virtual element methods, polygonal decompositions, patch test.

## 1. Introduction

The aim of this paper is to give some hints on a (brand new) technique, recently introduced in Scientific Computing, with the name of *Virtual Element Methods*. It is one of the many possible applications of the so-called Galerkin Method to approximate the solution of boundary value problems for Partial Differential Equations in variational form.

To give an idea of the Galerkin method in one of the simplest possible examples, assume that one wants to compute the approximate solution of the PDE  $-\Delta u = f$  in a given (say, polygonal, for hyper-simplicity) domain  $\Omega$ , with the boundary conditions  $u = 0$  on  $\partial\Omega$ . The *variational form* of this problem consists in looking for a function  $u \in V$  such that

$$\int_{\Omega} \mathbf{grad} u \cdot \mathbf{grad} v dx = \int_{\Omega} f v dx \quad \forall v \in V \quad (1.1)$$

where the space  $V$  is chosen as  $H_0^1(\Omega)$ , that is, the space of square integrable (classes of Lebesgue measurable) functions with square integrable derivatives (in  $\Omega$ ) that vanish on  $\partial\Omega$ .

The Galerkin method consists in choosing a finite dimensional subspace  $V_h \subset V$  and looking for  $u_h \in V_h$  such that

$$\int_{\Omega} \mathbf{grad} u_h \cdot \mathbf{grad} v_h dx = \int_{\Omega} f v_h dx \quad \forall v_h \in V_h. \quad (1.2)$$

It is then (in this toy-case) an easy exercise to show that such a  $u_h$  exists and is unique in  $V_h$ , together with the estimate

$$\int_{\Omega} |\mathbf{grad}(u - u_h)|^2 dx \leq \inf_{v_h \in V_h} \int_{\Omega} |\mathbf{grad}(u - v_h)|^2 dx \quad (1.3)$$

that connects the *error*  $\|u - u_h\|$  with the best approximation that could be given of the solution  $u$  within the subspace  $V_h$ .

More generally, the *mathematical analysis* of this type of procedures assumes that we are given a *sequence* of subspaces  $\{V_h\}_h$ , indexed by the parameter  $h$  (positive, and tending to zero). The target is to prove, under suitable assumptions on the sequence of decompositions, that the sequence of solutions  $\{u_h\}_h$  converges to the exact solution  $u$  when  $h$  tends to 0. As far as possible, one also tries to connect the *speed* of such a convergence in terms of suitable properties of the sequence  $\{V_h\}_h$ . See e.g. [39].

Many choices are available for the construction of such subspaces. One of the most common and most successful ones is that of Finite Elements: one decomposes the domain  $\Omega$  in small pieces and takes  $V_h$  as the space of functions that are piece-wise polynomials. The most classical case is that of decompositions in *triangles* (see two examples in Figure 1.1), in which one takes functions that are polynomials of degree  $\leq 1$  in each triangle. It is easy to see that each function of  $V_h$ , in this case, is characterized by its values at the vertices of the triangles, that will therefore become *the unknowns* of our approximate problems.

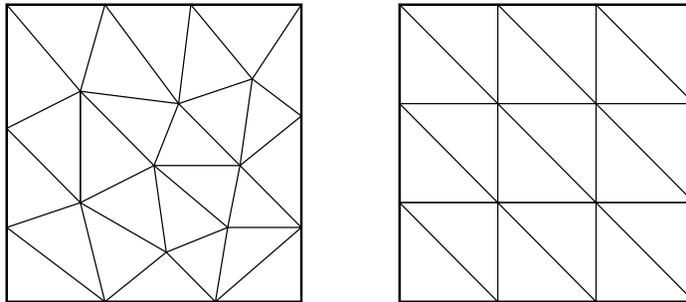


Figure 1.1. Triangulations of a rectangle: non-uniform or uniform

The most obvious generalization is obtained by taking, instead, polynomials of degree  $\leq 2$  in each triangle (and the unknowns will then be the values at the vertices and the values at the midpoint of each edge). And so on, using piecewise polynomials of degree  $\leq k$  with  $k = 1, 2, 3, \dots$  etc.

For the mathematical analysis one will then consider a sequence of decompositions  $\{\mathcal{T}_h\}_h$ , and, for a fixed  $k$ , connect the speed of convergence of  $u_h$  to  $u$  in terms of properties of the sequence. Typically, the parameter  $h$  will be connected to the biggest among the diameters of all the elements of the decomposition  $\mathcal{T}_h$ . Clearly, to let  $h \rightarrow 0$  will mean to consider finer and finer decompositions, and to measure *the speed of convergence* we look for estimates of the error  $\|u - u_h\|$  in terms of the *powers of  $h$*  (and of the degree  $k$ ). See again [39].

In three dimensions one uses, for instance, tetrahedra instead of triangles, and life is a bit more complicated. But already in two dimensions, as soon as we abandon the use of *triangles*, life becomes decidedly more complicated. Quadrilaterals (when we do not restrict ourselves to parallelograms) can already be a source of some practical (meaning: when we have to write the computer code!) headaches, and hexahedra are much worse. See for instance [6, 8, 22, 36, 44, 60, 63, 66], and the references therein.

Luckily, in the majority of applications the use of triangles/tetrahedra and or quadrilaterals/hexahedra is sufficient to give very effective practical methods.

There are however several types of problems where the use of much more general polygonal or polyhedral elements becomes highly desirable. The most relevant, so far, are the analysis of fractured materials and crack propagation (see, e.g. [1, 26, 40, 73, 74, 77, 78, 84], and the references therein), topology optimization (see, e.g. [3, 24, 25, 50, 56, 79, 87, 89, 98], and the references therein), computer graphics (see, e.g. [43, 45–47, 57, 61, 69, 72, 97]) and several other applications including fluid-structure interaction or two phase flows (see for instance [37, 38, 52, 64, 71], and the references therein). But their use for structured materials (see, e.g. [76, 79, 80, 83]) is also a promising direction, as well as for many other applications (see, e.g. [48, 76, 88, 90, 95] and the references therein).

The literature on these types of decompositions is quite wide, both from the Mathematical and the Engineering point of view. Here I just quote, in addition to the ones already mentioned: [5, 9–12, 23, 27, 29, 42, 53–55, 59, 62, 70, 75, 81, 82, 86, 91–94, 96], and the references therein.

In the last decade the use of Mimetic Finite Differences (a sort of finite differences, allowing very general decompositions, but not within the framework of Galerkin methods) underwent an impressive growth. I just mention, among the more recent papers, [4, 13, 17–20, 28, 30, 32–34, 41, 65].

The Virtual Element Methods (VEMs, in the title of the present paper) could be seen as an evolution of Mimetic Finite Differences, keeping their tremendous generality for the type of usable decompositions, but falling back into the simpler and more elegant realm of Galerkin approximations. See [2, 14–16, 21, 35, 49, 51, 68].

Here I want to describe, mostly for non-experts, the very basic features of the method, concentrating on a few very simple cases, and just giving hints and references to the more sophisticated (and practically much more interesting) developments of the last two years.

Here and there, I will do a certain amount of hand-waving, trying to trade precision for clarity. I apologize for that in advance. However, in these cases, I will always warn the readers, and address those that are interested in precise details to some papers already published or at least available on my web page.

An outline of the paper is as follows. In the next section, I will introduce some of the most commonly used functional spaces in the approximation of PDE's. In doing so, I will take, as toy-examples, some super-simplified problems in variational formulation (namely: Darcy flows, both in the primal and in the mixed formulation, and the magnetostatic problem). In the subsequent section I will try to give an idea on the classical Finite Element spaces used in the practice of Scientific Computing. Then, in Section 4 I will present the basic ideas on the construction of Virtual Element Spaces. Their main properties will be presented in the subsequent section, and their use in the approximation of PDE's will be briefly illustrated in Section 6. Some conclusions will be drawn in the final section, and a quite ample set of references will be in charge of (partly) heal the lack of details of the whole paper.

## 2. Typical model problems and functional spaces

In this section I will recall a few model problems of interest in applications, together with their variational formulations. To start with, I recall some of the most used functional spaces.

**2.1. The spaces most used in variational formulations.** Let  $\Omega$  be a Lipschitz continuous polyhedral domain. The following spaces are the common **bricks** used to deal with **PDEs**.

$$\begin{aligned}
L^2(\Omega) \text{ and } (L^2(\Omega))^3 &:= \text{square integrable (vector valued) functions on } \Omega. \\
H(\operatorname{div}; \Omega) &:= \{\boldsymbol{\tau} \in (L^2(\Omega))^3 \text{ s.t. } \operatorname{div} \boldsymbol{\tau} \in L^2(\Omega)\} \\
H(\operatorname{curl}; \Omega) &:= \{\boldsymbol{\varphi} \in (L^2(\Omega))^3 \text{ s.t. } \operatorname{curl} \boldsymbol{\varphi} \in (L^2(\Omega))^3\} \\
H(\operatorname{grad}; \Omega) &:= \{v \in L^2(\Omega) \text{ s.t. } \operatorname{grad} v \in (L^2(\Omega))^3\} \equiv H^1(\Omega)
\end{aligned}$$

**2.2. Primal formulation of Darcy problem.** We consider now the classical model problem of *Darcy flow* (fluid flow through a porous medium). We denote by  $p$  the pressure, by  $\mathbf{u}$  the velocities (actually, the *volumetric flow per unit area*), by  $f$  the source and by  $\mathbb{K}$  a material-depending tensor (representing the ratio between the permeability tensor and the viscosity coefficient). For the sake of simplicity, we also take the (totally) unrealistic choices:  $\mathbb{K} = \mathbb{I}$  (= identity) and  $p = 0$  at the boundary  $\partial\Omega$ . Taking also into account the physical laws:  $\mathbf{u} = -\mathbb{K}\nabla p = -\nabla p$  (Constitutive Equation), and  $\operatorname{div} \mathbf{u} = f$  (Conservation Equation) we end up with the model problem already considered in the introduction: *Find  $p \in H_0^1(\Omega)$  such that  $-\Delta p = f$  in  $\Omega$ .* As we already saw in the introduction, we can consider the variational formulation: *find  $p \in H_0^1(\Omega)$  such that:*

$$\int_{\Omega} \nabla p \cdot \nabla q \, dx = \int_{\Omega} f q \, dx \quad \forall q \in H_0^1(\Omega). \quad (2.1)$$

**2.3. Mixed formulation of Darcy problem.** There is however another variational formulation of the same problem, that in many practical cases is even more convenient than (2.1), and goes under the name of *mixed formulation*. It amounts to keep *both* unknowns  $\mathbf{u}$  and  $p$ , looking for  $p \in L^2(\Omega)$  and  $\mathbf{u} \in H(\operatorname{div}; \Omega)$  such that

$$\int_{\Omega} \mathbf{u} \cdot \mathbf{v} \, d\Omega = \int_{\Omega} p \operatorname{div} \mathbf{v} \, d\Omega \quad \forall \mathbf{v} \in H(\operatorname{div}; \Omega) \quad (2.2)$$

and

$$\int_{\Omega} \operatorname{div} \mathbf{u} q \, d\Omega = \int_{\Omega} f q \, d\Omega \quad \forall q \in L^2(\Omega), \quad (2.3)$$

where we see the spaces  $H(\operatorname{div}; \Omega)$  and  $L^2(\Omega)$  coming into the game (as spaces where we look for the solution, that therefore need to be discretized).

**2.4. Magnetostatic equations.** Another very simple model problem is given by the *magnetostatic equations*. Here, given a polyhedral domain  $\Omega$ , and given  $\mathbf{j} =$  (*divergence free*) current density vector and  $\mu =$  magnetic permeability constant, we consider the unknowns  $\mathbf{u} =$  vector potential with the gauge  $\operatorname{div} \mathbf{u} = 0$ ,  $\mathbf{H} = \mu^{-1} \operatorname{curl} \mathbf{u} =$  magnetic field, and  $\mathbf{B} =$  magnetic induction, together with the physical laws:  $\mathbf{B} = \mu \mathbf{H}$ ,  $\operatorname{curl} \mathbf{H} = \mathbf{j}$ , and  $\operatorname{div} \mathbf{B} = 0$  (that however has already been taken into account with the use of the vector potential  $\mathbf{u}$ , since  $\operatorname{div} \mathbf{B} = \operatorname{div} \mu \mathbf{H} = \operatorname{div} \operatorname{curl} \mathbf{u} = 0$ ). We supplement these equations with the (moderately realistic) boundary conditions  $\mathbf{u} \wedge \mathbf{n} = 0$  on  $\partial\Omega$ .

The classical magnetostatic equations can therefore be written now

$$\operatorname{curl} \mu^{-1} \operatorname{curl} \mathbf{u} = \mathbf{j} \quad \text{and} \quad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega \quad (2.4)$$

and we supplement them with the boundary conditions  $\mathbf{u} \wedge \mathbf{n} = 0$  on  $\partial\Omega$ . In order to reach a variational formulation of the problem, we define first

$$H_0(\operatorname{curl}; \Omega) := \{\boldsymbol{\varphi} \in H(\operatorname{curl}; \Omega) \text{ such that } \boldsymbol{\varphi} \wedge \mathbf{n} = 0 \text{ on } \partial\Omega\} \quad (2.5)$$

and we introduce a Lagrange multiplier  $p \in H_0^1(\Omega)$  to take into account the gauge  $\operatorname{div} \mathbf{u} = 0$ . Hence we can write the variational formulation as:

$$\left\{ \begin{array}{l} \text{Find } \mathbf{u} \in H_0(\mathbf{curl}; \Omega) \text{ and } p \in H_0^1(\Omega) \text{ such that :} \\ (\mu^{-1} \mathbf{curl} \mathbf{u}, \mathbf{curl} \mathbf{v}) - (\nabla p, \mathbf{v}) = (\mathbf{j}, \mathbf{v}) \quad \forall \mathbf{v} \in H_0(\mathbf{curl}; \Omega) \\ (\mathbf{u}, \nabla q) = 0 \quad \forall q \in H_0^1(\Omega), \end{array} \right. \quad (2.6)$$

showing an example of use for  $H(\mathbf{curl}; \Omega)$  and  $H_0^1(\Omega)$ .

**2.5. Continuity requirements for the basic spaces.** Before entering the details of the VEM approximations for these spaces, I will make a final consideration on the continuity requirements for each of them. Assume that we have, say, a piecewise smooth vector valued function  $\mathbf{v} : \Omega \rightarrow \mathbb{R}^3$ . Then, if you want to ensure that it belongs, globally, to  $(H^1(\Omega))^3$  you must require that *all the components* of  $\mathbf{v}$  are continuous at the inter-element boundaries. If instead you want to ensure that  $\mathbf{v}$  belongs, globally, to  $H(\mathbf{curl}; \Omega)$ , you must require that *its tangential components* are continuous at the inter-element boundaries, while for having  $\mathbf{v} \in H(\operatorname{div}; \Omega)$  you must require the continuity, at the inter-element boundaries, of *its normal component*. Finally, as natural, no continuity is required to ensure  $\mathbf{v} \in (L^3(\Omega))^3$ .

The knowledge of these continuity requirements is crucial in building approximations: roughly speaking, the quantities that are required to be continuous must be single-valued at the inter-element boundaries, and in practice one needs to prescribe them as degrees of freedom in the approximations.

### 3. Classical F.E. approximations

**3.1. Basic polynomial spaces.** To give the flavor of typical Finite Element approximations, let us see to simplest possible choices of polynomial spaces on a tetrahedron:

$$\begin{aligned} \mathbb{P}_0 &:= \{\text{constants}\} \quad (1 \text{ d.o.f.}) \\ RT_0 &:= \{\boldsymbol{\tau} = \mathbf{a} + c\mathbf{x}\} \text{ with } \mathbf{a} \in \mathbb{R}^3 \text{ and } c \in \mathbb{R} \quad (4 \text{ d.o.f.}) \\ N_0 &:= \{\varphi = \mathbf{a} + \mathbf{c} \wedge \mathbf{x}\} \text{ with } \mathbf{a} \in \mathbb{R}^3 \text{ and } \mathbf{c} \in \mathbb{R}^3 \quad (6 \text{ d.o.f.}) \\ \mathbb{P}_1 &:= \{v = a + \mathbf{c} \cdot \mathbf{x}\} \text{ with } a \in \mathbb{R} \text{ and } \mathbf{c} \in \mathbb{R}^3 \quad (4 \text{ d.o.f.}) \end{aligned}$$

A function in  $\mathbb{P}_1$  can obviously be individuated by its value at the four vertices of the tetrahedron, and a vector in  $(\mathbb{P}_1)^3$  will be individuated by the three values of its three components at each vertex. A vector valued function in  $N_0$  will be individuated by the (constant!) values of its tangential components along each of the six edges. Instead, a vector valued function in  $RT_0$  will be individuated by the values of its normal components on each of the four faces. It is an easy exercise to check that the normal component of an element of  $RT_0$ , on any plane, is always constant. Finally, a function in  $\mathbb{P}_0$  can obviously be individuated by its value, say, at the barycenter.

**3.2. Lowest order finite element spaces.** Let now  $\mathcal{T}_h$  be a decomposition of  $\Omega$  in tetrahedra. We consider the following finite element approximations.

$$L^2(\Omega) \sim \mathcal{L}_0^0 := \{q \in L^2(\Omega) \text{ such that } q|_T \in \mathbb{P}_0 \quad \forall T \in \mathcal{T}_h\},$$

$$\begin{aligned}
 H(\text{div}; \Omega) &\sim \mathcal{RT}_0 := \{\boldsymbol{\tau} \in H(\text{div}; \Omega) \text{ s.t. } \boldsymbol{\tau}|_T \in \mathcal{RT}_0 \quad \forall T \in \mathcal{T}_h\}, \\
 H(\text{curl}; \Omega) &\sim \mathcal{N}_0 := \{\boldsymbol{\varphi} \in H(\text{curl}; \Omega) \text{ s.t. } \boldsymbol{\varphi}|_T \in \mathcal{N}_0 \quad \forall T \in \mathcal{T}_h\}, \\
 H(\text{grad}; \Omega) &\sim \mathcal{L}_1^1 := \{v \in H(\text{grad}; \Omega) \text{ s.t. } v|_T \in \mathbb{P}_1 \quad \forall T \in \mathcal{T}_h\}.
 \end{aligned}$$

It is easy to see, from the previous discussion, that: i) a function in  $\mathcal{L}_0^0$  is individuated by its values at the barycenter of each tetrahedron of the decomposition, ii) a function in  $\mathcal{RT}_0$  is individuated by the values of its normal component at each face of the decomposition, iii) a function in  $\mathcal{N}_0$  is individuated by the values of its tangential component at each edge of the decomposition, and iv) a function in  $\mathcal{L}_1^1$  is individuated by its values at each vertex of the decomposition.

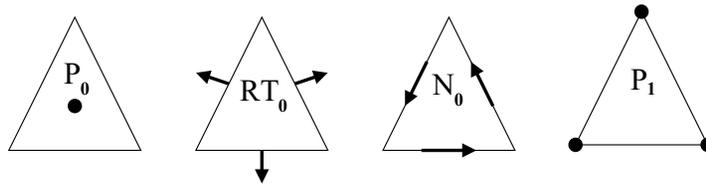


Figure 3.1. Degrees of freedom for the four polynomial spaces

All this is very elegant and, at the same time, very practical. This is not always the case. For instance, the most elegant available form for polynomial approximations (of degree  $k$ ) of  $H(\text{curl})$  in a cube like  $(-1, 1)^3$  is given by

$$\begin{aligned}
 \text{span} \left\{ &yz(w_2(x, z) - w_3(x, y)), \right. \\
 &zx(w_3(x, y) - w_1(y, z)), \\
 &\left. xy(w_1(y, z) - w_2(x, z)) \right\} \\
 &+ (\mathbb{P}_k)^3 + \text{grad } s(x, y, z)
 \end{aligned}$$

where each  $w_i$  ( $i = 1, 2, 3$ ) ranges over all polynomials (of 2 variables) of degree  $\leq k$  and  $s$  ranges over all polynomials of *superlinear degree*  $\leq k + 1$ , where the *superlinear degree* of a monomial is defined as “ordinary degree ignoring variables that appear linearly”, [7].

Clearly *nobody ever tried* to do something similar on a dodecahedron....

## 4. Virtual element spaces

**4.1. Polygonal and polyhedral elements.** There is a wide literature on Polygonal and Polyhedral Elements, with applications to several important fields in Engineering and Computer Sciences. See for instance [5, 27, 45, 58, 61, 67, 85, 93, 94], and the references therein.

In general, these methods present the members of the discrete subspace as the solutions of suitable problems within each element. These problems are then solved in an approximate way, to obtain their values at the nodes of a suitable numerical integration formula (that, in turn, is used in order to compute the integrals that appear in the variational formulation).

The Virtual Element Methods follow this path insofar as to use solutions of (systems) of PDE equations. However, they *do not* attempt an approximate solution of these equations

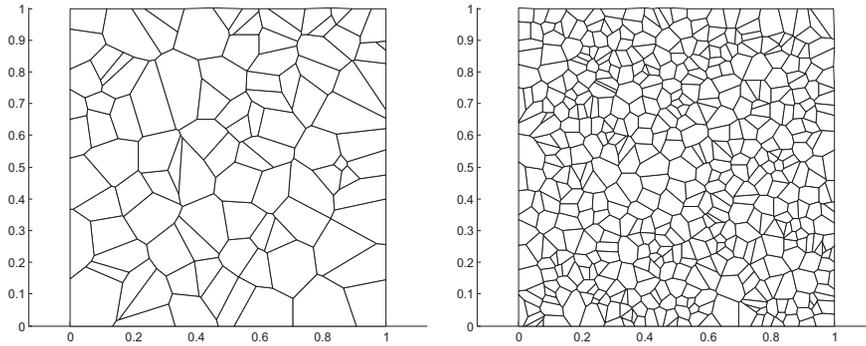


Figure 4.1. Voronoi tassellations: coarser and finer

(a most expensive step) and use instead suitable integrations by parts in order to compute the integrals (appearing in the variational formulations) *exactly*, at least when one of the two terms is a polynomial of a degree up to  $k$ , where  $k$  denotes the accuracy that has been chosen by the user (the higher is  $k$ , the most expensive is the computation). This ensures the full satisfaction of the so-called *patch-test of order  $k$* , that roughly requires that: if the solution of the original problem is, globally, a polynomial of degree  $\leq k$ , then the solution of the discretized problem coincides with the exact solution. A property that is considered as very important in the Engineering literature, and that is lost when using numerical integration.

Let us see how this can be done, on some toy problem.

Assume that we are given a sequence of decompositions  $\{\mathcal{T}_h\}_h$  of the computational domain  $\Omega$  into polygons or polyhedra. To fix the ideas, we just assume that the decomposition satisfies the following assumption

- **H0** For the 2-dimensional case, we assume that:  $H0_2$  - there exists a fixed real number  $\rho$  such that each polygon  $E$  is starshaped with respect to all the points of a ball of diameter  $\rho h_E$  and all its edges have a length  $\geq h_E$  (where  $h_E$  is the diameter of  $E$ ). In three dimensions, we assume that:  $H0_3$  - there exists a fixed real number  $\rho$  such that each polyhedron  $E$  is starshaped with respect to all the points of a ball of diameter  $\rho h_E$  and all its faces satisfy the two dimensional assumption  $H0_2$  with constant  $\rho$ .

Note that **H0** easily implies, among other things, that there exists an integer number  $N$ , depending only on  $\rho$ , such that the number of edges of each element is bounded by  $N$ .

**4.2. General features of VEM Spaces.** As for other methods, the trial and test functions inside each element are rather complicated (e.g. solutions of suitable PDE's or systems of PDE's).

However, contrary to other methods,

- i) they **do not** require the approximate evaluation of trial and test functions at the integration points.
- ii) In most cases they satisfy the *patch test exactly* (up to the computer accuracy).
- iii) We have a *whole family* of spaces (conforming and nonconforming approximations of all the main functional spaces)

**4.3. Approximations of  $H^1(\Omega) \equiv H(\text{grad}; \Omega)$ .** We consider first the two-dimensional case. For each element  $E$  that satisfies **H0**, and for each integer  $k \geq 1$  we consider the *local spaces*:

$$\mathcal{B}_{k,2}^{nodal}(\partial E) := \{g \mid g \in C^0(\partial E) \text{ and } g|_e \in \mathbb{P}_k(e) \text{ for all edge } e \in \partial E\}, \quad (4.1)$$

and

$$V_{k,2}^{nodal}(E) := \{v \mid v|_{\partial E} \in \mathcal{B}_{k,2}^{nodal}(\partial E) \text{ and } \Delta v \in \mathbb{P}_{k-2}(E)\}. \quad (4.2)$$

Then we define, in a very natural manner:

$$V_{k,2}^{nodal}(\Omega) := \{v \in H^1(\Omega) \mid v|_E \in V_{k,2}^{nodal}(E) \text{ for all } E \in \mathcal{T}_h\}. \quad (4.3)$$

We then consider the three dimensional case. For each element  $E$  that satisfies **H0**, and for each integer  $k \geq 1$  we consider first the *local spaces*:

$$\mathcal{B}_{k,3}^{nodal}(\partial E) = \{g \mid g \in C^0(\partial E) \text{ and } g|_f \in V_{k,2}^{nodal}(f) \text{ for all face } f \in \partial E\} \quad (4.4)$$

and

$$V_{k,3}^{nodal}(E) := \{v \mid v|_{\partial E} \in \mathcal{B}_{k,3}^{nodal}(\partial E) \text{ and } \Delta v \in \mathbb{P}_{k-2}(E)\}, \quad (4.5)$$

and then we define:

$$V_{k,3}^{nodal}(\Omega) := \{v \in H^1(\Omega) \mid v|_E \in V_{k,3}^{nodal}(E) \text{ for all } E \in \mathcal{T}_h\}. \quad (4.6)$$

We can now consider the global degrees of freedom (say, in three dimensions):

- The values of  $v$  at the vertices of  $\mathcal{T}_h$ ,
- $\int_e v q_{k-2} ds$  for all edge  $e \in \mathcal{T}_h$ ,  $\forall q_{k-2} \in \mathbb{P}_{k-2}(e)$ ,
- $\int_f v q_{k-2} df$  for all face  $f \in \mathcal{T}_h$ ,  $\forall q_{k-2} \in \mathbb{P}_{k-2}(f)$ ,
- $\int_E v q_{k-2} dE$  for all element  $E \in \mathcal{T}_h$ ,  $\forall q_{k-2} \in \mathbb{P}_{k-2}(E)$ ,

**4.4. Approximations of  $H(\text{div}; \Omega)$ .** In each element  $E$ , and for each integer  $k$ , we define

$$\mathcal{B}_{k,2}^{face}(\partial E) := \{g \mid g|_e \in \mathbb{P}_k \forall \text{ edge } e \in \partial E\} \text{ in 2d,}$$

$$\mathcal{B}_{k,3}^{face}(\partial E) := \{g \mid g|_f \in \mathbb{P}_k \forall \text{ face } f \in \partial E\} \text{ in 3d.}$$

The local spaces, in two dimensions, will then be

$$V_{k,2}^{face}(E) := \{\boldsymbol{\tau} \mid \boldsymbol{\tau} \cdot \mathbf{n} \in \mathcal{B}_{k,2}^{face}(\partial E), \text{ div } \boldsymbol{\tau} \in \mathbb{P}_{k-1}, \text{ rot } \boldsymbol{\tau} \in \mathbb{P}_{k-1}\},$$

and in 3 dimensions

$$V_{k,3}^{face}(E) := \{\boldsymbol{\tau} \mid \boldsymbol{\tau} \cdot \mathbf{n} \in \mathcal{B}_{k,3}^{face}(\partial E), \text{ div } \boldsymbol{\tau} \in \mathbb{P}_{k-1}, \mathbf{curl } \boldsymbol{\tau} \in (\mathbb{P}_{k-1})^3\}.$$

Finally, in all cases, the global spaces will be written as

$$V_{k,d}^{face}(\Omega) := \{\boldsymbol{\tau} \in H(\text{div}; \Omega) \mid \boldsymbol{\tau} \in V_{k,d}(E) \text{ for all } E \in \mathcal{T}_h\}. \quad (4.7)$$

Before describing the degrees of freedom, we define, on a generic domain  $\mathcal{O}$ , the space  $\mathcal{G}_k^\perp(\mathcal{O})$  as the subset of the  $\mathbf{g} \in (\mathbb{P}_k(\mathcal{O}))^3$  such that

$$\int_{\mathcal{O}} \mathbf{g} \cdot \mathbf{grad} q_{k+1} d\mathcal{O} = 0 \quad \forall q_{k+1} \in \mathbb{P}_{k+1}(\mathcal{O}).$$

Then we can choose the degrees of freedom in  $V_{k,d}^{face}(\Omega)$  as

- $\int_e \boldsymbol{\tau} \cdot \mathbf{n} q_k \, de \quad \forall q_k \in \mathbb{P}_k(e) \quad \forall \text{edge } e$
- $\int_E \boldsymbol{\tau} \cdot \mathbf{grad} q_{k-1} \, dE \quad \forall q_{k-1} \in \mathbb{P}_{k-1}(E) \quad \forall \text{element } E$
- $\int_E \boldsymbol{\tau} \cdot \mathbf{g}_k^\perp \, dE \quad \forall \mathbf{g}_k^\perp \in \mathcal{G}_k^\perp(E) \quad \forall \text{element } E$

in two dimensions, and

- $\int_f \boldsymbol{\tau} \cdot \mathbf{n} q_k \, df \quad \forall q_k \in \mathbb{P}_k(f) \quad \forall \text{face } f$
- $\int_E \boldsymbol{\tau} \cdot \mathbf{grad} q_{k-1} \, dE \quad \forall q_{k-1} \in \mathbb{P}_{k-1}(E) \quad \forall \text{element } E$
- $\int_E \boldsymbol{\tau} \cdot \mathbf{g}_k^\perp \, dE \quad \forall \mathbf{g}_k^\perp \in \mathcal{G}_k^\perp(E) \quad \forall \text{element } E$

in three dimensions.

**4.5. Approximations of  $H(\mathbf{curl}; \Omega)$ .** For the 2-dimensional case, we can think that  $H(\mathbf{curl}; \Omega)$  is obtained from  $H(\mathbf{div}; \Omega)$  by a simple rotation of  $\pi/2$ . With this, we can just think that also its discretization

$$V_{k,2}^{edge}(\Omega) \text{ is obtained by rotating } V_{k,2}^{face} \text{ of } \pi/2.$$

Namely, we can consider vector fields that on each edge have a tangential component in  $\mathbb{P}_k(e)$ , and whose divergence and rotation are in  $\mathbb{P}_{k-1}(e)$  for each element  $E$ . The corresponding degrees of freedom can also be easily obtained by rotating the corresponding ones for  $V_{k,2}^{face}(\Omega)$ .

We can therefore turn to the (more complex) discretizations of  $H(\mathbf{curl}; \Omega)$  in three dimensions.

In each element  $E$ , and for each integer  $k$ , we therefore set

$$\mathcal{B}_{k,3}^{edge}(\partial E) := \{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi}|_f \in V_{k,2}^{edge}(f) \forall \text{face } f \in \partial E \text{ and } \boldsymbol{\varphi} \cdot \mathbf{t}_e \text{ is single valued at each edge } e \in \partial E \}$$

where we denoted by  $\mathbf{t}_e$  the unit tangent vector to an edge  $e$ . Now we can set

$$V_{k,3}^{edge}(E) = \{ \boldsymbol{\varphi} \mid \boldsymbol{\varphi}|_t \in \mathcal{B}_{k,3}^{edge}(\partial E), \text{div} \boldsymbol{\varphi} \in \mathbb{P}_{k-1}, \mathbf{curl} \mathbf{curl} \boldsymbol{\varphi} \in (\mathbb{P}_{k-2})^3 \}$$

where  $\boldsymbol{\varphi}|_t$  is, on each face, the *tangential part* of  $\boldsymbol{\varphi}$ . We can therefore define the global space as:

$$V_{k,3}^{edge}(\Omega) := \{ \boldsymbol{\varphi} \in H(\mathbf{curl}; \Omega) \mid \boldsymbol{\varphi} \in V_{k,3}^{edge}(E) \text{ for all } E \in \mathcal{T}_h \}.$$

In  $V_{k,3}^{edge}(\Omega)$  we can take the following degrees of freedom:

- for every edge  $e$ :  $\int_e \boldsymbol{\varphi} \cdot \mathbf{t}_e q_k \, de \quad \forall q_k \in \mathbb{P}_k(e)$

- for every face  $f$ :

$$\int_f \boldsymbol{\varphi} \cdot \mathbf{rot} q_{k-1} df \quad \forall q_{k-1} \in \mathbb{P}_{k-1}(f)$$

$$\int_f \boldsymbol{\varphi} \cdot \mathbf{r}_{k,2}^\perp df \quad \forall \mathbf{r}_{k,2}^\perp \in \mathcal{R}_{k,2}^\perp(f)$$

where  $\mathcal{R}_{k,2}^\perp$  is the subset of the  $\mathbf{r} \in (\mathbb{P}_k(f))^3$  such that

$$\int_f \mathbf{r} \cdot \mathbf{rot} q_{k+1} df = 0 \quad \forall q_{k+1} \in \mathbb{P}_{k+1}(f)$$

- and for every element  $E$ :

$$\int_E \boldsymbol{\varphi} \cdot \mathbf{rot} q_{k-1} dE \quad \forall q_{k-1} \in (\mathbb{P}_{k-1}(E))^3$$

$$\int_E \boldsymbol{\varphi} \cdot \mathbf{r}_{k,3}^\perp dE \quad \forall \mathbf{r}_{k,3}^\perp \in \mathcal{R}_{k,3}^\perp(E)$$

where  $\mathcal{R}_{k,3}^\perp(E)$  is the subset of the  $\mathbf{r} \in (\mathbb{P}_k(E))^3$  such that

$$\int_E \mathbf{r} \cdot \mathbf{curl} \mathbf{q}_{k+1} dE = 0 \quad \forall \mathbf{q}_{k+1} \in (\mathbb{P}_{k+1}(E))^3$$

**4.6. Approximations of  $L^2(\Omega)$ .** The approximation of spaces as  $L^2(\Omega)$  or  $(L^2(\Omega))^d$  does not present any difficulties. As the space has no continuity requirements, we can just take piecewise polynomials discontinuous (vector valued) functions:

$$V_{k,d}^{volume}(\Omega) = \{q \mid q|_E \in \mathbb{P}_{k,d}(E) \text{ for all } E \in \mathcal{T}_h\}.$$

## 5. Useful properties

We observe that the classical differential operators  $grad$ ,  $curl$ , and  $div$  send these VEM spaces one into the other (up to the obvious adjustments for the polynomial degree). Indeed:

$$\mathbf{grad}(V_{k,d}^{nodal}) \subseteq V_{k-1,d}^{edge}; \quad \mathbf{curl}(V_{k,d}^{edge}) \subseteq V_{k-1,d}^{face}; \quad \mathbf{div}(V_{k,d}^{face}) \subseteq V_{k-1,d}^{volume}. \quad (5.1)$$

But possibly the most crucial feature common to all these choices is the possibility to construct (starting from the degrees of freedom, and without solving approximate problems in the element) an *approximate  $L^2$ -type scalar product*

$$[\mathbf{u}, \mathbf{v}]_h = \sum_{E \in \mathcal{T}_h} [\mathbf{u}, \mathbf{v}]_{h,E}, \quad (5.2)$$

with the following properties:

**P1**  $[\mathbf{p}_k, \mathbf{v}]_{h,E} = (\mathbf{p}_k, \mathbf{v})_{0,E} \quad \forall \mathbf{p}_k \in (\mathbb{P}_k(E))^d, \forall \mathbf{v}$  in the VEM space

(where  $(\mathbf{p}_k, \mathbf{v})_{0,E}$  represents the  $L^2(E)$  inner product, or the  $(L^2(E))^d$  inner product for vector valued functions), and

**P2**  $\exists \alpha^* \geq \alpha_* > 0$  independent of  $h$  such that

$$\alpha_* \|\mathbf{v}\|_{0,E}^2 \leq [\mathbf{v}, \mathbf{v}]_{h,E} \leq \alpha^* \|\mathbf{v}\|_{0,E}^2, \quad \forall \mathbf{v} \text{ in the VEM space,}$$

where obviously  $\|\mathbf{v}\|_{0,E}^2 := (\mathbf{v}, \mathbf{v})_{0,E}$ . In turn, properties **P1** and **P2** can be easily obtained, if we are able to compute the  $L^2$ -projections onto  $\mathbb{P}_k$  of the elements of the VEM spaces. Indeed, assume that for every  $v$  in the VEM space and for every polynomial  $p_k$  you can compute (up to computer precision) an element  $\Pi_k^0 v$  in  $\mathbb{P}_k$  such that

$$(v - \Pi_k^0 v, p_k)_{0,E} = 0 \quad \forall p_k \in \mathbb{P}_k \quad \forall v \text{ in the VEM space.} \quad (5.3)$$

Then you can set

$$[u, v]_{h,E} := (\Pi_k^0 u, \Pi_k^0 v) + S(u - \Pi_k^0 u, v - \Pi_k^0 v) \quad (5.4)$$

where  $S$  is “any” symmetric bilinear form that, roughly speaking, scales like the true  $L^2$  inner product (see [14], [35], or [16] for a precise definition, more details and examples).

Needless to say, these approximate  $L^2$ -type inner products depend on the type of Virtual Elements that we are dealing with. Hence, in what follows, we are going to use a different name for each of them. With obvious notation we will, therefore, have scalar products  $[u, v]_{VEM, nodal}$  and  $[u, v]_{VEM, volume}$  for scalar functions, together with  $[\mathbf{u}, \mathbf{v}]_{VEM, edge}$  and  $[\mathbf{u}, \mathbf{v}]_{VEM, face}$  for vector-valued functions.

## 6. VEM approximations of PDE's

Using the  $L^2$ -type projection operators, and, if needed, the properties (5.1) one can find an easy and systematic way to discretize PDE's by means of Virtual Element spaces. It should be pointed out, however, that on specific occasions alternative solutions could be more effective. Moreover, the discretization of the forcing terms requires some (minor) additional care that I do not discuss here. See for instance [14] or [31].

**6.1. VEM's for primal Darcy.** Remembering equation (2.1) we can now formulate the approximate problem as: *find*  $p_h \in V_{k,2}^{nodal}$  *such that*:

$$[\mathbf{grad} p_h, \mathbf{grad} q_h]_{VEM, edge} = [f, q_h]_{VEM, nodal}$$

for all  $q_h \in VEM_{k,2}^{nodal}$ .

**6.2. VEM's for mixed Darcy.** The approximate version of the mixed formulation (2.2)–(2.3) can now be written as: *find*  $p_h \in V_{k-1,d}^{volume}$  *and*  $\mathbf{u}_h \in V_{k,d}^{face}$  *such that*:

$$[\mathbf{u}_h, \mathbf{v}_h]_{VEM, face} = [p_h, \mathbf{div} \mathbf{v}_h]_{VEM, volume}$$

for all  $\mathbf{v}_h \in V_{k,d}^{face}$ , and

$$[\mathbf{div} \mathbf{u}_h, q_h]_{VEM, volume} = [f, q_h]_{VEM, volume}$$

for all  $q_h \in V_{k-1,d}^{volume}$ .

**6.3. VEM's for electromagnetic problems.** The VEM approximation of the magnetostatic problem (2.6), in turn, can be chosen as: *find*  $\mathbf{u}_h$  in  $V_{k,3}^{edge}$  and  $p_h$  in  $V_{k,3}^{nodal}$  such that:

$$\begin{aligned} [\mu^{-1} \mathbf{curl} \mathbf{u}_h, \mathbf{curl} \mathbf{v}_h]_{VEM,face} - [\nabla p_h, \mathbf{v}_h]_{VEM,edge} \\ = [\mathbf{j}, \mathbf{v}_h]_{VEM,edge} \quad \forall \mathbf{v}_h \in V_{k,3}^{edge} \end{aligned}$$

and

$$[\mathbf{u}, \nabla q_h]_{VEM,edge} = 0 \quad \forall q_h \in V_{k,3}^{nodal}.$$

**Remark 6.1.** To tell the truth, in order to set up the proof, one has to *think* that the Virtual Element space has been *tilted*, or, as we say (cfr. [2]), *enhanced*. This does not correspond to a change in the code, but it simplifies the proofs that, without it, would become more cumbersome. I decided not to enter these aspects, and to refer the interested readers to [2] and [16].

It has to be pointed out that these methods are extremely robust with respect to the choice of the geometry of the decomposition. To give the flavor of their capability, I report the results made on a *totally crazy* sequence of meshes going from  $4 \times 4$  to  $16 \times 16$  *winged horses*, clearly inspired by Escher. The results have been obtained with the primal and mixed formulation of Darcy problem, having  $p = \sin(2x) \cos(3y)$  as exact solution (courtesy of Alessandro Russo and Donatella Marini).

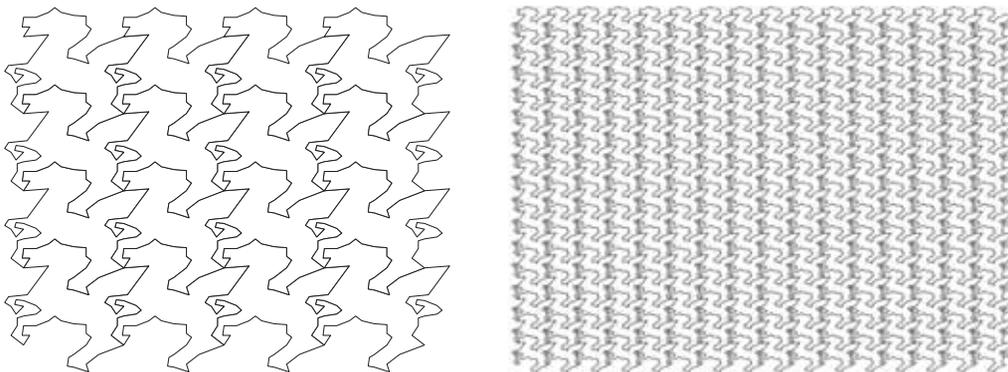


Figure 6.1. Winged horses:  $4 \times 4$  and  $16 \times 16$

## 7. Conclusions

Virtual Elements is a new method, and a lot of work is needed to assess its *pros* and *cons*. Its major interest is on polygonal and polyhedral elements, but its use on distorted quads, hexahedra, and the like, is also quite promising. For triangles and tetrahedra the interest seems to be concentrated in higher order continuity (e.g. [35]). The use of VEM mixed methods seems to be quite interesting, in particular for their connections with Finite Volumes and Mimetic Finite Differences.

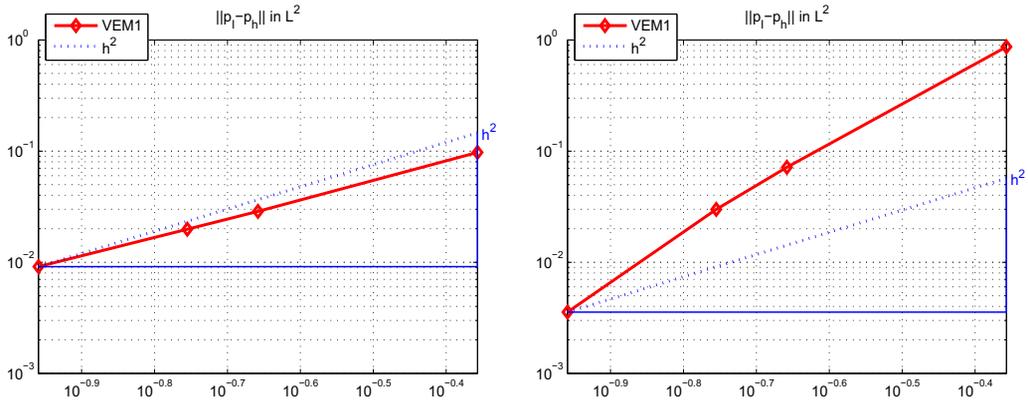


Figure 6.2.  $L^2$  error for primal (left) and mixed (right) formulations

## References

- [1] P. M. Adler, *Fractures and Fracture Networks*, Kluwer Academic, Dordrecht, 1999.
- [2] Ahmad, B., Alsaedi, A., Brezzi, F., Marini, L.D., and Russo, A., *Equivalent projectors for virtual element methods*, *Comput. Math. Appl.*, **66** (2013), 376–391.
- [3] Amir, O. and Sigmund, O., *On reducing computational effort in topology optimization: how far can we go?*, *Structural and Multidisciplinary Optimization* **44** (2011), 25–29.
- [4] Antonietti, P.F., Bigoni, N., and Verani, M., *Mimetic Discretizations of Elliptic Control Problems*, *J. Sci. Comput.* **56** (2013), 14–27.
- [5] Arroyo, M. and Ortiz, M., *Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods*, *Internat. J. Numer. Methods Engrg.* **65** (2006), 2167–2202.
- [6] Arnold, D. N., Boffi, D., and Falk, R. S., *Approximation by quadrilateral finite elements*, *Math. Comput.* **71** (2002), 909–922.
- [7] Arnold, D. N. and Awanou, G., *Finite element differential forms on cubical meshes*, submitted to *Math. Comput.*
- [8] Arnold, D. N., Boffi, D., and Falk, R. S., *Quadrilateral  $H(\text{div})$  finite elements*, *SIAM J. Numer. Anal.* **42** (2005), 2429–2451.
- [9] Babuška, I., Banerjee, U., and Osborn, J.E., *Survey of meshless and generalized finite element methods: a unified approach*, *Acta Numerica* **12** (2003), 1–125.
- [10] ———, *Generalized finite element methods: Main ideas, results and perspectives*, *Int. J. Comp. Meth.* **1** (2004), 67–103.
- [11] Babuška, I. and Melenk, J.M., *The partition of unity method*, *Internat. J. Numer. Methods Engrg.* **40** (1997), 727–758.
- [12] Babuška, I. and Osborn, J.E., *Generalized finite element methods: Their performance and their relation to mixed methods*, *SIAM J. Numerical Analysis* **20** (1983), 510–536.
- [13] Beirão da Veiga, L., *A mimetic discretization method for linear elasticity*, *M2AN Math. Model. Numer. Anal.* **44** (2010), 231–250.
- [14] Beirão Da Veiga, L., Brezzi, F., Cangiani, A., Manzini, G., Marini, L.D., and Russo,

- A., *Basic principles of virtual element methods*, Math. Models Methods Appl. Sci. **23** (2013), 199–214.
- [15] Beirão Da Veiga, L., Brezzi, F., and Marini, L.D., *Virtual Elements for linear elasticity problems*, SIAM J. Numer. Anal. **51** (2013), 794–812.
- [16] Beirão Da Veiga, L., Brezzi, F., Marini, L.D., and Russo, A., *Hitchhikers Guide to the VEM*, Math. Models Methods Appl. Sci. **24** (2014), 1541–1574.
- [17] Beirão Da Veiga, L., Gyrya, V., Lipnikov, K., and Manzini, G., *Mimetic finite difference method for the Stokes problem on polygonal meshes*, J. Comput. Phys. **228** (2009), 7215–7232.
- [18] Beirão Da Veiga, L., Lipnikov, K., and Manzini, G., *Error analysis for a mimetic discretization of the steady Stokes problem on polyhedral meshes*, SIAM J. Numer. Anal. **48** (2010), 1419–1443.
- [19] ———, *Arbitrary-order nodal mimetic discretizations of elliptic problems on polygonal meshes*, SIAM J. Numer. Anal. **49** (2011), 1737–1760.
- [20] ———, *The Mimetic Finite Difference Method*, volume 11 of Modeling, Simulations and Applications, Springer-Verlag, New York, I edition, 2013.
- [21] Beirão da Veiga, L., and Manzini, G., *A virtual element method with arbitrary regularity*, IMA J. Numer. Anal. in press, (2013), DOI:10.1093/imanum/drt018
- [22] Belytschko, T., Liu, W.K., and Kennedy, J.M., *Hourglass control in linear and nonlinear problems*, Comp. Meth. Appl. Mech. Engrg. **43** (1984), 251–276.
- [23] Belytschko, T., Lu, Y., and Gu, L., *Element free Galerkin methods*, Internat. J. Numer. Methods Engrg. **37** (1994), 229–256.
- [24] Belytschko, T., Xiao, S.P., and Parimi, C., *Topology optimization with implicit functions and regularization*, Internat. J. Numer. Methods Engrg. **57** (2003), 1177–1196.
- [25] Bendsoe, M.P., and Sigmund, O., *Topology Optimization—Theory, Methods and Applications*. Springer: New York, 2003.
- [26] Berrone, S., Pieraccini, S., and Scialò, S., *On simulations of discrete fracture network flows with an optimization-based extended finite element method*, SIAM J. Sci. Comput. **35** (2013), A908–A935.
- [27] Bishop, J. E., *Adisplacement-based finite element formulation for general polyhedra using harmonic shape functions*, Internat. J. Numer. Methods Engrg. **97** (2014), 1–31.
- [28] Bochev, P. and Hyman, J. M., Principle of mimetic discretizations of differential operators. In D. Arnold, P. Bochev, R. Lehoucq, R. Nicolaides, and M. Shashkov, editors, *Compatible discretizations. Proceedings of IMA hot topics workshop on compatible discretizations*, IMA Volume 142. Springer-Verlag, 2006.
- [29] Bonelle, J. and Ern, A., *Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes*, M2AN: Math. Model. Numer. Anal. **48** (2014), 553–581.
- [30] Brezzi, F., Buffa, A., and Lipnikov, K., *Mimetic finite differences for elliptic problems*, M2AN: Math. Model. Numer. Anal. **43** (2009), 277–295.
- [31] Brezzi, F., Falk, R.S., and Marini, L.D., *Basic Principles of Mixed Virtual Element Methods*, ESAIM Math. Model. Numer. Anal. to appear.
- [32] Brezzi, F., Lipnikov, K., and Shashkov, M., *Convergence of mimetic finite difference*

- method for diffusion problems on polyhedral meshes*, SIAM J. Num. Anal. **43** (2005), 1872–1896.
- [33] Brezzi, F., Lipnikov, K., Shashkov, M., and Simoncini, V., *A new discretization methodology for diffusion problems on generalized polyhedral meshes*, Comp. Meth. Appl. Mech. Engrg. **196** (2007), 3682–3692.
- [34] Brezzi, F., Lipnikov, K., Simoncini, V., *A family of mimetic finite difference methods on polygonal and polyhedral meshes*, Math. Models Methods Appl. Sci. **15** (2005), 1533–1553.
- [35] Brezzi, F. and Marini, L.D., *Virtual elements for plate bending problems*, Comp. Methods Appl. Mech. Engrg. **253** (2013), 455–462.
- [36] Cangiani, A., Manzini, G., Russo, A., and Sukumar, N., *Hourglass stabilization and the virtual element method*, submitted.
- [37] Chessa, J., Belytschko, T., *An extended finite element method for two-phase fluids*, ASME J. Appl. Mech. **70** (2003), 10–17.
- [38] Chessa, J., Smolinski, P., and Belytschko, T., *The extended finite element method (XFEM) for solidification problems*, Internat. J. Numer. Methods Engrg. **53** (2002), 1959–1977.
- [39] Ciarlet, P.G., *The finite element method for elliptic problems*, Studies in Mathematics and its Applications, Vol. 4. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.
- [40] Daux, C., Moës, N., Dolbow, J., Sukumar, N., and Belytschko, T., *Arbitrary branched and intersecting cracks with the extended finite element method*, Internat. J. Numer. Methods Engrg. **48** (2000), 1741–1760.
- [41] Droniou, J., Eymard, R., Gallouët, T. R., and Herbin, R., *A unified approach to Mimetic Finite Difference, Hybrid Finite Volume and Mixed Finite Volume methods*, Math. Models Methods Appl. Sci. **20** (2010), 265–295.
- [42] Duarte, C.A., Babuška, I., and Oden, J.T., *Generalized finite element methods for three-dimensional structural mechanics problems*, Computers & Structures **77** (2000), 215–232.
- [43] Edelsbrunner, H., Mücke, E.P., *Three-dimensional alpha shapes*, ACM Transactions on Graphics **13** (1994), 43–72.
- [44] Felippa, C.A., *Supernatural Quad4: A Template Formulation*, Comp. Methods Appl. Mech. Engrg. **195** (2006), 5316–5342.
- [45] Floater, M.S., *Mean value coordinates*, Comput. Aided Geom. Design **20** (2003), 19–27.
- [46] Floater, M.S., Hormann, K., and Kós, G., *A general construction of barycentric coordinates over convex polygons*, Adv. Comput. Math. **24** (2006), 311–331.
- [47] Floater, M. S., Kós, G., and Reimers, M., *Mean value coordinates in 3D*, Comput. Aided Geom. Design **22** (2005), 623–631.
- [48] Fries, T.P. and Belytschko, T., *The extended/generalized finite element method: An overview of the method and its applications*, Internat. J. Numer. Meth. Engrg. **84** (2010), 253–304.
- [49] Gain, A.L., *Polytope-based topology optimization using a mimetic-inspired method*,

- PhD thesis, University of Illinois at Urbana-Champaign (2013).
- [50] Gain A.L. and Paulino, G.H., *Phase-field based topology optimization with polygonal elements: A finite volume approach for the evolution equation*, Structural and Multi-disciplinary Optimization **46** (2012), 327–342,
  - [51] Gain, A. L., Talischi, C., and Paulino, G. H., *On the virtual element method for three dimensional elasticity problems on arbitrary polyhedral meshes* (2013), Submitted.
  - [52] Gerstenberger, A., and Wall, W.A., *An eXtended finite element method/Lagrange multiplier based approach for fluid-structure interaction*, Comp. Methods Appl. Mech. Engrg. **197** (2008), 1699–1714.
  - [53] Griebel, M. and Schweitzer, M.A., *A particle-partition of unity method for the solution of elliptic, parabolic and hyperbolic PDEs*, SIAM J. Sci. Comput. **22** (2000), 853–890.
  - [54] ———, *A particle-partition of unity method-part ii: Efficient cover construction and reliable integration*, SIAM J. Sci. Comput. **23** (2002), 1655–1682.
  - [55] ———, *A particle-partition of unity method-part v: boundary conditions*, In S. Hildebrandt and H. Karcher, editors, Geometric Analysis and Nonlinear Partial Differential Equations, Springer, Berlin, 2002, 517–540.
  - [56] Hemp, W.S., *Optimum Structures*, Clarendon Press, Oxford, 1973.
  - [57] Hormann, K. and Floater, M. S., *Mean value coordinates for arbitrary planar polygons*, ACM TRANSACTIONS ON GRAPHICS **25** (2006), 1424–1441.
  - [58] Hormann, K. and Sukumar, N., *Maximum entropy coordinates for arbitrary polytopes*, In Eurographics Symposium on Geometry Processing (2008), 1513–1520.
  - [59] Idelsohn, S.R., Onate, E., Calvo, N., and Del Pin, F., *The meshless finite element method*, Internat. J. for Numer. Methods Engrg. **58** (2003), 893–912.
  - [60] Jacquotte, O.-P. and Oden, J.T., *Analysis of hourglass instabilities and control in underintegrated finite element methods*, Comp. Methods Appl. Mech. Engrg. **44** (1984), 339–363.
  - [61] Joshi, P., Meyer, M., Derose, T., Green, B., and Sanocki, T., *Harmonic coordinates for character articulation*, ACM T. Graphics **26** (2007). Art. n. 71.
  - [62] Ju, T., Liepa, P., and Warren, J., *A general geometric construction of coordinates in a convex simplicial polytope*, Comput. Aided Geom. Design **24** (2007), 161–178.
  - [63] Koh, B.C. and Kikuchi, N., *New improved hourglass control for bilinear and trilinear elements in anisotropic linear elasticity*, Comp. Methods Appl. Mech. Engrg. **65** (1987), 1–46.
  - [64] Kuznetsov, Yu., Repin, S., *New mixed finite element method on polygonal and polyhedral meshes*, Russian J. Numer. Anal. Math. Modelling **18** (2003), 261–278.
  - [65] Lipnikov, K., Manzini, G. and Shashkov, M., *Mimetic finite difference method*, J. Comput. Phys. **257 B** (2014), 1163–1227.
  - [66] Liu, W.K. and Belytschko, T., *Efficient linear and nonlinear heat conduction with a quadrilateral element*, Internat. J. Numer. Meth. Engrg. **20** (1984), 931–948.
  - [67] Malsch, E. A., Lin, J. J., and Dasgupta, G., *Smooth two dimensional interpolants: a recipe for all polygons*, Journal of Graphics Tools **10** (2005), 27–39.
  - [68] Manzini, G., Russo, A., and Sukumar, N., *New perspectives on polygonal and polyhe-*

- dral finite element methods*, to appear in *Math. Models Methods Appl. Sci.* 2013, **24** (2014), 1665–1700.
- [69] Martin, S., Kaufmann, P., Botsch, M., Wicke, M., and Gross, M., *Polyhedral finite elements using harmonic basis functions*, *Comput. Graph. Forum* **27** (2008), 1521–1529.
- [70] Melenk, J.M. and Babuška, I., *The partition of unity finite element method: basic theory and applications*, *Comp. Methods Appl. Mech. Engrg.* **139** (1996), 289–314.
- [71] Merle, R. and Dolbow, J., *Solving thermal and phase change problems with the extended finite element method*, *Comput. Mech.* **28** (2002), 339–350.
- [72] Milbradt, P. and Pick, T., *Polytope finite elements*, *Internat. J. Numer. Meth. Engrg.* **73** (2008), 1811–1835.
- [73] Moes, N., Dolbow, J. and Belytschko, T., *A finite element method for crack growth without remeshing*, *Internat. J. Numer. Meth. Engrg.* **46** (1999), 131–150.
- [74] Mohammadi, S., *Extended finite element method for fracture analysis of structures*, Blackwell, Oxford, 2008.
- [75] Mousavi, S. E. and Sukumar, N., *Numerical integration of polynomials and discontinuous functions on irregular convex polygons and polyhedrons*, *Comput. Mech.* **47** (2011), 535–554.
- [76] Oswald, J., Gracie, R., Khare, R., and Belytschko, T., *An extended finite element method for dislocations in complex geometries: Thin films and nanotubes*, *Comp. Methods Appl. Mech. Engrg.* **198** (2009), 1872–1886.
- [77] Rabczuk, T., Bordas, S., and Zi, G., *On three-dimensional modelling of crack growth using partition of unity methods*, *Computers & Structures* **88** (2010), 1391–1411.
- [78] Rao, B.N. and Rahman, S., *An enriched meshless method for non-linear fracture mechanics*, *Internat. J. Numer. Methods Engrg.* **59** (2004), 197–223.
- [79] Sigmund, O., *Design of multiphysics actuators using topology optimization - Part II: Two-material structures*, *Comp. Methods Appl. Mech. Engrg.* **190** (2001), 6605–6627.
- [80] Smith, B.G., Vaughan, B.L., Jr., and Chopp, D.L., *The extended finite element method for boundary layer problems in biofilm growth*, *Comm. App. Math. and Comp. Sci.* **2** (2007), 35–56.
- [81] Sukumar, N., *Construction of polygonal interpolants: a maximum entropy approach*, *Internat. J. Numer. Methods Engrg.* **61** (2004), 159–2181.
- [82] Sukumar, N. and Malsch, E. A., *Recent advances in the construction of polygonal finite element interpolations*, *Arch. Comput. Methods Engrg.* **13** (2006), 129–163.
- [83] Sukumar, N., Chopp, D.L., Moës, N., and Belytschko, T., *Modeling holes and inclusions by level sets in the extended finite-element method*, *Comp. Methods Appl. Mech. Engrg.* **190** (2001), 6183–6200.
- [84] Sukumar, N., Moës, N., Moran, B., and Belytschko, T., *Extended finite element method for three-dimensional crack modelling*, *Internat. J. Numer. Methods Engrg.* **48** (2000), 1549–1570.
- [85] Sukumar, N., Moran, B., Semenov, A. Y., and Belikov, V. V., *Natural neighbor Galerkin methods*, *Int. J. Numer. Meth. Engrg.* **50** (2001), 1–27.
- [86] Sukumar, N. and Tabarraei, A., *Conforming polygonal finite elements*, *Int. J. Numer.*

- Meth. Engrg. **61** (2004), 2045–2066.
- [87] Sutradhar, A., Paulino, G. H., Miller, M. J., and Nguyen, T. H., *Topology optimization for designing patient-specific large craniofacial segmental bone replacements*, Proc. Natl. Acad. Sci. U.S.A. **107** (2010), 13222–13227.
- [88] Talischi, C. and Paulino, G.H., *Addressing integration error for polygonal finite elements through polynomial projections: A patch test connection*, Math. Models Methods Appl. Sci. **24** (2014), 1701–1728.
- [89] Talischi, C., Paulino, G. H., Pereira, A., and Menezes, I.F.M., *Polygonal finite elements for topology optimization: A unifying paradigm*, Int. J. Numer. Methods. Engrg. **82** (2010), 671–698.
- [90] Vigneron, L.M., Verly, J.G., and Warfield, S.K., *On extended finite element method (XFEM) for modelling of organ deformations associated with surgical cuts*, In S. Cotin and D. Metaxas, editors, Medical Simulation, volume 3078 of Lecture Notes in Computer Science. Springer, Berlin, 2004.
- [91] Vohralík, M., Maryška, J., and Severýn, O., *Mixed and nonconforming finite element methods on a system of polygons*, Appl. Numer. Math. **51** (2007), 176–193.
- [92] Vohralík, M. and Wohlmuth, B., *Mixed finite element methods: implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods*, Math. Models Methods Appl. Sci. **23** (2013), 803–838.
- [93] Wachspress, E., *A Rational Finite Element Basis*, Academic Press, New York, 1975.
- [94] ———, *Rational bases for convex polyhedra*, Comput. Math. Appl. **59** (2010), 1953–1956.
- [95] Wagner, G.J., Moës, N., Liu, W.K., and Belytschko, T., *The extended finite element method for rigid particles in Stokes flow*, Internat. J. Numer. Methods Engrg. **51** (2001), 293–313.
- [96] Warren, J., *Barycentric coordinates for convex polytopes*, Adv. Comput. Math. **6** (1996), 97–108.
- [97] Wicke, M., Botsch, M., and Gross, M., *A finite element method on convex polyhedra*, Comp. Graphics Forum, **26** (2007), 355–364.
- [98] Yip, M., Mohle, J., and Bolander, J.E., *Automated modeling of three-dimensional structural components using irregular lattices*, Computer Aided Civil and Infrastructure Engineering **20** (2005), 393–407.

IUSS and IMATI-CNR, Via Ferrata 5/A, I-27100 Pavia, Italy

E-mail: brezzi@imati.cnr.it

# Mathematics of sparsity (and a few other things)

Emmanuel J. Candès

**Abstract.** In the last decade, there has been considerable interest in understanding when it is possible to find structured solutions to underdetermined systems of linear equations. This paper surveys some of the mathematical theories, known as compressive sensing and matrix completion, that have been developed to find sparse and low-rank solutions via convex programming techniques. Our exposition emphasizes the important role of the concept of incoherence.

**Mathematics Subject Classification (2010).** Primary 00A69.

**Keywords.** Underdetermined systems of linear equations, compressive sensing, matrix completion, sparsity, low-rank-matrices,  $\ell_1$  norm, nuclear norm, convex programming, Gaussian widths.

## 1. Introduction

Many engineering and scientific problems ask for solutions to underdetermined systems of linear equations: a system is considered underdetermined if there are fewer equations than unknowns (in contrast to an overdetermined system, where there are more equations than unknowns). Examples abound everywhere but we immediately give two concrete examples that we shall keep as a guiding thread throughout the article.

- *A compressed sensing problem.* Imagine we have a signal  $x(t)$ ,  $t = 0, 1, \dots, n - 1$ , with possibly complex-valued amplitudes and let  $\hat{x}$  be the discrete Fourier transform (DFT) of  $x$  defined by

$$\hat{x}(\omega) = \sum_{t=0}^{n-1} x(t)e^{-i2\pi\omega t/n}, \quad \omega = 0, 1, \dots, n - 1.$$

In applications such as magnetic resonance imaging (MRI), it is often the case that we do not have the time to collect all the Fourier coefficients so we only sample  $m \ll n$  of them. This leads to an underdetermined system of the form  $y = Ax$ , where  $y$  is the vector of Fourier samples at the observed frequencies and  $A$  is the  $m \times n$  matrix whose rows are correspondingly sampled from the DFT matrix.<sup>1</sup> Hence, we would like to recover  $x$  from a highly incomplete view of its spectrum.

- *A matrix completion problem.* Imagine we have an  $n_1 \times n_2$  array of numbers  $x(t_1, t_2)$  perhaps representing users' preference for a collection of items as in the famous Netflix challenge; for instance,  $x(t_1, t_2)$  may be a rating given by user  $t_1$  (e.g. Emmanuel) for movie  $t_2$  (e.g. "The Godfather"). We do not get to see many ratings as only a few

---

<sup>1</sup>Proceedings of International Congress of Mathematicians, 2014, Seoul

entries from the matrix  $x$  are actually revealed to us. Yet we would like to correctly infer all the unseen ratings; that is, we would like to predict how a given user would rate a movie she has not yet seen. Clearly, this calls for a solution to an underdetermined system of equations.

In both these problems we have an  $n$ -dimensional object  $x$  and information of the form

$$y_k = \langle a_k, x \rangle, \quad k = 1, \dots, m, \quad (1.1)$$

where  $m$  may be far less than  $n$ . Everyone knows that such systems have infinitely many solutions and thus, it is apparently impossible to identify which of these candidate solutions is indeed the correct one without some additional information. In this paper, we shall see that if the object we wish to recover has a bit of structure, then exact recovery is often possible by simple convex programming techniques.

What do we mean by structure? Our purpose, here, is to discuss two types, namely, sparsity and low rank.

- *Sparsity.* We shall say that a signal  $x \in \mathbb{C}^n$  is sparse, when most of the entries of  $x$  vanish. Formally, we shall say that a signal is  $s$ -sparse if it has at most  $s$  nonzero entries. One can think of an  $s$ -sparse signal as having only  $s$  degrees of freedom (df).
- *Low-rank.* We shall say that a matrix  $x \in \mathbb{C}^{n_1 \times n_2}$  has low rank if its rank  $r$  is (substantially) less than the ambient dimension  $\min(n_1, n_2)$ . One can think of a rank- $r$  matrix as having only  $r(n_1 + n_2 - r)$  degrees of freedom (df) as this is the dimension of the tangent space to the manifold of rank- $r$  matrices.

The question now is whether it is possible to recover a sparse signal or a low-rank matrix—both possibly depending upon far fewer degrees of freedom than their ambient dimension suggests—from just a few linear equations. The answer is in general negative. Suppose we have a 20-dimensional vector  $x$  that happens to be 1-sparse with all coordinates equal to zero but for the last component equal to one. Suppose we have 10 equations revealing the first 10 entries of  $x$  so that  $a_k = e_k$ ,  $k = 1, \dots, 10$ , where throughout  $e_k$  is the  $k$ th canonical basis vector of  $\mathbb{C}^n$  or  $\mathbb{R}^n$  (here,  $n = 20$ ). Then  $y = 0$  and clearly no method whatsoever would be able to recover our signal  $x$ . Likewise, suppose we have a  $20 \times 20$  matrix of rank 1 with a first row equal to an arbitrary vector  $x$  and all others equal to zero. Imagine that we see half the entries selected completely at random. Then with overwhelming probability we would not see all the entries in the first row, and many completions would, therefore, be feasible even with the perfect knowledge that the matrix has rank exactly one.

These simple considerations demonstrate that structure is not sufficient to make the problem well posed. To guarantee recovery from  $y = Ax$  by any method whatsoever, it must be the case that the structured object  $x$  is not in the null space of the matrix  $A$ . We shall assume an *incoherence property*, which roughly says that in the sparse recovery problem, while  $x$  is sparse, the rows of  $A$  are *not*, so that each measurement  $y_k$  is a weighted sum of all the components of  $x$ . A different way to put this is to say that the *sampling vectors*  $a_k$  do not correlate well with the signal  $x$  so that each measurement contains a little bit of information about the nonzero components of  $x$ . In the matrix completion problem, however, the sampling elements are sparse since they reveal entries of the matrix  $x$  we care to infer, so clearly the matrix  $x$  cannot be sparse. As explained in the next section, the right notion of

---

<sup>1</sup>More generally,  $x$  might be a two- or three-dimensional image.

incoherence is that sparse subsets of columns (resp. rows) cannot be singular or uncorrelated with all the other columns (resp. rows). A surprise is that under such a general incoherence property as well as some randomness, solving a simple convex program usually recovers the unknown solution exactly. In addition, the number of equations one needs is—up to possible logarithmic factors—proportional to the degrees of freedom of the unknown solution. This paper examines this curious phenomenon.

## 2. Recovery by convex programming

The recovery methods studied in this paper are extremely simple and all take the form of a norm-minimization problem

$$\text{minimize } \|x\| \quad \text{subject to } y = Ax, \quad (2.1)$$

where  $\|\cdot\|$  is a norm promoting the assumed structure of the solution. In our two recurring examples, these are:

- The  $\ell_1$  norm for the compressed sensing problem. The  $\ell_1$  norm,  $\|x\|_{\ell_1} = \sum_i |x_i|$ , is a convex surrogate for the  $\ell_0$  counting ‘norm’ defined as  $|\{i : x_i \neq 0\}|$ . It is the best surrogate in the sense that the  $\ell_1$  ball is the smallest convex body containing all 1-sparse objects of the form  $\pm e_i$ .
- The *nuclear norm*, or equivalently, *Schatten-1 norm* for the matrix completion problem defined as the sum of the singular values of a matrix  $X$ . It is the best convex surrogate to the rank functional in the sense that the nuclear ball is the smallest convex body containing all rank-1 matrices with spectral norm at most equal to 1. This is the analogue to the  $\ell_1$  norm in the sparse recovery problem above since the rank functional simply counts the number of nonzero singular values.

In truth, there is much literature on the empirical performance of  $\ell_1$  minimization [72, 67, 66, 26, 73, 41] as well as some early theoretical results explaining some of its success [55, 35, 37, 34, 75, 40, 46]. In 2004, starting with [16] and then [32] and [20], a series of papers suggested the use of random projections as means to acquire signals and images with far fewer measurements than were thought necessary. These papers triggered a massive amount of research spanning mathematics, statistics, computer science and various fields of science and engineering, which all explored the promise of cheaper and more efficient sensing mechanisms. The interested reader may want to consult the March 2008 issue of the *IEEE Signal Processing Magazine* dedicated to this topic and [49, 39]. This research is highly active today. In this paper, however, we focus on modern mathematical developments inspired by the three early papers [16, 32, 20]: in the spirit of compressive sensing, the sampling vectors are, therefore, randomized.

Let  $F$  be a distribution of random vectors on  $\mathbb{C}^n$  and let  $a_1, \dots, a_m$  be a sequence of i.i.d. samples from  $F$ . We require that the ensemble  $F$  is complete in the sense that the covariance matrix  $\Sigma = \mathbb{E}aa^*$  is invertible (here and below,  $a^*$  is the adjoint), and say that the distribution is *isotropic* if  $\Sigma$  is proportional to the identity. The *incoherence parameter* is the smallest number  $\mu(F)$  such that if  $a \sim F$ , then

$$\max_{1 \leq i \leq n} |\langle a, e_i \rangle|^2 \leq \mu(F) \quad (2.2)$$

holds either deterministically or with high probability, see [14] for details. If  $F$  is the uniform distribution over scaled canonical vectors such that  $\Sigma = I$ , then the coherence is large, i.e.  $\mu = n$ . If  $x(t)$  were a time-dependent signal, this sampling distribution would correspond to revealing the values of the signal at randomly selected time points. If, however, the sampling vectors are spread as when  $F$  is the ensemble of complex exponentials (the rows of the DFT) matrix, the coherence is low and equal to  $\mu = 1$ . When  $\Sigma = I$ , this is the lowest value the coherence parameter can take on since by definition,  $\mu \geq \mathbb{E} |\langle a, e_i \rangle|^2 = 1$ .

**Theorem 2.1** ([14]). *Let  $x^*$  be a fixed but otherwise arbitrary  $s$ -sparse vector in  $\mathbb{C}^n$ . Assume that the sampling vectors are isotropic ( $\Sigma = I$ ) and let  $y = Ax^*$  be the data vector and the  $\ell_1$  norm be the regularizer in (2.1). If the number of equations obeys*

$$m \geq C_\beta \cdot \mu(F) \cdot \text{df} \cdot \log n, \quad \text{df} = s,$$

*then  $x^*$  is the unique minimizer with probability at least  $1 - 5/n - e^{-\beta}$ . Further,  $C_\beta$  may be chosen as  $C_0(1 + \beta)$  for some positive numerical constant  $C_0$ .*

Loosely speaking, Theorem 2.1 states that if the rows of  $A$  are diverse and incoherent (not sparse), then if there is an  $s$ -sparse solution, it is unique and  $\ell_1$  will find it. This holds as soon as the number of equations is on the order of  $s \cdot \log n$ . Continuing, one can understand the probabilistic guarantee as saying that most deterministic systems with diverse and incoherent rows have this property. Hence, Theorem 2.1 is a fairly general result with minimal assumptions on the sampling vectors, and which then encompasses many signal recovery problems frequently discussed in practice, see [14] for a non-exhaustive list.

Theorem 2.1 is also sharp in the sense that for any reasonable values of  $(\mu, s)$ , one can find examples for which *any* recovery algorithm would fail when presented with fewer than a constant times  $\mu(F) \cdot s \cdot \log n$  random samples [14]. As hinted, our result is stated for isotropic sampling vectors for simplicity, although there are extensions which do not require  $\Sigma$  to be a multiple of the identity; only that it has a well-behaved condition number [53].

Three important remarks are in order. The first, is that Theorem 2.1 extends the main result from [16], which established that a  $s$ -sparse signal can be recovered from about  $20 \cdot s \cdot \log n$  random Fourier samples via minimum  $\ell_1$  norm with high probability (or equivalently, from almost all sets with at least this cardinality). Among other implications, this mathematical fact motivated MR researchers to speed up MR scan acquisition times by sampling at a lower rate, see [56, 78] for some impressive findings. Moreover, Theorem 2.1 also sharpens and extends another earlier incoherent sampling theorem in [9]. The second is that other types of Fourier sampling theorems exist, see [43] and [79]. The third is that in the case the linear map  $A$  has i.i.d. Gaussian entries, it is possible to establish more precise sampling theorems. Section 5 is dedicated to describing a great line of research on this subject.

We now turn to the matrix completion problem. Here, the entries  $X_{ij}$  of an  $n_1 \times n_2$  matrix  $X$  are revealed uniformly at random so that the sampling vectors  $a$  are of the form  $e_i e_j^*$  where  $(i, j)$  is uniform over  $[n_1] \times [n_2]$  ( $[n] = \{1, \dots, n\}$ ). With this,

$$X_{ij} = \langle e_i e_j^*, X \rangle$$

where  $\langle \cdot, \cdot \rangle$  is the usual matrix inner product. Again, we have an isotropic sampling distribution in which  $\Sigma = (n_1 n_2)^{-1} I$ . We now need a notion of incoherence between the sampling vectors and the matrix  $X$ , and define the *incoherence parameter*  $\mu(X)$  introduced in [15],

which is the smallest number  $\mu(X)$  such that

$$\begin{aligned} \max_{1 \leq i \leq n_1} (n_1/r) \cdot \|\pi_{\text{col}(X)} e_i\|_{\ell_2}^2 &\leq \mu(X) \\ \max_{1 \leq j \leq n_2} (n_2/r) \cdot \|\pi_{\text{row}(X)} e_j\|_{\ell_2}^2 &\leq \mu(X), \end{aligned} \tag{2.3}$$

where  $r$  is the rank of  $X$  and  $\pi_{\text{col}(X)}$  (resp.  $\pi_{\text{row}(X)}$ ) is the projection onto the column (resp. row) space of  $X$ . The coherence parameter measures the overlap or correlation between the column/row space of the matrix and the coordinate axes. Since  $\sum_i \|\pi_{\text{col}(X)} e_i\|_{\ell_2}^2 = \text{tr}(\pi_{\text{col}(X)}) = r$ , we can conclude that  $\mu(X) \geq 1$ . Conversely, the coherence is by definition bounded above by  $\max(n_1, n_2)/r$ . A matrix with low coherence has column and row spaces away from the coordinate axes as in the case where they assume a uniform random orientation.<sup>2</sup> Conversely, a matrix with high coherence may have a column (or a row space) well aligned with a coordinate axis. As should become intuitive, we can only hope to recover ‘incoherent’ matrices; i.e. matrices with relatively low-coherence parameter values.

**Theorem 2.2.** *Let  $X^*$  be a fixed but otherwise arbitrary matrix of dimensions  $n_1 \times n_2$  and rank  $r$ . Let  $y$  in (2.1) be the set of revealed entries of  $X^*$  at randomly selected locations and  $\|\cdot\|$  be the nuclear norm. Then with probability at least  $1 - n^{-10}$ ,  $X^*$  is the unique minimizer to (2.1) provided that the number of samples obeys*

$$m \geq C_0 \cdot \mu(X) \cdot \text{df} \cdot \log^2(n_1 + n_2), \quad \text{df} = r(n_1 + n_2 - r),$$

for some positive numerical constant  $C_0$ .

We have adopted a formulation emphasizing the resemblance with the earlier sparse recovery theorem. Indeed just as before, Theorem 2.2 states that one can sample without any information loss the entries of a low-rank matrix at a rate essentially proportional to the coherence times its degrees of freedom. Moreover, the sampling rate is known to be optimal up to a logarithmic factor in the sense that for any reasonable values of the pair  $(\mu(X), \text{rank}(X))$ , there are matrices that cannot be recovered from fewer than a constant times  $\mu(X) \cdot \text{df} \cdot \log(n_1 + n_2)$  randomly sampled entries [21].

The role of the coherence in this theory is also very natural, and can be understood when thinking about the prediction of movie ratings. Here, we can imagine that the complete matrix of ratings has (approximately) low rank because users’ preferences are correlated. Now the reason why matrix completion is possible under incoherence is that we can exploit correlations and infer how a specific user is going to like a movie she has not yet seen, by examining her ratings and learning about her general preferences, and inferring how other users with such preferences have rated this particular item. Whenever we have users or small groups of users that are very singular in the sense that their ratings are orthogonal to those of all other users, it is not possible to correctly predict their missing entries. Such matrices have large coherence. (To convince oneself, consider situations where a few users enter ratings based on the outcome of coin tosses.) An amusing example of a low-rank and incoherent matrix may be the voting patterns of senators and representatives in the U. S. Congress.

A first version of this result appeared in [15], however, with one additional technical assumption concerning the approximate orthogonality between left- and right-singular vectors. This condition appears in all the subsequent literature except in unpublished work from Xi-aodong Li and the author and in [27], so that Theorem 2.2, as presented here, holds. Setting

---

<sup>2</sup>If the column space of  $X$  has uniform orientation, then for each  $i$ ,  $(n_1/r) \cdot \mathbb{E} \|\pi_{\text{col}(X)} e_i\|_{\ell_2}^2 = 1$ .

$n = \max(n_1, n_2)$ , [15] proved that on the order of  $\mu(X) \cdot n^{6/5} r \cdot \log n$  sampled entries are sufficient for perfect recovery, a bound which was lowered to  $\mu(X) \cdot nr \cdot \log^a n$  in [21], with  $a \leq 6$  and sometimes equal to 2. Later, David Gross [47], using beautiful and new arguments, demonstrated that the latter bound holds with  $a = 2$ . (Interestingly, all three papers exhibit completely different proofs.) For a different approach to matrix completion, please see [52].

One can ask whether matrix completion is possible from more general random equations, where the sampling matrices may not have rank one, and are still i.i.d. samples from some fixed distribution  $F$ . By now, one would believe that if the sampling matrices do not correlate well with the unknown matrix  $X$ , then matrix completion ought to be possible. This belief is correct. To give a concrete example, suppose we have an orthobasis of matrices  $\mathcal{F} = \{B_j\}_{1 \leq j \leq n_1 n_2}$  and that we select elements from this family uniformly at random. Then [47] shows that if

$$\begin{aligned} \max_{B \in \mathcal{F}} (n_1/r) \cdot \|\pi_{\text{col}(X)} B\|_F^2 &\leq \mu(X) \\ \max_{B \in \mathcal{F}} (n_2/r) \cdot \|B \pi_{\text{row}(X)}\|_F^2 &\leq \mu(X), \end{aligned}$$

( $\|\cdot\|_F$  is the Frobenius norm) holds along with another technical condition, Theorem 2.2 holds. Note that in the previous example where  $B = e_i e_j^*$ ,  $\|\pi_{\text{col}(X)} B\|_F^2 = \|\pi_{\text{col}(X)} e_i\|_{\ell_2}^2$  so that we are really dealing with the same notion of coherence.

### 3. Why does this work?

The results we have presented may seem surprising at first: why is it that with on the order of  $s \cdot \log n$  random equations,  $\ell_1$  minimization will find the unique  $s$ -sparse solution to the system  $y = Ax$ ? Our intent is to give an intuitive explanation of this phenomenon. Define the *cone of descent* of the norm  $\|\cdot\|$  at a point  $x$  as

$$\mathcal{C} = \{h : \|x + ch\| \leq \|x\| \text{ for some } c > 0\}. \quad (3.1)$$

This convex cone<sup>3</sup> is the set of non-ascent directions of  $\|\cdot\|$  at  $x$ . In the literature on convex geometry, this object is known as the tangent cone. Now it is straightforward to see that a point  $x$  is the unique solution to (2.1) if and only if the null space of  $A$  misses the cone of descent at  $x$ , i.e.  $\mathcal{C} \cap \text{null}(A) = \{0\}$ . A geometric representation of this fact is depicted in Figure 3.1. Looking at the figure, we also begin to understand why minimizing the  $\ell_1$  and nuclear norms recovers sparse and low-rank objects: indeed, as the figure suggests, the tangent cone to the  $\ell_1$  norm is ‘narrow’ at sparse vectors and, therefore, even though the null space is of small codimension  $m$ , it is likely that if  $m$  is large enough, it will miss the tangent cone. A similar observation applies to the nuclear ball, which also appears pinched at low-rank objects. As intuitive as it is, this geometric observation is far from accounting for the style of results introduced in the previous section. For instance, consider Theorem 2.1 in the setting of Fourier sampling: then we would need to show that a plane spanned by  $n - m$  complex exponentials selected uniformly at random misses the tangent cone. For matrix completion, the null space is the set of all matrices vanishing at the locations of the

<sup>3</sup>A cone is a set closed under positive linear combinations

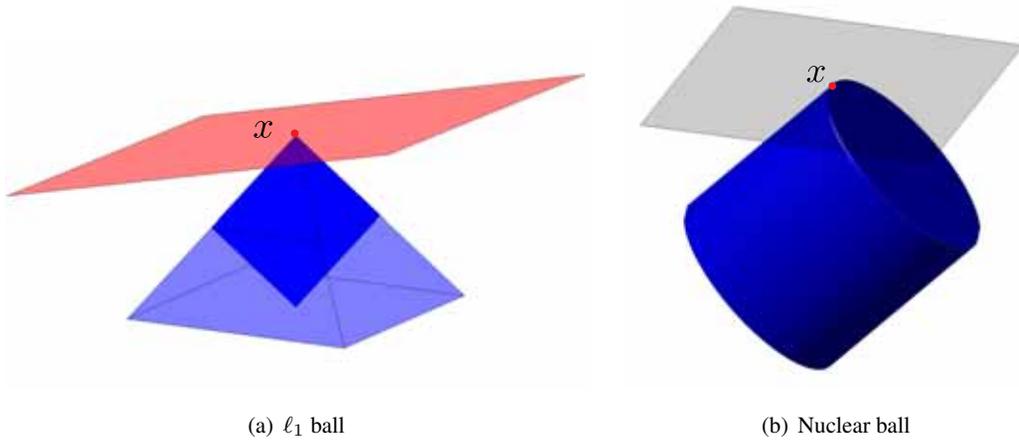


Figure 3.1. Balls associated with the  $\ell_1$  and nuclear norms together with the affine feasible set for (2.1). The ball in (b) corresponds to  $2 \times 2$  symmetric matrices—thus depending upon three parameters—with nuclear norm at most equal to that of  $x$ . When the feasible set is tangent to the ball, the solution to (2.1) is exact.

revealed entries. There, the null space misses the cone of the nuclear ball at low-rank objects, which are sufficiently incoherent. It does not miss the cone at coherent low-rank matrices since the exact recovery property cannot hold in this case. So how do we go about proving these things?

Introduce the subdifferential of  $\|\cdot\|$  at  $x$ , defined as the set of vectors

$$\partial\|x\| = \{w : \|x + h\| \geq \|x\| + \langle w, h \rangle \text{ for all } h\}. \tag{3.2}$$

Then  $x$  is a solution to (2.1) if and only if

$$\exists v \perp \text{null}(A) \text{ such that } v \in \partial\|x\|.$$

For the  $\ell_1$  norm, letting  $T$  be the linear span of vectors with the same support as  $x$  and  $T^\perp$  be its orthogonal complement (those vectors vanishing on the support of  $x$ ),

$$\partial\|x\|_{\ell_1} = \{\text{sgn}(x) + w : w \in T^\perp, \|w\|_{\ell_\infty} \leq 1\}, \tag{3.3}$$

where  $\text{sgn}(x)$  is the vector of signs equal to  $x_i/|x_i|$  whenever  $|x_i| \neq 0$  and to zero otherwise. If we would like  $x$  to be the unique minimizer, a sufficient (and almost necessary) condition is this:  $T \cap \text{null}(A) = \{0\}$  and

$$\exists v \perp \text{null}(A) \text{ such that } v = \text{sgn}(x) + w, w \in T^\perp, \|w\|_{\ell_\infty} < 1. \tag{3.4}$$

In the literature, such a vector  $v$  is called a *dual certificate*.

What does this mean for the Fourier sampling problem where we can only observe the Fourier transform of a signal  $x(t)$ ,  $t = 0, 1, \dots, n - 1$ , at a few random frequencies  $k \in \Omega \subset \{0, 1, \dots, n - 1\}$ ? The answer: a sparse candidate signal  $x$  is solution to the  $\ell_1$  minimization problem if and only if there exists a trigonometric polynomial with sparse coefficients  $P(t) = \sum_{k \in \Omega} c_k \exp(i2\pi kt/n)$  obeying  $P(t) = \text{sgn}(x(t))$  whenever  $x(t) \neq 0$

and  $|P(t)| \leq 1$  otherwise. If there is no such polynomial, (2.1) must return a different answer. Moreover, if  $T \cap \text{null}(A) = \{0\}$  and there exists  $P$  as above with  $|P(t)| < 1$  off the support of  $x$ , then  $x$  is the unique solution to (2.1).<sup>4</sup>

Turning to the minimum nuclear norm problem, let  $X = USV^*$  be a singular value decomposition. Then

$$\partial\|X\|_{S^1} = \{\text{sgn}(X) + W : W \in T^\perp, \|W\|_{S^\infty} \leq 1\};$$

here,  $\|\cdot\|_{S^1}$  and  $\|\cdot\|_{S^\infty}$  are the nuclear and spectral norms,  $\text{sgn}(X)$  is the matrix defined as  $\text{sgn}(X) = UV^*$ , and  $T^\perp$  is the set of matrices with both column and row spaces orthogonal to those of  $X$ . With these definitions, everything is as before and  $X$  is the unique solution to (2.1) if  $T \cap \text{null}(A) = \{0\}$  and, swapping the  $\ell_\infty$  norm for the spectral norm, (3.4) holds.

### 4. Some probability theory

We wish to show that a candidate solution  $x^*$  is solution to (2.1). This is equivalent to being able to construct a dual certificate, which really is the heart of the matter. Starting with [16], a possible approach is to study an ansatz, which is the solution  $v$  to:

$$\text{minimize } \|v\|_{\ell_2} \quad \text{subject to } v \perp \text{null}(A) \text{ and } P_T v = \text{sgn}(x^*),$$

where  $P_T$  is the projection onto the linear space  $T$  defined above. If  $\|\cdot\|^*$  is the norm dual to  $\|\cdot\|$ , then the property  $\|P_{T^\perp} v\|^* < 1$  would certify optimality (with the proviso that  $T \cap \text{null}(A) = \{0\}$ ). The motivation for this ansatz is twofold: first, it is known in closed form and can be expressed as

$$v = A^* A_T (A_T^* A_T)^{-1} \text{sgn}(x), \tag{4.1}$$

where  $A_T$  is the restriction of  $A$  to the subspace  $T$ ; please observe that  $A_T^* A_T$  is invertible if and only if  $T \cap \text{null}(A) = \{0\}$ . Hence, we can study this object analytically. The second reason is that the ansatz is the solution to a least-squares problem and that by minimizing its Euclidean norm we hope to make its dual norm small as well.

At this point it is important to recall the random sampling model in which the rows of  $A$  are i.i.d. samples from a distribution  $F$  so that

$$A^* A = \sum_{k=1}^m a_k a_k^*$$

can be interpreted as an empirical covariance matrix. When the distribution is isotropic ( $\Sigma = I$ ) we know that  $\mathbb{E} A^* A = m I$  and, therefore,  $\mathbb{E} A_T^* A_T = m I_T$ . Of course,  $A^* A$  cannot be close to the identity since it has rank  $m \ll n$  but we can nevertheless ask whether its restriction to  $T$  is close to the identity on  $T$ . It turns out that under the stated assumptions of the theorems,

$$\frac{1}{2} I_T \preceq \frac{1}{m} A_T^* A_T \preceq \frac{3}{2} I_T, \tag{4.2}$$

---

<sup>4</sup>The condition  $T \cap \text{null}(A) = \{0\}$  means that the only polynomial  $P(t) = \sum_{0 \leq k \leq n-1} c_k \exp(i2\pi kt/n)$ , with  $c_k = 0$  whenever  $k \in \Omega$  and support included in that of  $x$ , is the zero polynomial  $P = 0$ .

meaning that  $m^{-1}A_T^*A_T$  is reasonably close to its expectation. For our two running examples and presenting progress in a somewhat chronological fashion, [16] and [21] established this property by combinatorial methods, following a strategy originating in the work of Eugene Wigner [81]. The idea is to develop bounds on moments of the difference between the sampled covariance matrix and its expectation,

$$H_T = I_T - m^{-1}A_T^*A_T.$$

Controlling the growth of  $\mathbb{E} \operatorname{tr}(H_T^{2k})$  for large powers gives control of  $\|H_T\|_{S^\infty}$ . However, since the entries of  $A$  are in general not independent, it is not possible to invoke standard moment calculation methods, and this approach leads to delicate combinatorial issues involving statistics of various paths in the plane that can be interpreted as complicated variants of Dyck’s paths.

Next, to show that the ansatz (4.1) is indeed a dual certificate, one can expand the inverse of  $A_T^*A_T$  as a Neumann series and write it as

$$v = \sum_{j \geq 0} v_j, \quad v_j = m^{-1}A^*A_T H_T^j \operatorname{sgn}(x).$$

In the  $\ell_1$  problem, we would need to show that  $\|P_{T^\perp} v\|_{\ell_\infty} < 1$ ; that is to say, for all  $t$  at which  $x(t) = 0$ ,  $|v(t)| < 1$ . In [16], this is achieved by a combinatorial method bounding the size of each term  $v_j(t)$  by controlling an appropriately large moment  $\mathbb{E} |v_j(t)|^{2k}$ . This strategy yields the  $20 \cdot s \cdot \log n$  bound we presented earlier. In the matrix completion problem, each term  $v_j$  in the sum above is a matrix and we wish to bound the spectral norm of the random matrix  $P_{T^\perp} v$ . The combinatorial approach from [21] also proceeds by controlling moments of the form  $\mathbb{E} \operatorname{tr}(z_j^* z_j)^k$ , where  $z_j$  is the random matrix  $z_j = P_{T^\perp} v_j$ .

There is an easier way to show that the restricted sampled covariance matrix is close to its mean (4.2), which goes by means of powerful tools from probability theory such as the Rudelson selection theorem [64] or the operator Bernstein inequality [2]. The latter is the matrix-valued analog of the classical Bernstein inequality for sums of independent random variables and gives tail bounds on the spectral norm of a sum of mean-zero independent random matrices. This readily applies since both  $I - A^*A$  and its restriction to  $T$  are of this form. One downside is that these general tools are unfortunately not as precise as combinatorial methods. Also, this is only one small piece of the puzzle, and it is not clear how one would use this to show that  $\|P_{T^\perp} v\|^* < 1$ , although [15] made some headway. We refer to [61] for a presentation of these ideas in the context of signal recovery.

A bit later, David Gross [47] provided an elegant construction of an inexact dual certificate he called the *golfing scheme*, and we shall dedicate the remainder of this section to presenting the main ideas behind this clever concept. To fix things, we will assume that we are working on the minimum  $\ell_1$  problem although all of this extends to the matrix completion problem. Our exposition is taken from [14]. To begin with, it is not hard to see that if (4.2) holds, then the existence of a vector  $v \perp \operatorname{null}(A)$  obeying

$$\|P_T(v - \operatorname{sgn}(x))\|_{\ell_2} \leq \delta \quad \text{and} \quad \|P_{T^\perp} v\|_{\ell_\infty} < 1/2, \tag{4.3}$$

with  $\delta$  sufficiently small, certifies that  $x$  is the unique solution. This is interesting because by being a little more stringent on the size of  $v$  on  $T^\perp$ , we can relax the condition  $P_T v = \operatorname{sgn}(x)$  so that it only holds approximately. To see why this is true, take  $v$  as in (4.3) and consider the

perturbation  $v' = v - A^* A_T (A_T^* A_T)^{-1} P_T (\text{sgn}(x) - v)$ . Then  $v' \perp \text{null}(A)$ ,  $P_T v' = \text{sgn}(x)$  and

$$\|P_{T^\perp} v'\|_{\ell_\infty} \leq 1/2 + \|A_{T^\perp}^* A_T (A_T^* A_T)^{-1} P_T (\text{sgn}(x) - v)\|_{\ell_\infty}.$$

Because the columns of  $A$  have Euclidean norm at most  $\mu(F)\sqrt{m}$ , then (4.2) together with Cauchy-Schwarz give that the second term in the right-hand side is bounded by  $\delta \cdot \sqrt{2}\mu(F)$ , which is less than  $1/2$  if  $\delta$  is sufficiently small.

Now partition  $A$  into row blocks so that from now on,  $A_1$  are the first  $m_1$  rows of the matrix  $A$ ,  $A_2$  the next  $m_2$  rows, and so on. The  $\ell$  matrices  $\{A_j\}_{j=1}^\ell$  are independently distributed, and we have  $m_1 + m_2 + \dots + m_\ell = m$ . The golfing scheme then starts with  $v_0 = 0$ , inductively defines

$$v_j = \frac{1}{m_j} A_j^* A_j P_T (\text{sgn}(x) - v_{j-1}) + v_{j-1}$$

for  $j = 1, \dots, \ell$ , and sets  $v = v_\ell$ . Clearly,  $v$  is in the row space of  $A$ , and thus perpendicular to the null space. To understand this scheme, we can examine the first step

$$v_1 = \frac{1}{m_1} A_1^* A_1 P_T \text{sgn}(x),$$

and observe that it is perfect on the average since  $\mathbb{E} v_1 = P_T \text{sgn}(x) = \text{sgn}(x)$ . With finite sampling, we will not find ourselves at  $\text{sgn}(x)$  and, therefore, the next step should approximate  $P_T (\text{sgn}(x) - v_1)$ , and read

$$v_2 = v_1 + \frac{1}{m_2} A_2^* A_2 P_T (\text{sgn}(x) - v_1).$$

Continuing this procedure gives the golfing scheme, which stops when  $v_j$  is sufficiently close to the target. This reminds us of a golfer taking a sequence of shots to eventually put his ball in the hole, hence the name. This also has the flavor of an iterative numerical scheme for computing the ansatz (4.1), however, with a significant difference: at each step we use a fresh set of sampling vectors to compute the next iterate.

Set  $q_j = P_T (\text{sgn}(x) - v_j)$  and observe the recurrence relation

$$q_j = \left( I_T - \frac{1}{m_j} P_T A_j^* A_j P_T \right) q_{j-1}.$$

If the block sizes are large enough so that  $\|I_T - m_j^{-1} P_T A_j^* A_j P_T\|_{S^\infty} \leq 1/2$  (this is again the property that the empirical covariance matrix does not deviate too much from the identity, compare (4.2)), then we see that the size of the error decays exponentially to zero since it is at least halved at each iteration.<sup>5</sup> We now examine the size of  $v$  on  $T^\perp$ , that is, outside of the support of  $x$ , and compute

$$v = \sum_{j=1}^{\ell} \frac{1}{m_j} A_j^* A_j q_{j-1}.$$

<sup>5</sup>Writing  $H_j = I_T - m_j^{-1} P_T A_j^* A_j P_T$ , note that we do not require that  $\|H_j\|_{S^\infty} \leq 1/2$  with high probability, only that for a fixed vector  $z \in T$ ,  $\|H_j z\|_{\ell_2} \leq \|z\|_{\ell_2}/2$ , since  $H_j$  and  $q_{j-1}$  are independent. This fact allows for smaller block sizes.

The key point is that by construction,  $A_j^* A_j$  and  $q_{j-1}$  are stochastically independent. In a nutshell, conditioned on  $q_{j-1}$ ,  $A_j^* A_j q_{j-1}$  is just a random sum of the form  $\sum_k a_k \langle a_k, q_{j-1} \rangle$  and one can use standard large deviation inequalities to bound the size of each term as follows:

$$\frac{1}{m_j} \|P_T A_j^* A_j q_{j-1}\|_{\ell_\infty} \leq t_j \|q_{j-1}\|_{\ell_2}$$

for some scalars  $t_j > 0$ , with inequality holding with large probability. Such a general strategy along with many other estimates and ideas that we cannot possibly detail in a paper of this scope, eventually yield proofs of the two theorems from Section 2. Gross’ method is very general and useful, although it is generally not as precise as the combinatorial approach.

### 5. Gaussian models

The last decade has seen a considerable literature, which is impressive in its achievement, about the special case where the entries of the matrix  $A$  are i.i.d. real-valued standard normal variables. As a result of this effort, the community now has a very precise understanding of the performance of both  $\ell_1$ - and nuclear-norm minimization in this *Gaussian model*. We wish to note that [62] was the first paper to study the recovery of a low-rank matrix from Gaussian measurements, using ideas from restricted isometries.

The Gaussian model is very different from the Fourier sampling model or the matrix completion problem from Section 1. To illustrate this point, we first revisit the ansatz (4.1). The key point here is that when  $A$  is a Gaussian map,

$$P_{T^\perp} v = A_{T^\perp}^* q, \quad q = A_T (A_T^* A_T)^{-1} \text{sgn}(x),$$

where  $q$  and  $A_{T^\perp}^*$  are independent, no matter what  $T$  is [8].<sup>6</sup> Set  $d_T$  to be the dimension of  $T$  (this is the quantity we called degrees of freedom earlier on). Conditioned on  $q$ ,  $P_{T^\perp} v$  is then distributed as

$$\iota_{T^\perp} g,$$

where  $\iota_{T^\perp}$  is an isometry from  $\mathbb{R}^{n-d_T}$  onto  $T^\perp$  and  $g \sim \mathcal{N}(0, m^{-1} \|q\|_{\ell_2}^2 I)$ . In the sparse recovery setting, this means that conditioned on  $q$ , the nonzero components of  $P_{T^\perp} v$  are i.i.d.  $\mathcal{N}(0, m^{-1} \|q\|_{\ell_2}^2)$ . In addition,

$$\|q\|_{\ell_2}^2 = \langle \text{sgn}(x), (A_T^* A_T)^{-1} \text{sgn}(x) \rangle$$

and classical results in multivariate statistics assure us that up to a scaling factor,  $\|q\|_{\ell_2}^2$  is distributed as an inverse chi-squared random variable with  $m - d_T + 1$  degrees of freedom. From there, it is not hard to establish that just about  $2s \log n$  samples taken from a Gaussian map are sufficient for perfect recovery of an  $s$ -sparse vector. Also, one can show that just about  $3r(n_1 + n_2 - 5/3r)$  samples suffice for an arbitrary rank- $r$  matrix. We refer to [8] for details and results concerning other structured recovery problems.

This section is not about these simple facts. Rather it is about the fact that under Gaussian maps, there are immediate connections between our recovery problem and deep ideas from convex geometry: as we are about to see, these connections enable to push the theory

---

<sup>6</sup>In the Gaussian model,  $A_T^* A_T$  is invertible with probability one as long as  $m$  is greater or equal to the dimension of the linear space  $T$ .

very far. Recall from Section 3 that  $x$  is the unique solution to (2.1) if the null space of  $A$  misses the cone of descent  $\mathcal{C}$ . What makes a Gaussian map special is that its null space is uniformly distributed among the set of all  $(n - m)$ -dimensional subspaces in  $\mathbb{R}^n$ . It turns out that Gordon [45] gave a precise estimate of the probability that a random uniform subspace misses a convex cone. To state Gordon’s result, we need the notion of *Gaussian width* of a set  $\mathcal{K} \subset \mathbb{R}^n$  defined as:

$$w(\mathcal{K}) := \mathbb{E}_g \sup_{z \in \mathcal{K} \cap \mathbb{S}^{n-1}} \langle g, z \rangle,$$

where  $\mathbb{S}^{n-1}$  is the unit sphere of  $\mathbb{R}^n$  and the expectation is taken over  $g \sim \mathcal{N}(0, I)$ . To the best of the author’s knowledge, Rudelson and Vershynin [65] were the first to recognize the importance of Gordon’s result in this context.

**Theorem 5.1** (Gordon’s escape through the mesh lemma, [45]). *Let  $\mathcal{K} \subset \mathbb{R}^n$  be a cone and  $A$  be a Gaussian map. If*

$$m \geq (w(\mathcal{K}) + t)^2 + 1,$$

*then  $\text{null}(A) \cap \mathcal{K} = \{0\}$  with probability at least  $1 - e^{-t^2/2}$ .*

Hence, Gordon’s theorem allows to conclude that slightly more than  $w(\mathcal{C})$  Gaussian measurements are sufficient to recover a signal  $x$  whose cone of descent is  $\mathcal{C}$ . As we shall see later on, slightly fewer than  $w(\mathcal{C})$  would not do the job.

For Theorem 5.1 to be useful we need tools to calculate these widths. One popular way of providing an upper bound on the Gaussian width of a descent cone is via polarity [68, 60, 24, 3, 76]. The *polar cone* to  $\mathcal{C}$  is the set

$$\mathcal{C}^o = \{y : \langle y, z \rangle \leq 0 \text{ for all } z \in \mathcal{C}\},$$

see Figure 5.1 for a representation of a cone and its polar. For us, the cone polar to the cone of descent is the set of all directions  $t \cdot w$  where  $t > 0$  and  $w \in \partial\|x\|$ . With this, convex duality gives

$$w^2(\mathcal{C}) \leq \mathbb{E}_g \min_{z \in \mathcal{C}^o} \|g - z\|_{\ell_2}^2, \tag{5.1}$$

where, once again, the expectation is taken over  $g$ . In words, the right-hand side is the average squared distance between a random Gaussian vector and the cone  $\mathcal{C}^o$ , and is called the *statistical dimension* of the descent cone denoted by  $\delta(\mathcal{C})$  in [3]. (One can check that  $\delta(\mathcal{C}) = \mathbb{E}_g \|\pi_{\mathcal{C}}(g)\|_{\ell_2}^2$  where  $\pi$  is the projection onto the convex cone  $\mathcal{C}$ .) The particular inequality (5.1) appears in [24] but one can trace this method to the earlier works [68, 60].<sup>7</sup>

<sup>7</sup>There is an inequality in the other direction,  $w^2(\mathcal{C}) \leq \delta(\mathcal{C}) \leq w^2(\mathcal{C}) + 1$ .

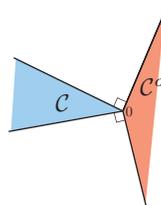


Figure 5.1. Schematic representation of the cone  $\mathcal{C}$  and its polar  $\mathcal{C}^o$ .

The point is that the statistical dimension of  $\mathcal{C}$  is often relatively easy to calculate for some usual norms such as the  $\ell_1$  and nuclear norms, please see [24, 3] for other interesting examples. To make this claim concrete, we compute the statistical dimension of an ‘ $\ell_1$  descent cone’.

Let  $x \in \mathbb{R}^n$  be an  $s$ -sparse vector assumed without loss of generality to have its first  $s$  components positive and all the others equal to zero. We have seen that  $\partial\|x\|_{\ell_1}$  is the set of vectors  $w \in \mathbb{R}^n$  obeying  $w_i = 1$ , for all  $i \leq s$  and  $|w_i| \leq 1$  for  $i > s$ . Therefore,

$$\delta(\mathcal{C}) = \inf_{t \geq 0} \left\{ \sum_{j \leq s} \mathbb{E}(g_j - t)^2 + \sum_{j > s} \mathbb{E}(|g_j| - t)_+^2 \right\}, \tag{5.2}$$

where  $a_+ := \max(a, 0)$ . Using  $t = 2 \log(n/s)$  in (5.2) together with some algebraic manipulations yield

$$\delta(\mathcal{C}) \leq 2s \log(n/s) + 2s$$

as shown in [24]. Therefore, just about  $2s \log(n/s)$  Gaussian samples are sufficient to recover an  $s$ -sparse signal by  $\ell_1$  minimization.

A beautiful fact is that the statistical dimension provides a sharp transition between success and failure of the convex program (2.1), as made very clear by the following theorem taken from Amelunxen, Lotz, McCoy and Tropp (please also see related works from Stojnic [71, 70, 69]).

**Theorem 5.2** (Theorem II in [3]). *Let  $x^* \in \mathbb{R}^n$  be a fixed vector,  $\|\cdot\|$  a norm, and  $\delta(\mathcal{C})$  be the cone of descent at  $x^*$ . Suppose  $A$  is a Gaussian map and let  $y = Ax^*$ . Then for a fixed tolerance  $\varepsilon \in (0, 1)$ ,*

$$\begin{aligned} m \leq \delta(\mathcal{C}) - a_\varepsilon \sqrt{n} &\implies (2.1) \text{ succeeds with probability } \leq \varepsilon; \\ m \geq \delta(\mathcal{C}) + a_\varepsilon \sqrt{n} &\implies (2.1) \text{ succeeds with probability } \geq 1 - \varepsilon. \end{aligned}$$

The quantity  $a_\varepsilon = \sqrt{8 \log(4/\varepsilon)}$ .

In other words, there is a phase transition of width at most a constant times root  $n$  around the statistical dimension. Later in this section, we discuss some history behind this result.

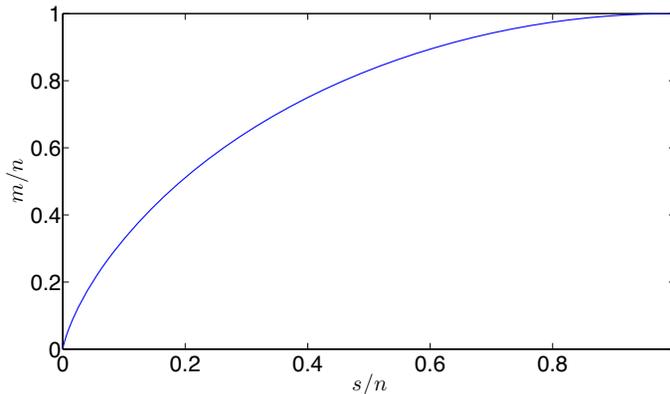


Figure 5.2. The curve  $\psi(\rho)$ .

It is possible to develop accurate estimates of the statistical dimension in case of the  $\ell_1$  norm and the nuclear norm. In the case of the  $\ell_1$  norm, [3, Proposition 4.5] shows that the statistical dimension of the  $\ell_1$  descent cone at an  $s$ -sparse point obeys

$$\psi(s/n) - \frac{2}{\sqrt{s \cdot n}} \leq \frac{\delta(\mathcal{C})}{n} \leq \psi(s/n).$$

The function  $\psi : [0, 1] \rightarrow [0, 1]$  is known and shown in Figure 5.2; we introduce it by making a connection to estimation theory. Let  $z \sim \mathcal{N}(\mu, 1)$  and consider the soft-thresholding rule defined as

$$\eta(z; \lambda) = \begin{cases} z - \lambda, & z > \lambda, \\ 0, & |z| \leq \lambda, \\ z + \lambda, & z < -\lambda. \end{cases}$$

Define its risk or mean-square error at  $\mu$  (when the mean of  $z$  is equal to  $\mu$ ) as

$$r(\mu, \lambda) = \mathbb{E}(z - \mu)^2.$$

Then with  $r(\infty, \lambda) = \lim_{\mu \rightarrow \infty} r(\mu, \lambda) = (1 + \lambda^2)$ ,

$$\psi(\rho) = \inf_{\lambda \geq 0} \{ \rho \cdot r(\infty, \lambda) + (1 - \rho) \cdot r(0, \lambda) \}.$$

Hence, the statistical dimension is nearly equal to the total mean-square error one would get by applying a coordinate-wise soft-thresholding rule, with the best parameter  $\lambda$ , to the entries of a Gaussian vector  $z \sim \mathcal{N}(\mu, I)$ , where  $\mu \in \mathbb{R}^n$  is structured as follows: it has a fraction  $\rho$  of its components set to infinity while all the others are set to zero. For small values of  $s$ , the statistical dimension is approximately equal to  $2s \log(n/s)$  and equal to the leading order term in the calculation from [24] we presented earlier. This value, holding when  $s$  is small compared to  $n$  is also close to the  $2s \log n$  bound given by the ansatz.

There has been much work over the last few years with the goal of characterizing as best as possible the phase transition from Theorem 5.2. As far as the author knows, the transition curve  $\psi$  first appears in the work of Donoho [33] who studied the recovery problem in an asymptotic regime, where both the ambient dimension  $n$  and the number of samples  $m$  tend to infinity in a fixed ratio. He refers to this curve as the *weak* threshold. Donoho's approach relies on the polyhedral structure of the  $\ell_1$  ball known as the cross-polytope in the convex geometry literature. A signal  $x$  with a fixed support of size  $s$  and a fixed sign pattern belongs to a face  $\mathcal{F}$  of dimension  $s - 1$ . The projection of the cross-polytope—its image through the Gaussian map—is a polytope and it is rather elementary to see that  $\ell_1$  minimization recovers  $x$  (and any signal in the same face) if the face is conserved, i.e. if the image of  $\mathcal{F}$  is a face of the projected polytope. Donoho [33] and Donoho and Tanner [31] leveraged pioneering works by Vershik and Sporyshev and by other authors on polytope-angle calculations to understand when low-dimensional faces are conserved; they established that the curve  $\psi$  asymptotically describes a transition between success and failure (we forgo some subtleties cleared in [3]). [31] as well as related works [38] also study projections conserving all low-dimensional faces (the *strong* threshold).

One powerful feature about the approach based on Gaussian process theory described above, is that it is not limited to polytopes. Stojnic [68] used Gordon's work to establish empirically sharp lower bounds for the number of measurements required for the  $\ell_1$ -norm problem. These results are asymptotic in nature and improve, in some cases, on earlier

works. Oymak and Hassibi [60] used these ideas to give bounds on the number of measurements necessary to recover a low-rank matrix in the Gaussian model, see also [63]. In the square  $n \times n$  case, for small rank, simulations in [60] show that about  $4nr$  measurements may suffice for recovery (recall that the ansatz gives a nonasymptotic bound of about  $6nr$ ). Chandrasekaran, Recht, Parrilo and Willsky [24] derived the first precise non-asymptotic bounds, and demonstrated how applicable the Gaussian process theory really is. Amelunxen et al. [3] bring definitive answers, and in some sense, this work represents the culmination of all these efforts, even though some nice surprises continue to come around, see [42] for example. Finally, heuristic arguments from statistical physics can also explain the phase transition at  $\psi$ , see [36]. These heuristics have been justified rigorously in [5].

## 6. How broad is this?

Applications of sparse signal recovery techniques are found everywhere in science and technology. These are mostly well known and far too numerous to review. Matrix completion is a newer topic, which also comes with a very rich and diverse set of applications in fields ranging from computer vision [74] to system identification in control [57], multi-class learning in data analysis [1], global positioning—e.g. of sensors in a network—from partial distance information [7], and quantum-state tomography [48]. The list goes on and on, and keeps on growing. As the theory and numerical tools for matrix completion develop, new applications are discovered, which in turn call for even more theory and algorithms... Our purpose in this section is not to review all these applications but rather to give a sense of the breadth of the mathematical ideas we have introduced thus far; we hope to achieve this by discussing two examples from the author's own work.

**Phase retrieval.** Our first example concerns the fundamental phase retrieval problem, which arises in many imaging problems for the simple reason that photographic plates, CCDs and other light detectors can only measure the intensity of an electromagnetic wave as opposed to measuring its phase. For instance, consider X-ray crystallography, which is a well-known technique for determining the atomic structure of a crystal: there, a collimated beam of X-rays strikes a crystal; these rays then get diffracted by the crystal or sample and the intensity of the diffraction pattern is recorded. Mathematically, if  $x(t_1, t_2)$  is a discrete two-dimensional object of interest, then to cut a long story short, one essentially collects data of the form

$$y(\omega_1, \omega_2) = \left| \sum_{t_1, t_2}^{n-1} x(t_1, t_2) e^{-i2\pi(\omega_1 t_1 + \omega_2 t_2)} \right|^2, \quad (\omega_1, \omega_2) \in \Omega, \quad (6.1)$$

where  $\Omega$  is a sampled set of frequencies in  $[0, 1]^2$ . The question is then how one can invert the Fourier transform from phaseless measurements. Or equivalently, how can we infer the phase of the diffraction pattern when it is completely missing? This question arises in many fields ranging from astronomical imaging to speech analysis and is, therefore, of significance.

While it is beyond the scope of this paper to review the immense literature on phase retrieval, it is legitimate to ask in which way this is related to the topics discussed in this paper. After all, the abstract formulation of the phase retrieval problem asks us to solve a

system of quadratic equations,

$$y_k = |\langle a_k, x \rangle|^2, \quad k = 1, \dots, m, \quad (6.2)$$

in which  $x$  is an  $n$ -dimensional complex or real-valued object; this is (6.1) with the  $a_k$ 's being trigonometric exponentials. This is quite different from the underdetermined linear systems considered thus far. In passing, solving quadratic equations is known to be notoriously difficult (NP-hard) [6, Section 4.3].

As it turns out, the phase retrieval problem can be cast as a matrix completion problem [10], see also [22] for a similar observation in a different setup. To see this, introduce the  $n \times n$  positive semidefinite Hermitian matrix variable  $X \in \mathcal{S}^{n \times n}$  equal to  $xx^*$ , and observe that

$$|\langle a_k, x \rangle|^2 = \text{tr}(a_k a_k^* x x^*) = \text{tr}(A_k X), \quad A_k = a_k a_k^*. \quad (6.3)$$

By lifting the problem into higher dimensions, we have turned quadratic equations into linear ones! Suppose that (6.2) has a solution  $x_0$ . Then there obviously is a rank-one solution to the linear equations in (6.3), namely,  $X_0 = x_0 x_0^*$ . Thus the phase retrieval problem is equivalent to finding a rank-one matrix from linear equations of the form  $y_k = \text{tr}(a_k a_k^* X)$ . This is a rank-one matrix completion problem! Since the nuclear norm of a positive definite matrix is equal to the trace, the natural convex relaxation called *PhaseLift* in [10] reads:

$$\text{minimize } \text{tr}(X) \quad \text{subject to } X \succeq 0, \quad \text{tr}(a_k a_k^* X) = y_k, \quad k \in [m]. \quad (6.4)$$

Similar convex relaxations for optimizing quadratic objectives subject to quadratic constraints are known as Schor's semidefinite relaxations, see [6, Section 4.3] and [44] on the MAXCUT problem from graph theory. The reader is also encouraged to read [80] to learn about another convex relaxation.

Clearly, whatever the sampling vectors might be, we are very far from the Gaussian maps studied in the previous section.<sup>8</sup> Yet, a series of recent papers have established that *PhaseLift* succeeds in recovering the missing phase of the data (and, hence, in reconstructing the signal) in various stochastic models of sampling vectors, ranging from highly structured Fourier-like models to unstructured Gaussian-like models. In fact, the next theorem shows an even stronger result than necessary for *PhaseLift*, namely, that there is only one matrix satisfying the feasibility conditions of (6.4) and, therefore, *PhaseLift* must recover  $x_0 x_0^*$  exactly with high probability.

**Theorem 6.1.** *Suppose the  $a_k$ 's are independent random vectors uniformly distributed on the sphere—equivalently, independent complex-valued Gaussian vectors—and let  $\mathcal{A} : \mathbb{C}^{n \times n} \rightarrow \mathbb{R}^m$  be the linear map  $\mathcal{A}(X) = \{\text{tr}(a_k a_k^* X)\}_{1 \leq k \leq m}$ . Assume that*

$$m \geq c_0 n, \quad (6.5)$$

where  $c_0$  is a sufficiently large constant. Then the following holds with probability at least  $1 - O(e^{-\gamma m})$ : for all  $x_0$  in  $\mathbb{C}^n$ , the feasibility problem

$$\{X : X \succeq 0 \text{ and } \mathcal{A}(X) = \mathcal{A}(x_0 x_0^*)\}$$

has a unique point, namely,  $x_0 x_0^*$ . Thus, with the same probability, *PhaseLift* recovers any signal  $x_0 \in \mathbb{C}^n$  up to a global sign factor.

---

<sup>8</sup>Under a Gaussian map, a sample is of the form  $\langle W, X \rangle$ , where  $W$  is a matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries.

This theorem states that a convenient convex program—a semidefinite program (SDP)—can recover any  $n$ -dimensional complex vector from on the order of  $n$  randomized quadratic equations. The first result of this kind appeared in [18]. As stated, Theorem 6.1 theorem can be found in [11], see also [29]. Such results are not consequences of the general theorems we presented in Section 2.

Of course the sampling vectors from Theorem 6.1 are not useful in imaging applications. However, a version of this result holds more broadly. In particular, [13] studies a physically realistic setup where one can modulate the signal of interest and then collect the intensity of its diffraction pattern, each modulation thereby producing a sort of *coded diffraction pattern*. To simplify our exposition, in one dimension we would collect the pattern

$$y(\omega) = \left| \sum_{t=0}^{n-1} x(t)d(t)e^{-i2\pi\omega t/n} \right|^2, \quad \omega = 0, 1, \dots, n-1, \quad (6.6)$$

where  $d := \{d(t)\}$  is a code or modulation pattern with random entries. This can be achieved by masking the object we wish to image or by modulating the incoming beam. Then [13] shows mathematically and empirically that if one collects the intensity of a few diffraction patterns of this kind, then the solution to PhaseLift is exact.

In short, convex programming techniques and matrix completion ideas can be brought to bear, with great efficiency, on highly nonconvex quadratic problems.

**Robust PCA.** We now turn our attention to a problem in data analysis. Suppose we have a family of  $n$  points belonging to a high-dimensional space of dimension  $d$ , which we regard as the columns of a  $d \times n$  matrix  $M$ . Many data analysis procedures begin by reducing the dimensionality by projecting each data point onto a lower dimensional subspace. *Principal component analysis* (PCA) [51] achieves this by finding the matrix  $X$  of rank  $k$ , which is closest to  $M$  in the sense that it solves:

$$\text{minimize } \|M - X\| \quad \text{subject to } \text{rank}(X) \leq k,$$

where  $\|\cdot\|$  is either the Frobenius or the usual spectral norm. The solution is given by truncating the singular value decomposition as to retain the  $k$  largest singular values. When our data points are well clustered along a lower dimensional plane, this technique is very effective.

In many real applications, however, many entries of the data matrix are typically either unreliable or missing: entries may have been entered incorrectly, sensors may have failed, occlusions in image data may have occurred, and so on. The problem is that PCA is very sensitive to outliers and few errors can throw the estimate of the underlying low-dimensional structure completely off. Researchers have long been preoccupied with making PCA robust and we cannot possibly review the literature on the subject. Rather, our intent is again to show how this problem fits together with the themes from this paper.

Imagine we are given a  $d \times n$  data matrix

$$M = L_0 + S_0,$$

where  $L_0$  has low rank and  $S_0$  is sparse. We observe  $M$  but  $L_0$  and  $S_0$  are hidden. The connection with our problem is that we have a low-rank matrix that has been corrupted in possibly lots of places but we have no idea about which entries have been tampered with.

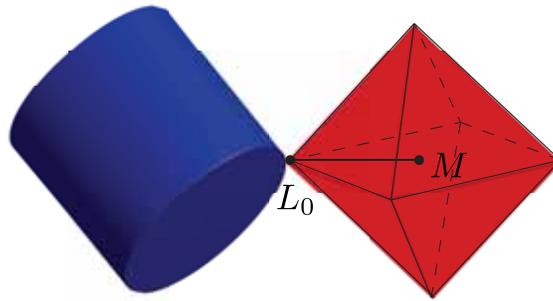


Figure 6.1. Geometry of the robust PCA problem. The blue body is the nuclear ball and the red the  $\ell_1$  ball (cross polytope). Since  $S_0 = M - L_0$ ,  $M - L_0$  is on a low-dimensional face of the cross polytope.

Can we recover the low-rank structure? The idea in [12] is to de-mix the low-rank and the sparse components by solving:

$$\text{minimize } \|L\|_{S_1} + \lambda \|S\|_{\ell_1} \quad \text{subject to } M = L + S; \quad (6.7)$$

here,  $\lambda$  is a positive scalar and abusing notation, we write  $\|S\|_{\ell_1} = \sum_{ij} |S_{ij}|$  for the  $\ell_1$  norm of the matrix  $S$  seen as an  $n \times d$  dimensional vector. Motivated by a beautiful problem in graphical modeling, Chandresakaran et al. proposed to study the same convex model [25], see also [23]. For earlier connections on  $\ell_1$  minimization and sparse corruptions, see [19, 82, 55]. The surprising result from [12] is that if the low-rank component is incoherent and if the nonzero entries of the sparse components occur at random locations, then (6.7) with  $\lambda = 1/\sqrt{\max(n, d)}$  recovers  $L_0$  and  $S_0$  perfectly! To streamline our discussion, we sketch the statement of Theorem 1.1 from [12].<sup>9</sup>

**Theorem 6.2** (Sketch of Theorem 1.1 in [12]). *Assume without loss of generality that  $n \geq d$ , and let  $L_0$  be an arbitrary  $n \times d$  matrix with coherence  $\mu(L_0)$  as defined in Section 2. Suppose that the support set of  $S_0$  is uniformly distributed among all sets of cardinality  $m$ . Then with probability at least  $1 - O(n^{-10})$  (over the choice of support of  $S_0$ ),  $(L_0, S_0)$  is the unique solution to (6.7) with  $\lambda = 1/\sqrt{n}$ , provided that*

$$\text{rank}(L_0) \leq C_0 \cdot d \cdot \mu(L_0)^{-1} (\log n)^{-2} \quad \text{and} \quad m \leq C'_0 \cdot n \cdot d. \quad (6.8)$$

Above,  $C_0$  and  $C'_0$  are positive numerical constants.

Hence, if a positive fraction of the entries from an incoherent matrix of rank at most a constant times  $d/\log^2 n$  are corrupted, the convex program (6.7) will detect those alterations and correct them automatically. In addition, the article [12] presents analog results when entries are both missing and corrupted but we shall not discuss such extensions here. For further results, see [50, 54] and [25] for a deterministic analysis.

Figure 6.1 shows the geometry underlying the exact de-mixing. The fact that we need incoherence should not be surprising. Indeed if  $L_0$  is a rank-1 matrix with one row equal to  $x$

<sup>9</sup>Technically, [12] requires the additional technical assumption discussed in Section 2 although it is probably un-necessary thanks to the sharpening from Li and the author, and from [27] discussed earlier.

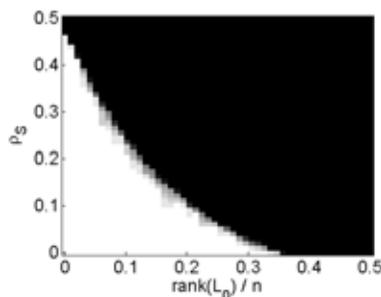


Figure 6.2. Fraction of correct recoveries across 10 trials, as a function of  $\text{rank}(L_0)$  (x-axis) and sparsity of  $S_0$  (y-axis). Here,  $n = d = 400$ . In all cases,  $L_0 = XY^*$  is a product of independent  $n \times r$  i.i.d.  $\mathcal{N}(0, 1/n)$  matrices, and  $\text{sgn}(S_0)$  is random. Trials are considered successful if  $\|\hat{L} - L_0\|_F / \|L_0\|_F < 10^{-3}$ . A white pixel indicates 100% success across trials, a black pixel 0% success, and a gray pixel some intermediate value.

and all the others equal to  $y$ , there is no way any algorithm can detect and recover corruptions in the  $x$  vector.

Finally, Figure 6.2 from [12] shows the practical performance of the convex programming approach to robust PCA on randomly generated problems: there is a sharp phase transition between success and failure. Looking at the numbers, we see that we can corrupt up until about 22.5% of the entries of a  $400 \times 400$  matrix of rank 40, and about 37.5% of those of a matrix of rank 20.

## 7. Concluding remarks

A paper of this length on a subject of this scope has to make some choices. We have certainly made some, and have consequently omitted to discuss other important developments in the field. Below is a partial list of topics we have not touched.

- We have not presented the theory based on the concept of restricted isometry property (RIP). This theory decouples the ‘stochastic part’ from the ‘deterministic part’. In a nutshell, in the sparse recovery problem, once a sampling matrix obeys a relationship called RIP in [19] of the form (4.2) for all subspaces  $T$  spanned by at most  $2s$  columns of  $A$ , then exact and stable recovery of all  $s$ -sparse signals occur [19, 17]; this is a deterministic statement. For random matrices, the stochastic part of the theory amounts to essentially showing that RIP holds [20, 4, 61]. For the matrix-completion analog, see [62].
- In almost any application the author can think of, signals are never exactly sparse, matrices do not have exactly low rank, and so on. In such circumstances, the solution to (2.1) continue to be accurate in the sense that if a signal is approximately sparse or a matrix has approximately low rank, then the recovered object is close. Ronald DeVore gave a plenary address at the 2006 ICM in Madrid on this topic as the theory started to develop. We refer to his ICM paper [30] as well as [28].
- For the methods we have described to be useful, they need to be robust to noise and measurement errors. There are noise aware variants of (2.1) with excellent empirical

and theoretical estimation properties—sometimes near-optimal. We have been silent about this, although many of the articles cited in this paper will actually contain results of this sort. To give two examples, Theorem 2.1 from Section 2 comes with variants enjoying good statistical properties, see [14]. The PhaseLift approach also comes with optimal estimation guarantees [11].

- We have not discussed algorithmic alternatives to convex programming. For instance, there are innovative greedy strategies, which can also have theoretical guarantees, e.g. under RIP see the works of Needell, Tropp, Gilbert and colleagues [77, 59, 58].

The author is thus guilty of a long string of omissions. However, he hopes to have conveyed some enthusiasm for this rich subject where so much is happening on both the theoretical and practical/empirical sides. Nonparametric structured models based on sparsity and low-rankedness are powerful and flexible and while they may not always be the best models in any particular application, they are often times surprisingly competitive.

**Acknowledgements.** The author gratefully acknowledges support from AFOSR under grant FA9550-09-1-0643, NSF under grant CCF-0963835 and from the Simons Foundation via the Math + X Award. He would like to thank Mahdi Soltanolkotabi and Rina Foygel Barber for their help in preparing this manuscript.

## References

- [1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert., *Low-rank matrix factorization with attributes*, Tech. Report N24/06/MM, Ecole des Mines de Paris, 2006.
- [2] R. Ahlswede and A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inf. Theory **48** (2002), no. 3, 569–579.
- [3] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, *Living on the edge: Phase transitions in convex programs with random data*, 2013.
- [4] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation **28** (2008), no. 3, 253–263.
- [5] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, Information Theory, IEEE Transactions on **57** (2011), no. 2, 764–785.
- [6] A. Ben-Tal and A. Nemirovski, *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, vol. 2, Society for Industrial and Applied Mathematics (SIAM), 2001.
- [7] P. Biswas, T-C. Lian, T-C. Wang, and Y. Ye, *Semidefinite programming based algorithms for sensor network localization*, ACM Trans. Sen. Netw. **2** (2006), no. 2, 188–220.
- [8] E. Candès and B. Recht, *Simple bounds for recovering low-complexity models*, Mathematical Programming **141** (2013), no. 1-2, 577–589.

- [9] E. Candès and J. Romberg, *Sparsity and incoherence in compressive sampling*, Inverse problems **23** (2007), no. 3, 969.
- [10] E. J. Candès, Y. C Eldar, T. Strohmer, and V. Voroninski, *Phase retrieval via matrix completion*, SIAM Journal on Imaging Sciences **6** (2013), no. 1, 199–225.
- [11] E. J. Candès and X. Li, *Solving quadratic equations via Phaselift when there are about as many equations as unknowns*, Foundations of Computational Mathematics (to appear) (2012).
- [12] E. J. Candès, X. Li, Y. Ma, and J. Wright, *Robust principal component analysis?*, Journal of the ACM (JACM) **58** (2011), no. 3, 11.
- [13] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval from coded diffraction patterns*, arXiv:1310.3240 (2013).
- [14] E. J. Candès and Y. Plan, *A probabilistic and riplless theory of compressed sensing*, IEEE Transactions on Information Theory **57** (2011), no. 11, 7235–7254.
- [15] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics **9** (2009), no. 6, 717–772.
- [16] E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, Information Theory, IEEE Transactions on **52** (2006), no. 2, 489–509.
- [17] ———, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on pure and applied mathematics **59** (2006), no. 8, 1207–1223.
- [18] E. J. Candès, T. Strohmer, and V. Voroninski, *Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics **66** (2013), no. 8, 1241–1274.
- [19] E. J. Candès and T. Tao, *Decoding by linear programming*, Information Theory, IEEE Transactions on **51** (2005), no. 12, 4203–4215.
- [20] E. J. Candès and T. Tao, *Near-optimal signal recovery from random projections: Universal encoding strategies?*, Information Theory, IEEE Transactions on **52** (2006), no. 12, 5406–5425.
- [21] ———, *The power of convex relaxation: Near-optimal matrix completion*, Information Theory, IEEE Transactions on **56** (2010), no. 5, 2053–2080.
- [22] A. Chai, M. Moscoso, and G. Papanicolaou, *Array imaging using intensity-only measurements*, Inverse Problems **27** (2011), no. 1.
- [23] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, *Latent variable graphical model selection via convex optimization*, Annals of Statistics **40** (2012), no. 4, 1935–1967.
- [24] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Foundations of Computational Mathematics **12** (2012), no. 6, 805–849.
- [25] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization **21** (2011), no. 2, 572–596.
- [26] S. S. Chen, D. L. Donoho, and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing **20** (1998), no. 1, 33–61.

- [27] Y. Chen, *Incoherence-optimal matrix completion*, arXiv:1310.0154 (2013).
- [28] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and best  $k$ -term approximation*, Journal of the American Mathematical Society **22** (2009), no. 1, 211–231.
- [29] L. Demanet and P. Hand, *Stable optimizationless recovery from phaseless linear measurements*, arXiv:1208.1803 (2012).
- [30] R. DeVore, *Optimal computation*, Proceedings of the International Congress of Mathematicians: Madrid, August 22–30, 2006: invited lectures, 2006, pp. 187–215.
- [31] D. Donoho and J. Tanner, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of the American Mathematical Society **22** (2009), no. 1, 1–53.
- [32] D. L. Donoho, *Compressed sensing*, Information Theory, IEEE Transactions on **52** (2006), no. 4, 1289–1306.
- [33] ———, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete & Computational Geometry **35** (2006), no. 4, 617–652.
- [34] D. L. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*, Information Theory, IEEE Transactions on **47** (2001), no. 7, 2845–2862.
- [35] D. L. Donoho and B. F. Logan, *Signal recovery and the large sieve*, SIAM Journal on Applied Mathematics **52** (1992), no. 2, 577–591.
- [36] D. L. Donoho, A. Maleki, and A. Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences **106** (2009), no. 45, 18914–18919.
- [37] D. L. Donoho and P. B. Stark, *Uncertainty principles and signal recovery*, SIAM Journal on Applied Mathematics **49** (1989), no. 3, 906–931.
- [38] D. L. Donoho and J. Tanner, *Counting the faces of randomly-projected hypercubes and orthants, with applications*, Discrete & Computational Geometry **43** (2010), no. 3, 522–541.
- [39] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, Ting Sun, K. F. Kelly, and R. G. Baraniuk, *Single-Pixel Imaging via Compressive Sampling*, Signal Processing Magazine, IEEE **25** (2008), no. 2, 83–91.
- [40] M. Elad and A. M. Bruckstein, *A generalized uncertainty principle and sparse representation in pairs of bases*, Information Theory, IEEE Transactions on **48** (2002), no. 9, 2558–2567.
- [41] M. Elad, J.-L. Starck, D. L. Donoho, and P. Querre, *Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)*, ACHA **19** (2005), no. 3, 340–358.
- [42] R. Foygel and L. Mackey, *Corrupted sensing: Novel guarantees for separating structured signals*, Information Theory, IEEE Transactions on **60** (2014), no. 2, 1223–1247.
- [43] A. Gilbert, S. Muthukrishnan, and M. Strauss, *Improved time bounds for near-optimal sparse Fourier representations*, Optics & Photonics 2005, International Society for Optics and Photonics, 2005, pp. 59141A–59141A.
- [44] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, Journal of the ACM (JACM) **42** (1995), no. 6, 1115–1145.

- [45] Y. Gordon, *On milman's inequality and random subspaces which escape through a mesh in  $\mathbb{R}^n$* , Springer, 1988.
- [46] R. Gribonval and M. Nielsen, *Sparse representations in unions of bases*, Information Theory, IEEE Transactions on **49** (2003), no. 12, 3320–3325.
- [47] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, Information Theory, IEEE Transactions on **57** (2011), no. 3, 1548–1566.
- [48] D. Gross, Y. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum-state tomography via compressed sensing*, Physical Review Letters **105** (2010), no. 15.
- [49] B. Hayes, *The best bits: A new technology called compressive sensing slims down data at the source*, American scientist **97** (2009), no. 4, 276.
- [50] D. Hsu, S. M. Kakade, and T. Zhang, *Robust matrix decomposition with sparse corruptions*, Information Theory, IEEE Transactions on **57** (2011), no. 11, 7221–7234.
- [51] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2005.
- [52] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from a few entries*, Information Theory, IEEE Transactions on **56** (2010), no. 6, 2980–2998.
- [53] R. Kueng and D. Gross, *Ripless compressed sensing from anisotropic measurements*, Linear Algebra and its Applications **441** (2014), 110–123.
- [54] X. Li, *Compressed sensing and matrix completion with constant proportion of corruptions*, Constructive Approximation **37** (2013), no. 1, 73–99.
- [55] B. F. Logan, *Properties of high-pass signals*, Ph.D. thesis, Columbia Univ., New York, 1965.
- [56] M. Lustig, D. L. Donoho, and J. M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*, Magn. Reson. Med. **58** (2007), no. 6, 1192–1195.
- [57] M. Mesbahi and G. P. Papavassilopoulos, *On the rank minimization problem over a positive semidefinite linear matrix inequality*, IEEE Transactions on Automatic Control **42** (1997), no. 2, 239–243.
- [58] D. Needell and J. A. Tropp, *Cosamp: Iterative signal recovery from incomplete and inaccurate samples*, Applied and Computational Harmonic Analysis **26** (2009), no. 3, 301–321.
- [59] D. Needell and R. Vershynin, *Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit*, Foundations of computational mathematics **9** (2009), no. 3, 317–334.
- [60] S. Oymak and B. Hassibi, *New null space results and recovery thresholds for matrix rank minimization*, arXiv preprint arXiv:1011.6326 (2010).
- [61] H. Rauhut, *Compressive sensing and structured random matrices*, Theoretical foundations and numerical methods for sparse recovery **9** (2010), 1–92.
- [62] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM review **52** (2010), no. 3, 471–501.
- [63] B. Recht, W. Xu, and B. Hassibi, *Null space conditions and thresholds for rank minimization*, Mathematical programming **127** (2011), no. 1, 175–202.

- [64] M. Rudelson, *Random vectors in the isotropic position*, J. Funct. Anal. **164** (1999), no. 1, 60–72.
- [65] M. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Communications on Pure and Applied Mathematics **61** (2008), no. 8, 1025–1045.
- [66] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena **60** (1992), no. 1, 259–268.
- [67] F. Santosa and W. W. Symes, *Linear inversion of band-limited reflection seismograms*, SIAM Journal on Scientific and Statistical Computing **7** (1986), no. 4, 1307–1330.
- [68] M. Stojnic, *Various thresholds for  $\ell$ -optimization in compressed sensing*, (2009).
- [69] ———, *A framework to characterize performance of lasso algorithms*, arXiv:1303.7291 (2013).
- [70] ———, *A performance analysis framework for socp algorithms in noisy compressed sensing*, arXiv:1304.0002 (2013).
- [71] ———, *Upper-bounding  $l_1$ -optimization weak thresholds*, arXiv:1303.7289 (2013).
- [72] H. L. Taylor, S. C. Banks, and J. F. McCoy, *Deconvolution with the  $\ell_1$  norm*, Geophysics **44** (1979), no. 1, 39–52.
- [73] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.
- [74] C. Tomasi and T. Kanade, *Shape and motion from image streams under orthography: a factorization method*, International Journal of Computer Vision **9(2)** (1992), 137–154.
- [75] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, Information Theory, IEEE Transactions on **52** (2006), no. 3, 1030–1051.
- [76] ———, *Convex recovery of a structured signal from independent random linear measurements*, To appear in Sampling Theory, a Renaissance (2014).
- [77] J. A. Tropp and A. C. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, Information Theory, IEEE Transactions on **53** (2007), no. 12, 4655–4666.
- [78] J. Trzasko and A. Manduca, *Highly undersampled magnetic resonance image reconstruction via homotopic-minimization*, Medical Imaging, IEEE Transactions on **28** (2009), no. 1, 106–121.
- [79] M. Vetterli, P. Marziliano, and T. Blu, *Sampling signals with finite rate of innovation*, Signal Processing, IEEE Transactions on **50** (2002), no. 6, 1417–1428.
- [80] I. Waldspurger, A. d’Aspremont, and S. Mallat, *Phase recovery, maxcut and complex semidefinite programming*, arXiv:1206.0102 (2012).
- [81] E. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. of Math. **62** (1955), 548–564.
- [82] J. Wright and Y. Ma, *Dense error correction via-minimization*, Information Theory, IEEE Transactions on **56** (2010), no. 7, 3540–3560.

Departments of Mathematics and of Statistics, Stanford University, CA 94305, USA

E-mail: candes@stanford.edu

# Hyperbolic P.D.E. and Lorentzian geometry

Demetrios Christodoulou

**Abstract.** Recent developments are discussed which deepen our understanding of the relationship between hyperbolic p.d.e. and Lorentzian geometry. These developments are connected with progress in the analysis of the Einstein equations of general relativity and in the analysis of the Euler equations of the mechanics of compressible fluids.

**Mathematics Subject Classification (2010).** Primary 35L72, 53C50; Secondary 83C35, 83C57, 35L67, 76L05.

**Keywords.** Hyperbolic partial differential equations, Lorentzian geometry, general relativity, fluid mechanics.

## 1. Introduction

As we all know, while Euclidean geometry was developed in ancient times as the science of physical space, it was only in the 18th century, after the development of differential calculus in the 17th century, that partial differential equations were introduced which were consistent with Euclidean geometry, being invariant under the Euclidean group, the simplest such equation being Laplace's equation, the simplest elliptic partial differential equation. In contrast, Lorentzian geometry, the geometry of physical spacetime, was introduced by Minkowski in the early 20th century as the culmination of a development called "special relativity" which started with Lorentz and in which the central role was played by Einstein, as the geometry the invariance group of which, the Lorentz group, is the invariance group of the linear wave equation, the simplest hyperbolic partial differential equation.

As we also all know, Riemann had already introduced, in the middle of the 19th century, Riemannian geometry, the geometry of curved space, inspired by the earlier work of Gauss on the intrinsic geometry of surfaces in Euclidean space. Einstein, after the contribution of Minkowski, seeking a new theory of gravity which would be consistent with special relativity, was led to the more general conception of Lorentzian geometry as the geometry of curved spacetime, bearing the same relation to Minkowski's geometry as Riemann's geometry bears to Euclidean geometry, the spacetime curvature corresponding to the Newtonian tidal gravitational force. Moreover Einstein formulated the laws of the new theory, called "general relativity", the Einstein equations. These constitute a nonlinear system of partial

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

differential equations of hyperbolic type, in fact the most fundamental such system as it governs the geometry of spacetime itself. In the absence of matter these equations express the vanishing of the Ricci curvature of the spacetime manifold.

In the present lecture I shall discuss recent developments which deepen our understanding of the relationship between hyperbolic p.d.e. and Lorentzian geometry. Methods for the analysis of hyperbolic p.d.e. have been developed in the recent past which fully exploit the associated Lorentzian geometry. These methods may thus be considered to be in the same spirit as that of geometric analysis of elliptic or parabolic p.d.e. in relation to Riemannian geometry. Since the vacuum Einstein equations of general relativity are completely geometric, it is not surprising that the methods in question would originate in the study of these equations. They in fact originated in my work with Sergiu Klainerman [9] on the stability of the Minkowski spacetime of special relativity the framework of general relativity. Thus, the first part of my lecture shall discuss progress in the analysis of the Einstein equations, while the second part shall discuss progress in the analysis of the Euler equations of compressible fluid flow. (For a treatment of hyperbolicity in the general framework of Euler-Lagrange systems of p.d.e. see [5].)

## 2. The Einstein equations of general relativity

Now, while the vacuum Einstein equations do constitute a p.d.e. system of hyperbolic type, this is only seen if proper account is taken of the action of the diffeomorphism group of the underlying differentiable manifold, that is two metrics must be considered equivalent if they are related by a diffeomorphism. Prior to the work [9] on the stability of Minkowski spacetime it seemed that the only way to proceed with the analysis is to introduce a coordinate system which reduces the equations to a p.d.e. system for the metric components which is hyperbolic in a standard analytic sense. The pioneering work of Y. Choquet-Bruhat [11] on existence and uniqueness of solutions to the Cauchy problem for the Einstein equations was based on the introduction of “harmonic” (more properly “wave”) coordinates, which reduce the Einstein equations to a system of wave equations for the metric components, in fact a system where the principal part is of diagonal form, the wave operator of the metric acting on each metric component considered as a scalar function. The analysis then proceeded through energy estimates associated to scalar wave equations.

In the work [9] we developed a different approach which, being geometric, does not depend on the introduction of a coordinate system. The approach consists of two methods. The first is a method which derives estimates for the spacetime curvature. Instead of considering the Einstein equations themselves, we considered the Bianchi identities in the form which they assume by virtue of the Einstein equations. We then introduced the general concept of a *Weyl field*  $W$  on a 4-dimensional Lorentzian manifold  $(M, g)$  to be a 4-covariant tensorfield with the algebraic properties of the Weyl or *conformal* curvature tensor. Given a Weyl field  $W$  one can define a left dual  ${}^*W$  as well as a right dual  $W^*$ , but as a consequence of the algebraic properties of a Weyl field the two duals coincide.

Moreover,  ${}^*W = W^*$  is also a Weyl field. A Weyl field is subject to equations which are

analogues of Maxwell's equations for the electromagnetic field. These are linear equations, in general inhomogeneous, which we call *Bianchi equations*. They are of the form:

$$\nabla^\alpha W_{\alpha\beta\gamma\delta} = J_{\beta\gamma\delta}$$

the right hand side  $J$ , or more generally any 3-covariant tensorfield with the algebraic properties of the right hand side, we call a *Weyl current*.

These equations seem at first sight to be the analogues of only half of Maxwell's equations, but it turns out that they are equivalent to the equations

$$\nabla_{[\alpha} W_{\beta\gamma]\delta\epsilon} = \epsilon_{\mu\alpha\beta\gamma} J^*{}^\mu{}_{\delta\epsilon}, \quad J^*{}_{\beta\gamma\delta} = \frac{1}{2} J_\beta{}^{\mu\nu} \epsilon_{\mu\nu\gamma\delta}$$

which are analogues of the other half of Maxwell's equations. Here  $\epsilon$  is the volume 4-form of  $(M, g)$ . The fundamental Weyl field is the Riemann curvature tensor of  $(M, g)$ ,  $(M, g)$  being a solution of the vacuum Einstein equations, and in this case the corresponding Weyl current vanishes, the Bianchi equations reducing to the Bianchi identities.

Given a vectorfield  $Y$  and a Weyl field  $W$  or Weyl current  $J$  there is a "variation" of  $W$  and  $J$  with respect to  $Y$ , a modified Lie derivative  $\tilde{\mathcal{L}}_Y W$ ,  $\tilde{\mathcal{L}}_Y J$ , which is also a Weyl field or Weyl current respectively. The modified Lie derivative commutes with duality. The Bianchi equations have certain conformal covariance properties which imply the following. If  $J$  is the Weyl current associated to the Weyl field  $W$  according to the Bianchi equations, then the Weyl current associated to  $\tilde{\mathcal{L}}_Y W$  is the sum of  $\tilde{\mathcal{L}}_Y J$  and a bilinear expression which is on one hand linear in  ${}^{(Y)}\tilde{\pi}$  and its first covariant derivative and other the other hand in  $W$  and its first covariant derivative. Here we denote by  ${}^{(Y)}\tilde{\pi}$  the *deformation tensor* of  $Y$ , namely the trace-free part of the Lie derivative of the metric  $g$  with respect to  $Y$ . This measures the rate of change of the conformal geometry of  $(M, g)$  under the flow generated by  $Y$ . From the fundamental Weyl field, the Riemann curvature tensor of  $(M, g)$ , and a set of vector fields  $Y_1, \dots, Y_n$  which we call *commutation fields*, derived Weyl fields of up to any given order  $m$  are generated by the repeated application of the operators  $\tilde{\mathcal{L}}_{Y_i} : i = 1, \dots, n$ . A basic requirement on the set of commutation fields is that it spans the tangent space to  $M$  at each point. The Weyl currents associated to these derived Weyl fields are then determined by the deformation tensors of the commutation fields.

Given a Weyl field  $W$  there is a 4-covariant tensorfield  $Q(W)$  associated to  $W$ , which is symmetric and trace-free in any pair of indices. It is a quadratic expression in  $W$ , analogous to the Maxwell energy-momentum-stress tensor for the electromagnetic field. We call  $Q(W)$  the *Bel-Robinson tensor* associated to  $W$ , because it coincides with the tensor discovered by Bel and Robinson (see [3]) in the case of the fundamental Weyl field, the Riemann curvature tensor of a solution of the vacuum Einstein equations. The Bel-Robinson tensor has a remarkable positivity property:  $Q(W)(X_1, X_2, X_3, X_4)$  is non-negative for any tetrad  $X_1, X_2, X_3, X_4$  of future directed non-spacelike vectors at a point. Moreover, the divergence of  $Q(W)$  is a bilinear expression which is linear in  $W$  and in the associated Weyl current  $J$ .

Given a Weyl field  $W$  and a triplet of future directed non-spacelike vectorfields  $X_1, X_2, X_3$ , which we call *multiplier fields* we define the *energy-momentum density* vectorfield

$P(W; X_1, X_2, X_3)$  associated to  $W$  and to the triplet  $X_1, X_2, X_3$  by:

$$P(W; X_1, X_2, X_3)^\alpha = -Q(W)_{\beta\gamma\delta}^\alpha X_1^\beta X_2^\gamma X_3^\delta$$

Then the divergence of  $P(W; X_1, X_2, X_3)$  is the sum of  $-(\operatorname{div}Q(W))(X_1, X_2, X_3)$  and a bilinear expression which is linear in  $Q(W)$  and in the deformation tensors of  $X_1, X_2, X_3$ .

The divergence theorem in spacetime, applied to a domain which is a development of part of the initial hypersurface, then expresses the integral of the 3-form dual to  $P(W; X_1, X_2, X_3)$  on the future boundary of this domain, in terms of the integral of the same 3-form on the past boundary of the domain, namely on the part of the initial hypersurface, and the spacetime integral of the divergence. The boundaries being *achronal* - that is, no pair of points on each boundary can be joined by a timelike curve - the integrals are integrals of non-negative functions, by virtue of the positivity property of  $Q(W)$ .

For the set of Weyl fields of order up to  $m$  which are derived from the fundamental Weyl field, the Riemann curvature tensor of  $(M, g)$ , the divergences are determined by the deformation tensors of the commutation fields and their derivatives up to order  $m$ , and by the deformation tensors of the multiplier fields. And the integrals on the future boundary give control of all the derivatives of the curvature up to order  $m$ . This is how estimates for the spacetime curvature are obtained, once a suitable set of multiplier fields and a suitable set of commutation fields have been provided.

This is precisely where a second method comes in. This method constructs the required sets of vectorfields by using the geometry of the two parameter foliation of spacetime by the level sets of two functions. These two functions, in the first realization of this method, where the *time function*  $t$ , the level sets of which are maximal spacelike hypersurfaces  $H_t$  of vanishing total linear momentum, and the *optical function*  $u$ , which we may think of as "retarded time", the level sets of which are outgoing null hypersurfaces  $C_u$ . These were chosen so that density of the foliation of each  $H_t$  by the traces of the  $C_u$ , that is, by the surfaces of intersection  $S_{t,u} = H_t \cap C_u$ , which are diffeomorphic to  $S^2$ , tends to 1 as  $t \rightarrow \infty$ .

It was clear that the two functions did not enter on equal footing. The optical function  $u$  played a much more important role. This is due to the fact that the problem involved outgoing waves reaching future null infinity, and it is the outgoing family of null hypersurfaces  $C_u$  which follow these waves. The role of the family of maximal spacelike hypersurfaces  $H_t$  was to obtain a suitable family of sections of each  $C_u$ , the family  $S_{t,u}$  corresponding to a given  $u$ , and to serve as a means by which, in the proof of the existence theorem, the method of continuity can be applied. The work [9] was the first study of the global geometry of null hypersurfaces and of the geometry of foliations by null hypersurfaces. The main features of this study, which is the essence of the second method, shall be discussed below in the connection with another, more recent, work. To complete the present discussion of the original approach, let it suffice to say that the geometric quantities describing the two parameter foliation of spacetime by the surfaces  $S_{t,u}$  were estimated in terms of the spacetime curvature. This yielded estimates for the deformation tensors of the multiplier fields and the commutation fields in terms of the spacetime curvature, thus connecting with the first method.

The main theorem of the work [9] on the stability of Minkowski spacetime asserted that any asymptotically flat initial data for the Einstein vacuum equations which is suitably close in a certain sense to the trivial initial data gives rise to a geodesically complete spacetime, solution of the vacuum Einstein equations, which tends to the Minkowski spacetime along any geodesic, as the affine parameter tends to infinity. Moreover the work gave a detailed analysis of the precise asymptotic behavior of the solutions at infinity (see also [6]). It should be noted that prior to that work the only known geodesically complete solution of the vacuum Einstein equations arising from asymptotically flat initial data had been the Minkowski spacetime itself. The stability theorem has since been extended to a more general class of initial data by L. Bieri, and to the Einstein-Maxwell equations by N. Zipser (see [4]).

I now turn to a more recent work in general relativity, my work [7] on the formation of trapped surfaces. The concept of a *trapped surface* was introduced by Penrose in 1965 (see [17]). He defined a trapped surface to be a closed spacelike surface in spacetime, such that an infinitesimal virtual displacement of the surface along either family of future-directed null geodesic normals to the surface leads to a pointwise decrease of the area element. On the basis of this concept, Penrose proved an *incompleteness* theorem. In the light of subsequent work, namely the uniqueness theorem of the maximal development of given initial data by Choquet-Bruhat and Geroch [12], and the work of Rendall [18] on the characteristic initial value problem, the incompleteness theorem of Penrose may be restated as follows:

*Let us be given regular characteristic initial data on a complete null geodesic cone  $C_o$  of a point  $o$ . Let  $(M^*, g)$  be the maximal future development of the data on  $C_o$ . Suppose that  $M^*$  contains a trapped surface. Then  $(M^*, g)$  is future null geodesically incomplete.*

Now, this theorem presupposes the presence of a trapped surface. A major challenge since the formulation of this theorem had been to develop methods of analysis which, even when the initial conditions are arbitrarily dispersed, hence arbitrarily far from already containing trapped surfaces, allow us to follow the long time evolution and show that, under suitable circumstances, trapped surfaces eventually form. This challenge was met by the work [7] which I shall now discuss. This work, like the work [9], is in the context of the vacuum Einstein equations.

The simplest to state version of the theorem which the monograph [7] establishes is the limiting version, where we have an asymptotic characteristic initial value problem with initial data at past null infinity. Denoting by  $\underline{u}$  the “advanced time”, it is assumed that the initial data are trivial for  $\underline{u} \leq 0$ .

*Let  $k, l$  be positive constants,  $k > 1$ ,  $l < 1$ . Let us be given smooth asymptotic initial data at past null infinity which is trivial for advanced time  $\underline{u} \leq 0$ . Suppose that the incoming energy per unit solid angle in each direction in the advanced time interval  $[0, \delta]$  is not less than  $k/8\pi$ . Then if  $\delta$  is suitably small, the maximal development of the data contains a trapped surface  $S$  which is diffeomorphic to  $S^2$  and has area*

$$\text{Area}(S) \geq 4\pi l^2$$

We remark that by virtue of the scale invariance of the vacuum Einstein equations, the

theorem holds with  $k$ ,  $l$ , and  $\delta$ , replaced by  $ak$ ,  $al$ , and  $a\delta$ , respectively, for any positive constant  $a$ .

The above theorem is obtained through a theorem in which the initial data is given on a complete future null geodesic cone  $C_o$ . The generators of the cone are parametrized by an affine parameter  $s$  measured from the vertex  $o$  and defined so that the corresponding null geodesic vectorfield has projection  $T$  at  $o$  along a fixed unit future-directed timelike vector  $T$  at  $o$ . It is assumed that the initial data are trivial for  $s \leq r_0$ , for some  $r_0 > 1$ . The boundary of this trivial region is then a round sphere of radius  $r_0$ . The advanced time  $\underline{u}$  is then defined along  $C_o$  by

$$\underline{u} = s - r_0$$

The formation of trapped surfaces theorem is similar in this case, the only difference being that the “incoming energy per unit solid angle in each direction in the advanced time interval  $[0, \delta]$ ”, a notion defined only at past null infinity, is replaced by the integral

$$\frac{r_0^2}{8\pi} \int_0^\delta e \, d\underline{u}$$

on the affine parameter segment  $[r_0, r_0 + \delta]$  of each generator of  $C_o$ . The function  $e$  is an invariant of the conformal intrinsic geometry of  $C_o$ , given by:

$$e = \frac{1}{2} |\hat{\chi}|_{\hat{g}}^2$$

where  $\hat{g}$  is the induced metric on the sections of  $C_o$  corresponding to constant values of the affine parameter, and  $\hat{\chi}$  is the *shear* of these sections, the trace-free part of their 2nd fundamental form relative to  $C_o$ .

The theorem for a cone  $C_o$  is established for any  $r_0 > 1$  and the smallness condition on  $\delta$  is independent of  $r_0$ . The domain of dependence, in the maximal development, of the trivial region in  $C_o$  is a domain in Minkowski spacetime bounded in the past by the trivial part of  $C_o$  and in the future by  $\underline{C}_e$ , the past null geodesic cone of a point  $e$  at arc length  $2r_0$  along the timelike geodesic  $\Gamma_0$  from  $o$  with tangent vector  $T$  at  $o$ . Considering then the corresponding complete timelike geodesic in Minkowski spacetime, fixing the origin on this geodesic to be the point  $e$ , the limiting form of the theorem is obtained by letting  $r_0 \rightarrow \infty$ , keeping the origin fixed, so that  $o$  tends to the infinite past along the timelike geodesic.

Almost all the monograph [7] is in fact devoted to establishing an *existence theorem* for a development of the initial data which extends far enough into the future so that trapped spheres have eventually a chance to form within this development. This existence theorem contains a wealth of information, which gives us full knowledge of the geometry of spacetime when trapped surfaces begin to form.

The work [7] uses three methods, two of which correspond to the two methods, outlined above, of [9], while the third method, which I call the *short pulse method*, was introduced in [7].

In connection with the work on the stability of Minkowski spacetime, a variant of the second method is obtained if we place in the role of the time function  $t$  another *optical*

function  $\underline{u}$ , which we may think of as “advanced time”, the level sets of which are incoming null hypersurfaces. A two parameter family of surfaces diffeomorphic to  $S^2$  are then obtained, namely the intersections of this incoming family with the outgoing family of null hypersurfaces.

In the work [7] on the formation of trapped surfaces, the roles of the two optical functions are reversed, because we are considering incoming rather than outgoing waves, and it is the incoming null hypersurfaces  $\underline{C}_{\underline{u}}$ , the level sets of  $\underline{u}$ , which follow these waves. However, in this work, taking the other function to be the conjugate optical function  $u$  is not merely a matter of convenience, but it is essential for what we wish to achieve. This is because the  $C_u$ , the level sets of  $u$ , are here, like the initial hypersurface  $C_o$  itself, future null geodesic cones with vertices on the timelike geodesic  $\Gamma_0$ , and the trapped spheres which eventually form are sections  $S_{\underline{u},u} = \underline{C}_{\underline{u}} \cap C_u$  of “late”  $C_u$ , everywhere along which those  $C_u$  have negative expansion.

The *short pulse method* is a method of treating the focusing of incoming waves, and like the second method, it is of wider application. Its point of departure resembles that of the short wavelength or geometric optics approximation, in so far as it depends on the presence of a small length, but thereafter the two approaches diverge. The short pulse method is a method which allows us to establish an existence theorem for a development of the initial data which is large enough so that interesting things have a chance to occur within this development, if a nonlinear system is involved. One may ask at this point: what does it mean for a length to be small in the context of the vacuum Einstein equations? For, the equations are scale invariant. Here *small* means *by comparison to the area radius of the trapped sphere to be formed*.

With initial data on a complete future null geodesic cone  $C_o$ , as explained above, which are trivial for  $s \leq r_0$ , we consider the restriction of the initial data to  $s \leq r_0 + \delta$ . In terms of the advanced time  $\underline{u}$ , we restrict attention to the interval  $[0, \delta]$ , the data being trivial for  $\underline{u} \leq 0$ . The retarded time  $u$  is set equal to  $u_0 = -r_0$  at  $o$  and therefore on  $C_o$ , which is then also denoted  $C_{u_0}$ . Also,  $u - u_0$  is defined along  $\Gamma_0$  to be one half the arc length from  $o$ . This determines  $u$  everywhere. The development whose existence we want to establish is that bounded in the future by the spacelike hypersurface  $H_{-1}$  where  $\underline{u} + u = -1$  and by the incoming null hypersurface  $\underline{C}_{\delta}$ . We denote this development  $M_{-1}$ .

We define  $L$  and  $\underline{L}$  to be the future directed null vectorfields the integral curves of which are the generators of the  $C_u$  and  $\underline{C}_{\underline{u}}$ , parametrized by  $\underline{u}$  and  $u$  respectively, so that

$$Lu = \underline{L}u = 0, \quad L\underline{u} = \underline{L}\underline{u} = 1$$

The flow  $\Phi_\tau$  generated by  $L$  defines a diffeomorphism of  $S_{\underline{u},u}$  onto  $S_{\underline{u}+\tau,u}$ , while the flow  $\underline{\Phi}_\tau$  generated by  $\underline{L}$  defines a diffeomorphism of  $S_{\underline{u},u}$  onto  $S_{\underline{u},u+\tau}$ .

The positive function  $\Omega$  defined by

$$g(L, \underline{L}) = -2\Omega^2$$

may be thought of as the inverse density of the double null foliation. The null geodesic fields  $L', \underline{L}'$  corresponding to  $L, \underline{L}$  are given by:

$$L'^\mu = -2(g^{-1})^{\mu\nu} \partial_\nu u, \quad \underline{L}'^\mu = -2(g^{-1})^{\mu\nu} \partial_\nu \underline{u}$$

These have the same integral curves as  $L, \underline{L}$  respectively, but are affinely parametrized. We have:

$$L' = \Omega^{-2}L, \quad \underline{L}' = \Omega^{-2}\underline{L}, \quad \text{so that } g(L', \underline{L}) = g(\underline{L}', L) = -2$$

We denote by  $\hat{L}$  and  $\hat{\underline{L}}$  the normalized future directed null vectorfields

$$\hat{L} = \Omega^{-1}L, \quad \hat{\underline{L}} = \Omega^{-1}\underline{L}, \quad \text{so that } g(\hat{L}, \hat{\underline{L}}) = -2$$

The *optical structure equations* are the equations satisfied by the *optical quantities*, namely the geometric quantities associated to the double null foliation, that is, the inverse density function  $\Omega$ , the metric  $\not{g}$  induced on the surfaces  $S_{u,u}$  and its Gauss curvature  $K$ , the second fundamental forms  $\chi', \underline{\chi}'$  of  $S_{u,u}$  relative to  $C_u, \underline{C}_u$  respectively, the torsion forms  $\eta, \underline{\eta}$  of  $S_{u,u}$  relative to  $C_u, \underline{C}_u$  respectively, and the functions  $\omega, \underline{\omega}$ , the derivatives of  $\log \Omega$  with respect to  $L, \underline{L}$  respectively. The second fundamental forms  $\chi'$  are defined relative to the geodesic null normals  $L', \underline{L}'$ . They are the symmetric 2-covariant  $S$  tensorfields given by:

$$\chi'(X, Y) = g(\nabla_X L', Y), \quad \underline{\chi}'(X, Y) = g(\nabla_X \underline{L}', Y)$$

for any pair of vectors  $X, Y$  tangent to  $S_{u,u}$  at a point. Then

$$\chi = \Omega\chi', \quad \underline{\chi} = \Omega\underline{\chi}'$$

are the corresponding second fundamental forms relative to the normalized null normals  $\hat{L}, \hat{\underline{L}}$ . By *torsion* of  $S_{u,u}$  we mean the connection of the normal bundle of  $S_{u,u}$ . Since the normal planes are timelike it is natural to refer to a null basis, namely a pair of future directed null vectors whose inner product is -2. The representation of the torsion depends on the choice of null basis sections of the normal bundle. The basis which is natural relative to  $C_u$  is  $(L', \underline{L})$  while the basis which is natural relative to  $\underline{C}_u$  is  $(\underline{L}', L)$ . The torsion forms are then the  $S$  1-forms given by:

$$\eta(X) = \frac{1}{2}g(\nabla_X L', \underline{L}), \quad \underline{\eta}(X) = \frac{1}{2}g(\nabla_X \underline{L}', L)$$

for any vector  $X$  tangent to  $S_{u,u}$  at a point. The torsion may also be represented in the normalized basis  $(\hat{L}, \hat{\underline{L}})$  as the  $S$  1-form  $\zeta$ :

$$\zeta(X) = \frac{1}{2}g(\nabla_X \hat{L}, \hat{\underline{L}}) = -\frac{1}{2}g(\nabla_X \hat{\underline{L}}, \hat{L})$$

We have:

$$\eta = \zeta + \not{d}\log \Omega, \quad \underline{\eta} = -\zeta + \not{d}\log \Omega$$

The curvature of the normal bundle of  $S_{u,u}$  is:

$$\text{curl} \zeta = \text{curl} \eta = -\text{curl} \underline{\eta}$$

The torsion form  $\zeta$  is also the obstruction to integrability of the distribution of timelike planes on the spacetime manifold which are orthogonal to the tangent planes to the  $S_{u,u}$ :

$$[\underline{L}, L] = 4\Omega^2\zeta^\sharp$$

where  $\zeta^\sharp$  is the  $S$  tangential vectorfield corresponding to the  $S$  1-form  $\zeta$  through the metric  $\not{g}$ .

The spacetime curvature corresponding to a solution of the vacuum Einstein equations decomposes relative to a double null foliation into the trace-free symmetric 2-covariant  $S$  tensorfields  $\alpha, \underline{\alpha}$ , given by:

$$\alpha(X, Y) = R(X, \hat{L}, Y, \hat{L}), \quad \underline{\alpha}(X, Y) = R(X, \underline{\hat{L}}, Y, \underline{\hat{L}})$$

for any pair of vectors  $X, Y$  tangent to  $S_{\underline{u}, u}$  at a point, the  $S$  1-forms  $\beta, \underline{\beta}$  given by:

$$R(X, \hat{L}, Y, Z) = \not{g}(X, Y)\beta(Z) - \not{g}(X, Z)\beta(X),$$

$$R(X, \underline{\hat{L}}, Y, Z) = \not{g}(X, Y)\underline{\beta}(Z) - \not{g}(X, Z)\underline{\beta}(Y)$$

for any triplet of vectors  $X, Y, Z$  tangent to  $S_{\underline{u}, u}$  at a point, and the functions  $\rho, \sigma$  given on each  $S_{\underline{u}, u}$  by:

$$R(X, Y, Z, W) = \not{\epsilon}(X, Y)\not{\epsilon}(Z, W), \quad \frac{1}{2}R(X, Y, \hat{L}, \hat{L}) = \sigma\not{\epsilon}(X, Y)$$

for any quadruplet of vectors  $X, Y, Z, W$  tangent to  $S_{\underline{u}, u}$  at a point,  $\not{\epsilon}$  being the area form of  $S_{\underline{u}, u}$ .

In the following, for a covariant  $S$  tensorfield  $\theta$  and  $V$  an arbitrary vectorfield, we denote by  $\not{\mathcal{L}}_V\theta$  the restriction to each  $TS_{\underline{u}, u}$  of  $\mathcal{L}_V\theta$ , another covariant  $S$  tensorfield of the same type as  $\theta$ . The optical structure equations consist of four sets of equations. The first set are the *propagation equations*, which consist of the basic first variation equations

$$\not{\mathcal{L}}_L\not{g} = 2\Omega\chi, \quad \not{\mathcal{L}}_{\underline{L}}\not{g} = 2\Omega\underline{\chi},$$

the second variation equations, which express  $\not{\mathcal{L}}_L\chi', \not{\mathcal{L}}_{\underline{L}}\underline{\chi}'$  in terms of  $\alpha, \underline{\alpha}$  respectively, equations which express  $\not{\mathcal{L}}_L\underline{\eta}, \not{\mathcal{L}}_{\underline{L}}\underline{\eta}$  in terms of  $\beta, \underline{\beta}$  respectively, and equations which express  $L\underline{\omega}, \underline{L}\omega$  in terms of  $\rho$ . The second set of equations are the Codazzi equations of  $S_{\underline{u}, u}$  considered as a section of  $C_u$  and  $\underline{C}_{\underline{u}}$ , systems of p.d.e. on  $S_{\underline{u}, u}$  satisfied by  $\chi'$  and  $\underline{\chi}'$  respectively, and involving  $\beta$  and  $\underline{\beta}$ , respectively. The third set of equations consists of the Gauss equation, which expresses  $K$ , the Gauss curvature of  $S_{\underline{u}, u}$  in terms of  $\rho$ , and an equation which expresses the curvature of the normal bundle of  $S_{\underline{u}, u}$  in terms of  $\sigma$ . Finally, the fourth set of equations are the cross-variation equations. These are equations expressing  $\not{\mathcal{L}}_{\underline{L}}(\Omega\chi)$  in terms of  $\nabla\eta + \tilde{\nabla}\eta$  and  $\rho$  and  $\not{\mathcal{L}}_L(\Omega\underline{\chi})$  in terms of  $\nabla\eta + \tilde{\nabla}\eta$  and  $\rho$ , and equations expressing  $\not{\mathcal{L}}_{\underline{L}}\underline{\eta}$  in terms of  $\not{g}\omega$  and  $\underline{\beta}$  and  $\not{\mathcal{L}}_L\underline{\eta}$  in terms of  $\not{g}\omega$  and  $\beta$ .

Now, the first step in the short pulse method is the analysis of the equations along the initial hypersurface  $C_{u_0}$ . The analysis is particularly clear and simple because of the fact that  $C_{u_0}$  is a null hypersurface, so we are dealing with the characteristic initial value problem and there is a way of formulating the problem in terms of free data which are not subject to any constraints. The full set of data which includes all the curvature components and their transversal derivatives, up to any given order, along  $C_{u_0}$ , is then determined by integrating ordinary differential equations along the generators of  $C_{u_0}$ . The free data may be

described as a 2-covariant symmetric positive definite tensor density  $m$ , of weight -1 and unit determinant, on  $S^2$ , depending on  $\underline{u}$ . This is of the form:

$$m = \exp \psi$$

where  $\psi$  is a 2-dimensional symmetric trace-free matrix valued function on  $S^2$ , depending on  $\underline{u} \in [0, \delta]$ , and transforming under change of charts on  $S^2$  in such a way so as to make  $m$  a 2-covariant tensor density of weight -1. The transformation rule is particularly simple if stereographic charts on  $S^2$  are used. Then there is a function  $O$  defined on the intersection of the domains of the north and south polar stereographic charts on  $S^2$ , with values in the 2-dimensional symmetric orthogonal matrices of determinant -1 such that in going from the north polar chart to the south polar chart or vice-versa,  $\psi \mapsto \tilde{O}\psi O$  and  $m \mapsto \tilde{O}mO$ .

The crucial ansatz of the short pulse method is the following. We consider an arbitrary smooth 2-dimensional symmetric trace-free matrix valued function  $\psi_0$  on  $S^2$ , depending on  $s \in [0, 1]$ , which extends smoothly by 0 to  $s \leq 0$ , and we set:

$$\psi(\underline{u}, \vartheta) = \frac{\delta^{1/2}}{|u_0|} \psi_0\left(\frac{\underline{u}}{\delta}, \vartheta\right), \quad (\underline{u}, \vartheta) \in [0, \delta] \times S^2 \tag{2.1}$$

The analysis of the equations along  $C_{u_0}$  then gives, for the components of the spacetime curvature along  $C_{u_0}$ :

$$\begin{aligned} \sup_{C_{u_0}} |\alpha| &\leq O_2(\delta^{-3/2}|u_0|^{-1}), \quad \sup_{C_{u_0}} |\beta| \leq O_2(\delta^{-1/2}|u_0|^{-2}) \\ \sup_{C_{u_0}} |\rho|, \sup_{C_{u_0}} |\sigma| &\leq O_3(|u_0|^{-3}) \\ \sup_{C_{u_0}} |\underline{\beta}| &\leq O_4(\delta|u_0|^{-4}), \quad \sup_{C_{u_0}} |\underline{\alpha}| \leq O_5(\delta^{3/2}|u_0|^{-5}) \end{aligned} \tag{2.2}$$

Here, the symbol  $O_k(\delta^p|u_0|^r)$  means the product of  $\delta^p|u_0|^r$  with a non-negative non-decreasing continuous function of the  $C^k$  norm of  $\psi_0$  on  $[0, 1] \times S^2$ . The pointwise magnitudes of tensors on  $S_{\underline{u}, u}$  are with respect to the induced metric  $\not{g}$ , which is positive definite, the surfaces being spacelike.

One should focus on the dependence on  $\delta$  of the right hand sides of 2.2. This displays what we may call the *short pulse hierarchy*. And this hierarchy is *nonlinear*. For, if only the linearized form of the equations was considered, a different hierarchy would be obtained: the exponents of  $\delta$  in the first two of 2.2 would be the same, but the exponents of  $\delta$  in the last three of 2.2 would instead be  $1/2, 3/2, 5/2$ , respectively. The short pulse hierarchy is the key to the existence theorem as well as to the trapped surface formation theorem. We must still outline however in what way do we establish that the short pulse hierarchy is preserved in evolution. This is of course the main step of the short pulse method. What we do is to reconsider the first two methods previously outlined in the light of the short pulse hierarchy.

Let us revisit the first method. We take as multiplier fields the vectorfields  $L$  and  $K$ , where

$$K = u^2 \underline{L}$$

As already mentioned above, we take the initial data to be trivial for  $\underline{u} \leq 0$  and as a consequence the spacetime region corresponding to  $\underline{u} \leq 0$  is a domain in Minkowski spacetime.

We may thus confine attention to the nontrivial region  $\underline{u} \geq 0$ . We denote by  $M'_{-1}$  this non-trivial region in  $M_{-1}$ .

For each of the Weyl fields to be specified below, we define the energy-momentum density vectorfields

$${}^{(n)}P(W) : n = 0, 1, 2, 3$$

where:

$$\begin{aligned} {}^{(0)}P(W) &= P(W; L, L, L) & {}^{(1)}P(W) &= P(W; K, L, L) \\ {}^{(2)}P(W) &= P(W; K, K, L) & {}^{(3)}P(W) &= P(W; K, K, K) \end{aligned}$$

As commutation fields we take  $L, S$ , defined by:

$$S = u\underline{L} + \underline{u}L,$$

and the three rotation fields  $O_i : i = 1, 2, 3$ . The latter are defined according to the second method as follows. In the Minkowskian region we introduce rectangular coordinates  $x^\mu : \mu = 0, 1, 2, 3$ , taking the  $x^0$  axis to be the timelike geodesic  $\Gamma_0$ . In the Minkowskian region, in particular on the sphere  $S_{0,u_0}$ , the  $O_i$  are the generators of rotations about the  $x^i : i = 1, 2, 3$  spatial coordinate axes. The  $O_i$  are then first defined on  $C_{u_0}$  by conjugation with the flow of  $L$  and then in spacetime by conjugation with the flow of  $\underline{L}$ .

The Weyl fields which we consider are, besides the fundamental Weyl field  $R$ , the Riemann curvature tensor, the following derived Weyl fields

$$\begin{aligned} \text{1st order: } & \tilde{\mathcal{L}}_L R, \tilde{\mathcal{L}}_{O_i} R : i = 1, 2, 3, \tilde{\mathcal{L}}_S R \\ \text{2nd order: } & \tilde{\mathcal{L}}_L \tilde{\mathcal{L}}_L R, \tilde{\mathcal{L}}_{O_i} \tilde{\mathcal{L}}_L R : i = 1, 2, 3, \tilde{\mathcal{L}}_{O_j} \tilde{\mathcal{L}}_{O_i} R : i, j = 1, 2, 3, \\ & \tilde{\mathcal{L}}_{O_i} \tilde{\mathcal{L}}_S R : i = 1, 2, 3, \tilde{\mathcal{L}}_S \tilde{\mathcal{L}}_S R \end{aligned}$$

We assign to each Weyl field the index  $l$  according to the number of  $\tilde{\mathcal{L}}_L$  operators in the definition of  $W$  in terms of  $R$ . We then define total 2nd order energy-momentum densities

$${}^{(n)}P_2 : n = 0, 1, 2, 3$$

as the sum of  $\delta^{2l} {}^{(n)}P(W)$  over all the above Weyl fields in the case  $n = 3$ , all the above Weyl fields except those whose definition involves the operator  $\tilde{\mathcal{L}}_S$  in the cases  $n = 0, 1, 2$ .

We then define the total 2nd order energies  ${}^{(n)}E_2(u)$  as the integrals on the  $C_u$  and the total 2nd order fluxes  ${}^{(n)}F_2(\underline{u})$  as the integrals on the  $\underline{C}_{\underline{u}}$ , of the 3-forms dual to the  ${}^{(n)}P_2$ . Of the fluxes only  ${}^{(3)}F_2(\underline{u})$  plays a role in the problem. Finally, with the exponents  $q_n : n = 0, 1, 2, 3$  defined by:

$$q_0 = 1, \quad q_1 = 0, \quad q_2 = -\frac{1}{2}, \quad q_3 = -\frac{3}{2},$$

according to the short pulse hierarchy, we define the quantities

$$\mathcal{E}_2^{(n)} = \sup_u \left( \delta^{2q_n} E_2^{(n)}(u) \right) : n = 0, 1, 2, 3; \quad \mathcal{F}_2^{(3)} = \sup_{\underline{u}} \left( \delta^{2q_3} F_2^{(3)}(\underline{u}) \right) \quad (2.3)$$

The objective then is to obtain bounds for these quantities in terms of the initial data.

This requires properly estimating the deformation tensor of  $K$ , as well as the deformation tensors of  $L, S$  and the  $O_i : i = 1, 2, 3$  and their derivatives of up to 2nd order. In view of the definitions of these vectorfields, this reduces to appropriately estimating the optical quantities by analyzing the optical structure equations. In doing this, the short pulse method meshes with the second method previously alluded to.

At this point I shall briefly describe a basic feature of the analysis of the optical structure equations which is at the heart of the second method. This is the consideration of systems consisting of *ordinary differential equations along the generators of the  $C_u$  or the  $\underline{C}_{\underline{u}}$  coupled to elliptic p.d.e. on their  $S_{\underline{u},u}$  sections*. This approach allows us to derive estimates for the optical quantities which are of one order of differentiation higher than those that can be derived on the basis of the propagation equations alone. What makes this possible is that the principal terms in the propagation equations for certain optical quantities vanish by virtue of the vacuum Einstein equations. In the case of  $\chi'$  and  $\underline{\chi}'$  these quantities are simply the traces  $\text{tr}\chi'$  and  $\text{tr}\underline{\chi}'$ , which satisfy the propagation equations:

$$L\text{tr}\chi' = -\Omega^2|\chi'|^2, \quad \underline{L}\text{tr}\underline{\chi}' = -\Omega^2|\underline{\chi}'|^2$$

while the Codazzi equations:

$$\text{div}\chi' - \not\partial\text{tr}\chi' + \chi' \cdot \eta^\sharp - \text{tr}\chi'\eta = \Omega^{-1}\beta,$$

$$\text{div}\underline{\chi}' - \not\partial\text{tr}\underline{\chi}' + \underline{\chi}' \cdot \underline{\eta}^\sharp - \text{tr}\underline{\chi}'\underline{\eta} = \Omega^{-1}\underline{\beta}$$

constitute elliptic systems for the trace-free parts  $\hat{\chi}'$  and  $\hat{\underline{\chi}}'$ , given the traces. In the case of  $\eta, \underline{\eta}$  the appropriate optical quantities are found at one order of differentiation higher. They are the *mass aspect functions*  $\mu$  and  $\underline{\mu}$ , defined by:

$$\mu = K + \frac{1}{4}\text{tr}\chi\text{tr}\chi - \text{div}\eta, \quad \underline{\mu} = K + \frac{1}{4}\text{tr}\chi\text{tr}\chi - \text{div}\underline{\eta}$$

These satisfy the propagation equations:

$$L\mu = -\Omega\text{tr}\chi \left( \mu + \frac{1}{2}\underline{\mu} \right) + \Omega \left( -\frac{1}{4}\text{tr}\chi|\hat{\chi}|^2 + \frac{1}{2}\text{tr}\chi|\underline{\eta}|^2 \right) + \text{div}j,$$

$$\underline{L}\underline{\mu} = -\Omega\text{tr}\chi \left( \underline{\mu} + \frac{1}{2}\mu \right) + \Omega \left( -\frac{1}{4}\text{tr}\chi|\hat{\underline{\chi}}|^2 + \frac{1}{2}\text{tr}\chi|\eta|^2 \right) + \text{div}\underline{j}$$

Here,  $j$  and  $\underline{j}$  are the  $S$  1-forms:

$$j = \Omega(2\hat{\chi} \cdot \eta^\sharp - \text{tr}\chi\eta), \quad \underline{j} = \Omega(2\hat{\underline{\chi}} \cdot \underline{\eta} - \text{tr}\chi\underline{\eta})$$

The elliptic systems are the Hodge systems on  $S_{\underline{u},u}$  obtained by substituting for  $K$  from the Gauss equation in the definitions of  $\mu$  and  $\underline{\mu}$  and adjoining the equation for the curvature of the normal bundle:

$$\begin{aligned} \mathfrak{d}\text{iv} \eta &= \rho + \frac{1}{2}(\hat{\chi}, \hat{\chi}) - \mu, \quad \mathfrak{c}\mathfrak{u}\mathfrak{r}\mathfrak{l} \eta = \sigma - \frac{1}{2}\hat{\chi} \wedge \hat{\chi} \\ \mathfrak{d}\text{iv} \underline{\eta} &= \rho + \frac{1}{2}(\hat{\chi}, \hat{\chi}) - \underline{\mu}, \quad \mathfrak{c}\mathfrak{u}\mathfrak{r}\mathfrak{l} \underline{\eta} = -\sigma + \frac{1}{2}\hat{\chi} \wedge \hat{\chi} \end{aligned}$$

In the case of  $\omega, \underline{\omega}$  the appropriate optical quantities are found at one order of differentiation still higher. They are the functions  $\omega, \underline{\psi}$ , defined by:

$$\psi = \Delta\omega - \mathfrak{d}\text{iv}(\Omega\beta), \quad \underline{\psi} = \Delta\underline{\omega} - \mathfrak{d}\text{iv}(\Omega\underline{\beta})$$

and the elliptic equations for  $\omega, \underline{\omega}$  are simply these definitions themselves.

Now, the estimates of the error integrals, namely the integrals of the absolute values of the divergences of the  $P_2^{(n)}$ , yield inequalities for the quantities 2.3. These inequalities contain, besides the initial data terms

$$D = \delta^{2q_n} E_2^{(n)}(u_0) : n = 0, 1, 2, 3,$$

terms of  $O(\delta^p)$  for some  $p > 0$ , which are innocuous, as they can be made less than or equal to 1 by subjecting  $\delta$  to a suitable smallness condition, *but they also contain terms of  $O(1)$  which are nonlinear in the quantities 2.3.* From such a nonlinear system of inequalities, no bounds can in general be deduced, because here, in contrast to the work on the stability of Minkowski spacetime, the initial data quantities are allowed to be arbitrarily large. However a fortunate circumstance occurs: our system of inequalities is *reductive*. That is, the inequalities, taken in proper sequence, reduce to a sequence of sublinear inequalities, thus allowing us to obtain the sought for bounds.

The existence and trapped surface formation theorems have since been extended by S. Klainerman and I. Rodnianski [14] to a more general class of initial data and by P. Yu [20] to the Einstein-Maxwell equations. Also, J. Li and P. Yu [16] have constructed a class of Cauchy data which is trivial inside a certain sphere, satisfies a short pulse ansatz in an annular region surrounding the former trivial region, and coincides with initial data for a stationary axially symmetric Kerr solution outside a larger spherical surface. The corresponding class of maximal developments then contain a trapped surface and have an exterior region isometric to a Kerr solution. In a most recent development [15], S. Klainerman, J. Luk, and I. Rodnianski, using the original existence theorem, have derived a result which considerably enhances the interest of the original trapped surface formation theorem, as it does not require a lower bound on the incoming energy in all directions. Considering the outer boundary  $\underline{C}_\delta$  of the existence domain  $M_{-1}$  they show that *if in some neighborhood in  $S^2$  of some direction the incoming energy in the directions corresponding to this neighborhood is sufficiently large depending on the angular size of the neighborhood in question, then  $\underline{C}_\delta$  contains a trapped surface.* In this case, even though none of the sections  $S_{\delta,u}$  may be trapped, they show that there is another section of  $\underline{C}_\delta$ , a surface represented as a graph  $u = G(\vartheta)$  over

$S^2$ , which is in fact trapped. This surface reaches large negative values of  $u$  in the directions corresponding to the neighborhood of large incoming energy, but comes near  $S_{\delta, -1+\delta}$ , the future boundary of  $\underline{C}_\delta$  in  $M_{-1}$ , in the antipodal directions.

### 3. The Euler equations of compressible fluid Flow

I now come to the second topic of my lecture, the Euler equations of compressible fluid flow. The mechanics of a perfect fluid are described in the framework of special relativity by a future-directed timelike vectorfield  $u$  of unit magnitude relative to the Minkowski metric  $g$ , the *fluid 4-velocity*, and two positive functions  $n$  and  $s$ , the *number of particles per unit volume* and the *entropy per particle*. The mechanical properties of a perfect fluid are determined once we give an *equation of state*, which expresses the mass-energy density  $\rho$  as a function of  $n$  and  $s$ :

$$\rho = \rho(n, s)$$

According to the laws of thermodynamics, the *pressure*  $p$  and the *temperature*  $\theta$  are then given by:

$$p = n \frac{\partial \rho}{\partial n} - \rho, \quad \theta = \frac{1}{n} \frac{\partial \rho}{\partial s}$$

The *particle current* is the vectorfield  $I$  given by:

$$I^\mu = nu^\mu$$

The *energy-momentum-stress tensor* is the symmetric 2-contravariant tensorfield  $T$  given by:

$$T^{\mu\nu} = (\rho + p)u^\mu u^\nu + p(g^{-1})^{\mu\nu}$$

and the *equations of motion* are the *differential conservation laws*:

$$\nabla_\mu I^\mu = 0, \quad \nabla_\nu T^{\mu\nu} = 0 \quad (3.1)$$

Let us define the *vorticity 2-form*  $\omega$  by:

$$\omega = d\beta \quad (3.2)$$

where  $\beta$  is the 1-form:

$$\beta_\mu = -\sqrt{\sigma}u_\mu, \quad u_\mu = g_{\mu\nu}u^\nu \quad (3.3)$$

with  $\sqrt{\sigma}$  the relativistic *enthalpy per particle*:

$$\sqrt{\sigma} = \frac{\rho + p}{n}$$

The differential energy-momentum conservation laws (2nd of 3.1) are then seen to be equivalent to the equation:

$$i_u \omega = -\theta ds \quad (3.4)$$

The *irrotational case* is the case where  $\beta = d\phi$  for some function  $\phi$ . In this case 3.4 implies that  $s$  is constant. Only the differential particle conservation (1st of 3.1) then remains, which takes the form of a *nonlinear wave equation*:

$$\nabla_{\mu}(G\partial^{\mu}\phi) = 0, \quad \partial^{\mu}\phi = (g^{-1})^{\mu\nu}\partial_{\nu}\phi \quad (3.5)$$

where

$$G = \frac{n}{\sqrt{\sigma}} = G(\sigma), \quad \sigma = -(g^{-1})^{\mu\nu}\partial_{\mu}\phi\partial_{\nu}\phi$$

Equation 3.5 derives from the Lagrangian

$$L = p = L(\sigma) \quad (3.6)$$

the pressure as a function of the squared enthalpy.

Returning to the general case, the sound speed  $\eta$  is defined by:

$$\left(\frac{dp}{d\rho}\right)_s = \eta^2 \quad (3.7)$$

it being assumed that the left hand side is positive. The *acoustical metric*  $h$  is another Lorentzian metric on  $M$  defined by:

$$h_{\mu\nu} = g_{\mu\nu} + (1 - \eta^2)u_{\mu}u_{\nu}, \quad u_{\mu} = g_{\mu\nu}u^{\nu} \quad (3.8)$$

The null cones of  $h$  are called *sound cones*. The condition  $\eta < 1$  is imposed, which means that the sound cones are contained within the null cones of  $g$ . What is important is the *conformal geometry* defined by  $h$ , which is equivalent to the *acoustical causal structure*.

Choosing a *time function*  $t$  in Minkowski spacetime, equal to the coordinate  $x^0$  of some rectangular coordinate system, we denote by  $\Sigma_t$  an arbitrary level set of the function  $t$ . The  $\Sigma_t$  are parallel spacelike hyperplanes relative to the Minkowski metric  $g$ .

Initial data for the equations of motion 3.1 is given on a domain in the hyperplane  $\Sigma_0$ , which may be the whole of  $\Sigma_0$ . To any given initial data there corresponds a unique *maximal classical development* of the equations of motion 3.1, or of the nonlinear wave equation 3.5 in the irrotational case. The notion of maximal development of given initial data is, in this context, the following. Given initial data, the local existence theorem asserts the existence of a *development* of this data, namely of a domain  $\mathcal{D}$  in Minkowski spacetime, whose past boundary is the domain of the initial data, and of a solution defined in  $\mathcal{D}$  and taking the given data at the past boundary, such that the following condition holds. If we consider any point  $p \in \mathcal{D}$  and any curve issuing at  $p$  with the property that its tangent vector at any point  $q$  belongs to the interior or the boundary of the past component of the sound cone at  $q$ , then the curve terminates at a point of the domain of the initial data. The local uniqueness theorem asserts that two developments of the same initial data, with domains  $\mathcal{D}_1$  and  $\mathcal{D}_2$  respectively, coincide in  $\mathcal{D}_1 \cap \mathcal{D}_2$ . It follows that the union of all developments of a given initial data set is itself a development, the unique maximal development of the initial data set.

In my monograph [8] I consider regular initial data on  $\Sigma_0$  for the general equations of motion 3.1 which outside a sphere coincide with the data corresponding to a constant state.

That is, outside that sphere  $n$  and  $s$  are constant and  $u$  coincides with the future-directed unit normal to  $\Sigma_0$ . I show that under a suitable restriction on the size of the departure of the initial data from those of the constant state, I can control the solution for a time interval of order  $1/\eta_0$ , where  $\eta_0$  is the sound speed in the surrounding constant state. I then show that at the end of this time interval a thick annular region has formed, bounded by concentric spheres, where the flow is irrotational, the constant state holding outside the outer sphere. I then study the maximal classical development of the restriction of the data at this time to the exterior of the inner sphere. Thus the global aspects of my work pertain to the irrotational case. I shall therefore confine myself in the remainder of this lecture to the case that the initial data are irrotational hence so is the maximal classical development. A simplified treatment in the non-relativistic case is given in the monograph [10]. Work in a similar context had been done earlier, starting with F. John (see [13]), by T. Sideris [19] and by S. Alinhac [1, 2].

Let  $O$  be the center of the sphere  $S_{0,0}$  in  $\Sigma_0$  outside which we have the constant state. Let  $u$  be a smooth function without critical points on  $\Sigma_0 \setminus O$  such that the restriction of  $u$  to the exterior of  $S_{0,0}$  is equal to minus the Euclidean distance from  $S_{0,0}$ . We extend  $u$  to the spacetime manifold by the condition that its level sets are outgoing null hypersurfaces relative to the acoustical metric  $h$ . We call  $u$  an *acoustical function* and we denote by  $C_u$  an arbitrary level set of  $u$ . Each  $C_u$  is generated by null geodesics of  $h$ . Let  $L$  be the tangent vectorfield to these geodesic generators parametrized not affinely but by  $t$ . We then define the surfaces  $S_{t,u}$  to be  $C_u \cap \Sigma_t$ . Finally we define the vectorfield  $T$  to be tangential to the  $\Sigma_t$  and so that the flow generated by  $T$  on each  $\Sigma_t$  is the normal, relative to the induced on  $\Sigma_t$  acoustical metric  $\bar{h}$ , flow of the foliation of  $\Sigma_t$  by the surfaces  $S_{t,u}$ . So  $T$  is the tangent vectorfield to the normal curves parametrized by  $u$ .

The geometry of a foliation of spacetime by the outgoing acoustically null hypersurfaces  $C_u$ , the level sets of  $u$ , plays a fundamental role in the problem. The most important geometric property of this foliation from the point of view of the study of shock formation is the density of the packing of its leaves  $C_u$ . One measure of this density is the *inverse spatial density*, that is, the inverse density of the foliation of each spatial hyperplane  $\Sigma_t$  by the surfaces  $S_{t,u}$ . This is simply the magnitude  $\kappa$  of the vectorfield  $T$  with respect to  $h$ . Another measure is the *inverse temporal density*, the function  $\mu$ , given in arbitrary coordinates by:

$$\frac{1}{\mu} = -(h^{-1})^{\mu\nu} \partial_\mu t \partial_\nu u$$

The two measures are related by:

$$\mu = \alpha \kappa$$

where  $\alpha$  is the inverse density, with respect to the acoustical metric  $h$ , of the foliation of spacetime by the hyperplanes  $\Sigma_t$ . The function  $\alpha$  is bounded above and below by positive constants. Consequently  $\mu$  and  $\kappa$  are equivalent measures of the density of the packing of the leaves of the foliation of spacetime by the  $C_u$ . Shock formation is characterized by the blow up of this density or equivalently by the vanishing of  $\kappa$  or  $\mu$ .

The domain of the maximal development being a domain in Minkowski spacetime, which by a choice of rectangular coordinates is identified with  $\mathbb{R}^4$ , inherits the subset topology and the standard differential structure induced by the rectangular coordinates  $x^\alpha$ . Choosing an

acoustical function  $u$  we introduce *acoustical coordinates*  $(t, u, \vartheta)$ ,  $\vartheta \in S^2$ , the coordinate lines corresponding to a given value of  $u$  and to constant values of  $\vartheta$  being the generators of  $C_u$ . The rectangular coordinates  $x^\alpha$  are smooth functions of the acoustical coordinates  $(t, u, \vartheta)$  and the Jacobian of the transformation is, up to a multiplicative factor which is bounded above and below by positive constants, the inverse temporal density function  $\mu$ . The acoustical coordinates induce another differential structure on the same underlying topological manifold. However since  $\mu > 0$  in the interior of the domain of the maximal development, the two differential structures coincide in this interior.

The main theorem of the monograph [8] asserts that relative to the differential structure induced by the acoustical coordinates the maximal classical development extends smoothly to the boundary of its domain. This boundary contains however a singular part  $B$  where the function  $\mu$  vanishes. The rectangular coordinates themselves extend smoothly to the boundary but the Jacobian vanishes on the singular part of the boundary. The mapping from acoustical to rectangular coordinates has a continuous but not differentiable inverse at  $B$ . As a result, the two differential structures no longer coincide when the singular boundary  $B$  is included. With respect to the standard differential structure the solution is continuous but not differentiable at  $B$ , the derivative  $\hat{T}^\mu \hat{T}^\nu \partial_\mu \partial_\nu \phi$  blowing up as we approach  $B$ . Here  $\hat{T} = \kappa^{-1}T$ , is the vectorfield of unit magnitude with respect to  $h$  corresponding to  $T$ . With respect to the standard differential structure, the acoustical metric  $h$  is everywhere in the closure of the domain of the maximal development non-degenerate and continuous, but it is not differentiable at  $B$ , while with respect to the differential structure induced by the acoustical coordinates  $h$  is everywhere smooth, but it is degenerate at  $B$ .

The starting point of the approach leading to the proof of this theorem, is the observation that any variation  $\psi$  of  $\phi$  through solutions of the nonlinear wave equation 3.5 is itself a solution of the linear wave equation

$$\square_{\tilde{h}} \psi = 0 \tag{3.9}$$

relative to the conformal acoustical metric  $\tilde{h} = \Omega h$ , where the conformal factor  $\Omega$  is the ratio of a function of  $\sigma$  to the value of this function in the surrounding constant state, thus  $\Omega$  is equal to unity in the constant state. It turns out moreover that  $\Omega$  is bounded above and below by positive constants.

Here the first order variations correspond to the one-parameter subgroups of the Poincaré group, the isometry group of Minkowski spacetime, extended by the one-parameter scaling or dilation group, which leave the surrounding constant state invariant. The higher order variations are generated from the first order variations by a set of vectorfields, the *commutation fields*, to be discussed below. These higher order variations satisfy inhomogeneous wave equations

$$\square_{\tilde{h}} \psi = \tilde{\rho}$$

the source functions  $\tilde{\rho}$  depending on the deformation tensors of the commutation fields.

Two *multiplier vectorfields* are used:

$$K_0 = (\eta_0^{-1} + \alpha^{-1}\kappa)L + \underline{L}, \quad \underline{L} = \alpha^{-1}\kappa L + 2T$$

and:

$$K_1 = (\omega/\nu)L$$

Here  $\nu$  is the mean curvature of the  $S_{t,u}$  relative to their null normal  $L$ , with respect to the conformal acoustical metric  $\tilde{h}$ . The function  $\omega$  is required to have linear growth in  $t$  and to be such that  $\square_{\tilde{h}}\omega$  is suitably bounded. To a variation  $\psi$ , of any order, and to each of  $K_0, K_1$ , an *energy current* associated, a vectorfield which is a quadratic expression in  $(d\psi, \psi)$ . These energy currents define the energies  $\mathcal{E}_0^u[\psi](t), \mathcal{E}_1^u[\psi](t)$ , and fluxes  $\mathcal{F}_0^t[\psi](u), \mathcal{F}_1^t[\psi](u)$ . For given  $t$  and  $u$  the energies are integrals over the exterior of the surface  $S_{t,u}$  in the hyperplane  $\Sigma_t$ , while the fluxes are integrals over the part of the outgoing null hypersurface  $C_u$  between the hyperplanes  $\Sigma_0$  and  $\Sigma_t$ . The energies and fluxes are positive-definite by virtue of the fact that the multiplier fields are non-spacelike future-directed relative to the acoustical metric  $h$ . It is these energy and flux integrals, together with a spacetime integral  $K[\psi](t, u)$  associated to  $K_1$ , to be discussed below, which are used to control the solution.

Evidently, the means by which the solution is controlled depend on the choice of the acoustical function  $u$ , the level sets of which are the outgoing null hypersurfaces  $C_u$ . The function  $u$  is determined by its restriction to the initial hyperplane  $\Sigma_0$ .

The divergence of the energy currents, which determines the growth of the energies and fluxes, itself depends on  $^{(K_0)}\tilde{\pi}$ , in the case of the energy current associated to  $K_0$ , and  $^{(K_1)}\tilde{\pi}$ , in the case of the energy current associated to  $K_1$ . Here for any vectorfield  $X$  in spacetime, we denote by  $^{(X)}\tilde{\pi}$  the Lie derivative of the conformal acoustical metric  $\tilde{h}$  with respect to  $X$ . This is what we call *deformation tensor* of  $X$  in the present context. In the case of higher order variations, the divergences of the energy currents depend also on the  $^{(Y)}\tilde{\pi}$ , for each of the commutation fields  $Y$  to be discussed below.

All these deformation tensors ultimately depend on the acoustical function  $u$ , or, what is the same, on the geometry of the foliation of spacetime by the outgoing null hypersurfaces  $C_u$ .

The other quantity, besides  $\mu$  which describes the geometry of the foliation by the  $C_u$  is the second fundamental form  $\chi$  of the  $C_u$ , which is a quantity intrinsic to  $C_u$  and in terms of the metric  $\mathbb{h}$  induced by  $h$  on the  $S_{t,u}$  sections, is given by the first variation equation

$$\mathcal{L}_L\mathbb{h} = 2\chi$$

The remaining *acoustical structure equations* are as follows. The second variation equation, which is a propagation equation for  $\chi$  along the generators of  $C_u$ . The Codazzi equation which expresses  $\mathcal{D}\text{iv}\chi$ , the divergence of  $\chi$  intrinsic to  $S_{t,u}$ , in terms of  $\mathcal{D}\text{tr}\chi$ , the differential on  $S_{t,u}$  of  $\text{tr}\chi$ , and a component of the acoustical curvature and of  $k$ , the second fundamental form of the  $\Sigma_t$  relative to  $h$ . The Gauss equation which expresses the Gauss curvature of  $(S_{t,u}, \mathbb{h})$  in terms of  $\chi$  and a component of the acoustical curvature and of  $k$ . The cross-variation equation, which expresses  $\mathcal{L}_T\chi$  in terms of the Hessian of the restriction of  $\mu$  to  $S_{t,u}$  and another component of the acoustical curvature and of  $k$ . These acoustical structure equations seem at first sight to contain terms which blow up as  $\mu$  tends to zero. The analysis of the acoustical curvature then shows that the terms which blow up as  $\mu$  tends to zero cancel.

The most important acoustical structure equation from the point of view of the formation of shocks is the propagation equation for  $\mu$  along the generators of  $C_u$ :

$$L\mu = m + \mu e$$

The function  $m$  given by:

$$m = \frac{1}{2}(\beta_L)^2 TH$$

and the function  $e$  depends only on the  $L\beta_\alpha$ , where the  $\beta_\alpha$  are the rectangular components of the 1-form  $\beta = d\phi$ . Note that the  $\beta_\alpha$  coincide with the first variations  $\psi_\alpha$  corresponding to the generators of translations  $\partial/\partial x^\alpha$  of the underlying Minkowski spacetime. It is the function  $m$  which determines shock formation, when being negative, causing  $\mu$  to decrease to zero.

A fact which is crucial to the whole approach is that *the derivatives of the rectangular coordinates  $x^\alpha$  with respect to the acoustical coordinates  $(t, u, \vartheta)$  can all be expressed in terms of the acoustical quantities  $\mu$  and  $\chi$* . Therefore estimates for the acoustical quantities yield estimates for these derivatives.

A theorem is first established, the fundamental energy estimate, which applies to a solution of the homogeneous wave equation in the acoustical spacetime, in particular to any first order variation. The proof of this theorem relies on certain bootstrap assumptions on the acoustical quantities. The most crucial of these assumptions concern the behavior of the function  $\mu$ . To give an idea of the nature of these assumptions, one of the assumptions required to obtain the fundamental energy estimate up to time  $s$  is:

$$\mu^{-1}(T\mu)_+ \leq B_s(t) : \text{for all } t \in [0, s]$$

where  $B_s(t)$  is a function such that:

$$\int_0^s (1+t)^{-2}[1 + \log(1+t)]^4 B_s(t) dt \leq C$$

with  $C$  a constant independent of  $s$ . Here  $T$  is the vectorfield defined above and we denote by  $f_+$  and  $f_-$ , respectively the positive and negative parts of an arbitrary function  $f$ .

The spacetime integral  $K[\psi](t, u)$  mentioned above, is essentially the integral of

$$-\frac{1}{2}(\omega/\nu)(L\mu)_- |\not{d}\psi|^2$$

in the spacetime exterior to  $C_u$  and bounded by  $\Sigma_0$  and  $\Sigma_t$ . Another assumption states that there is a positive constant  $C$  independent of  $s$  such that in the region below  $\Sigma_s$  where  $\mu < \eta_0/4$  we have:

$$L\mu \leq -C^{-1}(1+t)^{-1}[1 + \log(1+t)]^{-1}$$

In view of this assumption, the integral  $K[\psi](t, u)$  gives effective control of the derivatives of the variations tangential to the  $S_{t,u}$  in the region where shocks are to form. The same assumption also plays an essential role in the study of the singular boundary.

The final stage of the proof of the fundamental energy estimate is the analysis of system of integral inequalities in two variables  $t$  and  $u$  satisfied by the five quantities  $\mathcal{E}_0^u[\psi](t)$ ,  $\mathcal{E}_1^{tu}[\psi](t)$ ,  $\mathcal{F}_0^t[\psi](u)$ ,  $\mathcal{F}_1^{tu}[\psi](u)$ , and  $K[\psi](t, u)$ .

After this, the commutation fields  $Y$ , which generate the higher order variations, are defined. They are five: the vectorfield  $T$  which is transversal to the  $C_u$ , the field  $Q = (1+t)L$  along the generators of the  $C_u$  and the three rotation fields  $R_i : i = 1, 2, 3$  which are

tangential to the  $S_{t,u}$  sections. The latter are defined to be  $\Pi \overset{\circ}{R}_i : i = 1, 2, 3$ , where the  $\overset{\circ}{R}_i$   $i = 1, 2, 3$  are the generators of spatial rotations associated to the background Minkowskian structure, while  $\Pi$  is the  $h$ -orthogonal projection to the  $S_{t,u}$ . Expressions for the deformation tensors of the commutation fields are then derived, which show that these depend on the acoustical quantities  $\mu$  and  $\chi$ .

The source functions  $\tilde{\rho}$  which are associated to the higher order variations give rise to error integrals, that is to spacetime integrals of contributions to the divergence of the corresponding energy currents. The expressions for the source functions and the associated error integrals show that the error integrals corresponding to the energies of the  $n + 1$ st order variations contain the  $n$ th order derivatives of the deformation tensors, which in turn contain the  $n$ th order derivatives of  $\chi$  and  $n + 1$ st order derivatives of  $\mu$ . Thus to achieve closure, we must obtain estimates for the latter in terms of the energies of up to the  $n + 1$ st order variations. Now, the propagation equations for  $\chi$  and  $\mu$  give appropriate expressions for  $\not{L}\chi$  and  $L\mu$ . However, if these propagation equations, which may be thought of as ordinary differential equations along the generators of the  $C_u$ , are integrated with respect to  $t$  to obtain the acoustical quantities  $\chi$  and  $\mu$  themselves, and their spatial derivatives are then taken, a loss of one degree of differentiability would result and closure would fail.

The required regain of differentiability is accomplished in a manner similar to that in general relativity, that is by the consideration of *ordinary differential equations along the generators of the  $C_u$  coupled to elliptic p.d.e. on their  $S_{t,u}$  sections*. In the case of  $\chi$  we consider the propagation equation for  $\mu \text{tr}\chi$ . By virtue of a wave equation for  $\sigma$ , which follows from the wave equations satisfied by the first variations corresponding to the spacetime translations, the principal part on the right hand side of this propagation equation can be put into the form  $-L\tilde{f}$  of a derivative of a function  $-\tilde{f}$  with respect to  $L$ . This function is then brought to the left hand side and we obtain a propagation equation for  $\mu \text{tr}\chi + \tilde{f}$ . In this equation  $\hat{\chi}$ , the trace-free part of  $\chi$  enters, but the propagation equation in question is considered in conjunction with the Codazzi equation, which constitutes an elliptic system on each  $S_{t,u}$  for  $\hat{\chi}$ , given  $\text{tr}\chi$ . More precisely, the propagation equation which is considered at the same level as the Codazzi equation is a propagation equation for the  $S_{t,u}$  1-form  $\mu \not{d}\text{tr}\chi + \not{d}\tilde{f}$ , which is a consequence of the equation just discussed. To obtain estimates for the angular derivatives of  $\chi$  of order  $l$  we similarly consider a propagation equation for the  $S_{t,u}$  1-form:

$${}^{(i_1 \dots i_l)}x_l = \mu \not{d}(R_{i_1} \dots R_{i_l} \text{tr}\chi) + \not{d}(R_{i_1} \dots R_{i_l} \tilde{f})$$

In the case of  $\mu$  the regain of differentiability is accomplished by considering the propagation equation for  $\mu \not{\Delta}\mu$ , where  $\not{\Delta}\mu$  is the Laplacian of the restriction of  $\mu$  to the  $S_{t,u}$ . By virtue of a wave equation for  $T\sigma$ , which is a consequence of the wave equation for  $\sigma$ , the principal part on the right hand side of this propagation equation can again be put into the form  $L\tilde{f}'$  of a derivative of a function  $\tilde{f}'$  with respect to  $L$ . This function is then likewise brought to the left hand side and we obtain a propagation equation for  $\mu \not{\Delta}\mu - \tilde{f}'$ . In this equation  $\hat{D}^2\mu$ , the trace-free part of the Hessian of the restriction of  $\mu$  to the  $S_{t,u}$  enters, but the propagation equation in question is considered in conjunction with the elliptic equation on each  $S_{t,u}$  for  $\mu$ , which the specification of  $\not{\Delta}\mu$  constitutes. To obtain estimates of the spatial derivatives of  $\mu$  of order  $l + 2$  of which  $m$  are derivatives with respect to  $T$  we similarly consider a

propagation equation for the function:

$${}^{(i_1 \dots i_{l-m})} x'_{m,l-m} = \mu R_{i_{l-m}} \dots R_{i_1} (T)^m \not\Delta \mu - R_{i_{l-m}} \dots R_{i_1} (T)^m \check{f}'$$

This allows us to obtain estimates for the top order spatial derivatives of  $\mu$  of which at least two are angular derivatives. A remarkable fact is that the missing top order spatial derivatives do not enter the source functions, hence do not contribute to the error integrals.

The appearance of the factor of  $\mu$ , which vanishes where shocks originate, in front of  $\not\Delta R_{i_l} \dots R_{i_1} \chi$  and  $R_{i_{l-m}} \dots R_{i_1} (T)^m \not\Delta \mu$  in the definitions of  ${}^{(i_1 \dots i_l)} x_l$  and  ${}^{(i_1 \dots i_{l-m})} x'_{m,l-m}$  above, makes the analysis quite delicate. This is compounded with the difficulty of the slow decay in time which the addition of the terms  $-\not\Delta R_{i_l} \dots R_{i_1} \check{f}$  and  $R_{i_{l-m}} \dots R_{i_1} (T)^m \check{f}'$  forces. The analysis requires a precise description of the behavior of  $\mu$  itself, given by certain propositions, and a separate treatment of the condensation regions, where shocks are to form, from the rarefaction regions, the terms referring not to the fluid density but rather to the density of the stacking of the  $C_u$ . To overcome the difficulties the following weight function is introduced:

$$\bar{\mu}_{m,u}(t) = \min \left\{ \frac{\mu_{m,u}(t)}{\eta_0}, 1 \right\}, \quad \mu_{m,u}(t) = \min_{\Sigma_t^u} \mu$$

where  $\Sigma_t^u$  is the exterior of  $S_{t,u}$  in  $\Sigma_t$ , and the quantities  $\mathcal{E}_0^u[\psi](t)$ ,  $\mathcal{E}_1^u[\psi](t)$ ,  $\mathcal{F}_0^t[\psi](u)$ ,  $\mathcal{F}_1^t[\psi](u)$ , and  $K[\psi](t, u)$  corresponding to the highest order variations are weighted with a power,  $2a$ , of this weight function. The following lemma then plays a crucial role. Let:

$$M_u(t) = \max_{\Sigma_t^u} \{ -\mu^{-1} (L\mu)_- \}, \quad I_{a,u} = \int_0^t \bar{\mu}_{m,u}^{-a}(t') M_u(t') dt'$$

Then under certain bootstrap assumptions in the past of  $\Sigma_s$ , for any constant  $a \geq 2$ , there is a positive constant  $C$  independent of  $s, u$  and  $a$  such that for all  $t \in [0, s]$  we have:

$$I_{a,u}(t) \leq C a^{-1} \bar{\mu}_{m,u}^{-a}(t) \tag{3.10}$$

The acoustical assumptions on which the previous results depend are established, using the method of continuity, on the basis of the final bootstrap assumption, which consists only of pointwise estimates for the variations up to certain order.

The analysis of the structure of the terms containing the top order spatial derivatives of the acoustical quantities shows that these terms can be expressed in terms of the 1-forms  ${}^{(i_1 \dots i_l)} x_l$  and the functions  ${}^{(i_1 \dots i_{l-m})} x'_{m,l-m}$ . These contribute *borderline error integrals*, the treatment of which is the main source of difficulties in the problem.

I should make clear here that the only variations which are considered up to this point are the variations arising from the first order variations corresponding to the group of space-time translations of the underlying Minkowski spacetime. In particular the final bootstrap assumption involves only variations of this type, and each of the five quantities  $\mathcal{E}_{0,[n]}^u(t)$ ,  $\mathcal{F}_{0,[n]}^t(u)$ ,  $\mathcal{E}_{1,[n]}^u(t)$ ,  $\mathcal{F}_{1,[n]}^t(u)$ , and  $K_{[n]}(t, u)$ , which together control the solution, is defined to be the sum of the corresponding quantity  $\mathcal{E}_0^u[\psi](t)$ ,  $\mathcal{F}_0^t[\psi](u)$ ,  $\mathcal{E}_1^u[\psi](t)$ ,  $\mathcal{F}_1^t[\psi](u)$ , and  $K[\psi](t, u)$ , over all variations  $\psi$  of this type, up to order  $n$ .

To estimate the borderline integrals however, an additional assumption is introduced which concerns the first order variations corresponding to the scaling or dilation group and

to the rotation group, and the second order variations arising from these by applying the commutation field  $T$ . This assumption is later established through energy estimates of order 4 arising from these first order variations and derived on the basis of the final bootstrap assumption, just before the recovery of the final bootstrap assumption itself. It turns out that the borderline integrals all contain the factor  $T\psi_\alpha$ , and the additional assumption is used to obtain an estimate for  $\sup_{\Sigma_t^u} (\mu^{-1}|T\psi_\alpha|)$  in terms of  $\sup_{\Sigma_t^u} (\mu^{-1}|L\mu|)$ . Upon substituting this estimate, the borderline integrals are estimated using the inequality 3.10.

In proceeding to derive the energy estimates of top order,  $n = l + 2$ , the power  $2a$  of the weight  $\bar{\mu}_{m,u}(t)$  is chosen suitably large to allow us to transfer the terms contributed by the borderline integrals to the left hand side of the inequalities resulting from the integral identities associated to the multiplier fields  $K_0$  and  $K_1$ . The argument then proceeds along the lines of that of the fundamental energy estimate.

Once the top order energy estimates are established, I revisit the lower order energy estimates using at each order the energy estimates of the next order in estimating the error integrals contributed by the highest spatial derivatives of the acoustical quantities at that order. I then establish a descent scheme, which yields, after finitely many steps, estimates for the five quantities  $\mathcal{E}_{0,[n]}^u(t)$ ,  $\mathcal{F}_{0,[n]}^t(u)$ ,  $\mathcal{E}_{1,[n]}^{nu}(t)$ ,  $\mathcal{F}_{1,[n]}^{nt}(u)$ , and  $K_{[n]}(t, u)$ , for  $n = l + 1 - [a]$ , where  $[a]$  is the integral part of  $a$ , in which weights no longer appear.

It is these unweighted estimates which are used to close the bootstrap argument by recovering the final bootstrap assumption. This is accomplished by the method of continuity through the use of the isoperimetric inequality on the  $S_{t,u}$ , and leads to the main theorem.

The later part of the work is concerned with the structure of the boundary of the domain of the maximal development and the behavior of the solution at this boundary. The boundary consists of a regular part  $\underline{C}$  and a singular part  $B$ . Each component of  $\underline{C}$  is a regular incoming acoustically null hypersurface with a singular past boundary which coincides with the past boundary of an associated component of  $B$ . The union of these singular past boundaries we denote by  $\partial^- B$ . Each component of  $B$  is a hypersurface which is smooth relative to both differential structures and has the intrinsic geometry of a regular null hypersurface in a regular spacetime and, like the latter, is ruled by invariant curves of vanishing arc length. On the other hand, the extrinsic geometry of each component of  $B$  is that of an acoustically spacelike hypersurface which becomes acoustically null at its past boundary, an associated component of  $\partial^- B$ . This means that at each point  $q \in B$  the past null geodesic conoid of  $q$  does not intersect  $B$ . Each component of  $\partial^- B$  is an acoustically spacelike surface which is smooth relative to both differential structures. The main result of the last part of the work is the *trichotomy theorem*. According to this theorem, for each point  $q$  of the singular boundary, the intersection of the past null geodesic conoid of  $q$  with any  $\Sigma_t$  in the past of  $q$  splits into three parts, the parts corresponding to the outgoing and to the incoming sets of null geodesics ending at  $q$  being embedded discs with a common boundary, an embedded circle, which corresponds to the set of the remaining null geodesics ending at  $q$ . *All outgoing null geodesics ending at  $q$  have the same tangent vector at  $q$ .* This vector is then an invariant null vector associated to the singular point  $q$ . This is in fact the reason why the considerable freedom in the choice of the acoustical function does not matter in the end. For, considering the transformation from one acoustical function to another, I show that the foliations cor-

responding to different families of outgoing null hypersurfaces have equivalent geometric properties and degenerate in precisely the same way on the same singular boundary.

## References

- [1] Alinhac, S., *Blowup for Nonlinear Hyperbolic Equations*, Prog. Nonlinear Diff. Eq. and Appl. 17, Birkhäuser, 1995.
- [2] Alinhac, S., *Blowup of Small Data Solutions for a Class of Quasilinear Wave Equations in Two Space Dimensions II*, Acta Math. **182** (1999), 1–23.
- [3] Bel, L., *Introduction d'un tenseur du quatrième ordre*, C. R. Acad. Sci. Paris **248** (1959), 1297–1300.
- [4] Bieri, L. and Zipser, N., *Extensions of the Stability Theorem of the Minkowski Space in General Relativity*, Studies in Advanced Mathematics 45, American Mathematical Society and International Press, 2009.
- [5] Christodoulou, D., *The Action Principle and Partial Differential Equations*, Annals of Mathematics Studies 146, Princeton University Press, Princeton, NJ, 2000.
- [6] ———, *Nonlinear Nature of Gravitation and Gravitational Wave Experiments*, Phys. Rev. Lett. **67** (1991), 1486–1489.
- [7] ———, *The Formation of Black Holes in General Relativity*, EMS Monographs in Mathematics, EMS Publishing House, 2009.
- [8] ———, *The Formation of Shocks in 3-Dimensional Fluids*, EMS Monographs in Mathematics, EMS Publishing House, 2007.
- [9] ———, Klainerman, S., *The Global Nonlinear Stability of the Minkowski Space*, Princeton Mathematical Series 41, Princeton University Press, Princeton, NJ, 1993.
- [10] Christodoulou, D. and Miao, S., *Compressible Flow and Euler's Equations*, Higher Education Press and International Press, 2014.
- [11] Choquet-Bruhat, Y., *Theorem d'existence pour certain systems d'equations aux deriveés partielles nonlineaires*, Acta Mathematica **88** (1952), 141–225.
- [12] Choquet-Bruhat, Y. and Geroch, R. P., *Global aspects of the Cauchy problem in general relativity*, Commun. Math. Phys. **14** (1969), 329–335.
- [13] John, F., *Nonlinear Wave Equations, Formation of Singularities*, Lehigh University Lecture Series 2, American Mathematical Society, Providence, RI, 1990.
- [14] Klainerman, S. and Rodnianski, I., *On the Formation of Trapped Surfaces*, Acta Math. **208** (2012), 211–333.
- [15] Klainerman, S., Luk, J., and Rodnianski, I., *A Fully Anisotropic Mechanism for Formation of Trapped Surfaces*, Inventiones Mathematicae **195** (2014)
- [16] Li, J. and Yu, P., *Construction of Cauchy Data of Vacuum Einstein Field Equations Evolving to Black Holes*, <http://arxiv.org/abs/1207.3164>

- [17] Penrose, R., *Gravitational collapse and space-time singularities*, Phys. Rev. Lett. **14** (1965), 57–59.
- [18] Rendall, A. D., *Reduction of the characteristic initial value problem to the Cauchy problem and its applications to the Einstein equations*, Proc. Roy. Soc. Lond. A **427** (1990), 221–239.
- [19] Sideris, T., *Formation of Singularities in Three-Dimensional Compressible Fluids*, Commun. Math.Phys. **101** (1985), 475–485.
- [20] Yu, P., *Dynamical Formation of Black Holes Due to the Condensation of Matter Field*, <http://arxiv.org/abs/1105.5898>

HG G 48.2, ETH-Zentrum, CH-8092 Zürich, Switzerland

E-mail: demetri@math.ethz.ch

# Minimal surfaces: variational theory and applications

Fernando Codá Marques

**Abstract.** Minimal surfaces are among the most natural objects in Differential Geometry, and have been studied for the past 250 years ever since the pioneering work of Lagrange. The subject is characterized by a profound beauty, but perhaps even more remarkably, minimal surfaces (or minimal submanifolds) have encountered striking applications in other fields, like three-dimensional topology, mathematical physics, conformal geometry, among others. Even though it has been the subject of intense activity, many basic open problems still remain. In this lecture we will survey recent advances in this area and discuss some future directions. We will give special emphasis to the variational aspects of the theory as well as to the applications to other fields.

**Mathematics Subject Classification (2010).** Primary 53C42; Secondary 49Q05.

**Keywords.** Minimal surfaces, calculus of variations, conformal geometry, three-manifold topology.

## 1. Introduction and results

Minimal submanifolds are solutions of the most basic variational problem of submanifold geometry, that of extremizing the area. This was first considered by Lagrange (1762), who raised the question of existence of surfaces of least area having a given closed curve in three-space as the boundary. He derived the differential equation that must be satisfied by a function of two variables whose graph minimizes area among surfaces with a given contour. Later Meusnier discovered that this is equivalent to the vanishing of the mean curvature, and the study of the differential geometry of these surfaces was started. The theory of minimal surfaces (or minimal submanifolds) has been developed over the years by several outstanding mathematicians, and it is now extremely rich. It has been extended to other ambient geometries, and it is full of beautiful examples and deep theorems.

This paper attempts to give an overview and discuss some recent advances in the subject, with emphasis on the variational aspects and applications. Being such a large field, we do not have the pretension of being exhaustive. Part of the material of this article is also discussed in the contribution of André Neves [91].

Let us begin with a discussion of the first variation formula.

**1.1. First variation formula.** Let  $\Sigma$  be a two-dimensional oriented surface in  $\mathbb{R}^3$ , and let  $N$  denote a unit normal field. The local geometry of  $\Sigma$  at a point  $p$  can be understood in terms of the principal curvatures  $k_1, k_2$ , the maximum and minimum curvatures of the intersections

of the surface with normal planes passing through  $p$ . The classical notions of curvature of a surface in three-space are:

- the mean curvature  $H = (k_1 + k_2)/2$ ,
- the Gauss curvature  $K = k_1 \cdot k_2$ .

The Gauss curvature  $K$ , according to the Theorema Egregium of Gauss, is an intrinsic notion, i.e., depends only on measurements made in the surface  $\Sigma$  without reference to the space in which the surface is embedded. The mean curvature  $H$ , by contrast, is extrinsic and is naturally related to the area functional as follows.

Given a smooth variation  $F : (-\varepsilon, \varepsilon) \times \Sigma \rightarrow \mathbb{R}^3$  of  $\Sigma$ , with  $F(0, \cdot) = \text{id}$ ,  $\Sigma_t = F(t, \Sigma)$ , the *First Variation Formula* tells us that

$$\frac{d}{dt} \Big|_{t=0} \text{area}(\Sigma_t) = - \int_{\Sigma} \langle \vec{H}, X \rangle d\Sigma + \int_{\partial\Sigma} \langle \nu, X \rangle ds,$$

where  $\vec{H} = H \cdot N$  is the mean curvature vector of  $\Sigma$  in  $\mathbb{R}^3$ ,  $\nu$  is the outward unit conormal vector of  $\partial\Sigma$ , and  $X = \frac{\partial F}{\partial t}(0, \cdot)$  is the variational vector field.

The formula holds true in the more general setting of a  $k$ -dimensional submanifold  $\Sigma$  immersed in an  $n$ -dimensional Riemannian manifold  $M$ . It leads us to define a *minimal submanifold* as one for which the mean curvature vector vanishes ( $\vec{H} = 0$ ) or, equivalently, one for which the first derivative of area is zero with respect to any variation that keeps the boundary fixed ( $X = 0$  on  $\partial\Sigma$ ).

**1.2. Plateau's problem.** Minimal surfaces can be physically represented as soap films, which can be experimentally produced by dipping wire contours into soapy water. These experiments were systematically carried out by the physicist Joseph Plateau in the 19th century. The problem, raised by Lagrange, of finding the surface of least area with a given boundary in Euclidean space became known as the Plateau's Problem.

This became a central question in the field, inspiring the development of a great amount of mathematics since the time of Riemann, until it was independently solved in 1930 by Douglas [33] and Radó [105]. They considered two-dimensional surfaces that were given by mappings of the unit disk, and proved in particular that every smooth Jordan curve in Euclidean space is the boundary of a least area surface of disk type. Later Morrey [88] extended this existence theory to two-dimensional surfaces in  $n$ -dimensional Riemannian manifolds that are homogeneously regular, a condition satisfied by all closed manifolds.

The search for solving the Plateau's problem in greater generality, extending it to submanifolds of higher dimensions and of arbitrary topological type lead to the development of Geometric Measure Theory. In a seminal paper, Federer and Fleming [34] introduced the class of integral currents to model  $k$ -dimensional domains of integration. This class had the right compactness properties to allow the solution of an extremely general Plateau's problem by the direct method of the calculus of variations. The regularity of these  $k$ -area minimizing currents was the subject of much work later on ([6, 7, 14, 28, 30, 34, 35, 121]). In the case of codimension one, the area minimizing current is smooth outside a singular set of codimension 7.

An important source of area minimizing submanifolds comes from the calibration theory introduced by Harvey and Lawson [53]. Complex submanifolds in Kähler manifolds and special Lagrangian submanifolds in Calabi-Yau manifolds are calibrated, therefore area minimizing in their homology class.

Area minimizing submanifolds are in particular *stable*, i.e., the second variation of area is nonnegative for any variational normal vector field  $X$  with  $X = 0$  on  $\partial\Sigma$ .

**1.3. Minimizing in homotopy and isotopy classes.** The existence of incompressible minimal surfaces in Riemannian manifolds was proven by Schoen and Yau [117] and independently by Sacks and Uhlenbeck [111]. Let  $\Sigma_g$  be a compact Riemann surface of genus  $g$ , and suppose  $f : \Sigma_g \rightarrow M$  is a continuous map such that the action  $f_* : \pi_1(\Sigma_g) \rightarrow \pi_1(M)$  induced at the level of the fundamental groups is injective. Then there exists a branched minimal immersion  $h : \Sigma_g \rightarrow M$  that minimizes area among all maps  $h' : \Sigma_g \rightarrow M$  that satisfy  $h'_* = f_*$ . The idea is to first fix the conformal structure of  $\Sigma_g$  and minimize the Dirichlet energy  $E(f) = \int_{\Sigma_g} |df|^2 d\mu$ . This produces a family of harmonic maps, and the second step is to minimize their energies over the Teichmüller space. In case  $\dim(M) = 3$ , the works of Osserman [92] and Gulliver [49] establish that the map  $h$  has no branch points, i.e., it is a smooth immersion.

In [86], Meeks, Simon and Yau proved the existence of embedded minimal surfaces by minimizing area (instead of energy) in nontrivial isotopy classes. Their methods, together with Schoen's curvature estimates [112], are used to establish regularity in some treatments of min-max theory (see [22]). The embeddedness question and applications to three-dimensional topology were also the subject of [84, 85].

An existence theory that produces minimal two-spheres in every compact Riemannian manifold was developed in the article of Sacks and Uhlenbeck [110]. These minimal two-spheres are parametrized conformally by a harmonic map from  $S^2$  to  $M$  that is an immersion (not necessarily an embedding) outside finitely many branch points. Every harmonic map from  $S^2$  to  $M$  is also a conformal branched minimal immersion, hence these minimal spheres can be constructed by extremizing the Dirichlet energy:  $E(f) = \int_{S^2} |df|^2 d\mu$ ,  $f : S^2 \rightarrow M$ . A major difficulty arises from the facts that the energy  $E$  is conformally invariant and that the group of conformal transformations of  $S^2$  is noncompact. These issues are dealt with in [110] by approximating the energy  $E$  by a family of energy functionals  $E_\alpha$ ,  $\alpha > 1$ , that satisfy the Palais-Smale condition. The possible loss of compactness by concentration of energy in the limit, as  $\alpha \rightarrow 1$ , is treated by the introduction of the renormalization (or blow-up) technique.

**1.4. Scalar curvature and Positive Mass Conjecture.** A celebrated application of minimal hypersurfaces of minimizing type to mathematical physics is the proof of the Positive Mass Conjecture by Schoen and Yau [116, 118]. Later Witten [131] gave a different proof using harmonic spinors. The theorem establishes that the total mass of an isolated gravitational system, modeled by an asymptotically flat spacetime obeying the dominant energy condition, must be positive unless the spacetime is the Minkowski space (of zero mass). In the time-symmetric case, this reduces to showing that the mass of an asymptotically flat Riemannian three-manifold of nonnegative scalar curvature is positive unless the manifold is the Euclidean space.

The proof of Schoen and Yau is by contradiction. If the mass is negative, they prove that one can construct (by taking a limit of solutions to Plateau problems) a complete orientable area-minimizing (hence stable) minimal surface  $\Sigma$  in  $M$ . Curvature estimates for stable minimal submanifolds are needed in this process ([112], [114]). By the Second Variation

Formula, the stability condition gives that

$$\int_{\Sigma} (|\nabla f|^2 - (|A|^2 + Ric(N, N))f^2) d\Sigma \geq 0$$

for any smooth function  $f$  with compact support in  $\Sigma$ . Here  $A$  denotes the second fundamental form. The idea is to exploit the stability inequality and arrive at a contradiction with the Gauss-Bonnet Theorem.

The same type of argument works in the compact setting to prove that the torus  $T^3$  does not admit a metric of positive scalar curvature ([117]). The trick is to use the Gauss equation to make the ambient scalar curvature appear, rewriting the stability condition as:

$$\int_{\Sigma} \left( |\nabla f|^2 - \left( \frac{1}{2}R_M - K_{\Sigma} + \frac{1}{2}|A|^2 \right) f^2 \right) d\Sigma \geq 0$$

for  $f \in C_0^{\infty}(\Sigma)$ . We denote by  $R_M$  the scalar curvature of  $M$ . Since any Riemannian three-torus contains a stable minimal  $T^2$ , by minimization in a homotopy class, we obtain a contradiction between  $R_M > 0$  and the Gauss-Bonnet Theorem by choosing  $f \equiv 1$ .

Gromov and Lawson [48] used spinorial techniques to prove that the torus  $T^n$  does not carry a metric of positive scalar curvature, for any  $n$ . The argument of Schoen and Yau extends to any dimension  $3 \leq n \leq 7$ , and breaks down in higher dimensions because of the possibility of singularities in the solution to the Plateau problem. The proof of Witten of the positive mass theorem works in any dimension  $n \geq 3$  under the topological requirement that the manifold is spin. Despite recent approaches, the positive mass conjecture is still open for nonspin manifolds in high dimensions.

Finally, we point out that minimal surfaces also play a very important role in general relativity by modeling apparent horizons of black holes. The Penrose inequality (proven by Huisken and Ilmanen [60] and Bray [15]), for instance, gives a beautiful and sharp inequality between the total mass and the area of an outermost minimal sphere.

**1.5. Min-max methods.** So far we have discussed only minimization questions, but in general minimal submanifolds are critical points of saddle type. Poincaré [104] realized the importance of constructing these critical points, in the context of geodesics. He asked the foundational question of whether every Riemannian two-sphere contains a closed geodesic. Geodesics are, of course, examples of minimal submanifolds. This question had a tremendous impact in mathematics. Firstly, it can be interpreted from two different points of view: as the search for periodic orbits of the geodesic flow or for critical points of the length functional. Secondly, entirely new techniques and topological ideas had to be developed to answer it.

The first breakthrough was due to Birkhoff [13], who introduced min-max methods to the problem. He defined the notion of *sweepout*: a continuous family of closed curves  $\{c_t\}_{t \in [0,1]}$  in  $S^2$  that can be written as

$$c_t = f(\{x \in S^2 : x_3 = 1 - 2t\})$$

for some degree one map  $f : S^2 \rightarrow S^2$ . Given a Riemannian two-sphere  $(S^2, g)$ , one can consider the min-max invariant

$$L = \inf_f \sup_{t \in [0,1]} L(c_t),$$

where  $L(c)$  denotes the length of the curve  $c$ . Birkhoff proved that  $L > 0$  and that  $L = L(\gamma)$  for some smooth closed geodesic  $\gamma$ . The geodesic  $\gamma$  is obtained as a limit of curves of maximal length in a minimizing sequence of sweepouts. Therefore every Riemannian two-sphere  $(S^2, g)$  contains at least one closed geodesic. The fact that every compact Riemannian manifold  $M^n$  contains a closed geodesic was later established by Lusternik and Fet [77].

The work of Birkhoff inspired the development of Morse theory and Lusternik-Schnirelman theory, fundamental ideas in mathematics that brought together the fields of topology and the calculus of variations. For instance, Lusternik and Schnirelmann [78] introduced new topological methods that lead to a proof that every metric on a two-sphere admits at least three simple (embedded) closed geodesics. The search for a rigorous proof motivated a great amount of work (see [125]). The proof of Grayson [43] uses a parabolic partial differential equation, the curve shortening flow of curves. In the early 1990s, this activity culminated with the proof that every Riemannian two-sphere contains infinitely many smooth closed geodesics. This follows by combining the works of Franks [39] and Bangert [11], while Hingston [57] proved quantitative results.

It is natural to ask whether every closed Riemannian  $n$ -manifold  $M$  contains a  $k$ -dimensional closed minimal submanifold. This suggests looking for a Morse theory for minimal varieties, similar in spirit to the case of closed geodesics. The first step was done in [4], where Almgren (by suggestion of Federer) computes the homotopy groups of the space  $\mathcal{Z}_k(M)$  of  $k$ -dimensional integral cycles (integral currents with boundary zero) of  $M$ . Almgren proved that the  $l$ -dimensional homotopy group of  $\mathcal{Z}_k(M)$  is isomorphic to the  $(k+l)$ -dimensional homology group  $H_{k+l}(M, \mathbb{Z})$ . This theorem gives examples of one-parameter homotopically nontrivial families of cycles, so one could think of applying min-max methods just as in the case of Birkhoff's sweepouts. One major problem is that the mass functional (area functional) is only lower semicontinuous in the flat topology (the natural topology for currents). This is not a drawback in questions of minimization, but it could lead to loss of area in the limit, thereby preventing an unstable minimal surface from being detected by the min-max approach. (A study of the existence of unstable solutions to the two-dimensional Plateau problem had been done by Morse and Tompkins in [89].)

In [5], Almgren deals with this issue by considering the measure theoretic class of surfaces called varifolds. The natural topology of the space of varifolds allowed for both good compactness properties and continuity of the area functional. Almgren devised a very general min-max theory that worked in any dimension and codimension, and for families of cycles of any number of parameters. He was able to show that every closed Riemannian manifold  $M^n$  contains at least one stationary integral  $k$ -dimensional varifold, for each  $1 \leq k \leq n$ . (Recall that a varifold is stationary when the first variation of area is zero with respect to any smooth deformation of the ambient space.) This Morse-theoretic theorem left open the question of regularity of the min-max minimal variety. Note that a general stationary integral varifold contains an open dense set where it is a smooth submanifold, according to the regularity theorem of Allard [3].

In [102], Pitts improved considerably the theory of Almgren by showing that the stationary integral varifold can be chosen to satisfy an additional variational property, the almost minimizing condition. Roughly speaking, an almost minimizing varifold is one that can be arbitrarily approximated by integral currents that nearly minimize area. If the codimension is one, curvature estimates for stable minimal hypersurfaces can be used to prove regularity. Pitts employed the pointwise curvature estimates of Schoen, Simon and Yau [114] to prove that if  $3 \leq n \leq 6$  then the min-max minimal variety can be chosen to be an embedded

smooth closed minimal hypersurface. Schoen and Simon [113] then extended the regularity theory through different methods to higher dimensions, allowing singular sets of codimension 7 (see Wickramasekera [129] for a general regularity theory of stable hypersurfaces). By combining these results, the theorem is:

**Theorem 1.1.** *Let  $(M^n, g)$  be a compact Riemannian manifold, with  $n \geq 3$ . Then  $M$  contains a stationary integral varifold  $\Sigma$ , whose support is smooth outside a singular set of codimension 7. In particular, if  $n \leq 7$  then  $M$  contains a smooth embedded closed minimal hypersurface  $\Sigma^{n-1}$ .*

**Remark 1.2.** If  $M^n$  satisfies  $H_{n-1}(M, \mathbb{Z}) \neq 0$ , then the existence of  $\Sigma$  follows by direct minimization of area inside a nontrivial homology class  $\sigma \in H_{n-1}(M, \mathbb{Z})$ , together with the regularity theory for codimension one area minimizing currents mentioned before.

**Remark 1.3.** Although it is true, as mentioned before, that every compact surface contains a closed geodesic, the Almgren-Pitts min-max theory applied to that setting does not yield a smooth object. The three-legged starfish example (see [5]) suggests a situation in which the min-max curve has the shape of a figure eight. The Almgren-Pitts theory has been applied to the setting of one-cycles by Calabi and Cao [18], who proved that in two-spheres with nonnegative curvature the closed geodesics of shortest length are always embedded.

A beautiful application of higher index minimal two-spheres was given by Micallef and Moore [87], following the Sacks-Uhlenbeck existence approach. They introduced the positive isotropic curvature condition for Riemannian manifolds  $M^n$  of dimension  $n \geq 4$ , and proved that it implies  $\pi_2(M) = \dots = \pi_{[n/2]}(M) = 0$  when  $M$  is compact. If  $M$  is further assumed to be simply connected, it follows that it must be homeomorphic to a sphere. The method is reminiscent of Sygne's theorem in Riemannian geometry. The idea is to analyze the index of the Sacks-Uhlenbeck harmonic spheres (by applying Morse theory to the perturbed energy functionals) and compare with a lower bound for the index coming from the Riemann-Roch theorem. The trick is to use the complexified version of the second variation formula, previously considered by Siu and Yau [122] in their proof of the Frankel conjecture. For a general compact Riemannian manifold of positive isotropic curvature, the conjecture is that its fundamental group must contain a free subgroup of finite index (Fraser [40], Gromov [45]). In dimension four the conjecture follows from the complete classification of Chen, Tang and Zhu [19]. We also mention that minimal surfaces have been recently used by Liu [76] to classify complete three-manifolds of nonnegative Ricci curvature, thereby completing the work of Schoen and Yau [119].

**1.6. Recent applications.** We briefly describe some recent applications of min-max minimal surfaces that will be discussed in more detail in Sections 2, 3 and 4.

Colding and Minicozzi [23, 24] found a splendid application of min-max methods to the study of the topology of compact three-manifolds. Their contribution fits together with the celebrated proof of the Poincaré conjecture by Perelman [97–99], achieved by a profound analysis of Hamilton's Ricci flow. In [99], Perelman proved that the Ricci flow with surgeries starting at a homotopy three-sphere becomes extinct in finite time. The finite time extinction result was actually proven more generally for any closed orientable three-manifold whose prime decomposition contains only non-aspherical factors. In this case one can avoid the analysis of the longtime behavior of the flow, and conclude that  $M$  must be diffeomorphic to

a connected sum of copies of  $S^2 \times S^1$  and of spherical space forms  $S^3/\Gamma$ .

The proof of Perelman of the finite time extinction result was itself based in minimal surface techniques of a variational nature. For technical reasons it required in addition a regularized version of the curve shortening flow, due to Altschuler and Grayson [8]. In [23, 24], Colding and Minicozzi provided an alternative elegant argument inspired by the Sacks-Uhlenbeck approach. The evolution equation of the area of the min-max minimal surface produced by sweeping out a homotopy 3-sphere by  $S^2$ 's implies that the area eventually decreases to zero in finite time. We will say more about their argument in Section 2.

Recently, the author and A. Neves have been able to find a connection between the min-max theory of minimal surfaces in  $S^3$  and the Willmore conjecture (1965). The main new insights come from the analysis of the geometric and topological properties of a canonical 5-dimensional family of surfaces in the three-sphere, to which we apply min-max theory for the area functional.

The Willmore energy of a closed surface  $\Sigma$  immersed in Euclidean three-space is the total integral of the square of the mean curvature:

$$\mathcal{W}(\Sigma) = \int_{\Sigma} H^2 d\Sigma.$$

This functional is invariant under the action of any conformal transformation of  $\mathbb{R}^3$ , and it appears naturally in physical contexts. It was proposed in the 1800s by Sophie Germain [42] to describe elastic shells, and more modernly it appears in the Helfrich model [56] of mathematical biology as one of the terms that contribute to the energy of cell membranes.

Willmore proved that the round spheres (of energy  $4\pi$ ) minimize the Willmore energy among all closed surfaces in  $\mathbb{R}^3$ , and asked what is the optimal shape among surfaces of some fixed topological type. Motivated by the analysis of circular tori of revolution, Willmore made a conjecture for the case of genus one:

**Willmore conjecture** (1965, [130]). *The integral of the square of the mean curvature of a torus immersed in  $\mathbb{R}^3$  is at least  $2\pi^2$ .*

The torus  $\Sigma_{\sqrt{2}}$ , obtained by rotation of a circle of radius 1 with center at distance  $\sqrt{2}$  of the axis of revolution, satisfies  $\mathcal{W}(\Sigma_{\sqrt{2}}) = 2\pi^2$ .

The author and A. Neves proved the following theorem that implies the conjecture:

**Theorem 1.4** ([80]). *Let  $\Sigma \subset \mathbb{R}^3$  be a closed, embedded smooth surface with genus  $g \geq 1$ . Then  $\mathcal{W}(\Sigma) \geq 2\pi^2$ , and  $\mathcal{W}(\Sigma) = 2\pi^2$  if and only if  $\Sigma$  is a conformal image of  $\Sigma_{\sqrt{2}}$ .*

We were motivated by the problem of producing the Clifford torus  $\hat{\Sigma} = S^1(\frac{1}{\sqrt{2}}) \times S^1(\frac{1}{\sqrt{2}})$ , a minimal surface of area  $2\pi^2$  and index five in  $S^3$ , by min-max methods. As a byproduct of the construction, we proved:

**Theorem 1.5** ([80]). *Let  $\Sigma \subset S^3$  be a closed, embedded smooth surface with genus  $g \geq 1$ . If  $\Sigma$  is minimal then  $\text{area}(\Sigma) \geq 2\pi^2$ , and  $\text{area}(\Sigma) = 2\pi^2$  if and only if  $\Sigma$  is the image of the Clifford torus under an ambient isometry.*

Together with I. Agol and A. Neves, we used these ideas to solve another conformally invariant variational problem. Recall that a 2-component link in  $\mathbb{R}^3$  is a pair  $(\gamma_1, \gamma_2)$  of

rectifiable curves  $\gamma_i : S^1 \rightarrow \mathbb{R}^3$ ,  $i = 1, 2$ , such that  $\gamma_1(S^1) \cap \gamma_2(S^1) = \emptyset$ . The *Möbius cross energy* of the link  $(\gamma_1, \gamma_2)$  is defined to be

$$E(\gamma_1, \gamma_2) = \int_{S^1 \times S^1} \frac{|\gamma_1'(s)||\gamma_2'(t)|}{|\gamma_1(s) - \gamma_2(t)|^2} ds dt.$$

This energy was introduced in [41] and has the property, like the Willmore energy of surfaces, of conformal invariance.

It was conjectured by Freedman-He-Wang (1994, [41]) that the Möbius energy of any nontrivial link should be at least  $2\pi^2$ . The equality is attained by the stereographic projection of the so-called *standard Hopf link*:

$$\hat{\gamma}_1(s) = (\cos s, \sin s, 0, 0) \in S^3 \quad \text{and} \quad \hat{\gamma}_2(t) = (0, 0, \cos t, \sin t) \in S^3.$$

It follows from a result of He [54] that it suffices to prove the conjecture for links  $(\gamma_1, \gamma_2)$  that have linking number  $\text{lk}(\gamma_1, \gamma_2) = \pm 1$ . This is what we proved in [1]:

**Theorem 1.6** ([1]). *Let  $\gamma_i : S^1 \rightarrow \mathbb{R}^3$ ,  $i = 1, 2$ , be a 2-component link in  $\mathbb{R}^3$  with  $|\text{lk}(\gamma_1, \gamma_2)| = 1$ . Then  $E(\gamma_1, \gamma_2) \geq 2\pi^2$ . Moreover, if  $E(\gamma_1, \gamma_2) = 2\pi^2$  then there exists a conformal map  $F : \mathbb{R}^4 \rightarrow \mathbb{R}^4$  such that  $(F \circ \gamma_1, F \circ \gamma_2)$  describes the standard Hopf link up to orientation.*

We will give an idea about the proofs of these statements and what they have in common in Section 3.

Motivated by the results in the case of geodesics, Yau conjectured in [132] (first problem in the Minimal Surfaces section) that every compact Riemannian three-manifold admits an infinite number of smooth, closed, immersed minimal surfaces. In [81], we have been able to prove this conjecture in the positive Ricci curvature setting, or more generally, for manifolds  $(M, g)$  that satisfy the *embedded Frankel property*:

- any two smooth, closed, embedded minimal hypersurfaces of  $M$  intersect each other.

The author and A. Neves proved:

**Theorem 1.7** ([81]). *Let  $(M, g)$  be a compact Riemannian manifold of dimension  $(n + 1)$ , with  $2 \leq n \leq 6$ . Suppose that  $M$  satisfies the embedded Frankel property. Then  $M$  contains an infinite number of distinct smooth, closed, embedded, minimal hypersurfaces.*

Since manifolds of positive Ricci curvature satisfy the embedded Frankel property [38], we derive the following corollary:

**Corollary 1.8** ([81]). *Let  $(M, g)$  be a compact Riemannian  $(n + 1)$ -manifold with  $2 \leq n \leq 6$ . If the Ricci curvature of  $g$  is positive, then  $M$  contains an infinite number of distinct smooth, closed, embedded, minimal hypersurfaces.*

The proof of Theorem 1.7 uses the Almgren-Pitts min-max theory for the area functional, combined with ideas from Lusternik-Schnirelmann theory. The idea is to apply min-max theory to the multiparameter families of hypersurfaces (mod 2 cycles) studied by Gromov [44, 46, 47] and Guth [50].

In the case of generic metrics on three-manifolds, there is an alternative approach described by Kapouleas [65] to construct an infinite number of embedded minimal surfaces.

This works by either desingularizing two intersecting minimal surfaces or by doubling an existing unstable minimal surface. We also point out that Kahn and Markovic [64] proved the existence of infinitely many incompressible surfaces in compact hyperbolic three-manifolds. Hence, by minimization, every Riemannian metric on one of these manifolds admits infinitely many smooth immersed minimal surfaces.

An informal overview of the proof of Theorem 1.7 will be given in Section 4.

**1.7. Other advances.** Many other techniques have been employed in the study of minimal surfaces, like monotonicity formulas, the strong maximum principle, curvature estimates, Weierstrass representation and others. We describe recent and important advances in minimal surface theory that make use of some of these methods. We refer the reader to the book of Meeks and Pérez [82] and the references therein for more results of that type.

Colding and Minicozzi have developed a deep theory about the structure of arbitrary embedded minimal surfaces of bounded genus. As a consequence, they showed ([25]) that any complete embedded minimal surface with finite topology in  $\mathbb{R}^3$  must be proper, thereby proving that the Calabi-Yau conjectures for embedded surfaces are true. In particular, these results imply that the examples constructed by Jorge and Xavier [62] and by Nadirashvili [90] cannot be embedded.

An important theme in minimal surface theory is the search for classification results. Meeks and Rosenberg [83] have used the theory of Colding and Minicozzi to prove a great classification theorem: the plane and the helicoid are the only complete properly embedded simply connected minimal surfaces in  $\mathbb{R}^3$ . (The properness assumption can be removed because of [25].) This generalizes the classical Bernstein theorem: the only entire minimal graph in  $\mathbb{R}^3$  is the plane. (The Bernstein theorem holds true in  $\mathbb{R}^n$  for  $n \leq 8$ , and it is false if  $n \geq 9$ .) The Bernstein theorem has another beautiful extension, proven earlier and independently by do Carmo and Peng [32], Fischer-Colbrie and Schoen [37], and Pogorelov [103]: the only complete orientable stable minimal surface in  $\mathbb{R}^3$  is the plane. The nonorientable case has been recently settled by Ros [109], but the classification of stable minimal hypersurfaces in  $\mathbb{R}^n$  for  $n \geq 4$  is still an open problem.

Another important recent contribution to minimal surface theory fits perfectly into the classification theme. In [16], Brendle proved the Lawson conjecture ([73]): the only minimal embedded torus in  $S^3$  is the Clifford torus. The proof is based on the maximum principle technique, and it extends the work of Andrews [9] (on mean curvature flow) in an insightful way. The technique was used again by Andrews and Li [10] to confirm the classification of constant mean curvature tori in  $S^3$  conjectured by Pinkall and Sterling [101].

It is also fundamental to enrich the class of known examples of minimal surfaces. In the 1980s, Costa [27] discovered a very important one: the first finite topology properly embedded minimal surface in  $\mathbb{R}^3$  to be discovered after the plane, the catenoid and the helicoid. The embeddedness was proven by Hoffman and Meeks [58], who constructed an infinite family of similar examples. Recently, Hoffman, Traizet and White [59] have presented a construction, by variational means, of new minimal surfaces in  $\mathbb{R}^3$ : embedded helicoidal minimal surfaces of every genus  $g$ . These surfaces are limits of minimal surfaces in the homogeneous spaces  $S^2(r) \times \mathbb{R}$  as  $r \rightarrow \infty$ . There has been a lot of activity recently in the study of minimal and constant mean curvature surfaces in homogeneous three-manifolds (see Fernández and Mira [36] for a survey). Minimal surfaces in  $\mathbb{H} \times \mathbb{R}$ , for instance, have been used by Collin and Rosenberg [26] to construct harmonic diffeomorphisms from the complex plane  $\mathbb{C}$  onto the hyperbolic plane  $\mathbb{H}$ .

Minimal surfaces in  $S^3$  with arbitrary genus were constructed by Lawson [72]. Other examples were found by Karcher, Pinkall and Sterling [67]. This list has been recently enlarged with the examples of Kapouleas and Yang [66] (by gluing techniques) and of Choe and Soret [21] (by Lawson’s method). Minimal surfaces can also be constructed using degree arguments. In [128], White develops a mapping degree theory for closed minimal surfaces that implies every metric of positive Ricci curvature in  $S^3$  contains an embedded minimal torus and at least two embedded minimal spheres (the existence of one minimal sphere follows by min-max [124]).

Finally, we mention that there are several interesting connections between the theory of minimal hypersurfaces and the study of partial differential equations that come from phase transition theory. This is a very active field. We refer the reader to the article of Pacard [93] and the references therein.

## 2. Finite time extinction of Ricci flow

Let  $M^3$  be a closed orientable three-manifold that is prime and non-aspherical. In particular,  $\pi_3(M) \neq 0$  by standard topology. We consider the space  $\Omega$  of continuous maps

$$\varphi : S^2 \times [0, 1] \rightarrow M$$

such that  $\varphi(S^2 \times \{0\})$  and  $\varphi(S^2 \times \{1\})$  are points and so that

$$t \in [0, 1] \rightarrow \varphi(\cdot, t)$$

is a continuous map from  $[0, 1]$  to  $C^0 \cap W^{1,2}$ . Here  $W^{1,2}$  denotes the Sobolev space of maps  $f : S^2 \rightarrow M$  with first derivatives in  $L^2$ .

We denote by  $\Omega_\varphi$  the subspace of maps  $\psi \in \Omega$  that are homotopic to a given  $\varphi \in \Omega$  through maps in  $\Omega$ . A map  $\varphi \in \Omega$  naturally induces a continuous map  $\hat{\varphi} : S^3 \rightarrow M$ . We take  $\Omega_\varphi$  so that the corresponding map  $\hat{\varphi}$  represents a nontrivial element of  $\pi_3(M)$ . In that case we say that elements of  $\Omega_\varphi$  are sweepouts of  $M$ .

Given a Riemannian metric  $g$  on  $M$ , the *energy width* of  $\Omega_\varphi$  with respect to  $g$  is defined to be the min-max invariant:

$$W_E = W_E(\varphi, g) = \inf_{\psi \in \Omega_\varphi} \sup_{t \in [0,1]} E(\psi(\cdot, t)).$$

Recall that  $E(f) = \frac{1}{2} \int_{S^2} |df|^2 dv_{\bar{g}}$ , where  $\bar{g}$  is the standard metric of curvature one on  $S^2$ . If  $\varphi$  is a sweepout of  $M$ , then  $W_E > 0$ . We can define the *area width* of  $\Omega_\varphi$  with respect to  $g$  similarly by

$$W_A = W_A(\varphi, g) = \inf_{\psi \in \Omega_\varphi} \sup_{t \in [0,1]} \text{area}(\psi(\cdot, t)),$$

where  $\text{area}(f) = \int_{S^2} \text{Jac}(f) dv_{\bar{g}}$ .

If  $\{e_1, e_2\}$  is an orthonormal basis of  $T_p S^2$ , then

$$\text{Jac}(f)(p) = \sqrt{|df(e_1)|^2 |df(e_2)|^2 - \langle df(e_1), df(e_2) \rangle^2} \leq \frac{1}{2} (|df(e_1)|^2 + |df(e_2)|^2),$$

hence  $\text{area}(f) \leq E(f)$  with equality if and only if  $f$  is almost conformal, i.e.,  $\langle df(e_1), df(e_2) \rangle = 0$  and  $|df(e_1)| = |df(e_2)|$  almost everywhere. It follows immediately that  $W_A \leq W_E$ . By

conformally reparametrizing, through the Riemann mapping theorem for variable metrics (see [63]), the maps  $\varphi_i(\cdot, t)$  of a minimizing sequence  $\varphi_i \in \Omega_\varphi$  for  $W_A$ :

$$\lim_{i \rightarrow \infty} \sup_{t \in [0,1]} \text{area}(\varphi_i(\cdot, t)) = W_A,$$

Colding and Minicozzi showed that  $W_E \leq W_A$ . Hence  $W_E = W_A$ . In particular, any sequence of sweepouts  $\varphi_i \in \Omega_\varphi$  that is minimizing for the energy, i.e., such that

$$\lim_{i \rightarrow \infty} \sup_{t \in [0,1]} E(\varphi_i(\cdot, t)) = W_E,$$

is also minimizing for the area:  $\lim_{i \rightarrow \infty} \sup_{t \in [0,1]} \text{area}(\varphi_i(\cdot, t)) = W_A$ .

The theorem below establishes the existence of good minimizing sequences of sweepouts. Recall that any harmonic map  $u : S^2 \rightarrow M$  is a conformal branched minimal immersion.

**Theorem 2.1** ([24]). *Suppose that  $\varphi : S^2 \times [0, 1] \rightarrow M$  induces a homotopically nontrivial map  $\hat{\varphi} : S^3 \rightarrow M^3$ . For any Riemannian metric  $g$  on  $M$ , there exists a sequence of sweepouts  $\varphi_i \in \Omega_\varphi$  with*

$$\lim_{i \rightarrow \infty} \sup_{t \in [0,1]} E(\varphi_i(\cdot, t)) = W_E = W$$

and such that any sufficiently large slice (in area) is close in varifold sense to a union of branched minimal spheres. More precisely, for any  $\varepsilon > 0$  we can find  $i_0 \in \mathbb{N}$  and  $\delta > 0$  so that if  $i \geq i_0$  and  $s \in [0, 1]$  satisfy

$$\text{area}(\varphi_i(\cdot, s)) \geq W - \delta,$$

then

$$\mathbf{F}(\varphi_i(\cdot, s), \cup_k \{u_k\}) \leq \varepsilon$$

for some finite union  $\cup\{u_k\}$  of harmonic maps  $u_k : S^2 \rightarrow M$ .

In the above statement  $\mathbf{F}$  denotes the  $\mathbf{F}$ -metric, that gives the weak topology to the space of varifolds. We identify a map  $f : S^2 \rightarrow M$  with the two-dimensional rectifiable varifold  $f_\#(S^2)$  it induces in  $M$ .

The sequence  $\varphi_i$  is obtained from an arbitrary minimizing sequence  $\psi_i$  for the energy through harmonic replacement, a process similar in spirit to Birkhoff’s curve shortening deformation [13]. The convergence to the branched minimal spheres is inspired by compactness theorems for harmonic maps with bounded energy ([63], [94], [95], [122]). The fact that the convergence is in varifold sense implies, in particular, that there is no loss of area or energy in the limit.

Suppose now that we evolve the metric on  $M^3$  by Hamilton’s Ricci flow [51]:

$$\frac{\partial g}{\partial t} = -2\text{Ric}_{g(t)}.$$

Let  $[0, T)$  be the maximal time of smooth existence. If  $\varphi : S^2 \times [0, 1] \rightarrow M$  induces a homotopically nontrivial map  $\hat{\varphi} : S^3 \rightarrow M^3$ , the function  $W(t) = W_E(\varphi, g(t))$  satisfies  $W(t) > 0$  for all  $t \in [0, T)$ . The idea is to study the rate of change of the Lipschitz function

$W(t)$  and show that it must decrease to zero in finite time. Notice that before the work of Perelman it was not even known whether an arbitrary Ricci flow in a manifold like  $S^3$  would develop a singularity in finite time. It is quite remarkable that this can be proved using minimal surface techniques.

Let  $\Sigma$  be a branched minimal two-sphere in  $(M^3, g(t_0))$ . We denote by  $N$  its unit normal vector field. The scalar curvature of a metric  $g$  is denoted by  $R_g$ . We use  $K_\Sigma$  to indicate the Gauss curvature of the induced metric on  $\Sigma$ , and  $A$  to indicate its second fundamental form. The rate of change of the area of a minimal surface under Ricci flow was first considered by Hamilton [52]. We have

$$\begin{aligned} \frac{d}{dt}\Big|_{t=t_0} \text{area}_{g(t)}(\Sigma) &= \frac{1}{2} \int_{\Sigma} \text{tr}_{\Sigma}(-2\text{Ric}_{g(t_0)})d\Sigma \\ &= -\frac{1}{2} \int_{\Sigma} R_{g(t_0)}d\Sigma - \int_{\Sigma} K_{\Sigma}d\Sigma - \frac{1}{2} \int_{\Sigma} |A|^2d\Sigma \\ &\leq -\frac{1}{2} \int_{\Sigma} R_{g(t_0)}d\Sigma - 4\pi - 2\pi \sum_i b_i, \end{aligned}$$

where we have used the Gauss equation and the Gauss-Bonnet Theorem with finitely many branch points  $p_i$  of orders  $b_i > 0$ . Hence

$$\frac{d}{dt}\Big|_{t=t_0} \text{area}_{g(t)}(\Sigma) \leq -4\pi - \frac{1}{2} \int_{\Sigma} R_{g(t_0)}d\Sigma.$$

On the other hand, the scalar curvature of a Ricci flow  $g(t)$  satisfies the evolution equation:

$$\frac{\partial}{\partial t} R_{g(t)} = \Delta_{g(t)} R_{g(t)} + 2|\text{Ric}_{g(t)}|^2.$$

Therefore  $\frac{\partial}{\partial t} R_{g(t)} \geq \Delta_{g(t)} R_{g(t)} + \frac{2}{3} R_{g(t)}^2$ , and the maximum principle for parabolic equations implies

$$\min_M R_{g(t)} \geq \frac{1}{\frac{1}{\min_M R_{g(0)}} - \frac{2}{3}t}$$

for all  $t \in [0, T)$ .

Putting things together, we get that

$$\frac{d}{dt}\Big|_{t=t_0} \text{area}_{g(t)}(\Sigma) \leq -4\pi + \frac{3}{4(t_0 + C)} \text{area}_{g(t_0)}(\Sigma)$$

for any branched minimal sphere in  $(M, g(t_0))$ , where the constant  $C$  depends only on  $g(0)$ .

By comparison arguments and Theorem 2.1, Colding and Minicozzi proved:

**Theorem 2.2** ([23]). *Let  $M^3$  be a closed orientable three-manifold that is prime and non-spherical, and suppose that  $\varphi : S^2 \times [0, 1] \rightarrow M$  induces a homotopically nontrivial map  $\hat{\varphi} : S^3 \rightarrow M^3$ . Then, for any Ricci flow  $g(t)$  on  $M$ , we have*

$$\frac{d}{dt} W(t) \leq -4\pi + \frac{3}{4(t + C)} W(t)$$

in the sense of the limsup of forward difference quotients, where  $W(t) = W(\varphi, g(t))$  and the constant  $C > 0$  depends only on  $g(0)$ .

This implies

$$W(t)(t + C)^{-3/4} \leq -16\pi(t + C)^{1/4} + C'$$

for some constant  $C'$  independent of  $t$ . Since we always have  $W(t) > 0$ , the flow cannot exist for all time. This already implies that any Ricci flow in a prime non-aspherical orientable closed three-manifold must develop a singularity in finite time. By arguing as in Perelman [99], Theorem 2.2 holds for Ricci flow with surgeries as well, thereby implying that any Ricci flow with surgery on one of these manifolds must become extinct in finite time.

### 3. Conformally invariant variational problems

In this section we will describe some recent applications of the variational theory of minimal surfaces to global problems about the geometry of surfaces and links in three-space. We begin by describing the min-max procedure used to produce the minimal hypersurface  $\Sigma$  of Theorem 1.1. For simplicity, we restrict to the case  $n = 3$ .

A family of closed surfaces  $\{\Sigma(t)\}_{t \in [0,1]}$  of  $M^3$  (where closed surface here means a two-dimensional integral cycle) is called a *sweepout* if we can write  $\Sigma(t) = \partial\Omega(t)$  with the domains (integral 3-currents)  $\Omega(t)$  varying continuously and satisfying  $\Omega(0) = 0$  and  $\Omega(1) = M$ . In particular we have  $\Sigma(0) = \Sigma(1) = 0 \in \mathcal{Z}_2(M^3)$ . A standard example of a sweepout is obtained by choosing a Morse function  $f : M \rightarrow \mathbb{R}$ , with  $f(M) = [0, 1]$ , and considering

$$\begin{aligned} \Omega(t) &= \{x \in M : f(x) < t\} \\ \Sigma(t) &= \partial\Omega(t) = \{x \in M : f(x) = t\}. \end{aligned}$$

We denote by  $\Pi_1$  the class of all sweepouts  $\{\Sigma(t)\}_{t \in [0,1]}$  of  $M^3$ , and define the min-max invariant called the *width* of  $\Pi_1$ :

$$W(\Pi_1) = \inf_{\{\Sigma(t)\} \in \Pi_1} \sup_{t \in [0,1]} \text{area}(\Sigma(t)).$$

Since there will be some  $t_0 \in [0, 1]$  such that  $\text{vol}(\Omega(t_0)) = \text{vol}(M)/2$ , the isoperimetric inequality implies that  $W(\Pi_1) > 0$ . The minimal surface  $\Sigma$  produced by Theorem 1.1, which could have several connected components with integer multiplicities, is constructed to satisfy

$$\text{area}(\Sigma) = W(\Pi_1).$$

If the ambient manifold is the three-dimensional unit sphere  $S^3 \subset \mathbb{R}^4$ , the minimal surface produced by doing min-max over the class  $\Pi_1$  is, modulo rotations, the equator or great sphere:  $\bar{\Sigma} = S^3 \cap \{x_4 = 0\}$ . In other words,  $W(\Pi_1) = 4\pi$  and is achieved precisely by the great spheres.

As a consequence, we have:

**Theorem 3.1** ( $4\pi$  Theorem). *Let  $\Phi : I \rightarrow \mathcal{Z}_2(S^3)$  be a sweepout of  $S^3$ . Then there exists  $y \in [0, 1]$  such that*

$$\text{area}(\Phi(y)) \geq 4\pi.$$

The simplest minimal surface in  $S^3$  after the equator is the Clifford torus

$$\hat{\Sigma} = S^1\left(\frac{1}{\sqrt{2}}\right) \times S^1\left(\frac{1}{\sqrt{2}}\right),$$

with area  $2\pi^2$  and Morse index 5. In fact, the Clifford torus can be characterized by its index:

**Theorem 3.2** (Urbano, [127]). *Let  $\Sigma \subset S^3$  be a smooth closed minimal surface of genus  $g \geq 0$  and  $\text{index}(\Sigma) \leq 5$ . Then  $\Sigma$  is either a great sphere (with index 1) or the Clifford torus (with index 5), up to ambient isometries.*

The previous discussion implies that the great sphere can appear as a min-max minimal surface, by considering one-parameter sweepouts. The question we posed ourselves, and that it turned out to be key to the solution of well-known global problems in conformal geometry such as the Willmore conjecture, was the following:

**Question.** *Is it possible to produce the Clifford torus by min-max methods?*

We have answered this question affirmatively by discovering a natural class of five-parameter families of surfaces in  $S^3$  with interesting topological properties. The families we have discovered are parametrized by a map  $\Phi$  defined on the 5-cube  $I^5$ , and satisfy:

- (A1)  $\Phi(x, 0) = \Phi(x, 1) = 0$  (trivial surface) for any  $x \in I^4$ ,
- (A2)  $\{\Phi(x, t)\}_{t \in [0,1]}$  is the standard sweepout of  $S^3$  by oriented round spheres centered at some  $Q(x) \in S^3$ , for any  $x \in \partial I^4$ ,
- (A3)  $\Phi(x, 1/2) = \partial B_{\pi/2}(Q(x))$ , for any  $x \in \partial I^4$ ,
- (A4)  $\deg(Q) \neq 0$ ,
- (A5) there is no concentration of area:

$$\limsup_{r \rightarrow 0} \{\text{area}(\Phi(x) \cap B_r(p)) : p \in S^3, x \in I^5\} = 0.$$

The property (A4) above is saying that the restriction of the map  $\Phi$  to  $\partial I^4 \times \{1/2\}$  is a homotopically nontrivial map into the space of oriented great spheres, which is homeomorphic to  $S^3$ . This is the crucial topological condition that will rule out the possibility of producing great spheres by min-max over families homotopic to  $\Phi$ . The property (A5) is more of a technical nature. It is used in [80] in the proofs of the interpolation statements.

The min-max theory developed jointly with Neves in [80] implies:

**Theorem 3.3** ( $2\pi^2$  Theorem, [80]). *Let  $\Phi : I^5 \rightarrow \mathcal{Z}_2(S^3)$  be a continuous map in the flat topology satisfying the properties (A1)-(A5) above. Then there must exist  $y \in I^5$  with*

$$\text{area}(\Phi(y)) \geq 2\pi^2.$$

The surprising thing is that this is intimately related to the Willmore conjecture, a problem that we briefly discussed in Section 1.6. Let us describe this connection now.

First note that the torus  $\Sigma_{\sqrt{2}}$  of Section 1.6 is very special. In order to see this we need to recall that the stereographic projection  $\pi : S^3 \setminus \{p\} \rightarrow \mathbb{R}^3$ ,  $p \in S^3$ , is a conformal transformation. Since the Willmore energy is conformally invariant, we can formulate the

Willmore conjecture equivalently as a problem for surfaces in the three-sphere  $S^3$ . If  $\Sigma \subset S^3 \setminus \{p\}$ , we can calculate the energy of the projection  $\tilde{\Sigma} = \pi(\Sigma) \subset \mathbb{R}^3$ :

$$\int_{\tilde{\Sigma}} \tilde{H}^2 d\tilde{\Sigma} = \int_{\Sigma} (1 + H^2) d\Sigma,$$

where  $H$  now denotes the mean curvature of  $\Sigma$  with respect to the spherical geometry.

Hence we define the *Willmore energy* of  $\Sigma \subset S^3$  by

$$\mathcal{W}(\Sigma) = \int_{\Sigma} (1 + H^2) d\Sigma.$$

One advantage of considering the problem for surfaces in  $S^3$  is that a relation with the area functional becomes apparent. It follows immediately from the above definition that for  $\Sigma \subset S^3$  one always has  $\mathcal{W}(\Sigma) \geq \text{area}(\Sigma)$ , and  $\mathcal{W}(\Sigma) = \text{area}(\Sigma)$  if and only if  $\Sigma$  is a minimal surface. If  $\hat{\Sigma}$  denotes the Clifford torus  $S^1(1/\sqrt{2}) \times S^1(1/\sqrt{2}) \subset S^3$ , it is no coincidence that if we choose the right stereographic projection we will have  $\pi(\hat{\Sigma}) = \Sigma_{\sqrt{2}}$ .

The Willmore conjecture has a long history of partial results. We refer the reader to our paper [80] for an account. We were motivated by the work of Ros [108] on the antipodally symmetric case (see Topping [126] for a different proof), who used an area estimate we will mention in the sequel (Theorem 3.7).

A result of particular relevance to our approach is:

**Theorem 3.4** (Li and Yau, [75]). *If  $F : \Sigma \rightarrow S^3$  is an immersion and there exists  $p \in S^3$  such that  $\#F^{-1}(p) = k$ , then  $\mathcal{W}(\Sigma) \geq 4\pi k$ . In particular, if  $\Sigma$  is not embedded then  $\mathcal{W}(\Sigma) \geq 8\pi$ .*

Because of the result of Li and Yau, we may assume the surface is embedded. Together with Neves we proved the following theorem that implies the Willmore conjecture:

**Theorem 3.5** ([80]). *Let  $\Sigma \subset S^3$  be a closed, embedded smooth surface with genus  $g \geq 1$ . Then  $\mathcal{W}(\Sigma) \geq 2\pi^2$ , and  $\mathcal{W}(\Sigma) = 2\pi^2$  if and only if  $\Sigma$  is a conformal image of the Clifford torus.*

**Remark 3.6.** The theorem above is completely equivalent to Theorem 1.4 mentioned in Section 1.6.

The existence of a torus that minimizes the Willmore energy was established by Simon [120]. His work was later extended to surfaces of higher genus by Bauer and Kuwert [12] (see also [69]). Very little is known about these surfaces and their energies. It is known that the minimum energy  $\omega_g$  for orientable closed surfaces of genus  $g$  in  $\mathbb{R}^3$  is bigger than  $2\pi^2$ , less than  $8\pi$ , and converges to  $8\pi$  as  $g \rightarrow \infty$  [71].

It is also interesting to study general critical points of the Willmore energy, called Willmore surfaces. These are closed surfaces in  $\mathbb{R}^3$  that satisfy the fourth-order Euler-Lagrange equation:

$$\Delta H + 2(H^2 - K)H = 0,$$

where  $K$  denotes the Gauss curvature. The simplest examples are stereographic projections of minimal surfaces of  $S^3$ , but there are many more. Bryant [17] found and classified all critical points of genus zero, and Pinkall [100] constructed infinitely many embedded Willmore tori that are not the conformal image of a minimal surface. The understanding of the

analytical aspects of the Willmore equation has been greatly improved in recent years thanks to the work of Kuwert-Schätzle (e.g. [70]) and Rivière (e.g. [106]). The lecture notes of Rivière [107] provide a great introduction to the analysis behind conformally invariant variational problems. We point also that the Willmore energy appears in many other interesting contexts, like in general relativity (as the main term in the definition of the Hawking mass) and in relation with the renormalized area functional in the AdS/CFT correspondence, as recently studied by Alexakis and Mazzeo [2].

In what follows we give an idea of the proofs of the  $2\pi^2$  Theorem and Theorem 3.5, emphasizing their intimate relation. We start by constructing, for each embedded closed surface  $\Sigma \subset S^3$ , a *canonical family* of surfaces  $\Sigma_{(v,t)} \subset S^3$ , where  $(v,t) \in B^4 \times (-\pi, \pi)$ , with the properties that:

- $\Sigma_{(0,0)} = \Sigma$ ,
- $\text{area}(\Sigma_{(v,t)}) \leq \mathcal{W}(\Sigma)$  for every  $(v,t) \in B^4 \times (-\pi, \pi)$ .

Here  $B^4 \subset \mathbb{R}^4$  denotes the open unit ball. Each surface  $\Sigma_{(v,t)}$  is an equidistant surface of some conformal image of  $\Sigma$ .

The fact that  $\text{area}(\Sigma_{(v,t)}) \leq \mathcal{W}(\Sigma)$  for every  $(v,t)$  follows by combining the conformal invariance of the Willmore energy with the area estimate:

**Theorem 3.7.** *Let  $\Sigma_t, t \in (-\pi, \pi)$ , be an equidistant surface of an embedded closed surface  $\Sigma \subset S^3$ . Then*

$$\text{area}(\Sigma_t) \leq \mathcal{W}(\Sigma).$$

*Moreover, if  $\Sigma$  is not a geodesic sphere and*

$$\text{area}(\Sigma_t) = \mathcal{W}(\Sigma),$$

*then  $t = 0$  and  $\Sigma$  is a minimal surface.*

**Remark 3.8.** Theorem 3.7 is a particular case of more general estimates proved by Heintze and Karcher [55]. The statement above was used in connection with the Willmore problem by Ros [108].

The next step is to understand the geometric and topological properties of the canonical family, especially the behavior as we let the parameter  $(v,t)$  converge to the boundary of the parameter space  $\partial(\overline{B^4} \times [-\pi, \pi])$ . We prove that:

- for any sequence  $\{(v_i, t_i)\}$  that converges to  $\partial(\overline{B^4} \times [-\pi, \pi])$ , a subsequence of  $\{\Sigma_{(v_i, t_i)}\}$  converges in the flat topology to some round sphere (possibly trivial) in  $S^3$ .

The subtle case to consider is when  $v$  converges to a point  $p$  on the surface  $\Sigma$ , because then the limit is not unique and depends on the angle of convergence at which  $v$  approaches  $p$ . We perform a blow-up procedure along the surface  $\Sigma$  to solve this problem of nonuniqueness of the limit. After reparametrizing the family, and observing that  $\overline{B^4} \times [-\pi, \pi]$  is homeomorphic to  $I^5$ , we get a family  $\Phi : I^5 \rightarrow \mathcal{Z}_2(S^3)$  with properties (A1)-(A3) and (A5) above, and an explicit center map  $Q : \partial I^4 \rightarrow S^3$ .

The main topological ingredient is the discovery that the genus of the original surface  $\Sigma$  can be read off the topological properties of the canonical family at the boundary:

**Theorem 3.9** ([80]).  $\deg(Q) = \text{genus}(\Sigma)$ .

Summarizing, we obtain:

**Theorem 3.10** ([80]). *Let  $\Sigma \subset S^3$  be an embedded closed surface of genus  $g \geq 1$ . The map  $\Phi : I^5 \rightarrow \mathcal{Z}_2(S^3)$  is continuous in the flat topology, satisfies the properties (A1)-(A5) and*

$$\sup\{\text{area}(\Phi(x)) : x \in I^5\} \leq \mathcal{W}(\Sigma).$$

Informally, the min-max family  $\Phi$  can be thought of as an element of the relative homotopy group  $\pi_5(\mathcal{S}, \mathcal{G})$ , where  $\mathcal{S}$  denotes the space of two-surfaces in  $S^3$  and  $\mathcal{G}$  denotes the space of round spheres.

Given a smooth, embedded, closed surface  $\Sigma \subset S^3$  with genus  $g \geq 1$ , it follows from the above properties that the map  $\Phi : I^5 \rightarrow \mathcal{Z}_2(S^3)$  satisfies all the assumptions of the  $2\pi^2$  Theorem. Therefore there exists  $y \in I^5$  such that  $\text{area}(\Phi(y)) \geq 2\pi^2$ . Since  $\text{area}(\Phi(x)) \leq \mathcal{W}(\Sigma)$  for every  $x \in I^5$ , we get  $\mathcal{W}(\Sigma) \geq 2\pi^2$ . In particular this proves the Willmore conjecture. The rigidity statement can be derived from the particular structure of the canonical family and the equality case in the area estimate of Ros.

We apply the Almgren-Pitts min-max theory for the area functional in order to prove the  $2\pi^2$  Theorem. Given a family  $\Phi : I^5 \rightarrow \mathcal{Z}_2(S^3)$  with properties (A1)-(A5), we consider  $\Pi$  the homotopy class of  $\Phi$  relative to  $\partial I^5$ . Of course we have

$$\sup\{\text{area}(\Phi(x)) : x \in \partial I^5\} = 4\pi.$$

We first rule out great spheres as possible min-max surfaces for  $\Pi$ , by proving:

**Theorem 3.11** ([80]).  $L(\Pi) > 4\pi$ .

The proof is topological and goes by contradiction, by assuming  $L(\Pi) = 4\pi$ . In order to illustrate the idea, we suppose there is a map  $\tilde{\Phi} \in \Pi$  with

$$\sup_{x \in I^5} \text{area}(\tilde{\Phi}(x)) = 4\pi.$$

In particular, for any given continuous path  $\gamma : [0, 1] \rightarrow I^5$  with  $\gamma(0) \in I^4 \times \{0\}$  and  $\gamma(1) \in I^4 \times \{1\}$ ,  $\{\tilde{\Phi} \circ \gamma\}$  is optimal as a one-parameter sweepout of  $S^3$ . Therefore it must contain a great sphere. One could argue then that there must exist a 4-dimensional submanifold  $R \subset I^5$ , separating the top from the bottom of  $I^5$ , such that

- $\tilde{\Phi}(y)$  is a great sphere for any  $y \in R$ ,
- $\partial R = \partial I^4 \times \{1/2\}$ .

If  $\tilde{Q}(y)$  denotes the center of the great sphere  $\tilde{\Phi}(y)$ , for  $y \in R$ , and since  $\tilde{\Phi} = \Phi$  on  $\partial I^5$ , we get

$$[Q_{\#}(\partial I^4)] = [\tilde{Q}_{\#}(\partial R)] = [\partial \tilde{Q}_{\#}(R)] = 0 \in H_3(S^3, \mathbb{Z}).$$

But  $Q_{\#}(\partial I^4) = \deg(Q) \cdot S^3$ , and we reach a contradiction since  $\deg(Q) \neq 0$ . The complete proof is more involved and can be found in [80].

Once we know that  $L(\Pi) > 4\pi$ , here is how we prove that in reality  $L(\Pi) \geq 2\pi^2$ . By applying min-max theory to  $\Pi$ , we get the existence of a smooth embedded minimal surface  $\Sigma' \subset S^3$  such that

$$L(\Pi) = \text{area}(\Sigma') > 4\pi.$$

If  $\Sigma'$  has multiplicity bigger than one, then  $L(\Pi) \geq 8\pi$ . If the multiplicity is one, and since by Almgren [6] the only minimal spheres in  $S^3$  are the great spheres, we get  $\text{genus}(\Sigma') \geq 1$ .

The result follows once we show that the nonspherical minimal surface of lowest area  $\hat{\Sigma}$  in  $S^3$  is the Clifford torus. The idea is to use Urbano's theorem. If we had  $\text{index}(\hat{\Sigma}) \geq 6$ , we would be able to slightly perturb its canonical family and produce, by min-max theory and Theorem 3.11, a minimal surface whose area is strictly between  $4\pi$  and  $\text{area}(\hat{\Sigma})$ . Contradiction, hence  $\text{index}(\hat{\Sigma}) \leq 5$  and  $\hat{\Sigma}$  must be the Clifford torus by Urbano's theorem. Of course,  $\text{area}(\Sigma') \geq \text{area}(\hat{\Sigma})$ . This finishes a sketch of the proof that  $L(\Pi) \geq 2\pi^2$ , and this implies the  $2\pi^2$  Theorem.

**Remark 3.12.** Note that in the process of proving the  $2\pi^2$  Theorem we have showed Theorem 1.5.

**3.1. Links.** In order to prove Theorem 1.6, we will again use the  $2\pi^2$  Theorem. The basic observation is that if  $g$  denotes the Gauss map of a link  $(\gamma_1, \gamma_2)$  contained in  $S^3$ , i.e., the map  $g : S^1 \times S^1 \rightarrow S^3$  defined by

$$g(s, t) = \frac{\gamma_1(s) - \gamma_2(t)}{|\gamma_1(s) - \gamma_2(t)|},$$

then  $|\text{Jac } g|(s, t) \leq \frac{|\gamma_1'(s)||\gamma_2'(t)|}{|\gamma_1(s) - \gamma_2(t)|^2}$ . Hence  $\text{area}(g(S^1 \times S^1)) \leq E(\gamma_1, \gamma_2)$ .

By applying conformal transformations to the curves  $\gamma_1, \gamma_2$  of a link  $(\gamma_1, \gamma_2)$  in  $S^3 \subset \mathbb{R}^4$ , and considering the associated Gauss maps, we get a 5-parameter family of surfaces (parametrized tori) in  $S^3$ . The key again is to analyze the boundary behavior. After an extension, we get a family with the same basic properties of the canonical family for the Willmore problem, and such that the area of any surface in the family is bounded above by the Möbius energy of the link. We prove that if  $|\text{lk}(\gamma_1, \gamma_2)| = 1$  then the center map  $Q : \partial I^4 \rightarrow S^3$  associated with the family satisfies  $|\text{deg}(Q)| = 1$ . Therefore the  $2\pi^2$  Theorem applies and we conclude the existence of at least one surface in the family with area greater than or equal to  $2\pi^2$ . This establishes the inequality  $E(\gamma_1, \gamma_2) \geq 2\pi^2$ , and after some extra work one can also prove the rigidity part.

### 4. Further directions

In this section we describe some recent advances in min-max theory and discuss some future directions. We start with the problem of counting minimal hypersurfaces in Riemannian manifolds. We will give an informal overview of the proof of Theorem 1.7.

Let  $M^{n+1}$  be a Riemannian manifold as in Theorem 1.7. The homotopy groups of the space of modulo 2  $n$ -cycles in  $M$ ,  $\mathcal{Z}_n(M, \mathbb{Z}_2)$ , can be computed through the work of Almgren [4]. All homotopy groups vanish but the first one:  $\pi_1(\mathcal{Z}_n(M, \mathbb{Z}_2)) = \mathbb{Z}_2$ , just like for the topological space  $\mathbb{R}\mathbb{P}^\infty$ . Let  $\bar{\lambda} \in H^1(\mathcal{Z}_n(M, \mathbb{Z}_2), \mathbb{Z}_2)$  be the generator.

Gromov [44, 46, 47] and Guth [50] have studied continuous maps  $\Phi$  from a simplicial complex  $X$  into  $\mathcal{Z}_n(M, \mathbb{Z}_2)$  that detect  $\bar{\lambda}^p$ , in the sense that  $\Phi^*(\bar{\lambda}^p) \neq 0$ . Here  $\bar{\lambda}^p$  denotes the  $p$ -th cup power of  $\bar{\lambda}$ . We call these maps  $p$ -sweepouts. An example can be given by starting with a Morse function  $f : M \rightarrow \mathbb{R}$ . The open set  $\{x \in M : f(x) < t\}$  has finite perimeter for all  $t$ , hence we have a well-defined element

$$f^{-1}(t) = \partial\{x \in M : f(x) < t\} \in \mathcal{Z}_n(M; \mathbb{Z}_2).$$

For each  $a = (a_0, \dots, a_p) \in \mathbb{R}^{p+1}$ ,  $|a| = 1$ , we consider the polynomial  $P_a(t) = \sum_{i=0}^p a_i t^i$  and define the map

$$\Psi : \{a \in \mathbb{R}^{p+1} : |a| = 1\} \rightarrow \mathcal{Z}_n(M; \mathbb{Z}_2)$$

by

$$\Psi(a_0, \dots, a_p) = \partial \{x \in M : P_a(f(x)) < 0\}.$$

The fact that we are using  $\mathbb{Z}_2$  coefficients implies that  $\Psi(a) = \Psi(-a)$ , and therefore  $\Psi$  induces a map  $\Phi : \mathbb{R}\mathbb{P}^p \rightarrow \mathcal{Z}_n(M; \mathbb{Z}_2)$ . It satisfies  $\Phi^*(\bar{\lambda}^p) \neq 0$ .

We denote by  $\mathcal{P}_p$  the space of all maps that detect  $\bar{\lambda}^p$ , and define the min-max invariant:

$$\omega_p(M) := \inf_{\Phi \in \mathcal{P}_p} \sup_{x \in \text{dmn}(\Phi)} \text{area}(\Phi(x)),$$

where  $\text{dmn}(\Phi)$  stands for the domain of  $\Phi$ . The asymptotic behavior of the min-max  $n$ -volumes  $\omega_p(M)$  as  $p \rightarrow \infty$  has been studied previously by Gromov and Guth. The following result was proven by Gromov in [44, Section 4.2.B], and by Guth in [50] via an elegant bend-and-cancel argument.

**Theorem 4.1.** *There exists a constant  $C = C(M) > 0$  so that*

$$\omega_p(M) \leq Cp^{\frac{1}{n+1}}$$

for every  $p \in \mathbb{N}$ .

*Remark.* The lower bound  $\omega_p(M) \geq C'p^{\frac{1}{n+1}}$  also holds for some constant  $C' = C'(M) > 0$  (Gromov [46], Guth [50]).

We use Lusternik-Schnirelmann theory to show that if  $\omega_p = \omega_{p+1}$  then there are infinitely many embedded minimal hypersurfaces. This is inspired by the particular structure of the cohomology ring of a projective space. Details can be found in our paper [81].

We then prove Theorem 1.7 by contradiction, assuming that the set  $\mathcal{L}$  of all smooth, closed, embedded minimal hypersurfaces in  $M$  is finite. This implies that the sequence  $\{\omega_p(M)\}_{p \in \mathbb{N}}$  is strictly increasing. Since the support of the Almgren-Pitts min-max minimal surface is always embedded, the Frankel property implies that it must have the form  $k \cdot \Sigma$  for some  $\Sigma \in \mathcal{L}$ . We conclude, by applying min-max theory to the classes of  $p$ -sweepouts and Theorem 4.1, that

$$\begin{aligned} \#\{a = k|\Sigma| : k \in \mathbb{N}, \Sigma \in \mathcal{L}, k|\Sigma| \leq Cp^{\frac{1}{n+1}}\} \\ \geq \#\{\omega_k(M) : k = 1, \dots, p\} = p. \end{aligned}$$

Since  $\mathcal{L}$  is assumed to be finite, we get a contradiction for sufficiently large  $p$ .

**4.1. Open problems.** We describe some open problems related to the variational theory of minimal surfaces. Some of these questions are well-known and others arose from extensive discussions with André Neves.

The precise relation between the Morse index of the min-max minimal surface and the number of parameters is not known in the Almgren-Pitts min-max theory. In general, one should expect that  $\text{index}(\Sigma) \leq k$ , where  $k$  is the number of parameters. This is a subtle question, specially because of the phenomenon of multiplicity. In [133], Zhou proves this

for  $k = 1$  in the case of compact manifolds  $M^n$  of positive Ricci curvature, with  $3 \leq n \leq 7$ . This extends the results for  $n = 3$  of the author and Neves [79].

It is also natural to ask whether one can control the topology of  $\Sigma$  produced by Theorem 1.1. This problem has been studied in the three-dimensional case through variants of the Almgren-Pitts theory due to Simon-Smith [124] and Colding and De Lellis [22], in which the one-parameter sweepouts are made of actual smooth closed surfaces (perhaps up to finitely many singularities) with bounded genus. Simon and Smith proved in [124], by considering sweepouts of the form  $\Sigma(t) = \psi(t, \bar{\Sigma}(t))$ , where  $\bar{\Sigma}(t) = \{x \in S^3 : x_4 = 2t - 1\}$  is the standard foliation by round spheres and  $\psi : [0, 1] \times S^3 \rightarrow S^3$  is an ambient isotopy, that every Riemannian metric on  $S^3$  admits a minimal embedded two-sphere.

These theories can also deal with the case of sweepouts induced by Heegaard splittings. In the 1980s Pitts and Rubinstein made a number of conjectures about the index and genus of the min-max minimal surface obtained this way. They have conjectured, for instance, that after performing finitely many surgeries of controlled type to the Heegaard surfaces, each remaining component should be isotopic to a component of the min-max minimal surface or to a double cover of a component. A proof of this last conjecture, which in turn implies basically optimal genus bounds, has been recently presented by Ketover [68] (a previous result on genus bounds can be found in De Lellis and Pellandini [29]). It should be interesting to bound the Betti numbers of a min-max minimal hypersurface in higher dimensions. A shorter proof of Theorem 1.1, still in the setting of one-parameter sweepouts, was given by De Lellis and Tasnady in [31].

Are the Clifford hypersurfaces

$$\Sigma_{k+1,l+1} = S^k(\sqrt{\frac{k}{k+l}}) \times S^l(\sqrt{\frac{l}{k+l}}) \subset S^n(1),$$

$k + l = n - 1$ , the only non-equatorial minimal hypersurfaces of  $S^n$  with Morse index less than or equal to  $(n + 2)$ ? If  $n = 3$  this follows from Urbano [127] (Theorem 3.2). See Perdomo [96] for the antipodally symmetric case.

The minimal hypersurfaces  $\Sigma_{m,m}$  and  $\Sigma_{m,m+1}$  are conjectured (Solomon) to have least area among all non-equatorial minimal hypersurfaces of  $S^{2m-1}$  and  $S^{2m}$ , respectively. Similarly, the respective cones  $C_{m,m}$  and  $C_{m,m+1}$  over  $\Sigma_{m,m}$  and  $\Sigma_{m+1,m}$  should be the non-trivial minimal hypercones of least possible density. For  $m = 1$  these conjectures follow from the results of [80]. See also Ilmanen and White [61] for some partial results.

The general behavior of singularities of minimal submanifolds is very little understood. For instance, we do not know whether there exists a properly embedded minimal surface in  $B_1^3(0) \setminus \{0\}$  that does not extend smoothly to the origin. Another question (Schoen) is whether singularities of area-minimizing hypersurfaces are stable under small perturbations of the data. See N. Smale [123] for the case  $n = 8$ .

Gromov [44] has proposed to consider the sequence  $\{\omega_p(M)\}_{p \in \mathbb{N}}$  as a nonlinear analogue of the Laplace spectrum of  $M$ . In particular one can ask ([46, Section 8], [47, Section 5.2]) whether a Weyl Law holds for the sequence of numbers  $\{\omega_p(M)\}_{p \in \mathbb{N}}$ . More precisely, if

$$\lim_{p \rightarrow \infty} \omega_p(M) p^{-\frac{1}{n+1}} = a(n)(\text{vol}(M, g))^{\frac{n}{n+1}},$$

where  $a(n)$  is a constant that depends only on  $n$ .

In [81], we put this analogy forward by considering sweepouts whose surfaces are zero

sets of linear combinations of eigenfunctions. Note that a conjecture of Yau [132] states that

$$c^{-1}\sqrt{\lambda_p} \leq \mathcal{H}^n(\{\phi_p = 0\}) \leq c\sqrt{\lambda_p},$$

where  $c = c(M, g) > 0$ ,  $\phi_p$  is the  $p$ -th eigenfunction and  $\lambda_p$  is the  $p$ -th eigenvalue of the Laplacian. We conjecture that generically the minimal surfaces given by Theorem 1.7 should form a sequence  $\{\Sigma_p\}_{p \geq 1}$ , where  $\Sigma_p$  has index  $p$ , multiplicity one and area going to infinity. The first Betti number and the index should be linearly related, and by analogy with nodal sets of eigenfunctions, we conjecture that these minimal surfaces should become equidistributed in space.

Finally, we mention that there are many interesting questions in the higher codimension case, where the second variation formula becomes a complicated object. It would be interesting to construct calibrated submanifolds by variational methods (see Schoen and Wolfson [115]), and to understand better the influence of the ambient curvature. The only stable integral cycles of complex projective space are the formal sums of algebraic varieties (Lawson and Simons [74]).

**Acknowledgements.** The author is grateful to École Polytechnique, École Normale Supérieure, and Université Paris-Est (Marne-la-Vallée) for the hospitality during the writing of this paper.

## References

- [1] Agol, I., Marques, F.C., and Neves, A., *Min-max theory and the energy of links*, arXiv:1205.0825 [math.GT], (2012), 1–19.
- [2] Alexakis, S. and Mazzeo, R., *Renormalized area and properly embedded minimal surfaces in hyperbolic 3-manifolds*, *Comm. Math. Phys.* **297** (2010), no. 3, 621–651.
- [3] Allard, W.K., *On the first variation of a varifold*, *Ann. of Math. (2)* **95** (1972), 417–491.
- [4] Almgren, F., *The homotopy groups of the integral cycle groups*, *Topology* (1962), 257–299.
- [5] ———, *The theory of varifolds*, Mimeographed notes, Princeton University (1965).
- [6] ———, *Some interior regularity theorems for minimal surfaces and an extension of Bernstein’s theorem*, *Ann. of Math.* **84** (1966), 277–292.
- [7] Almgren, F.J., Jr., *Almgren’s big regularity paper*, World Scientific Monograph Series in Mathematics, 1. World Scientific Publishing Co., Inc., River Edge, NJ, 2000.
- [8] Altschuler, S. and Grayson, M., *Shortening space curves and flow through singularities*, *J. Differential Geom.* **35** (1992), no. 2, 283–298.
- [9] Andrews, B., *Noncollapsing in mean-convex mean curvature flow*, *Geom. Topol.* **16** (2012), no. 3, 1413–1418.
- [10] Andrews, B. and Li, H., *Embedded constant mean curvature tori in the three-sphere*, arXiv:1204.5007v3 [math.DG], (2012).
- [11] Bangert, V., *On the existence of closed geodesics on two-spheres*, *Internat. J. Math.* **4** (1993), 1–10.

- [12] Bauer, M. and Kuwert, E., *Existence of minimizing Willmore surfaces of prescribed genus*, Int. Math. Res. Not. (2003), 553–576.
- [13] Birkhoff, G., *Dynamical systems with two degrees of freedom*, Trans. Amer. Math. Soc. **18** (1917), 199–300.
- [14] Bombieri, E., De Giorgi, E., and Giusti, E., *Minimal cones and the Bernstein problem*, Invent. Math. **7** (1969), 243–268.
- [15] Bray, H., *Proof of the Riemannian Penrose inequality using the positive mass theorem*, J. Differential Geom. **59** (2001), no. 2, 177–267.
- [16] Brendle, S., *Embedded minimal tori in  $S^3$  and the Lawson conjecture*, Acta Math. **211** (2013), no. 2, 177–190.
- [17] Bryant, R., *A duality theorem for Willmore surfaces*, J. Differential Geom. **20** (1984), 23–53.
- [18] Calabi, E. and Cao, J.G., *Simple closed geodesics on convex surfaces*, J. Differential Geom. **36** (1992), no. 3, 517–549.
- [19] Chen, B.-L., Tang, S.-H., and Zhu, X.-P., *Complete classification of compact four-manifolds with positive isotropic curvature* J. Differential Geom. **91** (2012), no. 1, 41–80.
- [20] Choe, J. and Soret, M., *First eigenvalue of symmetric minimal surfaces in  $S^3$* , Indiana Univ. Math. J. **58** (2009), no. 1, 269–281.
- [21] ———, *New minimal surfaces in  $S^3$  desingularizing the Clifford tori*, arXiv:1304.3184v1 [math.DG], (2013)
- [22] Colding, T. and De Lellis, C., *The min-max construction of minimal surfaces*, Surveys in Differential Geometry VIII, International Press (2003), 75–107.
- [23] Colding, T. and Minicozzi, W., *Estimates for the extinction time for the Ricci flow on certain 3-manifolds and a question of Perelman*, J. Amer. Math. Soc. **18** (2005), 561–569.
- [24] ———, *Width and finite extinction time of Ricci flow*, Geom. Topol. **12** (2008), no. 5, 2537–2586.
- [25] Colding, T.H., Minicozzi, and W.P., II, *The Calabi-Yau conjectures for embedded surfaces*, Ann. of Math. (2) **167** (2008), no. 1, 211–243.
- [26] Collin, P. and Rosenberg, H., *Construction of harmonic diffeomorphisms and minimal graphs*, Ann. of Math. (2) **172** (2010), no. 3, 1879–1906.
- [27] Costa, C.J., *Example of a complete minimal immersion in  $R^3$  of genus one and three embedded ends*, Bol. Soc. Brasil. Mat. **15** (1984), no. 1-2, 47–54.
- [28] De Giorgi, E., *Frontiere orientate di misura minima*, Seminario di Matematica della Scuola Normale Superiore di Pisa, 1960–61, Editrice Tecnico Scientifica, Pisa (1961) 57 pp.
- [29] De Lellis, C. and Pellandini, F., *Genus bounds for minimal surfaces arising from min-max constructions*, J. Reine Angew. Math. **644** (2010), 47–99.
- [30] De Lellis, C. and Spadaro, E., *Regularity of area minimizing currents I: gradient  $L^p$  estimates*, arXiv:1306.1195 [math.DG], (2013).
- [31] De Lellis, C. and Tasnady, D., *The existence of embedded minimal hypersurfaces*, J.

- Differential Geom. **95** (2013), no. 3, 355–388.
- [32] do Carmo, M. and Peng, C.K., *Stable complete minimal surfaces in  $R^3$  are planes*, Bull. Amer. Math. Soc. (N.S.) **1** (1979), no. 6, 903–906.
- [33] Douglas, J., *Solution of the problem of Plateau*, Trans. Amer. Math. Soc. **33** (1931), no. 1, 263–321.
- [34] Federer, H. and Fleming, W., *Normal and integral currents*, Ann. of Math. **72** (1960), 458–520.
- [35] Federer, H., *The singular sets of area minimizing rectifiable currents with codimension one and of area minimizing flat chains modulo two with arbitrary codimension*, Bull. Amer. Math. Soc. **76** (1970), 767–771.
- [36] Fernández, I. and Mira, P., *Constant mean curvature surfaces in 3-dimensional Thurston geometries*, Proceedings of the International Congress of Mathematicians. Volume II, Hindustan Book Agency, New Delhi, (2010), 830–861.
- [37] Fischer-Colbrie, D. and Schoen, R., *The structure of complete stable minimal surfaces in 3-manifolds of nonnegative scalar curvature*, Comm. Pure Appl. Math. **33** (1980), no. 2, 199–211.
- [38] Frankel, T., *On the fundamental group of a compact minimal submanifold*, Ann. of Math. **83** (1966), 68–73.
- [39] Franks, J., *Geodesics on  $S^2$  and periodic points of annulus homeomorphisms*, Invent. Math. **108** (1992), 403–418.
- [40] Fraser, A.M., *Fundamental groups of manifolds with positive isotropic curvature*, Ann. of Math. (2) **158** (2003), no. 1, 345–354.
- [41] Freedman, M., He, Z.-X., and Wang, Z., *Möbius energy of knots and unknots*, Ann. of Math. (2) **139** no. 1 (1994), 1–50.
- [42] Germain, S., *Recherches sur la théorie des surfaces élastiques*, Paris (1921).
- [43] Grayson, M., *Shortening embedded curves*, Ann. Math **120** (1989), 71–112.
- [44] ———, *Dimension, nonlinear spectra and width*, Geometric aspects of functional analysis, Lecture Notes in Math. Springer, Berlin **1317** (1988), 132–184.
- [45] ———, *Positive curvature, macroscopic dimension, spectral gaps and higher signatures*, Functional analysis on the eve of the 21st century, Vol. II (New Brunswick, NJ, 1993), 1–213, Progr. Math. **132**, Birkhäuser Boston, Boston, MA, 1996.
- [46] ———, *Isoperimetry of waists and concentration of maps*, Geom. Funct. Anal. **13** (2003), 178–215.
- [47] ———, *Singularities, expanders and topology of maps. I. Homology versus volume in the spaces of cycles*, Geom. Funct. Anal. **19** (2009), 743–841.
- [48] Gromov, M. and Lawson, H.B., Jr., *Spin and scalar curvature in the presence of a fundamental group. I*, Ann. of Math. (2) **111** (1980), no. 2, 209–230.
- [49] Gulliver, R.D., II, *Regularity of minimizing surfaces of prescribed mean curvature*, Ann. of Math. (2) **97** (1973), 275–305.
- [50] Guth, L. *Minimax problems related to cup powers and Steenrod squares*, Geom. Funct. Anal. **18** (2009), 1917–1987.
- [51] Hamilton, R., *Three-manifolds with positive Ricci curvature*, J. Differential Geom. **17**

- (1982), no. 2, 255–306.
- [52] ———, *The formation of singularities in the Ricci flow*, *Surveys in Differential Geometry* **2** (1995), 7–136.
- [53] Harvey, R., Lawson, H.B., Jr., *Calibrated geometries*, *Acta Math.* **148** (1982), 47–157.
- [54] He, Z.-X., *On the minimizers of the Möbius cross energy of links*, *Experiment. Math.* **11** (2002), no. 2, 244–248.
- [55] E. Heintze and H. Karcher, *A general comparison theorem with applications to volume estimates for submanifolds*, *Annales scientifiques de l’E.N.S.* 11 (1978), 451–470.
- [56] Helfrich, W., *Elastic properties of lipid bilayers: Theory and possible experiments*, *Z. Naturforsch.* **28** (1973), 693–703.
- [57] Hingston, N., *On the growth of the number of closed geodesics on the two-sphere*, *Internat. Math. Res. Notices* (1993), 253–262.
- [58] Hoffman, D.A. and Meeks, W.H., III, *Complete embedded minimal surfaces of finite total curvature*, *Bull. Amer. Math. Soc. (N.S.)* **12** (1985), no. 1, 134–136.
- [59] Hoffman, D., Traizet, M., and White, B., *Helicoidal minimal surfaces of prescribed genus, II*, arXiv:1304.6180v1 [math.DG], (2013).
- [60] Huisken, G. and Ilmanen, T., *The inverse mean curvature flow and the Riemannian Penrose inequality*, *J. Differential Geom.* **59** (2001), no. 3, 353–437.
- [61] Ilmanen, T. and White, B., *Sharp Lower Bounds on Density of Area-Minimizing Cones*, arXiv:1010.5068v2 [math.DG], (2013)
- [62] Jorge, L.P. de M. and Xavier, F., *A complete minimal surface in  $R^3$  between two parallel planes*, *Ann. of Math. (2)* **112** (1980), no. 1, 203–206.
- [63] Jost, J., *Two-dimensional geometric variational problems*, J. Wiley and Sons, Chichester, N.Y. 1991.
- [64] Kahn, J. and Markovic, Vl., *Counting essential surfaces in a closed hyperbolic three-manifold*, *Geom. Topol.* **16** (2012), no. 1, 601–624.
- [65] Kapouleas, N., *Doubling and desingularization constructions for minimal surfaces*, *Surveys in geometric analysis and relativity*, *Adv. Lect. Math. (ALM)*, Int. Press, Somerville, MA **20** (2011), 281–325.
- [66] Kapouleas, N. and Yang, S.-D., *Minimal surfaces in the three-sphere by doubling the Clifford torus*, *Amer. J. Math.* **132** (2010), no. 2, 257–295.
- [67] Karcher, H., Pinkall, U., and Sterling, I., *New minimal surfaces in  $S^3$* , *J. Differential Geom.* **28** (1988), 169–185.
- [68] Ketover, D., *Degeneration of min-max sequences in 3-manifolds*, arXiv:1312.2666 [math.DG], (2013).
- [69] Kusner, R., *Estimates for the biharmonic energy on unbounded planar domains, and the existence of surfaces of every genus that minimize the squared-mean-curvature integral*, *Elliptic and parabolic methods in geometry*, A K Peters, (1996), 67–72.
- [70] Kuwert, E. and Schätzle, R., *Removability of point singularities of Willmore surfaces*, *Ann. of Math.* **160** (2004), 315–357.

- [71] Kuwert, E., Li, Y., and Schatzle, R., *The large genus limit of the infimum of the Willmore energy*, Amer. J. Math. **132** (2010), 37–51.
- [72] Lawson, B., *Complete minimal surfaces in  $S^3$* , Ann. of Math. **92** (1970), 335–374.
- [73] Lawson, H.B., Jr., *The unknottedness of minimal embeddings*, Invent. Math. **11** (1970), 183–187.
- [74] Lawson, H.B., Jr. and Simons, J., *On stable currents and their application to global problems in real and complex geometry*, Ann. of Math. (2) **98** (1973), 427–450.
- [75] Li, P. and Yau, S-T., *A new conformal invariant and its applications to the Willmore conjecture and the first eigenvalue of compact surfaces*, Invent. Math. **69** (1982), 269–291.
- [76] Liu, G., *3-manifolds with nonnegative Ricci curvature*, Invent. Math. **193** (2013), no. 2, 367–375.
- [77] Lyusternik, L.A. and Fet, A.I., *Variational problems on closed manifolds (Russian)*, Doklady Akad. Nauk SSSR (N.S.) **81**, (1951), 17–18.
- [78] Lusternik, L. and Schnirelmann, L., *Topological methods in variational problems and their application to the differential geometry of surfaces*, Uspehi Matem. Nauk (N.S.) **2** (1947), 166–217.
- [79] Marques, F.C. and Neves, A., *Rigidity of min-max minimal spheres in three-manifolds*, Duke Math. J. **161** (2012), no. 14, 2725–2752.
- [80] Marques, F.C. and Neves A., *Min-max theory and the Willmore conjecture*, Ann. of Math. **179** 2 (2014), 683–782.
- [81] ———, *Existence of infinitely many minimal hypersurfaces in positive Ricci curvature*, arXiv:1311.6501 [math.DG], (2013).
- [82] Meeks, W.H., III and Pérez, J., *A survey on classical minimal surface theory*, University Lecture Series, 60. American Mathematical Society, 2012.
- [83] Meeks, W. H., III and Rosenberg, H., *The uniqueness of the helicoid*, Ann. of Math. (2) **161** (2005), no. 2, 727–758.
- [84] Meeks, W.H., III and Yau, S.-T., *Topology of three-dimensional manifolds and the embedding problems in minimal surface theory*, Ann. of Math. (2) **112** (1980), no. 3, 441–484.
- [85] ———, *The classical Plateau problem and the topology of three-dimensional manifolds*, Topology **21** (1982), no. 4, 409–442.
- [86] Meeks, W., Simon and L., Yau, S-T., *Embedded minimal surfaces, exotic spheres, and manifolds with positive Ricci curvature*, Ann. of Math. **116** (1982), 621–659.
- [87] Micallef, M.J. and Moore, J.D., *Minimal two-spheres and the topology of manifolds with positive curvature on totally isotropic two-planes*, Ann. of Math. (2) **127** (1988), no. 1, 199–227.
- [88] Morrey, C.B., Jr., *The problem of Plateau on a Riemannian manifold*, Ann. of Math. (2) **49** (1948), 807–851.
- [89] Morse, M. and Tompkins, C., *The existence of minimal surfaces of general critical types*, Ann. of Math. (2) **40** (1939), no. 2, 443–472.
- [90] Nadirashvili, N., *Hadamard's and Calabi-Yau's conjectures on negatively curved and*

- minimal surfaces*, Invent. Math. **126** (1996), no. 3, 457–465.
- [91] Neves, A., *New applications of min-max theory*, Proceedings of the International Congress of Mathematicians, Seoul, Korea, 2014.
- [92] Osserman, R., *A proof of the regularity everywhere of the classical solution to Plateau's problem*, Ann. of Math. (2) **91** (1970), 550–569.
- [93] Pacard, F., *The role of minimal surfaces in the study of the Allen-Cahn equation*, Geometric analysis: partial differential equations and surfaces, Contemp. Math. **570** (2012), 137–163, Amer. Math. Soc., Providence, RI.
- [94] Parker, T.H., *Bubble tree convergence for harmonic maps*, J. Differential Geom. **44** (1996), no. 3, 595–633.
- [95] Parker, T.H. and Wolfson, J.G., *Pseudo-holomorphic maps and bubble trees*, J. Geom. Anal. **3** (1993), no. 1, 63–98.
- [96] Perdomo, O., *Low index minimal hypersurfaces of spheres*, Asian J. Math. **5** (2001), no. 4, 741–749.
- [97] Perelman, G., *The entropy formula for the Ricci flow and its geometric applications*, arXiv:math/0211159v1 [math.DG], (2002),
- [98] ———, *Ricci flow with surgery on three-manifolds*, arXiv:math/0303109v1 [math.DG], (2003).
- [99] ———, *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*, arXiv:math/0307245v1 [math.DG], (2003).
- [100] U. Pinkall, *Hopf tori in  $S^3$* , Invent. Math. **81** (1985), 379–386.
- [101] Pinkall, U. and Sterling, I., *On the classification of constant mean curvature tori*, Ann. of Math. (2) **130** (1989), no. 2, 407–451.
- [102] Pitts, J., *Existence and regularity of minimal surfaces on Riemannian manifolds*, Mathematical Notes 27, Princeton University Press, Princeton (1981).
- [103] Pogorelov, A.V., *On the stability of minimal surfaces* (Russian), Dokl. Akad. Nauk SSSR **260** (1981), no. 2, 293–295.
- [104] Poincaré, H. *Sur les lignes géodésiques des surfaces convexes*, Trans. Amer. Math. Soc. **6** (1905), 237–274.
- [105] Radó, T., *On Plateau's problem*, Ann. of Math. **31** (1930), no. 3, 457–469.
- [106] Rivière, T., *Analysis aspects of Willmore surfaces*, Invent. Math. **174** (2008), 1–45.
- [107] ———, *Conformally invariant variational problems*, arXiv:1206.2116 [math.AP] (2012).
- [108] Ros, A., *The Willmore conjecture in the real projective space*, Math. Res. Lett. **6** (1999), 487–493.
- [109] ———, *One-sided complete stable minimal surfaces*, J. Differential Geom. **74** (2006), no. 1, 69–92.
- [110] Sacks, J. and Uhlenbeck, K., *The existence of minimal immersions of 2-spheres*, Ann. of Math. (2) **113** (1981), no. 1, 1–24.
- [111] ———, *Minimal immersions of closed Riemann surfaces*, Trans. Amer. Math. Soc. **271** (1982), no. 2, 639–652.
- [112] Schoen, R., *Estimates for stable minimal surfaces in three-dimensional manifolds*,

- Seminar on minimal submanifolds, Ann. of Math. Stud., Princeton Univ. Press **103** (1983), 111–126.
- [113] Schoen, R. and Simon, L., *Regularity of stable minimal hypersurfaces*, Comm. Pure Appl. Math. **34** (1981), 741–797.
- [114] Schoen, R., Simon, L., and Yau, S.-T., *Curvature estimates for minimal hypersurfaces*, Acta Math. **134** (1975), no. 3-4, 275–288.
- [115] Schoen, R. and Wolfson, J., *Minimizing area among Lagrangian surfaces: the mapping problem*, J. Differential Geom. **58** (2001), no. 1, 1–86.
- [116] Schoen, R. and Yau, S.T., *On the proof of the positive mass conjecture in general relativity*, Comm. Math. Phys. **65** (1979), 45–76.
- [117] ———, *Existence of incompressible minimal surfaces and the topology of three dimensional manifolds of non-negative scalar curvature*, Ann. of Math. **110** (1979), 127–142.
- [118] ———, *Proof of the positive mass theorem. II*, Comm. Math. Phys. **79** (1981), no. 2, 231–260.
- [119] ———, *Complete three-dimensional manifolds with positive Ricci curvature and scalar curvature*, Seminar on Differential Geometry, Ann. of Math. Stud. **102**, pp. 209–228, Princeton Univ. Press, Princeton, N.J., 1982.
- [120] Simon, L., *Existence of surfaces minimizing the Willmore functional*, Comm. Anal. Geom. **1** no. 2, (1993), 281–326.
- [121] Simons, J., *Minimal varieties in riemannian manifolds*, Ann. of Math. (2) **88** (1968), 62–105.
- [122] Siu, Y.-T. and Yau, S.-T., *Compact Kähler manifolds of positive bisectional curvature*, Invent. Math. **59** (1980), no. 2, 189–204.
- [123] Smale, N., *Generic regularity of homologically area minimizing hypersurfaces in eight-dimensional manifolds*, Comm. Anal. Geom. **1** (1993), no. 2, 217–228.
- [124] Smith, *On the existence of embedded minimal 2D-spheres in the 3-sphere*, endowed with an arbitrary Riemannian metric, Ph.D. thesis, supervisor L. Simon, University of Melbourne (1982).
- [125] Taimanov, I., *Closed extremals on two-dimensional manifolds*, (Russian) Uspekhi Mat. Nauk **47** (1992), 143–185.
- [126] P. Topping, *Towards the Willmore conjecture*, Calc. Var. Partial Differential Equations **11** (2000), 361–393.
- [127] Urbano, F., *Minimal surfaces with low index in the three-dimensional sphere*, Proc. Amer. Math. Soc. **108** (1990), 989–992.
- [128] White, B., *The space of minimal submanifolds for varying Riemannian metrics*, Indiana Univ. Math. J. **40** (1991), no. 1, 161–200.
- [129] Wickramasekera, N., *A general regularity theory for stable codimension 1 integral varifolds*, Ann. of Math. (2) **179** (2014), no. 3, 843–1007.
- [130] Willmore, T.J., *Note on embedded surfaces*, An. Sti. Univ. “Al. I. Cuza” Iasi Sect. I a Mat. (N.S.) **11B** (1965), 493–496.
- [131] Witten, E., *A new proof of the positive energy theorem*, Comm. Math. Phys. **80** (1981),

381–402.

- [132] Yau, S.-T., *Problem section*, Seminar on Differential Geometry, Ann. of Math. Stud., Princeton Univ. Press, Princeton, N.J. **102** (1982), 669–706.
- [133] Zhou, X., *Min-max minimal hypersurface in  $(M^{n+1}, g)$  with  $Ric_g > 0$  and  $2 \leq n \leq 6$* , arXiv:1210.2112 [math.DG], (2012).

Instituto de Matemática Pura e Aplicada (IMPA), Estrada Dona Castorina 110, 22460-320 Rio de Janeiro, Brazil

E-mail: coda@impa.br

# Random Structures and Algorithms

Alan Frieze

**Abstract.** We provide an introduction to the analysis of random combinatorial structures and some of the associated computational problems.

**Mathematics Subject Classification (2010).** Primary 05C80; Secondary 68Q87.

**Keywords.** Random Graphs, Algorithms, Average Case.

## 1. Introduction

Our aim in this paper is to give a short survey on work on some probabilistic aspects of Combinatorics and their relation to algorithmic questions in Computer Science.

Combinatorics/Discrete Mathematics (in the main) concerns itself with certain properties of large, finite sets, with some defined structure.

Given such a set  $\Omega$ , (often a set of graphs) we have certain natural questions:

1. How big is  $\Omega$ : *Enumerative Combinatorics*.
2. How big is the greatest element of  $\Omega$ : *Extremal Combinatorics*.
3. What are the properties of a typical member of  $\Omega$ : *Probabilistic Combinatorics*.
4. What is the complexity of computational problems associated with the above topics.

This paper will concern itself with Items 3. and 4. of this list. We should not confuse Item 3 with the *Probabilistic Method* where we use probabilistic notions to prove the existence of certain objects. We will try to interweave the structural analysis of random structures with related algorithmic questions. We refer the reader to Stanley [109] for Item 1. and to Bollobás [21] or Jukna [71] or Lovasz [90] for Item 2. The probabilistic analysis of algorithms has (at least) two flavors. Flavor 1 is a very detailed analysis of simple algorithms. See Sedgewick and Flajolet [106] for details on this. This paper will restrict attention to flavor 2; more complex algorithms for which the level of detail is less than that achieved in flavor 1.

We will begin with the seminal work of Erdős and Rényi on the evolution of random graphs. This is the subject matter of Section 2. We will then choose some topics as further illustration of the diverse aspects of the area. We have chosen Graph Coloring (Section 3); Matchings in random graphs (Section 4); Hamilton cycles (Section 5); and some questions about the optimal value of random optimization problems (Section 6).

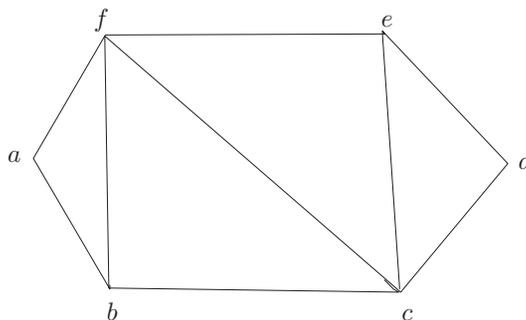
Our discussion of algorithms will focus on the probabilistic analysis of two NP-hard optimization problems. In the very early days of algorithmic analysis, it was considered sufficient to prove that an algorithm always terminated after a finite number of steps. This led to lexicographic versions of the simplex algorithm. When Integer Programming was realized to be important, Gomory's cutting plane algorithms [63] were considered to be a breakthrough. He showed that they would solve Integer Programming problems in finite time, but the bounds on running time were bad, and not stressed.

It was Edmonds [48] who pointed out in his seminal paper on matchings that "finite" can be very large and that we should try to find "good" algorithms i.e. those that always run in polynomial time. The search for polynomial time algorithms began in earnest. The most notable success in this quest being a polynomial time algorithm for Linear Programming, Khachiyan [80]. The optimism that most naturally occurring problems could be solved in polynomial time was crushed by the works of Cook [42], Levin [87] and Karp [74].

There have been several reactions to this negative state of affairs. One has been to see how well we can do in polynomial time, leading to the intense study of approximation algorithms. Another has been to focus on special cases that are quickly solvable. It should perhaps be pointed out that the cryptography community has turned this negative state of affairs into something positive.

To show that a problem can be difficult in the worst-case, one has to construct pathological examples. It is perhaps fortunate that in practice, the instances of NP-hard problems that are thrown up tend not to be pathological and large problems do get solved. Perhaps the best illustration of this comes from the success of researchers in solving large Traveling Salesperson Problems [9]. To explain this success, we have to define a "typical" problem and analyse the efficiency of algorithms on typical problems. For us, typical means drawn from some probability distribution and an algorithm will be considered efficient if its *expected* running time is polynomial in its input size or if it runs in polynomial time with high probability (w.h.p.)<sup>1</sup> We will describe some of the work in the area of analysing efficient algorithm for random NP-hard problems. In particular, we will discuss the Traveling Salesperson Problem in Section 7 and Random  $k$ -SAT in Section 8.

**1.1. Basic notions of graph theory.** A graph  $G = (V, E)$  consists of a set of vertices  $V$  together with a collection of edges  $E \subseteq \binom{V}{2}$  i.e.  $E$  is a set of 2-element subsets of  $V$ . It is useful to imagine  $G$  as drawn below.



<sup>1</sup>A sequence of events  $\mathcal{E}_n$  is said to occur w.h.p. if  $\Pr(\mathcal{E}_n) = 1 - o(1)$  as  $n \rightarrow \infty$ .

The *degree* of a vertex is the number of edges that it lies in. A *walk* is a sequence  $(w_1, w_2, \dots, w_k)$  where  $\{w_{i-1}, w_i\} \in E$  for  $1 \leq i < k$ . A *path* is a walk in which  $w_1, w_2, \dots, w_k$  are distinct. A *cycle* is a walk  $w_1, w_2, \dots, w_k = w_1$  where  $w_1, w_2, \dots, w_{k-1}$  are distinct. We can define an equivalence relation  $R$  on  $V$  where  $vRw$  iff  $G$  contains a path from  $v$  to  $w$ . The equivalence classes of this relation are called the *components* of  $G$  and a graph is *connected* if there is a unique component. Each component therefore forms a connected subgraph. A *tree* is a connected graph without cycles. If it has  $k$  vertices, then it necessarily has  $k - 1$  edges.

## 2. Evolution of a random graph

The analysis of random structures as a major field can be traced directly to the seminal paper [50] of Erdős and Rényi. Let us first establish notation.

One of the most important examples of a random structure is the random graph  $G_{n,m}$ , where  $0 \leq m \leq \binom{n}{2}$ . Here the vertex set is  $[n] = \{1, 2, \dots, n\}$  and the edge set  $E_{n,m}$  consists of  $m$  edges chosen uniformly at random. This is a subgraph of the complete graph  $K_n = \left([n], \binom{[n]}{2}\right)$ . The random graph  $G_{n,m}$  is intimately related to the random graph  $G_{n,p}$  where  $0 \leq p \leq 1$ . In this graph, also a subgraph of  $K_n$ , each edge of  $K_n$  is independently included as an edge with probability  $p$ . If  $m \sim \binom{n}{2}p$  then  $G_{n,p}$  and  $G_{n,m}$  have “similar” properties.

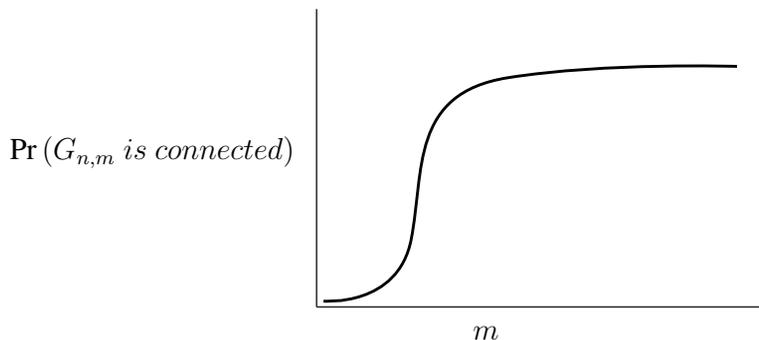
The paper [50] describes the typical properties of  $G_{n,m}$  for various values of  $m$  and as  $n \rightarrow \infty$ . The component structure of  $G = G_{n,m}$  is summarised by the following.

- (a) If  $m = o(n^{1/2})$  then w.h.p.  $G$  consists of isolated edges and vertices.
- (b) If  $n^{\frac{k-1}{k}} \ll m = o(n^{\frac{k}{k+1}})$  then w.h.p. the components are trees with  $1 \leq j \leq k + 1$  vertices. Each possible such tree appears.
- (c) If  $m = cn$  for some constant  $0 < c < \frac{1}{2}$  then almost all of the components are trees. There will be a few unicyclic components i.e. components containing a unique cycle. The maximum component size is  $O(\log n)$ .
- (d) If  $m \sim \frac{1}{2}n$  then the component structure is more complicated. The maximum component size is of order  $n^{2/3}$ . It has been the subject of intensive study e.g. Janson, Knuth, Łuczak and Pittel [69].
- (e) If  $m = \frac{1}{2}cn$  where  $c > 1$  then w.h.p. there is a unique “giant” component of size  $\sim \gamma(c)n$ . The remaining components are almost all trees. The second largest component is of size  $O(\log n)$ .  $\gamma(c)$  is the probability that a branching process where each particle has a Poisson, mean  $c$ , number of descendants, does not become extinct.
- (f) If  $m = \frac{1}{2}n(\log n + c_n)$  then [49],

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr(G_{n,m} \text{ is connected}) &= \begin{cases} 0 & c_n \rightarrow -\infty \\ e^{-e^{-c}} & c_n \rightarrow c \\ 1 & c_n \rightarrow +\infty \end{cases} \quad (2.1) \\ &= \lim_{n \rightarrow \infty} \Pr(\delta(G_{n,m}) \geq 1) \end{aligned}$$

where  $\delta$  denotes minimum degree.

Notice the sharp transition from being disconnected to connected, as claimed in (g). This is shown pictorially in the diagram below:



This leads us to the notion of *thresholds*. A function  $\tau(n)$  is a threshold for a graph property  $\mathcal{P}$  if

$$\lim_{n \rightarrow \infty} \Pr(G_{n,m} \text{ has } \mathcal{P}) = \begin{cases} 0 & \frac{m}{\tau(n)} \rightarrow 0 \\ 1 & \frac{m}{\tau(n)} \rightarrow \infty \end{cases}.$$

Thus  $n \log n$  is the threshold for connectivity. Of course (2.1) claims something much stronger. What we have here is a *sharp* threshold. One of the main quests in the theory of random graphs is for the precise thresholds for important graph properties.

Before leaving this section, we remark on the proof technique introduced by Erdős and Rényi. There are  $M = \binom{n}{m}$  distinct graphs with vertex set  $[n]$  and  $m$  edges. The probability that  $G_{n,m}$  is connected is then simply  $M_c/M$  where  $M_c$  is the number of connected graphs with vertex set  $[n]$  and  $m$  edges. One gets nowhere if one tries to estimate the probability of connectivity by trying to evaluate  $M_c$ .

One key insight of Erdős and Rényi is to identify events that are very unlikely to occur. If one identifies the correct collection, then one can often reach the goal of estimating the probability of occurrence of the event that you are really interested in. As an example, we let  $p = \frac{c \log n}{n}$  where  $c > 1$  is constant. According to (2.1), we should be able to prove that  $G_{n,p}$  is connected w.h.p. for this value of  $p$ .

Let  $X$  denote the number of components with at most  $n/2$  vertices. We observe that a graph with  $n$  vertices is connected iff  $X = 0$ . Then

$$\Pr(X \neq 0) \leq \mathbf{E}(X) \leq \sum_{k=1}^{n/2} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)} \leq \frac{n}{c \log n} \sum_{k=1}^{n/2} \left( \frac{ce \log n}{n^{c(1-k/n)}} \right)^k \rightarrow 0. \tag{2.2}$$

**Explanation:** For a fixed  $k$ , there are  $\binom{n}{k}$  choices for the vertex set of a component  $C$ . To ensure it is connected we choose a spanning tree  $T$  of  $C$ , in  $k^{k-2}$  ways. Then we multiply by the probability  $p^{k-1}$  that  $T$  exists in  $G_{n,p}$  and then by the probability  $(1-p)^{k(n-k)}$  that there are no edges between  $C$  and the rest of the graph.

**2.1. Hitting Times.** Consider the sequence  $G_0, G_1, \dots, G_m, \dots$ , where  $G_{i+1}$  is  $G_i$  plus a random edge.

Let  $m_k$  denote the minimum  $m$  for which the minimum vertex degree  $\delta(G_m) \geq k$ . These are important “times” for the occurrence of important properties.

**Theorem 2.1** (Erdős and Rényi [49]). *W.h.p.  $m_1$  is the time when  $G_m$  first becomes connected.*

Observe that the largest term in the sum in (2.2) is for  $k = 1$ .

For other important properties, we need a couple of definitions. A *matching* in a graph  $G$  is a set of vertex disjoint edges. The matching is *perfect* if every vertex is covered by an edge of the matching. This is impossible if  $|V|$  is odd, in which case we allow one uncovered vertex.

**Theorem 2.2** (Erdős and Rényi [51]). *W.h.p.  $m_1$  is the “time” when  $G_m$  first has a perfect matching.*

A *Hamilton cycle* in a graph  $G$  is a cycle that passes through each vertex exactly once.

**Theorem 2.3** (Ajtai, Komlós, Szemerédi [4], Bollobás [20]). *W.h.p.  $m_2$  is the time when  $G_m$  first has a Hamilton cycle.*

In general there will be many distinct Hamilton cycles at time  $\tau_1$ .

**Theorem 2.4** (Cooper and Frieze [43]). *W.h.p. at “time”  $m_2$ ,  $G_m$  has  $(\log n)^{n-o(n)}$  distinct Hamilton cycles.*

This was recently improved to

**Theorem 2.5** (Glebov and Krivelevich [62]). *W.h.p. at time  $m_2$ ,  $G_m$  has  $n!p^n e^{-o(n)}$  distinct Hamilton cycles.*

One can also ask for edge disjoint Hamilton cycles. Let Property  $\mathcal{A}_k$  denote the existence of  $\lfloor k/2 \rfloor$  disjoint Hamilton cycles plus a disjoint perfect matching if  $k$  is odd.

**Theorem 2.6** (Bollobás and Frieze [28]). *W.h.p. at time  $m_k, k = O(1)$ ,  $G_m$  has property  $\mathcal{A}_k$ .*

We believed that the  $k = O(1)$  bound was unnecessary. This has recently been verified.

**Theorem 2.7** (Knox, Kühn and Osthus [82]). *W.h.p.  $G_m$  has property  $\mathcal{A}_\delta$  for  $n \log^{50} n \leq m \leq \binom{n}{2} - o(n^2)$ .*

**Theorem 2.8** (Krivelevich and Samotij [85]). *W.h.p.  $G_m$  has property  $\mathcal{A}_\delta$  for  $\frac{1}{2}n \log n \leq m \leq n^{1+\epsilon}$ .*

### 3. Graph Coloring

A *proper  $k$ -coloring* of a graph  $G = (V, E)$  is a map  $f : V \rightarrow [k]$  such that if  $\{v, w\}$  is an edge of  $G$  then  $f(v) \neq f(w)$ . The chromatic number  $\chi(G)$  is the smallest  $k$  for which there is a proper  $k$ -coloring.

A set of vertices  $S \subseteq V$  is *independent* if  $v, w \in S$  implies that  $\{v, w\}$  is not an edge. In a proper  $k$ -coloring, each color class is an independent set. The *independence number*  $\alpha(G)$  is the size of a largest independent set.

### 3.1. Dense random Graphs.

**Theorem 3.1** (Matula [93]). *W.h.p.*

$$\alpha(G_{n,1/2}) = 2 \log_2 n - 2 \log_2 \log_2 n + O(1).$$

Finding an independent set of size  $\sim \log_2 n$  in polynomial time is easy.

**Greedy Algorithm:**

Start with  $I = \{1\}$ .

Repeatedly add  $v \in [n] \setminus (I \cup N(I))$  until no such  $v$  can be found.

(Here  $N(I)$  is the set of neighbors of  $I$  i.e.  $\{w \notin I : \exists v \in I \text{ s.t. } \{v, w\} \in E\}$ ).

After  $k$  successful steps we find that the number of choices for  $v$  is distributed as the binomial

$\text{Bin}(n - k, 2^{-k})$ . If  $k \leq \log_2 n - 2 \log_2 \log_2 n$  then this is non-zero with probability  $1 - o(1/\log n)$ . So, w.h.p. the algorithm succeeds in finding an independent set of size at least  $(1 - o(1)) \log_2 n$ .

Surprisingly, no-one has been able to find a polynomial time algorithm that w.h.p. finds an independent set of size  $(1 + \epsilon) \log_2 n$  for any positive constant  $\epsilon > 0$ .

Indeed, it may not be possible to find such an independent set in polynomial time w.h.p. Deciding the truth of this is a challenging problem in complexity theory.

It follows from Matula's result that w.h.p.  $\chi(G_{n,1/2}) \geq \frac{n}{2 \log_2 n}$

**Theorem 3.2** (Bollobás and Erdős [25], Grimmett and McDiarmid [64]). *W.h.p. a simple greedy algorithm uses  $\sim \frac{n}{\log_2 n}$  colors.*

Given the fact that no-one knows how to find a large independent set in polynomial time, no-one knows how to find a coloring with at most  $(1 - \epsilon)n/\log_2 n$  colors in polynomial time.

It may even be NP-hard to find such a coloring in polynomial time w.h.p.

For a long time, no-one could prove an upper bound  $\chi(G_{n,1/2}) \leq (1 + o(1)) \frac{n}{2 \log_2 n}$ .

The "discovery" of Martingale Concentration Inequalities was a great help. Let  $Z = Z(X_1, \dots, X_N)$  where  $X_1, \dots, X_N$  are independent. Suppose that changing one  $X_i$  only changes  $Z$  by  $\leq 1$ . Then

$$\Pr(|Z - \mathbf{E}(Z)| \geq t) \leq e^{-2t^2/N}. \quad (3.1)$$

They were discovered by Shamir and Spencer [107] and by Rhee and Talagrand [99]. They have had a profound effect on the area. Further concentration inequalities by Talagrand [111] and Kim and Vu [81] have been extremely useful.

Concentration inequalities are extremely useful. They enable some random variables to be treated more or less like constants. The inequality (3.1) is a special case of what has become known as the Azuma-Hoeffding concentration inequality.

**Theorem 3.3** (Bollobás [22]).  $\chi(G_{n,1/2}) \sim \frac{n}{2 \log_2 n}$ .

*Proof.* Let  $Z$  be the maximum number of independent sets in a collection  $S_1, \dots, S_Z, |S_i| \sim 2 \log_2 n$  and  $|S_i \cap S_j| \leq 1$ .

$$\mathbf{E}(Z) = n^{2-o(1)} \text{ and changing one edge changes } Z \text{ by } \leq 1$$

So,

$$\Pr(\exists S \subseteq [n] : |S| \geq \frac{n}{(\log_2 n)^2} \text{ and } S \text{ doesn't contain a } (2 - o(1)) \log_2 n \text{ independent set}) \leq 2^n e^{-n^{2-o(1)}} = o(1).$$

So, we color  $G_{n,1/2}$  with color classes of size  $\sim 2 \log_2 n$  until there are  $\leq n/(\log_2 n)^2$  vertices uncolored and then give each remaining vertex a new color.  $\square$

**3.2. Sparse Random Graphs.** There has recently been a lot of research concerning the chromatic number of sparse random graphs viz.  $G_{n,p}$ ,  $p = d/n$  where  $d = O(1)$ .

**Conjecture.** *There exists a sequence  $d_k : k \geq 2$  such that w.h.p.*

$$\chi(G_{n,d/n}) = k \text{ for } d_{k-1} < d < d_k.$$

Friedgut [53] and Achlioptas and Friedgut [1] came close to proving this. Friedgut [53] (with an appendix by Bourgain) characterised properties which had a sharp threshold in terms of the non-existence of small local obstructions. He showed that if there are no small obstructions then for each  $n$  there is a value  $\tau_n$  such that the property is likely to occur close to  $\tau_n$ . Unfortunately, at the moment there is no general proof that the sequence  $(\tau_n)$  tends to a limit. Friedgut's prime example was for  $k$ -SAT, described later, and together with Achlioptas, he showed the existence of such a sequence  $(\tau_n)$  in the case of  $k$ -colorability.

It was soon established that w.h.p. the chromatic number only took one of two values:

**Theorem 3.4** (Łuczak [91]). *W.h.p.  $\chi(G_{n,d/n})$  takes one of two values.*

Surprisingly, using Chebyshev's inequality we get

**Theorem 3.5** (Achlioptas and Naor [2]). *Let  $k_d$  be the smallest integer  $k \geq 2$  such that  $d < 2k \log k$  then w.h.p.  $\chi(G_{n,d/n}) \in \{k_d, k_d + 1\}$ .*

We find this surprising as we would have expected the variance of the number of  $k$ -colorings to be too large to apply.

If  $X$  denotes the number of  $k$ -colorings of  $G_{n,d/n}$  then

$$\Pr(X > 0) \geq \frac{\mathbf{E}(X)^2}{\mathbf{E}(X^2)} = \Omega(1)$$

for  $d < (k - 1) \log(k - 1)$ .

This shows that  $\Pr(X > 0)$  is bounded below by a constant. We can now use Achlioptas and Friedgut.

The idea is straightforward. The difficulty lies in estimating the ratio  $\mathbf{E}(X)^2/\mathbf{E}(X^2)$ .

Achlioptas and Naor showed that for approximately half of the possible values for  $d$ ,  $\chi(G_{n,d/n})$  is determined w.h.p.

**Theorem 3.6** (Achlioptas and Naor [2]). *If  $d \in ((2k - 1) \log k, 2k \log k)$  then w.h.p.*

$$\chi(G_{n,d/n}) = k + 1.$$

This has been improved so that we now have

**Theorem 3.7** (Coja-Oghlan and Vilenchik [40]). *Let  $\kappa_d$  be the smallest integer  $k \geq 2$  such that  $d < (2k - 1) \log k$ . Then  $\chi(G_{n,d/n}) = \kappa_d$  for  $d \in \mathcal{A}$  where  $\mathcal{A}$  has density one in  $R_+$ .*

Furthermore, Coja-Oghlan has also improved on the naive first moment upper bound.

**Theorem 3.8** (Coja-Oghlan [35]).

$$d_k \leq 2k \log k - \log k - 1 + o_k(1).$$

Now for large  $k$ , the value of  $d_k$  is known within an interval of length less than 0.39.

There is still a factor of two gap between what can be proved existentially and what can be proved to be constructible in polynomial time.

## 4. Matchings

The seminal paper of Edmonds [48] showed that a matching of maximum size in a graph can be found in polynomial time. The algorithm is relatively complicated. Karp and Sipser [78] proposed the following greedy algorithm for finding a large matching:

### KSGREEDY

```

begin
   $M \leftarrow \emptyset$ ;
  while  $E(G) \neq \emptyset$  do
    begin
      A1: If  $G$  has a vertex of degree one, choose one,  $x$  say, randomly.
        Let  $e = \{x, y\}$  be the unique edge of  $G$  incident with  $x$  Endif;
      A2: Else, (no vertices of degree one) choose
         $e = \{x, y\} \in E$  randomly Endelse;

       $G \leftarrow G \setminus \{x, y\}$ ;
       $M \leftarrow M \cup \{e\}$ 
    end;
  Output  $M$ 
end

```

Note that this algorithm never makes a ‘mistake’ when executing command **A1**: If the graph  $G$  has a degree 1 vertex  $x$  then there is a maximum matching that contains the unique edge that contains  $x$ .

Karp and Sipser analysed the algorithms performance on  $G_{n,p}$ , where  $p = c/n$ . The random graph  $G_{n,p}$  with  $p = c/n$  will have a linear number of vertices of degree 1 w.h.p. The Karp Sipser Algorithm will therefore have an initial phase, which we call Phase 1, in which it executes command A1 in every step. After all degree 1 vertices are exhausted (for the first time) we move to Phase 2. During this phase both A1 and A2 are performed.

**Theorem 4.1** (Karp and Sipser [78]). *If  $c < e$  then w.h.p. Phase 1 ends with a graph with  $o(n)$  vertices. If  $c \geq e$  then w.h.p. Phase 2 isolates  $o(n)$  vertices.*

**Theorem 4.2** (Aronson, Frieze and Pittel [10]). *If  $c < e$  then w.h.p. Phase 1 ends with a graph consisting of a few vertex disjoint cycles. (The expected number of vertices on these cycles is  $O(1)$ ). If  $c > e$  then w.h.p. Phase 2 isolates  $\Theta(n^{1/5} \log^{O(1)} n)$  vertices.*

*Sketch of Proof.* The proof of Theorem 4.2 illustrates the use of differential equations in the analysis of algorithms. We now sketch a proof.

For the graph  $G_i$  remaining after  $i$  steps of the algorithm, let

- $\mathbf{v}_1$  = the number of vertices of degree one
- $\mathbf{v}$  = the number of vertices of degree at least two
- $\mathbf{m}$  = the number of edges

It is not hard to show that if we condition on the values of  $\mathbf{v}$ ,  $\mathbf{v}_1$  and  $\mathbf{m}$  then  $G$  is uniformly distributed on the set of graphs with these parameters. We use this fact to determine the likely evolution of the remaining graph. We can show that  $\mathbf{v}$ ,  $\mathbf{v}_1$  and  $\mathbf{m}$  are very likely to follow their expected trajectories. These trajectories are given by the solution of a set of differential equations in variables,  $v, v_1, m$ .

Let  $v_k$  be the number of vertices of degree  $k$  in  $G_i$ . One can show that  $v_k \approx v_k$  where

$$v_k = \frac{v z^k}{k!(e^z - 1 - z)}$$

where  $z$  is the solution to

$$\frac{2m - v_1}{v} = \frac{z(e^z - 1)}{f}, \quad f = f(z) = e^z - 1 - z.$$

The differential equations arise from the consideration of the expected change in these parameters in one step.

**One step transitions:** If  $v'_1, v', m'$  denote the values of the parameters after one step of the algorithm then, given  $v_1, v, m$

$$\begin{aligned} \mathbf{E}[v'_1 - v_1] &\approx -1 - \frac{v_1}{2\mathbf{m}} + \frac{v^2 z^4 e^z}{(2\mathbf{m}f)^2} - \frac{v_1 v z^2 e^z}{(2\mathbf{m})^2 f}, \\ \mathbf{E}[v' - v] &\approx -1 + \frac{v_1}{2\mathbf{m}} - \frac{v^2 z^4 e^z}{(2\mathbf{m}f)^2}, \\ \mathbf{E}[m' - m] &\approx -1 - \frac{v z^2 e^z}{2\mathbf{m}f}. \end{aligned}$$

$v_1, v, m$  closely follow the trajectory of a set of differential equations. These equations model the one step transitions.

$$\frac{dv_1}{dt} = -1 - \frac{v_1}{2m} + \frac{v^2 z^4 e^z}{(2mf)^2} - \frac{v_1 v z^2 e^z}{(2m)^2 f}; \quad \frac{dv}{dt} = -1 + \frac{v_1}{2m} - \frac{v^2 z^4 e^z}{(2mf)^2}; \quad \frac{dm}{dt} = -1 - \frac{v z^2 e^z}{2mf}.$$

Their solution is, where  $\beta e^{c\beta} = e^z$ ,

$$2m = \frac{n}{c} z^2; \tag{4.1}$$

$$\begin{aligned}
 v &= n(1 - e^{-z}(1 + z))\beta; \\
 v_1 &= \frac{n}{c} [z^2 - zc\beta(1 - e^{-z})]; \\
 t &= \frac{n}{c} \left[ c(1 - \beta) - \frac{1}{2} \log^2 \beta \right].
 \end{aligned}$$

**Sub-critical case:**  $c < e$

Let  $h(z) = \frac{n}{c} [z^2 - zc\beta(1 - e^{-z})]$ . Let  $z^*$  be the largest nonnegative root of  $h(z) = 0$ . If  $c < e$  then  $z^* = 0$ . Because the process closely follows (4.1) we see that  $z = o(1)$  w.h.p. at the end of Phase 1. This gives  $m = o(n)$  (actually  $m = o(n^{0.9})$ ) and then a careful first moment calculation yields the conclusion of Theorem 4.2.

**Super-critical case:**  $c > e$

In this case we end Phase 1 with  $z = z^* > 0$ .

We have observed that

$$\mathbf{E}[v'_0 - v_0] = O\left(\frac{v_1}{m}\right) \text{ --- expected increase in unmatched vertices.}$$

It is enough to show that w.h.p.  $v_1 = \tilde{O}(n^{1/5})$  throughout the algorithm. Then we can argue that w.h.p. there are

$$\tilde{O}\left(n^{1/5} \sum_{m=1}^{cn} \frac{1}{m}\right)$$

vertices left unmatched in Phase 2.

**Controlling  $v_1$ :** We first observe that

$$v_1 > 0 \text{ implies } \mathbf{E}[v'_1 - v_1] \leq -\min\left(\frac{z^2}{200}, \frac{1}{20000}\right)$$

**Early Phase:**  $z \geq n^{-1/100}$ .

**Whp**  $v_1$  stays  $\tilde{O}(z^{-2})$ .

**Middle Phase:**  $n^{-1/100} \geq z \geq n^{-1/5}$

The graph is very sparse, most vertices are of degree two.

When  $v_1 > 0$  most vertices of degree one are at end of a long path. Removing such a vertex and its edge does not change  $v_1$  i.e.

$$\mathbf{Pr}(v'_1 = v_1 \mid v_1 > 0) = 1 - z + O(z^2).$$

**Whp**  $v_1$  stays  $\tilde{O}(z^{-1})$ .

**Final Phase:**  $z \leq n^{-1/5}$

We start this phase with

$$v \sim v_2 \sim Cnz^2 = \tilde{O}(n^{3/5})$$

Only  $\tilde{O}(n^{3/5}z) = \tilde{O}(n^{2/5})$  moves made in the “ $v_1$  walk” and so  $v_1$  can only move by the square root of this.

This completes our outline analysis of the Karp-Sipser algorithm. □

The Karp-Sipser algorithm runs in  $O(n)$  time and makes  $\tilde{O}(n^{1/5})$  mistakes.

Chebolu, Frieze and Melsted [31] show that these mistakes can be corrected i.e one can find a true maximum matching in  $O(n)$  time w.h.p., for  $c$  sufficiently large.

The key idea of this paper is that even though we have “looked” at all of the edges while running KSGREEDY, there is enough “residual randomness” to use to find a true maximum matching. We have to examine the output of the algorithm for unused edges whose distribution can be properly understood.

### 5. Hamilton Cycles

Determining whether or not a graph has a Hamilton cycle is NP-hard, Karp [74].

The threshold problem was solved in

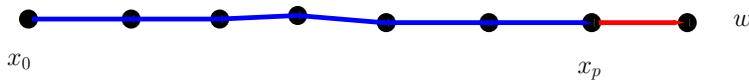
**Theorem 5.1** (Komlós and Szemerédi [83]). *Suppose that  $m = \frac{1}{2}n(\log n + \log \log n + c_n)$ . Then*

$$\lim_{n \rightarrow \infty} \Pr(G_{n,m} \text{ is Hamiltonian}) = \begin{cases} 0 & c_n \rightarrow -\infty \\ e^{-e^{-c}} & c_n \rightarrow c \\ 1 & c_n \rightarrow \infty \end{cases}$$

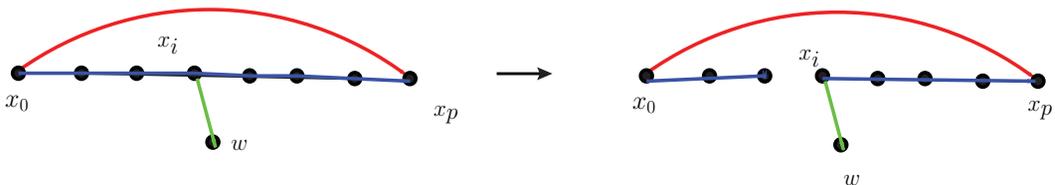
We will describe an algorithm that runs in polynomial time and finds a Hamilton cycle w.h.p. for the case  $c_n = \omega \rightarrow \infty$ .

**5.1. Pósa Rotations.** We can start our algorithm with any path  $P = (x_0, x_1, \dots, x_p)$ . Each round of the algorithm tries to replace  $P$  by a path  $Q$  of length one more i.e.  $p + 1$ . If successful,  $Q$  replaces  $P$  for the next round. This continues until we have a path of length  $n - 1$  i.e. a Hamilton path. After this, we attempt to create a Hamilton cycle.

If there is an edge joining  $x_0$  or  $x_p$  to a vertex  $w$  not in  $P$ , then we can extend the path to a longer one. With the edge  $\{x_p, w\}$  we have the longer path  $Q = (x_0, x_1, \dots, x_p, w)$ . We call this a *simple extension*.

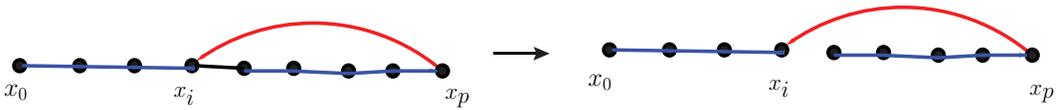


There is an alternative way of extending a path:



If the edge  $\{x_0, x_p\}$  exists then so will the edge  $\{x_i, w\}$ , for some  $i$ , and for some  $w \notin P$ , unless  $P + \{x_0, x_p\}$  is a hamilton cycle. This is under the assumption that our graph is connected. We then have the longer path  $(w, x_i, x_{i+1}, \dots, x_p, x_0, \dots, x_{i-1})$ . We call this a *cyclic extension*.

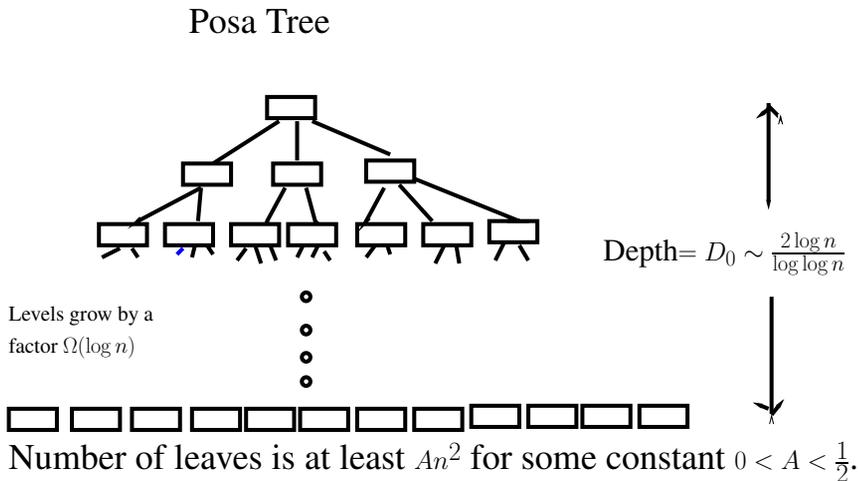
If there is no extension of either type, then we *rotate* the path:



Given the edge  $\{x_p, x_i\}$  where  $0 < i < p - 1$ , we create the new path  $(x_0, \dots, x_i, x_p, x_{p-1}, \dots, x_{x+1})$ , of the same length as  $P$ . By doing this we gain a new opportunity for an extension.

Pósa [98] used these rotations in a breakthrough paper that found the threshold for Hamiltonicity. He showed that if  $m \geq Kn \log n$  for some constant  $K > 0$  then  $G_{n,m}$  is hamiltonian w.h.p.

Let  $m = \frac{1}{2}n(\log n + \log \log n + \omega)$  and  $m_2 = \omega n/2$  and let  $m_1 = m - m_2$ . In the diagram below, labeled “Posa Tree”, each rectangle  $R$  represents a path  $P(R)$  of the same length as the path represented by the root rectangle. Rectangle  $R_1$  is the child of rectangle  $R_0$  iff  $P(R_1)$  is obtained from  $P(R_0)$  by a single rotation.



The algorithm we have in mind can be described as follows: At the beginning of a round we have a path  $P$ . We build a Pósa tree as indicated in the diagram, only using the first  $m_1$  edges. If at any stage we can extend one of the paths that we have constructed, then we extend it and go to the next round. Otherwise, we can show that w.h.p. we will end up with at least  $An^2$  paths, each path having a distinct pair of endpoints. Here  $A$  is a positive constant. We denote the set of pairs of endpoints by BOOST. We then go to our  $m_2$  edges and check one by one to see if any of them lies in BOOST. If so, we can make a cycle extension and move to the next round. We have enough edges so that we need not use an edge in more than one round.

The probability this algorithm fails can therefore be bounded by

$$\Pr(\text{Bin}(\omega_2 n, A) < n) = o(1).$$

It is polynomial time, because w.h.p. the Pósa tree has an  $O(\log n)$  branching factor at each vertex.

This algorithm has an unintuitive feel. Why should we restrict the roles of the edges to either tree building or cycle closing? We now describe the algorithm of Bollobás, Fenner

and Frieze [26] which is similar to what we have described, but makes no partition of the edges and furthermore is deterministic. It proceeds in rounds and in each round it uses all of the  $m$  available edges.

In a successful construction of a Pósa tree there are  $O(\log n / \log \log n)$  edges that are “vital”. These are the edges that are involved in the successful sequence of rotations that lead to an extension.

Let  $W$  be the set of vital edges picked out this way in all successful rotations, or incident to vertices of “low” degree. Then  $|W| = O(n \log n / \log \log n)$ .

We organise our algorithm deterministically so that the following is true. If the algorithm fails on  $G$  and we delete a set of edges  $X$  from  $G$ , where

$$(i) X \cap W = \emptyset \text{ and } (ii) X \text{ forms a matching}$$

then re-running the algorithm on  $G - X$  leads to failure at exactly the same stage.

Suppose  $|X| = \omega = \log n$  and that after removing  $X$ , any Posa tree has at least  $An^2$  leaves. This is true w.h.p.

Consider the  $\{0, 1\}$  function  $\psi(M, X)$  where  $M$  ranges over  $\binom{[n]}{m}$  and  $X$  ranges over  $\binom{[m]}{\omega}$ .

$\psi(M, X) = 1$  iff  $X$  satisfies (i), (ii) and our algorithm fails on the graph  $G = ([n], M)$ .

Observe that

$$\sum_X \psi(M, X) > 0 \text{ implies } \sum_X \psi(M, X) \geq \binom{m}{\omega} \times \left(1 - O\left(\frac{\log \log n}{\log n}\right)\right)^\omega.$$

So,

$$\Pr(\text{Algorithm fails}) \leq \frac{\sum_{M,X} \psi(M, X)}{\binom{m}{\omega} \times \left(1 - O\left(\frac{\log \log n}{\log n}\right)\right)^\omega} \times \frac{1}{\binom{N}{m}}$$

where  $N = \binom{[n]}{2}$ .

But,

$$\sum_{M,X} \psi(M, X) \leq \binom{N}{m-\omega} \binom{N-m+\omega-An^2}{\omega} \leq \binom{N}{m} \binom{m}{\omega} (1-A)^\omega.$$

This is because having chosen  $M \setminus X$  in  $\binom{N}{m-\omega}$  ways, we can't choose an edge that lies in BOOST, if we want to have  $\psi(M, X) = 1$ .

Therefore

$$\Pr(\text{Algorithm fails}) \leq \frac{(1-A)^\omega}{\left(1 - O\left(\frac{\log \log n}{\log n}\right)\right)^\omega} = o(1).$$

□

**5.2. Sparse random graphs.** With the threshold problem solved, existentially and constructively, we can consider other models of a random graph. In particular to those models where a minimum degree condition is automatically satisfied.

Let  $G(n, r)$  denote a random  $r$ -regular graph chosen uniformly from the set of all graphs with vertex set  $[n]$ . (Regular means that all vertices have the same degree)

**Theorem 5.2.**

$$\lim_{n \rightarrow \infty} \Pr(G(n, r) \text{ is Hamiltonian}) = 1, \quad r \geq 3.$$

$r = O(1)$  was proved by Robinson and Wormald [101], [102].

$r \rightarrow \infty$  was proved by Krivelevich, Sudakov, Vu, Wormald [86] and Cooper, Frieze, Reed [46].

If each vertex independently chooses  $k$  random neighbors then we have the random graph  $G_{k-out}$ .

**Theorem 5.3** (Bohman and Frieze [19]).

$$\lim_{n \rightarrow \infty} \Pr(G_{k-out} \text{ is Hamiltonian}) = 1, \quad k \geq 3.$$

This is not implied by the previous results on random regular graphs.

Let  $G_{n,m;k}$  be sampled uniformly from all graphs with vertex set  $[n]$  that have  $m$  edges and minimum degree at least  $k$ . In this way the minimum degree condition is obtained directly by conditioning.

**Theorem 5.4** (Bollobás, Fenner and Frieze [27]). *Let  $m = \frac{1}{6}n(\log n + \log \log n + c_n)$  then*

$$\lim_{n \rightarrow \infty} \Pr(G_{n,m;2} \text{ is Hamiltonian}) = \begin{cases} 0 & c_n \rightarrow -\infty \\ e^{-f(c)} & c_n \rightarrow c \\ 1 & c_n \rightarrow +\infty \end{cases}$$

for some explicit function  $f(c)$ .

Now consider conditioning on minimum degree at least 3. Let

$$L_c = \lim_{n \rightarrow \infty} \Pr(G_{n,cn;3} \text{ is Hamiltonian})$$

**Conjecture:**  $L_c = 1$  for all  $c \geq 3/2$ .

The conjecture is true for  $c = 3/2$ . In this case we are dealing with random 3-regular graphs.

**Theorem 5.5** (Bollobás, Cooper, Fenner, Frieze [24]).  $L_c = 1$  for  $c \geq 128$ .**Theorem 5.6** (Frieze [55]).  $L_c = 1$  for  $c \geq 10$ .

The conjecture is true for  $c \geq 3$ , assuming a numerical solution of some differential equations.

*Sketch of Proof.* The proof for the case  $c \geq 10$  is based on the analysis of a greedy algorithm for finding a good 2-matching in the random graph  $G_{n,cn;3}$ . A 2-matching is a set of edges  $M$  where no vertex meets more than two edges of  $M$  (as opposed to one for a matching). By “good” we mean that as a spanning subgraph of  $[n]$ , it has  $O(\log n)$  components. The greedy algorithm can be thought of as a natural generalisation of the Karp-Sipser algorithm for matchings.

Given  $M$ , one can convert it to a Hamilton cycle, relatively easily. This being based on the relatively few components that it has.  $\square$

Following the approach of [31] to finding a perfect matching, Frieze and Haber [56] devised a fast algorithm to find the promised Hamilton cycle. It is based on first finding a good 2-matching and uses the “residual randomness” left by the greedy algorithm.

**Theorem 5.7** (Frieze and Haber [56]). *If  $c$  is sufficiently large then w.h.p. a hamilton cycle can be found in  $G_{n,cn;3}$  in  $O(n^{1+o(1)})$  time.*

## 6. Edge Weighted Graphs

In this section we consider some results for the expected optimal objective value of some classical well solved problems in Combinatorial Optimization when the costs are random.

**6.1. Spanning Trees.** Every edge  $e$  of the complete graph  $K_n$  is given a random length  $X_e$ . The edge lengths are independently uniform  $[0, 1]$  distributed.  $Z_n$  is the minimum total length of a spanning tree. A spanning tree is a connected subgraph of  $K_n$  with no cycles. It has  $n - 1$  edges.

The conceptually simplest algorithm for finding a minimum spanning tree in a an edge weighted graph  $G = (V, E)$  is the greedy algorithm. We order the edges of  $G$  as  $e_1, e_2, \dots, e_m$  where  $\ell(e_i) \leq \ell(e_{i+1})$ , here  $\ell$  denotes edge length. Let  $G_i = (V, \{e_1, e_2, \dots, e_i\})$ .

### GREEDY

```

I ← ∅.
For i = 1, 2, . . . , m do
    begin
        If  $e_i$  joins two disitinct components of  $G_{i-1}$  then  $I \leftarrow I \cup \{e_i\}$ .
    end;
Output I
end
    
```

The property we need from this is the following. Fix  $p$  and let  $G_p = (V, E_p)$  where  $E_p = \{e \in E : \ell(e) \leq p\}$  and let  $I_p = I \cap E_p$ . Suppose that the graph  $G_p$  has  $\tau(G_p)$  components. Then  $|I \setminus I_p| = \tau - 1$ .

Let  $T$  be the minimum spanning tree. In the following,  $G_p$  has the same distribution as  $G_{n,p}$ . Then,

$$\begin{aligned}
 Z_n = \ell(T) &= \sum_{e \in T} X_e \\
 &= \sum_{e \in T} \int_{p=0}^1 1_{(p < X_e)} dp \\
 &= \int_{p=0}^1 \sum_{e \in T} 1_{(p < X_e)} dp \\
 &= \int_{p=0}^1 |\{e \in T : p < X_e\}| dp \\
 &= \int_{p=0}^1 ((G_p) - 1) dp.
 \end{aligned}
 \tag{6.1}$$

Equation (6.1) follows from  $x = \int_{p=0}^1 1_{(p \leq x)} dp$  for any  $x \in [0, 1]$ .

**Fact.**  $p \geq 6 \log n/n$  implies that  $G_{n,p}$  is connected with sufficiently high probability.

**Fact.** Almost all of the integral is accounted for by small isolated tree components.

So,

$$\mathbf{E}(Z_n) = \int_{p=0}^1 (\mathbf{E}(\langle G_p \rangle) - 1) dp \tag{6.2}$$

$$\begin{aligned} &\sim \int_{p=0}^{6 \log n/n} \mathbf{E}(\# \text{ small isolated trees in } G_{n,p}) dp \\ &\sim \int_{p=0}^{6 \log n/n} \left( \sum_{k=1}^{\log^2 n} \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k) + \binom{k}{2} - k + 1} \right) dp \end{aligned} \tag{6.3}$$

$$\begin{aligned} &\sim \sum_{k=1}^{\log^2 n} \frac{n^k}{k!} k^{k-2} \frac{k!(k(n-k)!}{(k(n-k+1)!} \\ &\sim \sum_{k=1}^{\log^2 n} \frac{1}{k^3} \\ &\sim \zeta(3). \end{aligned}$$

**Expansion for (6.3):** We choose the vertices  $S$  of the small tree in  $\binom{n}{k}$  ways. We choose a spanning tree  $T$  of  $S$  in  $k^{k-2}$  ways. The probability that  $T$  exists is  $p^{k-1}$  and the probability that there are no other edges in  $S$  is  $(1-p)^{\binom{k}{2} - k + 1}$  and the probability that  $T$  is not connected to the rest of  $G_p$  is  $(1-p)^{k(n-k)}$ .

The above is most of the proof of the following:

**Theorem 6.1** (Frieze [54]).

$$Z_n \sim \zeta(3) \quad w.h.p.$$

□

The original proof was not so “clean”: The remarkable integral formula is due to Janson [67].

With more work we have

**Theorem 6.2** (Cooper, Frieze, Ince, Janson, Spencer [45]).

$$\mathbf{E}(Z_n) = \zeta(3) + \frac{c_1}{n} + \frac{c_2 + o(1)}{n^{4/3}}.$$

The constants  $c_1, c_2$  are made explicit in [45], but they are not “pretty”.

If we give random weights to an arbitrary  $r$ -regular graph  $G$  then under some mild expansion assumptions

**Theorem 6.3** (Beveridge, Frieze, McDiarmid [17]).

$$Z_n \sim \frac{n}{r} (\zeta(3) + \epsilon_r) \quad w.h.p.$$

where  $\epsilon_r \rightarrow 0$  as  $r \rightarrow \infty$ .

For example, if  $G$  is the complete bipartite graph  $K_{n/2, n/2}$  then  $Z_n \sim 2\zeta(3)$  w.h.p.

**6.2. Shortest Paths.** Here we consider the following problem. Every edge  $e$  of the complete graph  $K_n$  is given a random length  $X_e$ . The edge lengths are independently distributed with an exponential distribution, mean one, i.e.  $\Pr(X_e \geq t) = e^{-t}$  for all  $t \geq 0$ .

We let  $D(i, j)$  denote the shortest distance between vertex  $i$  and  $j$  i.e. the minimum length of a path from  $i$  to  $j$ .

**Theorem 6.4** (Janson [68]). *Let  $D_{i,j}$  be the shortest distance between  $i, j$  in the above model. Then*

$$D_{1,2} \sim \frac{\log n}{n}; \quad \max_j D_{1,j} \sim \frac{2 \log n}{n}; \quad \max_{i,j} D_{i,j} \sim \frac{3 \log n}{n}.$$

The proof of this is based on an analysis of the well know Dijkstra algorithm for finding shortest paths from a fixed vertex  $s$  to all other vertices in a non-negatively edge weighted graph.

After several iterations of this algorithm there is a tree  $T$ , rooted at  $s$ , such that if  $v$  is a vertex of  $T$  then the tree path from  $s$  to  $v$  is a shortest path. Let  $d(v)$  be its length. For  $x \notin T$  let  $d(x)$  be the minimum length of a path  $P$  that goes from  $s$  to  $v$  to  $x$  where  $v \in T$  and the sub-path of  $P$  that goes to  $v$  is the tree path from  $s$  to  $v$ . Then if  $d(y) = \min \{d(x) : x \notin T\}$  then  $d(y)$  is the length of a shortest path from  $s$  to  $y$  and  $y$  can be added to the tree.

Suppose that vertices are added to the tree in the order  $v_1, v_2, \dots, v_n$  and that  $Y_j = \text{dist}(v_1, v_j)$  for  $j = 1, 2, \dots, n$ . It follows from the memoryless property of the exponential distribution that

$$Y_{k+1} = \min_{\substack{i=1,2,\dots,k \\ j=k+1,\dots,n}} [Y_i + X_{v_i,v_j}] = Y_k + E_k$$

where  $E_k$  is exponential with rate  $k(n - k)$  and is independent of  $Y_k$ . This is because  $X_{\{v_i,v_j\}}$  is distributed as an independent exponential  $X$  conditioned on  $X \geq Y_k - Y_i$ .

Hence

$$\mathbf{E}Y_n = \sum_{k=1}^{n-1} \frac{1}{k(n - k)} = \frac{1}{n} \sum_{k=1}^{n-1} \left( \frac{1}{k} + \frac{1}{n - k} \right) = \frac{2}{n} \sum_{k=1}^{n-1} \frac{1}{k} = \frac{2 \log n}{n} + O(n^{-1}).$$

The Chebyshev inequality can be used to show concentration around the mean. This yields the second part of Theorem 6.4. The first part comes from the fact that vertex 2 is on average the  $\sim n/2$ th vertex added to the tree.

**6.3. Assignment Problem.** The assignment problem is the name given to the problem of finding a minimum weight perfect matching in a complete bipartite graph  $K_{n,n}$  where each edge is given a weight. ( $K_{n,n}$  has vertices  $V_1, V_2$  where  $V_1, V_2$  can be thought of as disjoint copies of  $[n]$ . There is an edge  $\{i, j\}$  for every  $i \in V_1, j \in V_2$ .)

In the random assignment problem, each edge  $e$  of  $K_{n,n}$  is given an independent random edge weight from some distribution  $\mathcal{D}$ . Let  $A_n$  denote the expected value of the minimum weight matching. Walkup [113] proved that  $\mathbf{E}(A_n) \leq 3$  when  $\mathcal{D}$  is the uniform  $[0, 1]$  distribution. Karp [77] improved this to  $\mathbf{E}(A_n) \leq 2$ . At this point there was no proof that  $\lim_{n \rightarrow \infty} \mathbf{E}(A_n)$  existed or not. Aldous [5] proved the limit existed and in a follow up [6] proved that  $\lim_{n \rightarrow \infty} \mathbf{E}(A_n) = \zeta(2) = \frac{\pi^2}{6}$ . Here the distribution  $\mathcal{D}$  is the exponential with mean one. Parisi [97] conjectured that the following is true.

$$\mathbf{E}(A_n) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \dots + \frac{1}{n^2}. \tag{6.4}$$

**Theorem 6.5** (Linusson and Wästlund [89], Nair, Prabhakar and Sharma [94]).  
*Equation (6.4) holds.*

It took quite a deal of effort to prove this theorem. Wästlund however, [114] finally gave a short proof of it.

### 7. Traveling Salesperson Problem- TSP

Having seen some results about polynomially time solvable problems. We now discuss two important NP-hard problems. We begin with the TSP. This is important in the history of the average case analysis of algorithms in that the paper by Karp [75] was an influential paper that proposed average case analysis as an antidote to the negative consequences of NP-completeness.

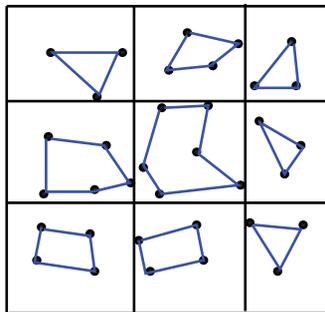
**7.1. Euclidean Version.** We begin with the Euclidean version of the TSP. Let  $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$  be chosen independently and uniformly from  $[0, 1]^2$ .

**Theorem 7.1** (Beardwood, Halton and Hammersley [13]). *There exists an absolute constant  $\beta > 0$  such*

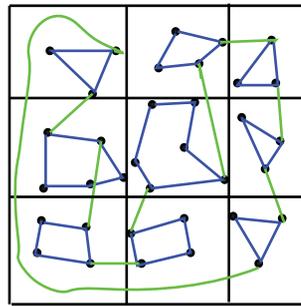
$$\frac{Z}{n^{1/2}} \rightarrow \beta \text{ with probability 1}$$

The precise value of  $\beta$  is unknown to this day.

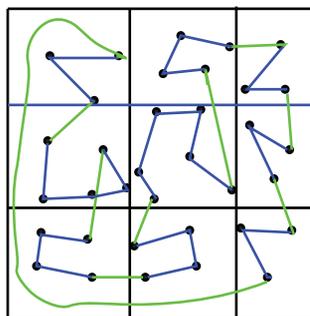
Karp [75] described a heuristic that runs in polynomial time and w.h.p. finds a tour of length within  $o(n^{1/2})$  of the minimum. Here is a simplified version of Karp's algorithm. First divide the unit square up into  $n/\log n$  subsquares of side  $(\log n/n)^{1/2}$ . Each subsquare will w.h.p. contain  $O(\log n)$  points of  $\mathcal{X}$ . We can then use Dynamic Programming [66] to solve these problems in polynomial time.



Solve the individual problems in each sub-square.



Connect up the smaller tours as shown



Now remove edges to create a tour as shown in the diagram

It is not difficult to prove that the difference between the tour found by this heuristic and the optimal tour is bounded by a constant times the total length of the lines used to create the grid i.e.  $O((n/\log n)^{1/2})$ .

**7.2. Independent Costs Version.** We are given an  $n \times n$  matrix  $[c_{i,j}]$  where we assume that the  $c_{i,j}$  are independent uniform  $[0, 1]$  variables.

The aim is to compute

$$T(C) = \min \left\{ \sum_{i=1}^n c_{i,\pi(i)} : \pi \text{ is a \textbf{cyclic} permutation of } [n] \right\}$$

We compare this with the Assignment Problem already discussed. This can be re-formulated as to compute

$$A(C) = \min \left\{ \sum_{i=1}^n c_{i,\pi(i)} : \pi \text{ is a permutation of } [n] \right\}.$$

Here

$$\frac{\pi^2}{6} \sim A(C) \leq T(C) \leq A(C) + o(1) \text{ w.h.p.}$$

The RHS is due to Karp [76].

The TSP can then be thought of as finding a minimum weight cycle cover of the complete digraph in which there is only one cycle. Here a cycle cover is a set of vertex disjoint oriented cycles that cover every vertex.

Karp's Patching Algorithm:

- Solve the associated assignment problem.
- *Patch* the cycles together to get a tour. Karp observed that if  $C$  is a matrix with i.i.d. costs then the optimal permutation is uniformly distributed and so w.h.p. the number of cycles is  $\sim \log n$  – *Key Observation*.
- Karp showed that the cost of patching is  $o(1)$  w.h.p.

To patch vertex disjoint oriented cycles  $C_1, C_2$  we choose edges  $e_i = (x_i, y_i) \in C_i, i = 1, 2$ . We delete  $e_1, e_2$  from  $C_1, C_2$  and then add the edges  $(x_1, y_2), (x_2, y_1)$  to create a single cycle covering the vertices in  $C_1$  or  $C_2$ . In this way, Karp proved

**Theorem 7.2** (Karp [76]). *W.h.p.*  $GAP = T(C) - A(C) = o(1)$ .

This was improved:

**Theorem 7.3** (Karp and Steele [79]). *W.h.p.*  $GAP = T(C) - A(C) = O(n^{-1/2})$ .

By making the cycles large before doing the patching we have

**Theorem 7.4** (Dyer and Frieze [47]). *W.h.p.*  $GAP = T(C) - A(C) = o\left(\frac{\log^4 n}{n}\right)$ .

With more care we get

**Theorem 7.5** (Frieze and Sorkin [58]). *W.h.p.*  $GAP = T(C) - A(C) = O\left(\frac{\log^2 n}{n}\right)$ .

The main tool in the improvements to Karp and Steele comes from cheaply transforming the cycle cover so that each cycle has length at least  $n_0 = n \log \log n / \log n$ .

Having increased the cycle size to  $n_0 = n \log \log n / \log n$  we patch the cycles together using short edges. Each patch will cost  $O(\log n/n)$  and so the patching cost is  $o(\log^2 n/n)$ . The assumption here is that after making all the cycles of the permutation large, there are many edges of length  $O(\log n/n)$  that can be used to patch cycles together. Furthermore, the distribution of these edges is sufficiently nice so that we can claim: The probability we cannot patch a pair of cycles is at most

$$\left(1 - \Omega\left(\frac{\log^2 n}{n^2}\right)\right)^{\Omega(n_0^2)} = e^{-\Omega(\log^2 \log n)} = o(1/\log n).$$

In the same paper, Frieze and Sorkin observed that

**Theorem 7.6.** *W.h.p. the TSP can be solved exactly in  $2^{O(n^{1/2} \log^{O(1)} n)}$  time.*

Having solved the assignment relaxation as a linear program, we search for a set of non-basic variables to increase from zero to one. We then argue that the distribution of the reduced costs are independent and uniform in  $[o(1), 1]$ .

Let

$$I_k = \frac{\log^2 n}{n} [2^{-k}, 2^{-k+1}].$$

Chernoff bounds for the binomial distribution imply that w.h.p. there are  $\leq c_1 2^{-(k-1)} n \log n$  non-basic variables with reduced cost in  $I_k, 1 \leq k \leq k_0 = \frac{1}{2} \log_2 n$  and  $\leq 2c_1 \sqrt{n} \log n$  non-basic variables with reduced cost  $\leq c_1 \frac{(\log n)^2}{n^{3/2}}$ .

Thus w.h.p. we need only check the following number of sets:

$$2^{2c_1 \sqrt{n} \log n} \prod_{k=1}^{k_0} \sum_{t=1}^{2^k} \binom{c_1 2^{-(k-1)} n \log n}{t} = e^{O(\sqrt{n} \log^{O(1)} n)}$$

## 8. Random $k$ -SAT

This is the name given to random instances of a version of the Satisfiability problem for Boolean formulae in conjunctive normal form. More precisely, we have a set of *variables*  $V = \{x_1, x_2, \dots, x_n\}$ . The associated set of *literals* is  $L = \{x_1, \bar{x}_1, x_2, \bar{x}_2, \dots, x_n, \bar{x}_n\}$ . A *clause* is a subset of  $L$ .

An instance  $I$  of  $k$ -SAT is defined as follows: Clauses  $C_1, C_2, \dots, C_m$  where  $C_i \subseteq L, |C_i| = k, i = 1, 2, \dots, m$ .

A *truth assignment* is a map  $\phi : L \rightarrow \{0, 1\}$  such that  $\phi(\bar{x}_j) = 1 - \phi(x_j)$  for  $j = 1, 2, \dots, n$ .

$\phi$  satisfies  $I$  if  $1 \in \phi(C_i)$  for  $i = 1, 2, \dots, m$ .

The  $k$ -SAT problem: Determine whether or not there is a satisfying assignment for  $I$ . It is solvable in polynomial time for  $k \leq 2$ . It is NP-hard for  $k \geq 3$ .

For a random instance  $I$ , we choose literals  $\ell_1, \ell_2, \dots, \ell_k$  independently and uniformly for each  $C_i$ , without replacement.

**8.1. Bounds on the threshold.** Let  $m = cn$ . Then for a fixed  $\phi$ ,

$$\Pr(\phi \text{ satisfies } I) = \left(1 - \frac{1}{2^k}\right)^m.$$

So, if  $Z$  is the number of satisfying assignments,

$$\Pr(\exists \phi \text{ satisfying } I) \leq \mathbf{E}(Z) = 2^n \left(1 - \frac{1}{2^k}\right)^m = \left(2 \left(1 - \frac{1}{2^k}\right)^c\right)^n.$$

So  $I$  is unsatisfiable w.h.p. if  $c > 2^k \log 2$ . The hard question now is what is the smallest value of  $c$  for which  $I$  is unsatisfiable w.h.p. It is natural to make the following conjecture.

**Conjecture:**  $\exists c_k$  such that if  $m = cn$  then

$$\lim_{n \rightarrow \infty} \Pr(I \text{ is satisfiable}) = \begin{cases} 1 & c < c_k \\ 0 & c > c_k \end{cases}$$

Friedgut [53] has come close to proving this, in the same sense that he came close to proving the existence of a sharp threshold for graph coloring.

The conjecture is true for  $k = 2$ . It is known that  $c_2 = 1$ . Now if  $m = cn$  and

$$\Pr(Z > 0) \geq \frac{\mathbf{E}(Z)^2}{\mathbf{E}(Z^2)} \tag{8.1}$$

and if the RHS here is bounded below then Friedgut's result will imply that  $c \leq c_k$ .

However, with  $Z$  equal to the number of satisfying assignments, the second moment method fails in the sense that the RHS of (8.1) tends to zero. Achlioptas and Peres [3] replace  $Z$  by

$$Z_1 = \sum_{\phi \text{ satisfies } I} \gamma^{H(\phi, I)}$$

where  $H(\phi, I) = \# \text{ true literals} - \# \text{ false literals}$  in  $I$  for  $\phi$ .

Using a careful choice of  $0 < \gamma < 1$  they proved

**Theorem 8.1.** *If*

$$c < 2^k \log 2 - (k + 1) \frac{\log 2}{2} - 1 - o_k(1)$$

*then  $I$  is satisfiable w.h.p.*

Using a more complicated random variable and doing more conditioning, but still using the second moment method, Coja-Oghlan and Panagiotou [39] proved that if

**Theorem 8.2.**

$$c < 2^k \log 2 - 3 \frac{\log 2}{2} - 1 - o_k(1)$$

*then  $I$  is satisfiable w.h.p.*

Very recently, Coja-Oghlan [34] tightened this to

**Theorem 8.3.**

$$c < 2^k \log 2 - \frac{1 + \log 2}{2} - o_k(1)$$

*then  $I$  is satisfiable w.h.p.*

Finding  $c_k$  for  $k = O(1)$  is a major open problem. If we allow  $k$  to grow then things become simple: Coja-Oghlan and Frieze [37] proved

**Theorem 8.4.** *Suppose that  $k - \log_2 n \rightarrow \infty$  and that  $m = 2^k(n \ln 2 + c)$  for an absolute constant  $c$ . Then,*

$$\lim_{n \rightarrow \infty} \Pr(I_m \text{ is satisfiable}) = 1 - e^{-e^{-c}}$$

## 8.2. Algorithms.

We first consider *Greedy Algorithms*:

These start with no values assigned to the variables.

Then, they repeatedly, choose a random clause  $C$  and assign a value to a variable of  $C$  to satisfy it.

The number of variables in the problem goes down by one.

Some clauses get satisfied and disappear from the problem, others shrink in size by one.

**Caveat:** If there are “small” clauses be careful. For example if there is a clause of size one, then one is forced to assign a particular value to the variable it contains. Of course, this might create an empty clause, if the current set of clauses contains a pair  $\{x_j\}, \{\bar{x}_j\}$ .

Repeat until all of the clauses are satisfied (**success**) or there is a clause remaining that is empty (**fail**).

Most of these greedy algorithms find a satisfying assignment w.h.p. provided there are at most  $\frac{c2^k}{k}n$  clauses, for small enough  $c$ .

A notable exception is the algorithm of Coja-Oghlan [33] which finds a satisfying assignment w.h.p. provided there are at most  $\frac{(1-\epsilon)2^k \log k}{k}n$  clauses.

We now consider *Walksat*

Start with the “all true” assignment:  $\phi(x_j) = 1$  for  $j = 1, 2, \dots, n$

### Repeat

Choose an unsatisfied clause  $C$

Choose a random variable from  $C$  and change its assigned value

**Until** instance is satisfied.

If the instance is satisfiable, then Walksat will *eventually* find a solution.

**Theorem 8.5** (Papadimitriou [96]). *For arbitrary cases of 2-SAT, the expected time to finish is polynomial.*

For random instances of 3-SAT:

**Theorem 8.6** (Alekhnovich and Ben-Sasson [7]). *Walksat solves a random instance of 3-SAT in polynomial time w.h.p. for  $m < 1.67n$ .*

The argument in [7] does not give a very good result for large  $k$ .

**Theorem 8.7** (Coja-Oghlan, Feige, Frieze, Krivelevich and Vilenchik [36]). *For large  $k$ , Walksat solves a random instance of  $k$ -SAT in polynomial time w.h.p. for  $m/n \leq c2^k/k^2$ .*

This was subsequently improved to

**Theorem 8.8** (Coja-Oghlan and Frieze [38]). *For large  $k$ , Walksat solves a random instance of  $k$ -SAT w.h.p. for  $m/n \leq c2^k/k$ .*

**Outline proof of Theorem 8.7.**

$$A_0 = \bigcup_{C \subseteq \bar{V}} C; \quad A_i = \left\{ x : \exists C \ni \bar{x} \text{ and } C \cap V \subseteq \bigcup_{j \leq i-1} A_j \right\}; \quad A = \bigcup_{i \geq 0} A_i.$$

**Fact 1** Walksat only changes the truth value of variables in  $A$ .

**Fact 2** So if  $C \setminus A \neq \emptyset$ , then  $C$  remains satisfied throughout.

**Fact 3**  $\forall C \subseteq A \exists L_C \subseteq C, |L_C| = 2k/3$  such that  $C \neq C'$  implies  $L_C \cap L_{C'} = \emptyset$ .

Now define assignment  $\sigma_A$ :

$$\sigma_A(x) = \begin{cases} 1 & x \notin A \text{ or } x \in \bigcup_{C \subseteq A} L_C \\ 0 & \bar{x} \in \bigcup_{C \subseteq A} L_C \\ 1 & \text{Otherwise} \end{cases}$$

$\sigma_A$  is a satisfying assignment. Now consider the Hamming distance between the current assignment  $\sigma_W$  of Walksat and  $\sigma_A$ . An iteration of Walksat reduces this by one with probability at least  $2/3$  and so by properties of simple random walk, this distance becomes zero in  $O(n)$  time w.h.p., (unless another satisfying assignment is found). This is similar to the idea used by Papadimitriou [96] for 2-SAT.

**8.3. Unsatisfiable instances.** When the number of clauses is greater than the (conjectured) satisfiability threshold then one is interested in the time taken to prove that such an instance is unsatisfiable. A seminal paper by Chvátal and Szemerédi [32] showed that if  $c > 0.7 \times 2^k$  is constant then w.h.p. any resolution proof of unsatisfiability of  $k$ -SAT must be exponentially long. A nice exposition of this and related results is given in Ben-Sasson and Wigderson [15].

**9. Conclusions**

We have hopefully shown a glimpse of an interesting area of research at the intersection of Combinatorics, Probability and Computing. Because of space limitations, we have omitted many things that we might have discussed and that the reader might find worth pursuing. As examples we have

- (i) Perfect matchings in random hypergraphs, Johansson, Kahn and Vu [70];
- (ii) Graph Property Resilience e.g. Sudakov and Vu [110], Ben-Shimon, Krivelevich and Sudakov [16];
- (iii) Ramsey properties of random graphs e.g. Rödl and Ruciński [103];
- (iv) Extremal properties of random graphs e.g. Conlon and Gowers [41], Schacht [105], Balogh, Morris and Samotij [11] and Saxton and Thomason [104];
- (v) Random subgraphs of arbitrary graphs of large degree e.g. Krivelevich, Lee and Sudakov [84], Riordan [100], Frieze and Krivelevich [57];
- (vi) Smoothed Analysis of the Simplex Algorithm for Linear Programs, Spielman and Teng [108] or Vershynin [112];

- (vii) Integer Feasibility of Random Polytopes, Chandrasekaran and Vempala [30];
- (viii) Nash Equilibria in Random Matrix Games, Bárány, Vempala and Vetta [12];
- (ix) Random Knapsack Problems, Lueker [92], Goldberg and Marchetti-Spaccemela [61], Beier and B. Vöcking [14];
- (x) The random graph  $G_{K,p} = ([n], E_p)$  where  $K$  is a down monotone convex body in the non-negative orthant of  $\mathbf{R}^n$  and  $E_p = \left\{ e \in \binom{[n]}{2} : X_e \leq p \right\}$  where  $X$  is chosen uniformly at random from  $K$ , see Frieze, Vera and Vempala [60];
- (xi) Models of real world networks e.g. Bollobás and Riordan [29], Cooper and Frieze [44];
- (xii) Random Simplicial Complexes e.g. Linial and Meshulam [88] or Kahle [72];
- (xiii) Achlioptas processes e.g. Bohman and Frieze [18] or Kang, Perkins and Spencer [73];
- (xiv) Random groups e.g. Gromov [65], Ollivier [95] or Antoniuk, Friedgut and Łuczak [8].

We conclude by mentioning some open questions.

- P1** Find a polynomial time algorithm that w.h.p. finds a planted clique of size  $o(n^{1/2})$  in  $G_{n,1/2}$ . (Negative results on the planted clique problem are given in [52]).
- P2** Find the precise threshold for the  $k$ -colorability of the random graph  $G_{n,p}$ . Find a polynomial time algorithm that optimally colors  $G_{n,p}$  w.h.p. or prove that this is impossible under some accepted complexity conjecture.
- P3** Find the precise threshold for the satisfiability of random  $k$ -SAT. Find a polynomial time algorithm that determines the satisfiability of random  $k$ -SAT w.h.p. or prove that this is impossible under some accepted complexity conjecture.
- P4** Prove that  $\lim_{n \rightarrow \infty} \Pr(G_{n,cn;3} \text{ is Hamiltonian}) = 1$  for  $c > 3/2$ .
- P5** Determine whether or not solving random asymmetric TSPs with independent costs by branch and bound runs in polynomial time w.h.p. when the bound used is the assignment problem value.
- P6** Analyse the *ordinary* simplex algorithm on random instances.
- P7** Let  $M$  be randomly chosen from the set of  $n \times n$  symmetric  $\{0, 1\}$  matrices with  $r \geq 3$  ones in each row and column. Prove that  $M$  is non-singular w.h.p.
- P8** Find a heuristic for the TSP in the unit square that w.h.p. comes with  $n^\alpha$  of the optimum, where  $0 < \alpha < 1/2$  is constant.
- P9** Determine the constant  $\beta$  in Theorem 7.1.
- P10** Determine the asymptotics for the value of a random multi-dimensional assignment problem and find asymptotically optimal heuristics, see Frieze and Sorkin [59].
- P11** Determine the threshold for a random subgraph of the  $n$ -cube to be Hamiltonian. See Bollobás [23] for the existence of a perfect matching.

**Acknowledgements.** The author was supported in part by NSF grants DMS0753472 and CCF1013110.

## References

- [1] D. Achlioptas and E. Friedgut, *A sharp threshold for  $k$ -colorability*, Random Structures and Algorithms **14** (1999), 63–70.
- [2] D. Achlioptas and A. Naor, *The two possible values of the chromatic number of a random graph*, Annals of Mathematics **162** (2005), 1333–1349.
- [3] D. Achlioptas and Y. Peres, *The threshold for random  $k$ -SAT is  $2^k \ln 2 - O(k)$* , Journal of the AMS **17** (2004), 947–973.
- [4] M. Ajtai, J. Komlós, and E. Szemerédi, *The first occurrence of Hamilton cycles in random graphs*, Annals of Discrete Mathematics **27** (1985), 173–178.
- [5] D. Aldous, *Asymptotics in the random assignment problem*, Probability Theory and Related Fields **93** (1992), 507–534.
- [6] ———, *The  $\zeta(2)$  limit in the random assignment problem*, Random Structures and Algorithms **18** (2001), 381–418.
- [7] M. Alekhnovich and E. Ben-Sasson, *Linear upper bounds for random walk on small density random 3-cnfs*, SIAM Journal on Computing **36** (2007), 1248–1263.
- [8] S. Antoniuk, E. Friedgut, and T. Łuczak, *A sharp threshold for collapse of the random triangular group*.
- [9] D.L. Applegate, R.E. Bixby, V. Chvátal, and W.J. Cook, *The Traveling Salesman Problem: A Computational Study*, Princeton University Press 2006.
- [10] J. Aronson, A. Frieze, and B. Pittel, *Maximum matchings in sparse random graphs: Karp-Sipser revisited*, Random Structures and Algorithms **12** (1998), 111–178.
- [11] J. Balogh, R. Morris, and W. Samotij, *Independent sets in hypergraphs*.
- [12] I. Bárány, S. Vempala, and A. Vetta, *Nash Equilibria in Random Games*, Random Structures and Algorithms **31** (2007), 391–405.
- [13] J. Beardwood, J. H. Halton, and J. M. Hammersley, *The shortest path through many points*, Proceedings of the Cambridge Philosophical Society **55** (1959), 299–327.
- [14] R. Beier and B. Vöcking, *Random knapsack in expected polynomial time*, Journal of Computer and System Science **69** (2004), 306–329.
- [15] E. Ben-Sasson and A. Wigderson, *Short proofs are narrow-resolution made simple*, Journal of the ACM **48** (2001), 149–169.
- [16] S. Ben-Shimon, M. Krivelevich, and B. Sudakov, *Local resilience and Hamiltonicity Maker-Breaker games in random regular graph*, Combinatorics, Probability and Computing **20** (2011), 173–211.
- [17] A. Beveridge, A.M. Frieze, and C. McDiarmid, *Random minimum length spanning trees in regular graphs*, Combinatorica **18**, 311–333.
- [18] T. Bohman and A.M. Frieze, *Avoiding a giant component*, Random Structures and Algorithms **19** (2001), 75–85.
- [19] ———, *Hamilton cycles in 3-out*, Random Structures and Algorithms **35** (2010), 393–417.
- [20] B. Bollobás, *The evolution of sparse graphs*, In Graph Theory and Combinatorics,

- Proceedings of a Cambridge Combinatorial Conference in honour of Paul Erdős (Bollobás, B., Ed.). Academic Press (1984), 35–57.
- [21] ———, *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability*, Cambridge University Press, 1986.
- [22] ———, *The chromatic number of random graphs*, *Combinatorica* **8** (1988), 49–56.
- [23] ———, *Complete matchings in random subgraphs of the cube*, *Random Structures and Algorithms* **1** (1990), 95–104.
- [24] B. Bollobás, C. Cooper, T.I. Fenner, and A.M. Frieze, *On Hamilton cycles in sparse random graphs with minimum degree at least  $k$* , *Journal of Graph Theory* **34** (2000), 42–59.
- [25] B. Bollobás and P. Erdős, *Cliques in random graphs*, *Mathematical Proceedings of the Cambridge Philosophical Society* **80** (1976), 419–427.
- [26] B. Bollobás, T.I. Fenner, and A.M. Frieze, *An algorithm for finding Hamilton paths and cycles in random graphs*, *Combinatorica* **7** (1987), 327–341.
- [27] ———, *Hamilton cycles in random graphs with minimal degree at least  $k$* , in *A tribute to Paul Erdős*, edited by A. Baker, B. Bollobás, and A. Hajnal (1990), 59–96.
- [28] B. Bollobás and A.M. Frieze, *On matchings and hamiltonian cycles in random graphs*, *Annals of Discrete Mathematics* **28** (1985), 23–46.
- [29] B. Bollobás and O. Riordan, *Mathematical results on scale-free random graphs*, in *Handbook of graphs and networks: from genome to the internet*, S. Bornholdt and H.G. Schuster eds. (2002), 1–34.
- [30] K. Chandrasekaran and S. Vempala, *Integer Feasibility of Random Polytopes*.
- [31] P. Chebolu, A.M. Frieze, and P. Melsted, *Finding a Maximum Matching in a Sparse Random Graph in  $O(n)$  Expected Time*, *Journal of the ACM* (2010), 161–172.
- [32] V. Chvátal and E. Szemerédi, *Many Hard Examples for Resolution*, *Journal of the ACM* **35** (1988), 759–768.
- [33] A. Coja-Oghlan, *A better algorithm for random  $k$ -SAT*, *Proceedings of the 36th IICALP Conference* (2009), 292–303.
- [34] ———, *The asymptotic  $k$ -SAT threshold*.
- [35] ———, *Upper-bounding the  $k$ -colorability threshold by counting covers*, *Electronic Journal of Combinatorics* 20:P32 (2013).
- [36] A. Coja-Oghlan, U. Feige, A.M. Frieze, M. Krivelevich, and D. Vilenchik, *On smoothed  $k$ -CNF formulas and the Walksat algorithm*, *Proceedings of the 20th ACM-SIAM Conference on Discrete Algorithms* (2009), 451–460.
- [37] A. Coja-Oghlan and A.M. Frieze, *Random  $k$ -SAT: the limiting probability for satisfiability for moderately growing  $k$* , *Electronic Journal of Combinatorics* 15:N2 (2008).
- [38] ———, *Analyzing Walksat on random formulas*, *Proceedings of the 9th ANALCO* (2012), 48–55.
- [39] A. Coja-Oghlan and K. Panagiotou, *Going after the  $k$ -SAT threshold*, *Proceedings of the 45th ACM Symposium on the Theory of Computing* (2013), 705–714.
- [40] A. Coja-Oghlan and D. Vilenchik, *Chasing the  $k$ -colorability threshold*, *Proceedings of the 54th IEEE Symposium on the Foundations of Computing* (2013), 380–389.

- [41] D. Conlon and T. Gowers, *Combinatorial theorems in sparse random sets*.
- [42] S. Cook, *The complexity of theorem proving procedures*, Proceedings of the Third Annual ACM Symposium on Theory of Computing (1971), 151–158.
- [43] C. Cooper and A.M. Frieze, *On the number of hamilton cycles in a random graph*, Journal of Graph Theory **13** (1989), 719–735.
- [44] ———, *On a general model of web graphs*, Random Structures and Algorithms **22**, (2003), 311–335.
- [45] C. Cooper, A.M. Frieze, N. Ince, S. Janson, and J. Spencer, *On the length of a random minimum spanning tree*.
- [46] C. Cooper, A.M. Frieze, and B. Reed, *Random regular graphs of non-constant degree: connectivity and Hamilton cycles*, Combinatorics, Probability and Computing **11**, 249–262.
- [47] M.E. Dyer and A.M. Frieze, *On patching algorithms for random asymmetric traveling salesman problems*, Mathematical Programming **46** (1990), 361–378.
- [48] J. Edmonds, *Paths, trees, and flowers*, Canadian Journal of Mathematics **17** (1965), 125–130.
- [49] P. Erdős and A. Rényi, *On random graphs I*, Publ. Math. Debrecen **6** (1959), 290–297.
- [50] ———, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci. **5** (1960), 17–61.
- [51] ———, *On the existence of a factor of degree one of a connected random graph*, Acta. Math. Acad. Sci. Hungar. **17** (1966), 359–368.
- [52] V. Feldman, E. Grigorescu, L. Reyzin, and Y. Xiao, *Statistical Algorithms and a lower bound for detecting planted cliques*.
- [53] E. Friedgut, *Sharp thresholds for graph properties, and the  $k$ -sat problem, with an appendix by Jean Bourgain*, Journal of the American Mathematical Society **12** (1999), 1017–1054.
- [54] A.M. Frieze, *On the value of a random minimum spanning tree problem*, Discrete Applied Mathematics **10** (1985), 47–56.
- [55] ———, *On a Greedy 2-Matching Algorithm and Hamilton Cycles in Random Graphs with Minimum Degree at Least Three*.
- [56] A.M. Frieze and S. Haber, *An almost linear time algorithm for finding Hamilton cycles in sparse random graphs with minimum degree at least three*.
- [57] A. Frieze and M. Krivelevich, *On the non-planarity of a random subgraph*, Combinatorics, Probability and Computing **22** (2013), 722–732.
- [58] A.M. Frieze and G. Sorkin, *The probabilistic relationship between the assignment and asymmetric traveling salesman problems*, SIAM Journal on Computing **36** (2007), 1435–1452.
- [59] ———, *Efficient algorithms for three-dimensional axial and planar random assignment problems*.
- [60] A.M. Frieze, S. Vempala, and J. Vera, *Logconcave Random Graphs*, Electronic Journal of Combinatorics (2010).
- [61] A.V. Goldberg and A. Marchetti-Spaccemela, *On finding the exact solution of a 0,1*

- knapsack problem*, Proceedings of the 16th Annual ACM Symposium on the Theory of Computing (1984), 359–368.
- [62] R. Glebov and M. Krivelevich, *On the number of Hamilton cycles in sparse random graphs*, SIAM Journal on Discrete Mathematics **27** (2013), 27–42.
- [63] R. Gomory, *Outline of an algorithm for integer solutions to linear programs*, Bulletin of the American Mathematical Society **64** (1958), 275–278.
- [64] G.R. Grimmett and C.J.H. McDiarmid, *On colouring random graphs*, Proceedings of the Cambridge Philosophical Society **77** (1975), 313–324.
- [65] M. Gromov, *Asymptotic invariants of infinite groups. Geometric Group Theory*, London Mathematical Society Lecture Notes Series **182** (1993), 1–295.
- [66] M. Held and R.M Karp, *A Dynamic Programming Approach to Sequencing Problems*, SIAM Journal of Applied Mathematics **10** (1962), 196–210.
- [67] S. Janson, *The minimal spanning tree in a complete graph and a functional limit theorem for trees in a random graph*, Random Structures Algorithms **7** (1995), 337–355.
- [68] ———, *One, two and three times  $\log n/n$  for paths in a complete graph with random weights*, Combinatorics Probability and Computing **8** (1999), 347–361.
- [69] S. Janson, D.E. Knuth, T. Łuczak and B. Pittel, *The Birth of the Giant Component*, Random Structures and Algorithms **4** (1993), 233–358.
- [70] A. Johansson, J. Kahn, and V. Vu, *Factors in Random Graphs*, Random Structures and Algorithms **33** (2008), 1–28.
- [71] S. Jukna, *Extremal Combinatorics: With Applications in Computer Science*, Springer 2011.
- [72] M. Kahle, *Topology of random simplicial complexes: a survey*.
- [73] M. Kang, W. Perkins, and J. Spencer, *The Bohman-Frieze process near criticality*, Random Structures and Algorithms **43** (2013), 221–250.
- [74] R.M. Karp, *Reducibility Among Combinatorial Problems*, In Complexity of Computer Computations, R.E. Miller and J.W. Thatcher Eds. (1972), 85–103.
- [75] ———, *Probabilistic Analysis of Partitioning Algorithms for the Traveling-Salesman Problem in the Plane*, Mathematics of Operations Research **2** (1977), 209–244.
- [76] ———, *A patching algorithm for the non-symmetric traveling salesman problem*, SIAM Journal on Computing **8** (1979), 561–573.
- [77] ———, *An upper bound on the expected cost of an optimal assignment*, Discrete Algorithms and Complexity: Proceedings of the Japan-US Joint Seminar (D. Johnson et al., eds.) Academic Press (1987), 1–4.
- [78] R.M. Karp and M. Sipser, *Maximum matchings in sparse random graphs*, Proceedings of the 22nd IEEE Symposium on the Foundations of Computer Science (1981), 364–375.
- [79] R.M. Karp and J.M. Steele, *Probabilistic analysis of heuristics*, in The traveling salesman problem: a guided tour of combinatorial optimization, E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan and D.B. Shmoys Eds., (1985), 181–205.
- [80] L.G. Khacian, *A polynomial algorithm in Linear Programming*, Soviet Mathematics

- Doklady **20** (1979,) 191–194.
- [81] J. Kim and V. Vu, *Concentration of multi-variate polynomials and its applications*, *Combinatorica* **20** (2000,) 417–434.
- [82] F. Knox, D. Kühn, and D. Osthus, *Edge-disjoint Hamilton cycles in random graphs, to appear in Random Structures and Algorithms*.
- [83] J. Komlós and E. Szemerédi, *Limit distributions for the existence of Hamilton circuits in a random graph*, *Discrete Mathematics* **43** (1983), 55–63.
- [84] M. Krivelevich, C. Lee, and B. Sudakov, *Long paths and cycles in random subgraphs of graphs with large minimum degree*.
- [85] M. Krivelevich and W. Samotij, *Optimal packings of Hamilton cycles in sparse random graphs*, *SIAM Journal on Discrete Mathematics* **26** (2012), 964–982.
- [86] M. Krivelevich, B. Sudakov, V. H. Vu and N. Wormald, *Random regular graphs of high degree*, *Random Structures and Algorithms* **18** (2001), 346–363.
- [87] L. Levin, *Universal search problems*, *Problems of Information Transmission* **9** (1973), 115–116.
- [88] N. Linial and R. Meshulam, *Homological connectivity of random 2-dimensional complexes*, *Combinatorica* **26** (2006), 475–487.
- [89] S. Linusson and J. Wästlund, *A proof of Parisi’s conjecture on the random assignment problem*, *Probability Theory and Related Fields* (2004), 419–440.
- [90] L. Lovász, *Combinatorial problems and exercises*, Akadémiai Kiadó - North Holland, Budapest, 1979.
- [91] T. Łuczak, *A note on the sharp concentration of the chromatic number of random graphs*, *Combinatorica* **11** (1991), 295–297.
- [92] G.S. Lueker, *On the average distance between the solutions to linear and integer knapsack problems*, *Applied Probability - Computer Science, The Interface* **1** (1982), 489–504.
- [93] D.W. Matula, *The largest clique size in a random graph*, Technical Report, Department of Computer Science, Southern Methodist University, Dallas, Texas, 1976.
- [94] C. Nair, B. Prabhakar, and M. Sharma, *Proofs of the Parisi and Coppersmith-Sorkin random assignment conjectures*, *Random Structures and Algorithms* **27** (2005), 413–44.
- [95] Y. Ollivier, *Sharp phase transition theorems for hyperbolicity of random groups*, *Geometry and Functional Analysis* **14** (2004), 595–679.
- [96] C.H. Papadimitriou, *On selecting a satisfying truth assignment*, *Proceeding of the 32nd IEEE Symposium on the Foundations of Computing* (1991), 163–169.
- [97] G. Parisi, *A Conjecture on Random Bipartite Matching*, *Physics e-Print archive* (1998). <http://xxx.lanl.gov/ps/cond-mat/9801176>
- [98] L. Pósa, *Hamiltonian circuits in random graphs*, *Discrete Mathematics* **14** (1976), 359–364.
- [99] W. Rhee and M. Talagrand, *Martingale Inequalities and NP-Complete Problems*, *Mathematics of Operations Research* **12** (1987), 177–181.
- [100] O. Riordan, *Long cycles in random subgraphs of graphs with large minimum degree*.

- [101] R.W. Robinson and N.C. Wormald, *Almost all cubic graphs are Hamiltonian*, Random Structures and Algorithms **3** (1992), 117–126.
- [102] \_\_\_\_\_, *Almost all regular graphs are Hamiltonian*, Random Structures and Algorithms **5** (1994), 363–374.
- [103] V. Rödl and A. Ruciński, *Threshold functions for Ramsey properties*, Journal of the American Mathematical Society **8** (1995), 917–942.
- [104] D. Saxton and A. Thomason, *Hypergraph containers*.
- [105] M. Schacht, *Extremal results for random discrete structures*.
- [106] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Pearson Education, 2013.
- [107] E. Shamir and J. Spencer, *Sharp concentration of the chromatic number on random graphs  $G_{n,p}$* , Combinatorica **7** (1987), 121–129.
- [108] D. Spielman and S-H. Teng, *Smoothed Analysis of Algorithms: Why The Simplex Algorithm Usually Takes Polynomial Time*, Journal of the ACM **51** (2004), 38–463.
- [109] R. Stanley, *Enumerative Combinatorics, Volumes 1 and 2*, Cambridge University Press, 1999.
- [110] B. Sudakov and V. Vu, *Local resilience of graphs*, Random Structures and Algorithms **33** (2008), 409–433.
- [111] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Publications Mathématiques de l’I.H.E.S. **81** (1995), 73–205.
- [112] R. Vershynin, *Beyond Hirsch Conjecture: walks on random polytopes and smoothed complexity of the simplex method*, Proceedings of the 47th Annual Symposium on Foundations of Computer Science (2006), 133–142.
- [113] D.W. Walkup, *On the expected value of a random assignment problem*, SIAM Journal on Computing **8** (1979), 440–442.
- [114] J. Wästlund, *An easy proof of the  $\zeta(2)$  limit in the random assignment problem*, Electronic Communications in Probability (2009), 261–269.

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh PA15213, USA.

E-mail: alan@random.math.cmu.edu

# Approximate algebraic structure

Ben Green

**Abstract.** We discuss a selection of recent developments in arithmetic combinatorics having to do with “approximate algebraic structure” together with some of their applications.

**Mathematics Subject Classification (2010).** Primary 00A05; Secondary 00B10.

**Keywords.** Approximate group, Gowers norm, nilsequence, arithmetic combinatorics, additive combinatorics.

## 1. Introduction

Given an inequality, an extremely natural question to ask is

When does equality occur?

If a satisfactory answer to this is available, one might then ask

When does equality *almost* occur?

To be a little more precise, suppose that we have some family of functions  $\mathcal{F}$  and some map (functional)  $v : \mathcal{F} \rightarrow \mathbb{R}$ . The inequality we are considering might then be of the form

$$v(f) \leq M \quad \text{for all } f \in \mathcal{F}.$$

To give an example, the well-known isoperimetric inequality on  $\mathbb{R}^n$  may be stated in this form, with  $\mathcal{F}$  being the set of all functions  $1_A$  where  $A \subset \mathbb{R}^n$  is bounded and open (say),  $v(1_A)$  being the isoperimetric ratio  $\frac{|A|}{|\partial A|^{n/(n-1)}}$ , and  $M$  being the isoperimetric ratio of any Euclidean ball.

The first natural question, the *equality question*, is then

For which  $f \in \mathcal{F}$  do we have  $v(f) = M$ ?

In the case of the isoperimetric inequality, it is well-known (and invariably stated as part of the inequality) that  $f = 1_A$  with  $A$  being a Euclidean ball.

The second natural question is

For which  $f \in \mathcal{F}$  do we have  $v(f) \approx M$ ?

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

Of course, to make this precise we must specify what is meant by  $\approx$ . We further distinguish between what might be called the *stability question*, which asks

$$\text{For which } f \in \mathcal{F} \text{ do we have } v(f) \geq (1 + o(1))M?$$

and what I shall term the *robustness question*, which asks

$$\text{For which } f \in \mathcal{F} \text{ do we have } v(f) \geq \frac{1}{100}M \text{ (say)?}$$

Most of this article will be concerned with the robustness question for two particular inequalities, an instance of *Young’s inequality for convolutions* and an inequality concerning the *Gowers norms*. In both situations the equality cases are easily established and are highly algebraic in nature (essentially they characterise finite groups and polynomial phases respectively). In both cases study of the robustness question has proven to be surprisingly subtle and has led to diverse applications in areas as different as group theory, additive prime number theory and theoretical computer science.

The stability question for these same inequalities is much better understood, though it is still nontrivial and has many applications. For want of space, we will not say a great deal about it. As it turns out the stability question for the isoperimetric inequality and related inequalities such as the Brunn-Minkowski inequality is the subject of much current research, not entirely unrelated to the topics discussed in this article: see for example [26, 27].

## 2. Approximate groups

**2.1. Young’s inequality.** Let  $G$  be a group with identity element  $\text{id}_G$ , and let  $\mathcal{F}$  be the collection of all finitely-supported functions  $f : G \rightarrow [0, \infty)$  with  $\sum_{x \in G} f(x) = 1$ ,  $f(x) = f(x^{-1})$  for all  $x$  and  $f(\text{id}_G) > 0$ . One may think of  $f$  as a probability measure on  $G$ , the measure of a set  $A \subset G$  being  $\sum_{x \in A} f(x)$ . A particular (rather simple) case of a well-known inequality of Young [101] for convolutions is the bound

$$v(f) \leq 1 \quad \text{for all } f \in \mathcal{F},$$

where

$$v(f) = \frac{\|f * f\|_2^2}{\|f\|_2^2} = \frac{\sum_{x \in G} (\sum_{y \in G} f(y)f(yx))^2}{\sum_{x \in G} f(x)^2}.$$

Let us give the proof, which follows in a couple of lines using the Cauchy-Schwarz inequality: for each  $x \in G$  we have

$$\sum_{y \in G} f(y)f(yx) \leq \left(\sum_{y \in G} f(y)^2\right)^{1/2} \left(\sum_{y \in G} f(yx)^2\right)^{1/2} = \sum_{y \in G} f(y)^2, \tag{2.1}$$

and thus

$$\left(\sum_{y \in G} f(y)f(yx)\right)^2 \leq \left(\sum_{y \in G} f(y)f(yx)\right) \sum_{y \in G} f(y)^2. \tag{2.2}$$

Summing over  $x \in G$  and using the fact that  $\sum_{t \in G} f(t) = 1$ , we obtain the result.

Let us address the equality question, that is to say let us characterise those  $f \in \mathcal{F}$  for which  $v(f) = 1$ . For this to happen, we must have equality in (2.2) for every  $x$ . For a

given  $x$  this means that either  $\sum_y f(y)f(yx) = 0$ , or else equality occurs in (2.1). The first case implies that for all  $y$  at least one of  $f(y)$  and  $f(yx)$  is zero. The second case may be analysed using the well-known criterion for equality in the Cauchy-Schwarz inequality. This implies that there is some  $\lambda(x)$  such that  $f(y) = \lambda(x)f(yx)$  for all  $y$ ; using the fact that  $\sum_{t \in G} f(t) = 1$ , it follows that  $\lambda(x) = 1$  and therefore  $f(y) = f(yx)$  for all  $y$ .

Thus  $v(f) = 1$  if and only if for all  $x \in G$  we have one of the following two mutually exclusive options:

1. For all  $y \in G$ , either  $f(y)$  or  $f(yx)$  is zero;
2. For all  $y \in G$ ,  $f(y) = f(yx)$ .

It follows immediately from this that  $f$  cannot take more than one non-zero value, and therefore  $f(x) = \frac{1}{|A|}1_A(x)$  for some (finite) symmetric set  $A \subset G$  containing the identity. The above two properties then tell us that for all  $x \in G$  we have one of the following two mutually exclusive options:

1.  $A$  and  $Ax$  are disjoint;
2.  $A = Ax$ .

The set of  $x$  for which (2) is satisfied is a subgroup of  $G$  (the stabiliser of  $A$  when  $G$  acts on finite subsets of itself by right multiplication). Call this group  $H$ . If  $x \in A$  then, since  $\text{id}_G \in A$ , the sets  $A$  and  $Ax$  are not disjoint and so  $x \in H$ ; thus  $A \subset H$ . On the other hand if  $x \in H$  then  $A = Ax$  and so in particular, since  $\text{id}_G \in A$ , we have  $x \in A$  and so  $H \subset A$ . It follows that  $A = H$  is a subgroup of  $G$ . Observations equivalent to these may be found in Hardy-Littlewood [58].

Now let us think about the stability and robustness questions. To do this, let us introduce a parameter  $K \geq 1$ , and let us ask what may be said about those  $f \in \mathcal{F}$  for which  $v(f) \geq \frac{1}{K}$ . This includes both the stability question (where  $K \approx 1$ ) and the robustness question (where  $K$  is somewhat larger, for example  $K \sim 100$ ). To spell it out, we are asking for a description of the finitely-supported, symmetric probability measures  $f : G \rightarrow [0, \infty)$  for which

$$\|f * f\|_2^2 \geq \frac{1}{K} \|f\|_2^2. \tag{2.3}$$

To get a feel for this question, let us specialise to the case  $f(x) = \frac{1}{|A|}1_A(x)$ , for some finite, symmetric set  $A \subset G$  containing the identity. We saw above that only this case is relevant for discussion of the equality question, and in fact the analysis of the stability and robustness questions may be reduced to this case by fairly routine technical arguments [7], [15, Appendix A]. In this case one may check that  $\|f\|_2^2 = |A|^{-1}$  and

$$\|f * f\|_2^2 = |A|^{-4} \#\{(a_1, a_2, a_3, a_4) \in A \times A \times A \times A : a_1 a_2 = a_3 a_4\}.$$

Thus (2.3) holds if and only if we have

$$|A|^{-3} \#\{(a_1, a_2, a_3, a_4) \in A \times A \times A \times A : a_1 a_2 = a_3 a_4\} \geq \frac{1}{K}. \tag{2.4}$$

The quantity on the left here is usually called the *multiplicative energy*  $E(A)$  of the set  $A$ . As can be seen, it records coincidences amongst products of elements of  $A$ . Young's inequality implies that  $E(A) \leq 1$ , and we showed above that equality occurs if and only if  $A$  is a

subgroup. That  $E(A) \leq 1$  can in fact be established easily and directly by noting that if  $a_1 a_2 = a_3 a_4$  then  $a_4$  is uniquely determined by  $a_1, a_2$  and  $a_3$ .

When, then, does (2.4) hold? Here we split the discussion of the stability question ( $K \approx 1$ ) and the robustness question ( $K \gg 1$ ), making just a few remarks about the former. In the stability case it turns out that  $A$  must be “almost” a subgroup; in fact there is a subgroup  $H$  such that the symmetric difference of  $A$  and  $H$  is very small. Results of this type are certainly very interesting and may be dated to work of Freiman [29] and Fournier [28] amongst others. Among the diverse applications are the analysis of certain algorithms for sampling at random from finite groups [19, 22, 42] and the solution of the Dirac-Motzkin conjecture in combinatorial geometry connected with point-line configurations having few ordinary lines [51].

Our main focus here, however, is on the robustness regime  $K \gg 1$ , where the flavour and the applications are somewhat different. We begin by observing that (2.4) is implied by a condition which is perhaps easier to understand, that of *small doubling*. We say that a set  $A \subset G$  has doubling at most  $K$  if

$$|A^2| \leq K|A|, \tag{2.5}$$

where  $A^2 = \{a_1 a_2 : a_1, a_2 \in A\}$ . To see that (2.5) implies (2.4), write  $r(x)$  for the number of representations of pairs  $(a_1, a_2) \in A \times A$  with  $a_1 a_2 = x$ . Then  $r(x) = 0$  for  $x \notin A^2$  and so by the Cauchy-Schwarz inequality we have

$$E(A) = \sum_x r(x)^2 \geq \frac{1}{|A^2|} \left( \sum_x r(x) \right)^2 = \frac{|A|^4}{|A^2|^2} \geq \frac{1}{K} |A|^3.$$

We have shown that (2.5) implies (2.4), and so if  $f(x) = \frac{1}{|A|} 1_A(x)$  for a symmetric set  $A$  satisfying (2.5) then indeed  $v(f) \geq \frac{1}{K}$ . Thus an analysis of the robustness question for Young’s inequality necessarily involves studying sets  $A$  satisfying (2.5). It is not at all obvious that such a study is sufficient for that task, because we have not shown that (2.4) implies (2.5). In fact, it does not, as be easily seen by taking  $A$  to be  $H \cup X$ , where  $H$  is a subgroup of  $G$  and  $X$  is an arbitrary symmetric set of the same size, disjoint from  $H$ . Then (2.4) holds with  $K = 8$ , since we may take all quadruples  $(a_1, a_2, a_3, a_4)$  with  $a_1, \dots, a_4 \in H$  and  $a_1 a_2 = a_3 a_4$ . However, there is absolutely no reason to suppose that (2.5) holds, and indeed  $A^2$  contains  $X^2$  which could have size as large as  $c|X|^2$ . We leave it to the reader to provide an explicit example in a suitable group  $G$ . Remarkably, however, the large multiplicative energy condition (2.4) does imply a weak version of (2.5): specifically, (2.5) is true after passing from  $A$  to a large subset  $A'$  and replacing (2.5) by a somewhat weaker condition  $|A'^2| \leq K'|A'|$  with  $K' \sim K^{10}$ , say. This result is known as the Balog-Szemerédi-Gowers theorem, because in the case  $G$  abelian it was established by Gowers [35] in the course of his seminal work on Szemerédi’s theorem, an earlier result of a qualitatively similar form but with the bound on  $K'$  being vastly weaker having previously been established by Balog and Szemerédi [2] by different means. It was shown by Tao [94] that the assumption that  $G$  is abelian could be dropped.

**2.2. Approximate groups.** We have discussed the relationship between the robustness question for Young’s inequality and the study of finite sets  $A$  satisfying the small doubling condition  $|A^2| \leq K|A|$ . Since subgroups of  $G$  provide equality in Young’s inequality, this provides some justification for thinking of such  $A$  as “approximate groups”. Moreover, the small doubling condition visibly suggests that  $A$  is somehow almost closed under multiplica-

tion, surely a property we would expect from any sensible notion of an approximate group. As it turns out, it has been found convenient to introduce a slightly different but closely related notion.

**Definition 2.1** (Approximate group). Let  $A$  be a subset of a group  $G$ . Then we say that  $A$  is a  $K$ -approximate group if  $A$  is symmetric, contains the identity, and if  $A^2 \subset XA$  for some set  $X$  of size at most  $K$ .

This definition was introduced by Tao [94] and has certain advantages such as behaving well under homomorphisms, making sense for infinite sets  $A$  as well as finite ones, and immediately implying further conditions on  $A$  such the tripling bound  $|A^3| \leq K^2|A|$ .

Note that if  $A$  is a  $K$ -approximate group then  $A$  automatically satisfies the small doubling condition (2.5), and hence the large multiplicative energy condition (2.4). The reverse direction is less clear, and the situation is much the same as before: a set satisfying (2.5) need not be a  $K$ -approximate group, but there is a closely related set  $A'$  which is a  $K'$ -approximate group for some  $K' \sim K^{10}$ . This deduction is essentially due to Ruzsa, who laid the foundation for the whole theory in a series of works. For a precise statement and further references, §4 of [14] may be consulted.

**2.3. Examples.** We now give some examples of approximate groups. The first example is fairly trivial.

**Example 1.** If  $A \subset G$  is a subgroup then of course  $A$  is symmetric,  $\text{id}_G \in A$  and  $A^2 = A$ . Thus  $A$  is a 1-approximate group.

Thus far, we have not pointed out that there are in fact nontrivial examples of approximate groups. The simplest is a geometric progression.

**Example 2.** If  $P$  is the geometric progression

$$P = P(u; N) := \{u^n : 0 \leq n < N\}$$

for some element  $u \in G$  and if  $A = P \cup P^{-1}$  then  $A$  is a 2-approximate group. Indeed  $A = \{u^n : -N + 1 \leq n \leq N - 1\}$ ,  $A^2 = \{u^n : -2N + 2 \leq n \leq 2N - 2\}$  and so  $A^2 \subset XA$  where  $X = \{u^{N-1}, u^{-N+1}\}$ .

Less obviously, there are multidimensional generalisations of the preceding example.

**Example 3.** If  $P$  is the multidimensional geometric progression

$$P = P(u_1, \dots, u_d; N_1, \dots, N_d) := \{u_1^{n_1} u_2^{n_2} \dots u_d^{n_d} : 0 \leq n_i < N_i\}$$

for some commuting elements  $u_1, \dots, u_d \in G$  and integers  $N_1, \dots, N_d > 0$  and if  $A = P \cup P^{-1}$  then  $A$  is a  $2^d$ -approximate group. We leave the confirmation of this to the reader.

The commuting assumption was very important in the previous example (otherwise we cannot simplify a product  $u_1^{n_1} \dots u_d^{n_d} u_1^{n'_1} \dots u_d^{n'_d}$  to  $u_1^{n_1+n'_1} \dots u_d^{n_d+n'_d}$ ). However, it can be replaced by the weaker condition of nilpotence, as the following example shows.

**Example 4.** Let  $N_1, N_2, N_{1,2}$  be positive integers with  $N_{1,2} \geq N_1 N_2$ , let  $G = \begin{pmatrix} 1 & \mathbb{R} & \mathbb{R} \\ 0 & 1 & \mathbb{R} \\ 0 & 0 & 1 \end{pmatrix}$  be the Heisenberg group and let  $A \subset G$  be the following set of matrices. Let

$$u_1 := \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad u_2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

and take  $A = P \cup P^{-1}$  where  $P = P(u_1, u_2, [u_1, u_2]; N_1, N_2, N_{1,2})$  is the set

$$\{u_1^{n_1} u_2^{n_2} [u_1, u_2]^{n_{1,2}} : 0 \leq n_1 < N_1, 0 \leq n_2 < N_2, 0 \leq n_{1,2} < N_{1,2}\}.$$

Here,  $[u_1, u_2]$  is the commutator given by

$$[u_1, u_2] := u_1 u_2 u_1^{-1} u_2^{-1} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It may be straightforwardly checked that

$$u_1^{n_1} u_2^{n_2} [u_1, u_2]^{n_{1,2}} \cdot u_1^{n'_1} u_2^{n'_2} [u_1, u_2]^{n'_{1,2}} = u_1^{n_1+n'_1} u_2^{n_2+n'_2} [u_1, u_2]^{n_{1,2}+n'_{1,2}-n'_1 n_2}. \quad (2.6)$$

Hence

$$P^{-1} \subset \{u_1^{n_1} u_2^{n_2} [u_1, u_2]^{n_{1,2}} : -N_1 < n_1 \leq 0, -N_2 < n_2 \leq 0, -2N_{1,2} < n_{1,2} \leq 0\}$$

and

$$A^2 \subset \{u_1^{n_1} u_2^{n_2} [u_1, u_2]^{n_{1,2}} : |n_1| < 2N_1, |n_2| < 2N_2, |n_{1,2}| < 5N_{1,2}\}.$$

Now for any  $n'_1, n'_2, n'_{1,2}$  in (2.6) we may choose (unique) integers  $k_1, k_2, k_{1,2}$  such that

$$u_1^{k_1 N_1} u_2^{k_2 N_2} [u_1, u_2]^{k_{1,2} N_{1,2}} \cdot u_1^{n'_1} u_2^{n'_2} [u_1, u_2]^{n'_{1,2}} \in P.$$

Indeed we have  $k_1 = -\lfloor n'_1/N_1 \rfloor$ ,  $k_2 = -\lfloor n'_2/N_2 \rfloor$  and  $k_{1,2} = -\lfloor (n'_{1,2} - n'_1 k_2 N_2)/N_{1,2} \rfloor$ .

Thus if  $u_1^{n'_1} u_2^{n'_2} [u_1, u_2]^{n'_{1,2}} \in A^2$  then  $|k_1| \leq 1$ ,  $|k_2| \leq 1$  and  $|k_{1,2}| \leq 6$ . Hence

$$A^2 \subset XP \subset XA,$$

where

$$X = \{u_1^{k_1 N_1} u_2^{k_2 N_2} [u_1, u_2]^{k_{1,2} N_{1,2}} : |k_1| \leq 1, |k_2| \leq 1, |k_{1,2}| \leq 6\}$$

is a set of size 117. That is,  $A$  is a 117-approximate group. (A smaller constant could be obtained with a more careful analysis.)

Example 4 is an example of a *nilprogression*. The key feature of the Heisenberg group  $G$  relevant to this example is the fact that it is nilpotent of class 2, which means that commutators of order 3 or higher are all equal to the identity, or equivalently that  $[u_1, u_2]$  commutes with everything else. Similar examples can be constructed in more general nilpotent groups of arbitrary class  $s$ , though the constant  $K$  (117 in Example 4) will generally grow with  $s$ . We will not give the details here, and refer the reader instead to [14, Definition 2.1]. The nilprogression in Example 4 is said to have rank 2 and class 2 (the rank being the number of generators  $u_i$  and the class being the nilpotency class of the group generated by  $u_1, u_2$ ).

Different instances of the above constructions may be combined to create new examples. For example, it is easy to see that the direct product of a  $K_1$ -approximate group and a  $K_2$ -approximate group is a  $(K_1 K_2)$ -approximate group. There are also combinations of the above examples which are not direct products, for example the following example of Helfgott [60].

**Example 5.** Let  $p$  be a large prime, let  $r, s, t \in \mathbb{F}_p$  be fixed generators of  $\mathbb{F}_p^*$ , let  $N_1, N_2, N_3$  be positive integers, and define  $A$  to be a set of  $3 \times 3$  matrices over  $\mathbb{F}_p$  as follows. Set  $A = HP \cup (HP)^{-1}$ , where

$$H := \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} : x, y, z \in \mathbb{F}_p \right\}$$

and

$$P = P(u_1, u_2, u_3; N_1, N_2, N_3) := \{u_1^{n_1} u_2^{n_2} u_3^{n_3} : 0 \leq n_i < N_i\}$$

with

$$u_1 := \begin{pmatrix} r & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, u_2 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{pmatrix}, u_3 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & t \end{pmatrix}$$

for some  $r, s, t \in \mathbb{F}_p^*$ . as in Example 3 above. It is quite easy to check that

$$A = \left\{ \begin{pmatrix} r^{n_1} & x & z \\ 0 & s^{n_2} & y \\ 0 & 0 & t^{n_3} \end{pmatrix} : x, y, z \in \mathbb{F}_p, -N_i < n_i < N_i \right\}$$

and hence

$$A^2 \subset \left\{ \begin{pmatrix} r^{n'_1} & x & z \\ 0 & s^{n'_2} & y \\ 0 & 0 & t^{n'_3} \end{pmatrix} : x, y, z \in \mathbb{F}_p, -2N_i < n'_i < 2N_i \right\},$$

from which it follows that  $A$  is an 8-approximate group.

In Example 5,  $HP$  was an example of a *coset nilprogression*, in this case of rank 3 and step 1. The general form of a coset progression is  $HP$  where  $P$  is a nilprogression (a notion we did not define in full generality) and  $H$  is a subgroup normal in the group  $\langle P \rangle$  generated by  $P$ . In fact, all five of our examples were of the form  $A = (HP) \cup (HP)^{-1}$  for some coset progression  $HP$  (in Examples 2, 3 and 4 the subgroup  $H$  was trivial, whilst in Example 1 the nilprogression  $P$  was trivial). Conversely, every  $A$  of this form is a  $K$ -approximate group, where  $K$  is bounded as a function of the rank  $r$  and the class  $s$  of  $P$ . Once again we refer the reader to [14] for more information.

**2.4. Theorems about approximate groups.** Given the discussion of the last section, it is natural to ask whether every  $K$ -approximate group is of the form  $(HP) \cup (HP)^{-1}$  for some coset nilprogression  $HP$  (of rank and step bounded in terms of  $K$ ). The answer to this is, strictly speaking, negative, as the following example of a set  $A \subset \mathbb{Z}$  shows. Here, we use additive notation for the group operation on  $\mathbb{Z}$  and so our interest is in  $2A = A + A$  rather than  $A^2$ . Define  $A$  to be  $\{0\} \cup \bigcup_{j=1}^N \{2j - \varepsilon_j, -2j + \varepsilon_j\}$ , where the  $\varepsilon_j$  are independent  $\{0, 1\}$ -valued random variables. Then  $2A \subset [-4N, 4N]$ . However,  $\{-1, 0, 1\} + A \supset [-2N, 2N]$ , and so  $\{-2N, 2N\} + \{-1, 0, 1\} + A \supset [-4N, 4N] \supset 2A$ . It follows that  $A$  is a 6-approximate group. However, for a typical choice of the  $\varepsilon_j$ ,  $A$  does not have nearly so much structure as a progression (though it is *syndetic*, that is to say has bounded gaps, which is what makes this construction work).

However we do have the following recent result of Breuillard, Tao and the author [13].

**Theorem 2.2.** *Let  $A$  be a  $K$ -approximate subgroup of a group  $G$ . Then there is a coset nilprogression  $B = HP$  of rank and class bounded as functions of  $K$ , where such that  $|B| \leq K'|A|$  and there is a set  $X \subset G$  with  $|X| \leq K'$  such that  $A \subset (XB) \cap (BX)$ . Here,  $K'$  may be bounded as a function of  $K$  only.*

We say that  $A$  is  $K'$ -controlled by  $B$ . In the example preceding the theorem, we may take  $B = \{0, \dots, N - 1\}$ . The reader is encouraged not to dwell too lengthily on the notion of “control” and read the above theorem as follows: every approximate group is roughly a coset nilprogression.

- For many specific types of group  $G$ , statements equivalent to Theorem 2.2 had previously been established, often with good quantitative control over the parameter  $K'$  as well as the rank and class. When  $G = \mathbb{Z}$ , this is essentially the celebrated Freiman-Ruzsa theorem [30, 80]. The general abelian case was handled by Ruzsa and the author [44], building on earlier work of Ruzsa [81]. Various matrix groups  $G$  were handled in work of (in chronological order) Elekes-Király [25], Chang [18] and Helfgott [59, 60], the latter handling  $\mathrm{SL}_2(k)$  and  $\mathrm{SL}_3(k)$  with  $k = \mathbb{F}_p$  or  $k = \mathbb{C}$ , amongst others.
- Hrushovski [65], in a very important 2009 breakthrough, dealt with  $G = \mathrm{GL}_n(\mathbb{C})$  (though with some dependence on  $n$ ). His argument was model-theoretic and a key ingredient of it was his “Lie model theorem”, also a key ingredient in the proof of Theorem 2.2.
- The proof of Theorem 2.2 additionally requires arguments related to the solution of Hilbert’s Fifth Problem (every locally compact group is locally an inverse limit of Lie groups), specifically lemmas due to Gleason from the 1950s. It also makes use of a lemma in additive combinatorics of a type developed by Sanders [84] and Croot-Sisask [21].
- Theorem 2.2 is in fact valid when  $G$  is a “local group” rather than a *bona fide* group. Moreover, it was necessary in [13] to work in this larger category, although Hrushovski and van den Dries have since managed to arrange the argument so that, at the expense of proving a slightly weaker result, one need only work in genuine groups.
- Theorem 2.2 rather easily implies Gromov’s famous theorem [56] on groups of polynomial growth. However, it does not really provide a new proof of Gromov’s theorem as all the deep ingredients Gromov developed (the notion of an asymptotic cone, and the application of the solution to Hilbert’s Fifth Problem) are also required here in some form.

Whilst Theorem 2.2 is definitive from the qualitative point of view, for many applications more quantitative statements are required. Unfortunately, a crucial use of ultrafilters in the proof of Theorem 2.2 means that no quantitative dependence of  $K'$  on  $K$  is currently known.<sup>1</sup> In the next section we will discuss perhaps the most substantial application of the theory of approximate groups so far, to the study of rapidly mixing random walks on groups (expanders), which find further application in the “affine sieve”. For these applications much more quantitative statements are required, but in more restricted settings.

A celebrated result of the type we have in mind is the theorem of Helfgott [59].

**Theorem 2.3** (Helfgott). *Let  $K \geq 2$ . Suppose that  $A \subset G$  is a  $K$ -approximate group, where  $G = \mathrm{PSL}_2(\mathbb{F}_p)$ , and that  $A$  generates  $G$ . Then either  $|A| \leq K^{C_2}$  or  $|A| \geq K^{-C_2}|G|$ , where  $C_2$  is an absolute constant.*

This theorem was generalised to  $\mathrm{PSL}_3(\mathbb{F}_p)$  by Helfgott in a subsequent paper [60], and then to  $\mathrm{PSL}_n(\mathbb{F}_p)$  (with  $C_2$  replaced by an exponent  $C_n$  depending on  $n$ ) and other finite simple groups of Lie type in independent works of Pyber-Szabó [79] and Breuillard, Tao and the author [11], the former paper containing a slightly more general result than the latter.

It is worth remarking that Helfgott’s arguments made substantial use of the theory of *approximate fields* or “sum-product theory”, in particular a result of Bourgain, Katz and Tao

---

<sup>1</sup>In principle one could be obtained by quantifier elimination but this would be a huge amount of effort and the bound would be desperately weak.

[10]. This is an important topic in arithmetic combinatorics and it has links to the theory of approximate groups as well as other substantial applications, perhaps most notably estimates for the additive Fourier transform of multiplicative subgroups of  $\mathbb{F}_p^\times$  due to Bourgain, Glibichuk and Konyagin [9]. The subsequent works [11, 79] do not make explicit use of this theory, and in fact it was noted in [11] that, conversely, results about approximate subgroups of  $SL_2(k)$  imply results about approximate subfields of  $k$ . Sadly we do not have the space to discuss these aspects any further here.

It is also of interest to note that [11] made use of an analogue for approximate groups of an argument of Larsen and Pink [72], which gives a self-contained and relatively concise proof of certain statements which follow from the Classification of Finite Simple Groups (CFSG).

**2.5. Applications.** Several applications of Theorem 2.2 are given in the paper [13]. In addition to certain refinements of Gromov’s theorem they include a result about the virtual nilpotence of the fundamental group of almost negatively-curved Riemannian manifolds, and a generalisation of a lemma of Margulis stating that the “almost stabiliser” of a point  $x$  in a finite-dimensional metric space  $X$  under the action of a discrete group of isometries is virtually nilpotent. Here, however, we wish to discuss an appealing application, due to Bourgain and Gamburd [7] of Helfgott’s result, Theorem 2.3.

The result concerns a property of a generating set  $S$  of a finite group  $G$  known as *expansion*. This property has several equivalent characterisations, details of which may be found in the survey [62]. For our purposes here, however, it is convenient to define expansion in terms of the rapid mixing of the random walk on generators  $S \cup S^{-1}$ . For the sake of illustration suppose that  $|S| = 2$ , write  $S = \{a, b\}$ , and imagine  $G$  being quite large. Then we perform a random walk of  $m$  steps, the end result of which is a product  $x_m = g_1 \dots g_m$  where each  $g_i$  is selected independently at random from the set  $\{a, b, a^{-1}, b^{-1}\}$ . Note that if  $G = \mathbb{Z}^2$  and  $S = \{(1, 0), (0, 1)\}$  (and if additive notation is used) then this is precisely the classical random walk on the plane  $\mathbb{R}^2$ .

Now  $x_1$  takes values in a set of size 4, tiny in comparison to  $|G|$ , and by a trivial induction  $x_j$  takes values in a set of size at most  $4^j$  (in fact by an almost-as-trivial induction one may reduce this to  $4 \cdot 3^{j-1}$ ). Thus if  $j = c \log |G|$  for some small value of  $c$  then  $x_j$  takes values in a set of size at most  $|G|^{c'}$ , and in particular is nowhere near to equidistributed on  $G$ . However in certain situations it turns out to be the case that  $x_j$  is highly equidistributed not much later than this time, say for  $j \geq C \log |G|$ , for some  $C$ . By “highly-equidistributed” let us (slightly arbitrarily) say that we mean  $\mathbb{P}(x_j = g) = \frac{1}{|G|} + O(\frac{1}{|G|^{10}})$  for all  $g \in G$ . The situation just described is one possible definition of what it means for  $S$  to be an expander (the precise definition must include the parameter  $C$ ).

**Theorem 2.4** (Bourgain-Gamburd). *Let  $G = \text{PSL}_2(\mathbb{F}_p)$  and suppose that  $S = \{(\begin{smallmatrix} 1 & 3 \\ 0 & 1 \end{smallmatrix}), (\begin{smallmatrix} 1 & 0 \\ 3 & 1 \end{smallmatrix})\}$ . Then the random walk on generating set  $S \cup S^{-1}$  becomes highly equidistributed in time at most  $C \log |G|$  for some absolute constant  $C$ , independent of  $p$ .*

The only reason we have put “3” in the matrices here is that the result was already known with “1” and “2” by different methods. Bourgain and Gamburd actually proved a far more general result, but we fix on this special case for the sake of illustration. To describe the proof, we note that the result can be reformulated in terms of convolution powers of  $\mu = \frac{1}{4}(\delta_a + \delta_b + \delta_{a^{-1}} + \delta_{b^{-1}})$ , that is to say the function  $\mu : G \rightarrow [0, \infty)$  taking values  $\frac{1}{4}$  at

$x = a, b, a^{-1}, b^{-1}$  and zero elsewhere. The probability that  $x_j = g$  is then precisely  $\mu^{(j)}(g)$ , where  $\mu^{(j)} = \mu * \dots * \mu$  and the convolution is repeated  $j$  times, and where we are defining  $\nu_1 * \nu_2(x) = \sum_y \nu_1(y)\nu_2(xy^{-1})$ . Note that  $\sum_x \mu^{(j)}(x) = 1$ .

We are interested in how quickly  $\mu^{(j)}$  tends towards the constant function  $\frac{1}{|G|}$ . To study this, we follow the progress of  $\mu^{(j)}$  in three stages:

- The early stage in which  $j \leq \frac{1}{10} \log |G|$ ;
- The middle stage in which  $\frac{1}{10} \log |G| \leq j \leq \frac{C}{10} \log |G|$ ;
- The end stage in which  $\frac{C}{10} \log |G| \leq j \leq C \log |G|$ .

The early stage is relatively easy to analyse. This is because the elements  $a, b$  behave as though they were generators of a free group, or in other words the random walk does not intersect itself nontrivially, so long as words of length at most  $\frac{1}{10} \log |G|$  are being considered. As a consequence, the size  $|\text{Supp}(\mu^{(j)})|$  of the support of  $\mu^{(j)}$  at the end of the early stage is somewhat large, of size at least about  $|G|^{0.01}$ , say. In fact the support is not quite the most sensible thing to look at, because  $\mu^{(j)}$  may take on several different values. A more nuanced quantity is  $\|\mu^{(j)}\|_2^{-2}$ , which we will call the *weighted support*. This would equal  $|\text{Supp}(\mu^{(j)})|$  if  $\mu^{(j)}$  did happen to be constant on its support, as is easily checked.

The theory of approximate groups is applied to analyse the middle stage. If  $j_1$  is the end of the early stage, we look at iterates  $\mu^{(j_1)}, \mu^{(2j_1)}, \dots, \mu^{(j_2)}$  where  $j_2 = 2^\ell j_1$  for some  $\ell$ . If  $\ell$  is somewhat large, the weighted support of  $\mu^{(2^t j_1)}$  cannot always increase substantially as we change  $t$  to  $t + 1$ , and so there must be some  $t$  for which the weighted supports of  $\mu^{(2^t j_1)}$  and of  $\mu^{(2^{t+1} j_1)}$  are roughly the same. Since  $\mu^{(2^{t+1} j_1)} = \mu^{(2^t j_1)} * \mu^{(2^t j_1)}$ , the only way that this can happen is if  $f = \mu^{(2^t j_1)}$  satisfies (2.3) for some fairly small value of  $K$ , that is to say  $\|f * f\|_2^2 \geq \frac{1}{K} \|f\|_2^2$  or  $v(f) \geq \frac{1}{K}$ . This is precisely the robustness question for Young's inequality that we have been studying. As we discussed, this situation implies, very roughly speaking<sup>2</sup>, that  $f \sim \frac{1}{|A|} 1_A(x)$  where  $A$  is a  $K$ -approximate group. Here of course we are concerned with the particular case  $G = \text{PSL}_2(\mathbb{F}_p)$ , so by Helfgott's Theorem 2.3 there are three possibilities: (i)  $A$  is tiny, (ii)  $A$  is almost all of  $G$  and (iii)  $A$  does not generate  $G$ . Case (i) cannot occur, because at the end of the early stage the weighted support of  $\mu^{(j_1)}$  was quite large. It turns out that (iii) also cannot occur, because of the particular structure of proper subgroups of  $\text{PSL}_2(\mathbb{F}_p)$ : they are all soluble and so satisfy the law  $[[x_1, x_2], [x_3, x_4]] = \text{id}_G$ , quite at odds with the free behaviour exhibited during the early stage. We are left, then with possibility (ii), which implies that the weighted support of  $\mu^{(2^t j_1)}$  is almost  $|G|$ . By further applications of Young's inequality the same is true of  $\mu^{(2^\ell j_1)} = \mu^{(j_2)}$ . That is to say, at the end of the middle stage  $\mu^{(j_2)}$  fills out a large portion of  $G$  in a fairly uniform way.

The analysis of the end stage involves still different ideas – an application of representation theory having its origin in a paper of Sarnak and Xue [88]. The crucial input is the fact that all nontrivial representations of  $G = \text{PSL}_2(\mathbb{F}_p)$  have dimension at least  $\frac{1}{2}(p - 1)$

<sup>2</sup>The  $\sim$  notation here hides quite a few technicalities.

and in particular at least  $|G|^c$  for some constant  $c$ . (In the language of Gowers [37],  $G$  is an example of a “quasirandom” group.) Further details may, of course, be found in the original paper [7].

The “Bourgain-Gamburd expansion machine” just described and modifications of it have found many further applications. One is the following variant of Theorem 2.4 due to Breuillard, Guralnick, Tao and the author [15].

**Theorem 2.5.** *Let  $G$  be any finite simple group of Lie type and suppose that  $S = \{a, b\}$  where  $a, b$  are chosen uniformly at random from  $G$ . Then, with probability at least  $1 - O(|G|^{-c})$ , the random walk on generating set  $S \cup S^{-1}$  becomes highly equidistributed in time at most  $C \log |G|$ . Here  $c, C > 0$  depend only on the rank of  $G$ .*

For example, this theorem holds with  $G = \text{PSL}_n(\mathbb{F}_q)$  and with  $C$  depending only on  $n$  and not on  $q$ . The proof of this theorem relies on the Bourgain-Gamburd expansion machine but with the work of Pyber-Szabó [79] and Breuillard, Tao and the author [11] in place of Helfgott’s theorem. It also requires several other ingredients, including two different *ad hoc* analyses in two particular families of groups (the symplectic groups  $\text{Sp}_4(k)$  in characteristic 3 and the triality Groups  ${}^3D_4(q)$ ). A different particular case, that in which  $G$  is a Suzuki group  $\text{Sz}(q)$ , had been handled in an earlier paper [12] of the authors. This was of a certain amount of interest because it completed the proof of the following theorem of Lubotzky, Kassabov and Nikolov [67].

**Theorem 2.6.** *There are absolute constants  $k, C$  with the following property. For any non-abelian finite simple group  $G$ , there is a set  $S \subset G$  of size at most  $k$  such that the random walk on generating set  $S \cup S^{-1}$  becomes highly equidistributed in time at most  $C \log |G|$ .*

The proof of this theorem depends on CFSG and the most impressive ingredient is, in my view, Kassabov’s proof [66] in the case  $G = A_n$ . It appears to be unknown whether or not Theorem 2.5 holds uniformly for all finite simple groups  $G$  with  $C$  an absolute constant, even in (especially in?) the case  $G = A_n$ .

Perhaps of greater interest for applications than results such as Theorem 2.5, however, are generalisations of the original Bourgain-Gamburd theorem, where the groups under consideration range over a family such as  $G = \text{PSL}_n(\mathbb{F}_p)$ ,  $p$  prime and the set  $S$  is obtained by reduction of a *fixed* set of integer matrices, rather than by random selection for each  $p$ . In the Bourgain-Gamburd theorem as stated above,  $S = \{(\begin{smallmatrix} 1 & 3 \\ 0 & 1 \end{smallmatrix}), (\begin{smallmatrix} 1 & 0 \\ 3 & 1 \end{smallmatrix})\}$ . The crucial property of this set of generators for rapid mixing of the random walk is that, considered as a subset of  $\text{SL}_2(\mathbb{Z})$ , the subgroup they generate is *Zariski dense* (not contained in any proper algebraic subvariety). That this condition is sufficient was established by Bourgain-Gamburd for the family  $\text{PSL}_2(\mathbb{F}_p)$ ,  $p$  prime. Varjú [97] obtained the same result for  $\text{PSL}_n(\mathbb{F}_p)$ , and moreover for  $\text{PSL}_n(\mathbb{Z}/q\mathbb{Z})$  where  $q$  is squarefree but may well be composite. (Such results had already been established in the case  $n = 2$  by Bourgain, Gamburd and Sarnak [8] by a more complicated method based in part on Helfgott’s arguments, necessitating in particular a foray into the tricky territory of *approximate subrings* of  $\mathbb{Z}/q\mathbb{Z}$ .) This last result is a crucial ingredient in the so-called *affine sieve* of Bourgain, Gamburd and Sarnak which finds almost primes in the matrix entries of orbits in matrix groups. Any serious discussion of this would take us too far afield, so we refer the reader to [82] for the state of the art and to the very nice exposition [87] for a (somewhat outdated) introduction. See also [41], again rather outdated.

**2.6. Open questions.** There are many open questions concerning the quantitative aspects of the theory described above. For example, no version of Theorem 2.2 in which the pa-

parameter  $K'$  is given quantitatively in terms of  $K$  is known, and nor does it seem prudent at this stage to speculate on what might be true in this regard. One tempting line of enquiry would be to look at Kleiner's alternative proof [68] of Gromov's theorem in the context of approximate groups, but this has not so far been successful.

Even in the case  $G = \mathbb{Z}$  there are unsolved problems connected with approximate groups. As previously noted, Theorem 2.2 in this case is due to Freiman [30] and Ruzsa [80]. In  $\mathbb{Z}$ , there are no interesting finite subgroups and, of course, all nilprogressions are automatically abelian progressions as in Example 3. Writing the group operation on  $\mathbb{Z}$  using addition as usual, the Freiman-Ruzsa theorem may be stated as follows.

**Theorem 2.7** (Freiman-Ruzsa). *Suppose that  $A \subset \mathbb{Z}$  is a  $K$ -approximate group, that is to say  $2A \subset X + A$  for some set  $X \subset \mathbb{Z}$  with  $|X| \leq K$ . Then there is a proper<sup>3</sup> progression  $P = P(u_1, \dots, u_d; N_1, \dots, N_d) := \{n_1 u_1 + \dots + n_d u_d : 0 \leq n_i < N_i\}$  which  $K'$ -controls  $A$ . Here,  $d$  and  $K'$  are bounded as functions of  $K$  only.*

The definition of "control" here is the same as in Theorem 2.2.

The optimal bounds on  $d$  and  $K'$  are not known. Following a sequence of developments by Chang [17] and Schoen [89], the state of the art is contained in a breakthrough paper of Sanders [85]. Sanders shows that we may take  $d \sim (\log K)^C$  and  $K' \sim e^{(\log K)^C}$  for some reasonable value of  $C$  (such as  $C = 4$ ). A key open question, known as the *Polynomial Freiman-Ruzsa conjecture*, asks whether one could in fact take  $d \sim \log K$  and  $K' \sim K^C$ . The bound  $d \sim \log K$  is significant as if  $P$  is a progression of this dimension then  $P \cup -P$  is itself a  $K^{C'}$ -approximate group. If one is prepared to sacrifice  $K'$  then bounds of this strength are known due to work of Freiman-Bilu [5] and Tao and the author [46]. For much greater depth on the quantitative issues surrounding Theorem 2.7, the recent survey of Sanders [86] may be consulted.

A solution to the Polynomial Freiman-Ruzsa conjecture ought to have serious applications in additive number theory – perhaps, for example, to questions about bases such as Waring's problem. However, no definite deductions of this type have so far been made.

Another abelian setting has attracted a lot of interest, and that is the case  $G = \mathbb{F}_2^{\mathbb{Z}}$ . In this group, where we have  $2 \cdot x = 0$  for every  $x$ , there are no interesting nilprogressions and one is left only with subgroups. Theorem 2.2 in this case is due to Ruzsa [81], and it may be stated as follows.

**Theorem 2.8** (Ruzsa). *Suppose that  $A \subset \mathbb{F}_2^{\mathbb{Z}}$  is a  $K$ -approximate group, that is to say  $2A \subset X + A$  for some set  $X \subset \mathbb{F}_2^{\mathbb{Z}}$  with  $|X| \leq K$ . Then there is a subgroup  $H \subset \mathbb{F}_2^{\mathbb{Z}}$  which  $K'$ -controls  $A$ . Here  $K'$  is bounded as a function of  $K$  only.*

The question of whether  $K'$  may be taken to be polynomial in  $K$  is also known as the Polynomial Freiman-Ruzsa conjecture, and it has attracted much attention. Ruzsa [81] attributes it to Katalin Marton. Once again the best results are due to Sanders [85], who shows that we may take  $K' \sim e^{(\log K)^C}$ . Ruzsa (unpublished, but see [39]) offers several equivalent formulations, of which the following is perhaps particularly appealing.

**Conjecture 2.9.** *Let  $V$  be a finite-dimensional vector space in characteristic 2. Suppose that  $f : V \rightarrow V$  satisfies the "approximate homomorphism" condition*

$$\{f(x+y) - f(x) - f(y) : x, y \in V\} \subset S.$$

<sup>3</sup>This means that all the sums  $n_1 u_1 + \dots + n_d u_d$  under consideration are distinct.

Then there is a linear map  $\tilde{f} : V \rightarrow V$  and a set  $\tilde{S}$  with  $|\tilde{S}| \ll |S|^C$  such that

$$\{f(x) - \tilde{f}(x) : x \in V\} \subset \tilde{S}.$$

There is an extremely extensive literature on the closely-related notion of a *quasimorphism* in contexts arising in geometric group theory; see [70] for a brief introduction. At present there seems to be little connection between that context, where the concern is usually with quasimorphisms on infinite groups, and ours.

### 3. Approximate polynomials

**3.1. Gowers norms and polynomial phases.** We turn now to the discussion of a different inequality. If  $f : \mathbb{Z} \rightarrow \mathbb{C}$  is a function and  $h \in \mathbb{Z}$  then we define the multiplicative derivative  $\Delta_h f$  by  $\Delta_h f(x) = f(x)\overline{f(x+h)}$ . Let  $k \geq 2$  be a fixed integer, and suppose that  $N$  is large in terms of  $k$ . Write  $[N] = \{1, \dots, N\}$ . Then we define the *Gowers  $U^k[N]$ -norm* of  $f$  by

$$\|f\|_{U^k[N]} = \left(\mathbb{E}_{x, h_1, \dots, h_k} \Delta_{h_1} \dots \Delta_{h_k} f(x)\right)^{1/2^k}.$$

Here, the average  $\mathbb{E}$  is over all  $x, h_1, \dots, h_k$  for which  $x + \omega_1 h_1 + \dots + \omega_k h_k \in [N]$  for all  $\omega_i \in \{0, 1\}$ ; this means that the Gowers norm depends only on the values taken by  $f$  on  $[N]$ . In taking  $2^k$ th roots we make use of the not completely obvious fact that  $\mathbb{E}_{x, h_1, \dots, h_k} \Delta_{h_1} \dots \Delta_{h_k} f(x)$  is real and non-negative. This is not too hard to prove by induction: see for example [96]. The basic theory of Gowers norms was originally developed in [36].

The Gowers norms satisfy the following rather trivial inequality: if  $\mathcal{F}$  is the set of all functions  $f : [N] \rightarrow \mathbb{C}$  with  $\|f\|_\infty \leq 1$ ,  $v(f) = \|f\|_{U^k[N]}$ , then

$$v(f) \leq 1. \tag{3.1}$$

(The inequality is indeed trivial – bound every instance of  $f(\cdot)$  in the definition of the Gowers  $U^k$ -norm by 1).

When does equality occur, that is to say for which  $f$  do we have  $v(f) = 1$ ? For this to happen, we must have<sup>4</sup>

$$\Delta_{h_1} \dots \Delta_{h_k} f(x) = 1 \quad \text{for all } x, h_1, \dots, h_k. \tag{3.2}$$

This implies that  $|f(x)| = 1$  for all  $x$ , and so we may write  $f(x) = e^{2\pi i \phi(x)}$  for some phase function  $\phi : \mathbb{Z} \rightarrow \mathbb{R}/\mathbb{Z}$ . The condition (3.2) then becomes

$$\partial_{h_1} \dots \partial_{h_k} \phi(x) = 0 \quad \text{for all } x, h_1, \dots, h_k, \tag{3.3}$$

where  $\partial_h$  is the additive difference operator defined by  $\partial_h \psi(x) = \psi(x) - \psi(x+h)$ .

The condition (3.3) is satisfied if and only if  $\phi$  is a polynomial of degree at most  $k - 1$ . The “if” direction of this assertion may be established by induction on the degree, since if

---

<sup>4</sup>Here and in what follows we ignore the restriction that  $x + \omega_1 h_1 + \dots + \omega_k h_k \in [N]$ ; this has little bearing on the argument.

$\phi$  is a polynomial of degree  $d$  then, for fixed  $h$ ,  $\Delta_h \phi$  is a polynomial of degree  $d - 1$ . Then “only if” direction can then be established by taking  $h_1 = \dots = h_k = 1$  in (3.3), which tells us that  $\phi(x + k)$  is uniquely determined as a function of  $\phi(x), \phi(x + 1), \dots, \phi(x + k - 1)$ . Therefore  $\phi$  is uniquely determined by its values at  $0, 1, \dots, k - 1$ , and hence coincides with the unique polynomial of degree at most  $k - 1$  which agrees with it at those points.

The stability question, that is to say the characterisation of those  $f$  for which  $v(f) \geq 1 - o(1)$ , is already interesting. It turns out that  $f$  must be closely approximated by a polynomial phase  $e^{2\pi i \phi(x)}$ . A precise statement and proof of this result may be found in [24, Theorem 1.2]. The argument there is analogous to an earlier argument [1] in a finite field setting, which has applications to property testing in theoretical computer science.

As with Young’s inequality, however, our main focus here will be on the robustness question: for which  $f$  do we have  $v(f) \geq \frac{1}{K}$ ? This is known as the *inverse question for the Gowers norms*. When  $k = 2$ , all such  $f$  are at least somewhat related to exponentials of linear phases (the solutions to the equality question  $v(f) = 1$ ).

**Theorem 3.1.** *Suppose that  $f : [N] \rightarrow \mathbb{C}$  is a function with  $|f(x)| \leq 1$  for all  $x$ , and that  $\|f\|_{U^2[N]} \geq \frac{1}{K}$ . Then there is some  $\theta \in \mathbb{R}/\mathbb{Z}$  such that*

$$\frac{1}{N} \left| \sum_{x \in N} f(x) e^{-2\pi i \theta x} \right| \geq \frac{1}{K^2}.$$

The proof of this is an exercise in Fourier analysis, given in detail in [40, Proposition 8.2]. When  $k \geq 3$ , however, the situation is different. There are examples of functions  $f : [N] \rightarrow \mathbb{C}$  with  $|f(x)| \leq 1$  for all  $x$ ,  $\|f\|_{U^3[N]} \geq \frac{1}{K}$ , but for which

$$\frac{1}{N} \left| \sum_{x \in N} f(x) e^{-2\pi i \phi(x)} \right| \ll N^{-c} \tag{3.4}$$

for all quadratic phases  $\phi : \mathbb{Z} \rightarrow \mathbb{R}/\mathbb{Z}$ . It is actually rather easy to give an example of such a function, though considerably less easy to prove rigorously that it *is* an example: take  $f(x) = e^{2\pi i \alpha x \{\beta x\}}$ , with  $\alpha, \beta \in \mathbb{R}$  sufficiently irrational numbers such as  $\alpha = \sqrt{2}$  and  $\beta = \sqrt{3}$ . Here  $\{t\}$  denotes fractional part. The “reason” this function  $f$  has large  $U^3[N]$ -norm is that the phase  $\phi(x) = \alpha x \{\beta x\}$ , whilst it does not satisfy the derivative condition (3.3) exactly, does satisfy this condition for a positive proportion of  $x, h_1, h_2, h_3$ : in fact whenever  $\{\beta x\}, \{\beta h_1\}, \{\beta h_2\}, \{\beta h_3\} \in [-\frac{1}{10}, \frac{1}{10}]$ . Establishing (3.4) rigorously is quite tricky.

A more natural way to construct such functions is as *nilsequences*. These objects should be thought of as “higher-order characters” generalising the linear exponentials  $\Phi(n) = e^{2\pi i \theta n}$ . To explain this generalisation we write  $\Phi$  in the form

$$\Phi(n) = F(p(n)) \tag{3.5}$$

where

- $p(n) = T^n 0$ , where  $T : \mathbb{R} \rightarrow \mathbb{R}$  is the translation map  $Tx = x + \theta$ ;
- $F(x) = e^{2\pi i x}$ . Note that this function is  $\mathbb{Z}$ -periodic.

A nilsequence corresponds to a generalisation of this in which  $\mathbb{R}$  is replaced by a simply-connected nilpotent Lie group  $G$  and  $\mathbb{Z}$  is replaced by a lattice  $\Gamma \subset G$ . With this setup, a nilsequence is of the form (3.5) with

- $p(n) = T^n \text{id}_G$ , where  $T : G \rightarrow G$  is a *nilrotation*, that is to say a map of the form  $Tx = xg$  for some  $g \in G$ ;
- $F : G \rightarrow \mathbb{C}$  is smooth and  $\Gamma$ -*automorphic*, which means that  $F(\gamma x) = F(x)$  for all  $x \in G$  and  $\gamma \in \Gamma$ .

For example, we could take  $G$  to be the Heisenberg group  $\begin{pmatrix} 1 & \mathbb{R} & \mathbb{R} \\ 0 & 1 & \mathbb{R} \\ 0 & 0 & 1 \end{pmatrix}$  and  $\Gamma$  to be the lattice  $\begin{pmatrix} 1 & \mathbb{Z} & \mathbb{Z} \\ 0 & 1 & \mathbb{Z} \\ 0 & 0 & 1 \end{pmatrix}$ . In fact for various reasons one usually considers a generalisation of this in which  $p(n)$  is a *polynomial sequence* on the group  $G$ . We will not discuss this important issue here, save to remark that it leads to essentially the same concept in the end due to a lifting argument of Furstenberg [31, p. 31] (see also [54, Appendix C]). We say that  $\Phi$  is an  $s$ -step nilsequence if the underlying nilpotent group  $G$  has nilpotency class  $s$ , that is to say if the lower central series of  $G$  is

$$G = G_1 \supset G_2 \supset \dots \supset G_s \supset G_{s+1} = \{\text{id}_G\},$$

with  $G_s$  nontrivial. For the Heisenberg group we have  $s = 2$ .

To give a specific example in the Heisenberg case, we need to specify  $g$  and an automorphic function  $F$ . The element  $g$  can of course be specified just by choosing a matrix. Non-trivial automorphic functions can be defined by hand (define  $F$  to be a smooth bump function supported on the interior a fundamental domain for  $\Gamma \backslash G$  and extend by automorphy). In the Heisenberg case there is a construction, pointed out in [64] for example, using the *Jacobi  $\theta$ -function*  $\theta(u, z) := \sum_n e^{\pi i z n^2 + 2\pi i n u}$  by defining  $F\left(\begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}\right) = e(z)e^{-\pi x^2} \theta(y + ix)$ .

Functions of the form  $e^{2\pi i \phi(x)}$  with  $\phi$  quadratic (that is to say, the functions for which  $\|f\|_{U^3[N]} = 1$ ) “morally” arise as nilsequences on the Heisenberg group by taking  $g = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & \alpha \\ 0 & 0 & 1 \end{pmatrix}$  and  $F(x, y, z) = e^{2\pi i(z-y[x])}$ . Indeed it may be checked by a computation that in this case we have  $\Phi(n) = e^{2\pi i \phi(n)}$  with  $\phi(n) = -\frac{1}{2}\alpha n(n+1)$ . The slight technical issue here is that, although  $F$  is automorphic (as may be confirmed by a computation) it is only piecewise smooth.

The following turns out to be true.

**Theorem 3.2.** *Suppose that  $\Phi(n)$  is an  $s$ -step nilsequence with<sup>5</sup>  $\|\Phi\|_2 = 1$ . Then  $\|\Phi\|_{U^{s+1}[N]} \geq \frac{1}{K}$ , where  $K$  is bounded above in terms of  $s$  and the “complexity” of  $\Phi$ .*

Giving a proper definition of the complexity is a rather tedious matter; it must take account of various parameters associated with  $G, \Gamma$  and the smoothness of the automorphic function  $F$ . The appendices of [49] go into considerable further detail.

The key to the proof is the observation that the multiplicative derivative  $\Delta_h \Phi(n) = \Phi(n)\overline{\Phi(n+h)}$  is an  $(s-1)$ -step nilsequence, which allows us to proceed inductively. In

---

<sup>5</sup>Here  $\|\Phi\|_2^2 = \frac{1}{N} \sum_{n \leq N} |\Phi(n)|^2$ . Some condition is needed to ensure that we do not have  $\Phi(n) = 0$  identically.

fact, this is not quite true, but it is true if the automorphic function  $F$  has the additional transformation property

$$F(g_s x) = \xi(g_s)F(x) \tag{3.6}$$

for every  $g_s$  in  $G_s$ , the last nontrivial subgroup in the lower central series of  $G$ , for some character  $\xi : G_s \rightarrow \mathbb{C}$  invariant under  $\Gamma$ . One may reduce to this case by a Fourier expansion on cosets of  $G_s$ . It is in effecting this Fourier expansion that the complexity of  $\Phi$ , and in particular the smoothness properties of  $F$ , comes into play. Suppose now that we do have the transformation property (3.6). For fixed  $h$  we have

$$\Delta_h \Phi(n) = F(T^n \text{id}_G) \overline{F(T^{n+h} \text{id}_G)} = \tilde{F}(T^n \text{id}_G)$$

where  $\tilde{F}(x) = F(x) \overline{F(T^h x)}$ . The function  $\tilde{F}$  is easily seen to be  $\Gamma$ -automorphic, and moreover it is invariant under  $G_s$ :

$$\tilde{F}(g_s x) = F(g_s x) \overline{F(T^n g_s x)} = \xi(g_s) F(x) \overline{\xi(g_s) F(T^n x)} = F(x) \overline{F(T^n x)} = \tilde{F}(x).$$

Here we used the fact that  $G_s$  is central in  $G$  to commute  $T$  with multiplication by  $g_s$ . As a consequence,  $\tilde{F}$  descends to an automorphic function on  $G/G_s$ , a nilpotent Lie group of class  $s - 1$ . (Unfortunately the preceding discussion was actually quite a serious oversimplification, as in the definition of Gowers norm  $h$  is not fixed but can vary over  $[-N, N]$ . With the argument just described, various smoothness norms of  $\tilde{F}$  depend heavily on  $h$  and to get around this a more complicated construction is required. Such a construction is given in [49, Section 7].)

We have seen that functions supplying equality in (3.1), the inequality  $v(f) \leq 1$ , are polynomial phases. Theorem 3.2 states that nilsequences of step  $k - 1$  and suitably bounded complexity are solutions to the corresponding robustness problem  $v(f) \geq \frac{1}{K}$ , and so we think of them as ‘‘approximate polynomials’’ (or, more accurately, approximate polynomial phases). The discussion of the previous paragraph, where we saw that the multiplicative derivatives of nilsequences (with an additional invariance property) are nilsequences of lower step, adds further weight to this philosophy.

It is very far from true that every solution to the robustness problem is a nilsequence. Indeed<sup>6</sup> if  $f : [N] \rightarrow \mathbb{C}$  and if  $\varepsilon : [N] \rightarrow \{-1, 1\}$  is a random  $\pm 1$ -valued function then almost surely  $v(f) \approx v(f + \varepsilon)$ . However, it is true that every solution is somewhat related to a nilsequence.

**Theorem 3.3.** *Suppose that  $f : [N] \rightarrow \mathbb{C}$  is a function with  $|f(n)| \leq 1$  and that  $\|f\|_{U^k[N]} \geq \frac{1}{K}$ . Then there is a  $(k - 1)$ -step nilsequence  $\Phi(n)$  with  $|\Phi(n)| \leq 1$  for all  $n$  such that*

$$\frac{1}{N} \left| \sum_{n \in N} f(n) \overline{\Phi(n)} \right| \geq \frac{1}{K'},$$

where  $K'$  and the complexity of  $\Phi$  are bounded in terms of  $K$  and  $k$  only.

Note that Theorem 3.3 is a generalisation of Theorem 3.1, which was essentially the case  $k = 2$ . This result is due to Tao, Ziegler and the author [54] and is known as the *Inverse*

---

<sup>6</sup>Passing from  $f$  to  $f + \varepsilon$  may destroy the property  $|f(x)| \leq 1$ , but we ignore this for the sake of illustration.

*Theorem for the Gowers norms.* A weaker “local” version of it was obtained by Gowers (in [35] for  $k = 3$ , and in [36] for general  $k$ ). The case  $k = 3$  was established by Tao and the author [45], and the case  $k = 4$  by Tao, Ziegler and the author [52]. It should most certainly be mentioned that the relevance of nilpotent Lie groups in this general arena first became apparent in the context of ergodic theory in works of Conze, Lesigne, Furstenberg and Weiss [20, 32–34]. A result which may be thought of as an “ergodic analogue” of Theorem 3.3 was obtained by Host and Kra [63] (see also independent work of Ziegler [102]). The notion of nilsequence itself, as well as the word, was introduced by Bergelson, Host and Kra [4]. See [71] for a nice introduction to these connections.

The conjecture to which Theorem 3.3 is a solution, together with potential applications of it to prime numbers, was formulated by Tao and the author [47] about four years before it was finally proved. The proof is unfortunately inordinately long and complicated. For a summary in about 20 pages, see [53]. An alternative approach has been developed by Szegedy [93] and Camarena-Szegedy [16], based in part on the work of Host and Kra, but these papers are not an easy read either.

The converse of Theorem 3.3 is also true, with appropriate changes to the constants. The proof is relatively straightforward and goes along very similar lines to the sketch of the proof of Theorem 3.2 we gave above: see [52, Appendix G] for further details.

Although we do not plan to discuss it much here, there has also been a good deal of work on finite field analogues of Theorem 3.3, which have applications in theoretical computer science. In addition to work by various subsets of the authors named above, we note that Samorodnitsky [83] established the case  $k = 3$  of Theorem 3.3 in the particularly interesting setting where  $[N]$  is replaced by a vector space in characteristic 2.

**Theorem 3.4.** *Suppose that  $V$  is a finite-dimensional vector space in characteristic 2 and that  $f : V \rightarrow \mathbb{C}$  is a function with  $|f(x)| \leq 1$  for all  $x \in V$ . Suppose that  $\|f\|_{U^3(V)} \geq \frac{1}{K}$ . Then there is a function  $\Phi : V \rightarrow \mathbb{C}$  of the form  $\Phi(x) = (-1)^{\psi(x)}$ , where  $\psi : V \rightarrow \mathbb{F}_2$  is a quadratic form, such that*

$$\frac{1}{|V|} \left| \sum_{x \in V} f(x)\Phi(x) \right| \geq \frac{1}{K'}.$$

Here  $K'$  is bounded in terms of  $K$  only.

The definition of the  $U^3(V)$ -norm is entirely analogous to that of the  $U^3[N]$  norm, except that the average  $\mathbb{E}$  is now simply taken over all  $x, h_1, h_2, h_3 \in V$ . Samorodnitsky obtained a bound of the form  $K' \sim e^{K^C}$ , but by combining his methods with the work of Sanders [85] one could improve this to  $K' \sim e^{(\log K)^C}$ . The similarity of these bounds to those stated in conjunction with Theorem 2.8 is no coincidence. Indeed a close relationship between the structure theory of approximate subgroups of  $\mathbb{F}_2^{\mathbb{Z}}$  and Theorem 3.4 was discovered by Tao and the author [48] and independently by Lovett [76]. In particular, it is known that the Polynomial Freiman-Ruzsa conjecture for finite fields, which is equivalent to Conjecture 2.9, is also equivalent to having a bound of shape  $K' \ll K^C$  in Theorem 3.4.

A similar equivalence between bounds in Theorem 2.7 and the case  $k = 3$  of Theorem 3.3 was developed in [48]: in other words the theories of approximate subgroups of  $\mathbb{Z}$  and of approximate quadratic polynomials are in a sense the same. I have often informally advanced the speculation that looking for a more effective proof of Theorem 3.3 may be a way of

attacking the Polynomial Freiman-Ruzsa Conjecture, though without any convincing ideas about how this might be achieved.

**3.2. Applications.** The theory of Gowers norms as described in the last section was for the most part developed to study arithmetic progressions. Gowers himself was interested in Szemerédi’s theorem, and Tao and the author were subsequently concerned with arithmetic progressions of primes. In [47] it was observed that the theory applied to a fairly wide class of “linear” problems, including questions about linear configurations of primes. Since this theory was discussed<sup>7</sup> in my 2006 ICM lecture [40] and is described in the article of Ziegler in these *Proceedings*, we restrict ourselves to a very brief account.

The connection of the Gowers norms to linear configurations comes from results called *generalised von Neumann inequalities*, which have the form

$$|T(f_1, \dots, f_t)| \ll \inf_{i=1, \dots, t} \|f_i\|_{U^{s+1}[N]}. \quad (3.7)$$

Here,  $f_1, \dots, f_t : [N] \rightarrow [-1, 1]$  are functions and

$$T(f_1, \dots, f_t) = \mathbb{E}_{(n_1, \dots, n_d) \in S} f_1(\psi_1(n_1, \dots, n_d)) \dots f_t(\psi_t(n_1, \dots, n_d)),$$

where the  $\psi_i : \mathbb{Z}^d \rightarrow \mathbb{Z}$  are affine-linear forms and  $S$  is a “nice” set (for example a convex set). For any system of forms  $\psi_1, \dots, \psi_d$  which is not degenerate in a certain sense, there is a value of  $s$  for which (3.7) holds. For example, if  $d = 2$ ,  $t = 3$  and  $\psi_1(n_1, n_2) = n_1$ ,  $\psi_2(n_1, n_2) = n_1 + n_2$ ,  $\psi_3(n_1, n_2) = n_1 + 2n_2$  (3-term arithmetic progressions) then we may take  $k = 2$ , whilst if  $d = 2$ ,  $t = 4$  and  $\psi_1(n_1, n_2) = n_1$ ,  $\psi_2(n_1, n_2) = n_1 + n_2$ ,  $\psi_3(n_1, n_2) = n_1 + 2n_2$ ,  $\psi_4(n_1, n_2) = n_1 + 3n_2$  (4-term arithmetic progressions) then we may take  $k = 3$ . The degenerate configurations are those in which some two of the  $\psi_i$  have equal homogeneous part, up to scalar equivalence: thus for example we cannot take  $\psi_1(n_1) = n_1$  and  $\psi_2(n_1) = n_1 + 2$ . The proof of any generalised von Neumann inequality is conceptually quite easy, involving only several applications of the Cauchy-Schwarz inequality, but notationally a little unpleasant. A general form of (3.7) was obtained in [47, Appendix D]. Furthermore the inequality was established there under a weaker condition on the  $f_i$  than boundedness, namely that  $|f_i(x)| \leq \nu(x)$  pointwise for some “pseudorandom measure”  $\nu$ . This is crucial for applications to the primes.

Ignoring a few technicalities, the manner in which (3.7) is applied to the primes is as follows. For technical convenience the primes are weighted using the von Mangoldt function  $\Lambda$ , defined by  $\Lambda(n) = \log p$  if  $n = p^k$  is a prime power and  $\Lambda(n) = 0$  otherwise. We are interested in  $T(\Lambda, \dots, \Lambda)$ , which counts how often the linear forms

$$\psi_1(n_1, \dots, n_d), \dots, \psi_t(n_1, \dots, n_d)$$

all take prime values as  $(n_1, \dots, n_d)$  ranges over a set  $S$ . To estimate this we split  $\Lambda$  in a certain manner as

$$\Lambda = \Lambda^\sharp + \Lambda^\flat, \quad (3.8)$$

where  $\Lambda^\sharp$  is “structured” and  $\Lambda^\flat$  is “unstructured”. Since  $T$  is multilinear, we may split  $T(\Lambda, \dots, \Lambda)$  as a sum of  $T(\Lambda^\sharp, \dots, \Lambda^\sharp)$  plus  $2^t - 1$  other terms, each of which involves at

<sup>7</sup>Naturally, however, this account is quite out of date and in particular predates the general case of Theorem 3.3.

least one copy of  $\Lambda^b$ . The first term provides the main term in the asymptotic formula for  $T(\Lambda, \dots, \Lambda)$ , and the aim is then to show that the other  $2^t - 1$  terms are all small. By (3.7), this may be accomplished if it can be shown that

$$\|\Lambda^b\|_{U^{s+1}[N]} = o(1).$$

By the inverse theorem for the Gowers norms, Theorem 3.3 (in the contrapositive), it is enough to establish that

$$\frac{1}{N} \left| \sum_{n \leq N} \Lambda^b(n) \overline{\Phi(n)} \right| = o(1) \tag{3.9}$$

for every  $s$ -step nilsequence  $\Phi(n)$  of bounded complexity. At least, this would be so were it not for the restriction  $|f(x)| \leq 1$  in Theorem 3.3: a large part of [47] is devoted to removing this restriction, showing that Theorem 3.3 implies a more general version of itself in which we only assume that  $|f(x)| \leq \nu(x)$  for some pseudorandom measure  $\nu$ .

The actual decomposition (3.8) we choose is based on the formula

$$\Lambda(n) = \sum_{d|n} \mu(d) \log(n/d),$$

where  $\mu$  is the Möbius function. It transpires that the task of establishing (3.9) may be further reduced to establishing that

$$\frac{1}{N} \left| \sum_{n \leq N} \mu(n) \overline{\Phi(n)} \right| \ll_A \log^{-A} N \tag{3.10}$$

for every  $A > 0$ . This statement was formerly known as the ‘‘Möbius and nilsequences conjecture’’, but it is now a theorem of Tao and the author [50]. Although the paper [50] is relatively short, it depends crucially on the much longer paper [49], in which various properties of nilsequences are established, in particular with regard to the distribution of finite orbit segments  $(T^n \text{id}_G)_{n \leq N}$  in  $\Gamma \backslash G$ . This work, like other material in this section, was motivated by earlier developments in the ergodic theory community, in particular work of Leon Green [55] and papers of Leibman of both an algebraic [73] and an ergodic-theoretic [74] nature.

**3.3. Open questions.** For me the key open question is to find the ‘‘right’’ proof of the inverse conjecture for the Gowers norms. At the moment the proofs are unsatisfactory on a conceptual level (the notion of a nilsequence is extremely natural, so it would be disappointing if it genuinely required 100+ pages to explain its role in Theorem 3.3). Furthermore, these proofs provide rather poor bounds for the complexity of the nilsequence  $\Phi$ , particularly when  $k \geq 4$  (in fact for  $k \geq 5$  the proofs provide no explicit bounds at all due to the use of ultrafilter arguments, though once again an explicit bound could in principle be extracted via quantifier elimination). As noted above it would be particularly interesting, in view of the link to approximate subgroups of  $\mathbb{Z}$ , to find a new approach to the inverse theorem when  $k = 3$ .

A more specific question is whether there is some smaller ‘‘natural’’ class of nilsequences. The space  $C^\infty(\Gamma \backslash G)$  of automorphic functions is extremely large, but we know for example

that in the case  $\Gamma = \mathbb{Z}$ ,  $G = \mathbb{R}$  the exponentials  $e^{2\pi i x}$  have a special role. Eigenfunctions of Laplacians are one natural avenue of enquiry. Furthermore the space of all simply-connected nilpotent Lie groups  $G$  together with lattices  $\Gamma$  is also extremely large and complicated, and it may be natural to focus on some subclass (for example free nilpotent Lie groups).

## 4. Other directions

To conclude this article I want to mention a personal selection of a few other inequalities where the equality, stability and robustness questions may hide interesting algebraic or somewhat algebraic structure. In some cases there is at least a tenuous connection to the main sections of the article, and in others less so.

**4.1. Inverse questions for the large sieve.** Let  $\mathcal{A}$  be a set of natural numbers with the property that  $|\mathcal{A} \pmod{p}| \leq \frac{1}{2}(p+1)$  for all sufficiently large primes  $p$ . The large sieve guarantees that  $|\mathcal{A} \cap [N]| \ll N^{1/2}$  for all  $N$ . This is sharp up to a multiplicative constant, as is shown by taking  $\mathcal{A}$  to be the set of squares (or the set of integer values of an arbitrary quadratic with rational coefficients).

It may well be the case that a very strong robustness assertion holds: if there is some  $K$  such that  $|\mathcal{A} \cap [N]| \geq \frac{1}{K}N^{1/2}$  for all sufficiently large  $N$  then  $\mathcal{A}$  is contained, up to a finite set, in the set of values of a rational quadratic. See [43] for evidence in this direction. This type of question was first raised by Helfgott and Venkatesh [61]; see also [99, 100].

**4.2. Point-line configurations.** Let  $\mathcal{P} \subset \mathbb{R}^2$  be a set of  $n$  points, no four on a line.<sup>8</sup> Write  $T(\mathcal{P})$  for the number of pairs  $(x, y) \in \mathcal{P}$  of distinct points for which there is a third distinct point  $z \in \mathcal{P}$  on the line  $\overline{xy}$ . Trivially,  $T(\mathcal{P}) \leq n(n-1)$ . Less obviously, equality cannot occur: this follows from a famous result known as the Sylvester–Gallai theorem.

Almost-equality can occur: we can obtain  $T(\mathcal{P}) = n^2 - O(n)$  by taking  $\mathcal{P}$  to be a suitable set of points on a suitable cubic curve (for example a coset of a subgroup on an elliptic curve, although there are singular examples too). This was noted by Sylvester in the 1860s [92]. Conversely, it was recently shown by Tao and the author [51] that there is a strong converse to this statement.

It would be very interesting to have an understanding of those  $\mathcal{P}$  for which  $T(\mathcal{P}) = n^2(1 - o(1))$  (the stability question) or, more ambitiously,  $T(\mathcal{P}) \geq \frac{n^2}{K}$  (the robustness question). The paper [51] only covers the extreme end of the stability region. It is possible that cubic structure is responsible for all such  $\mathcal{P}$ . There are links here to the theory of approximate groups: for example, finite approximate subgroups of elliptic curve groups are a source of examples of such sets  $\mathcal{P}$ .

An interesting nontrivial result in higher dimensions is [3], motivated by applications in theoretical computer science.

**4.3. The Littlewood Problem.** Suppose that  $A \subset \mathbb{Z}$  is a set of  $n$  integers. Then it was established 30 years ago by Konyagin [69] and McGehee-Pigno-Smith [77], answering a

---

<sup>8</sup>This condition is included here for simplicity, but can probably be relaxed.

question of Littlewood [75], that

$$\int_0^1 \left| \sum_{a \in A} e^{2\pi i \theta a} \right| d\theta \gg \log n.$$

Earlier results had been obtained by Paul Cohen and others. This is sharp up to the constant, as is shown by taking  $A$  to be an arithmetic progression of length  $n$ . (In fact, this example may also provide the sharp constant, a conjecture known as the *Strong Littlewood Conjecture*.) Very little is known about the robustness question, that is to say about the structure of those  $A$  for which

$$\int_0^1 \left| \sum_{a \in A} e^{2\pi i \theta a} \right| d\theta \leq K \log n.$$

It is possible that such  $A$  are very close to being unions of a few arithmetic progressions. If so, this would have applications to questions in combinatorial number theory about sum-free sets due to a connection established by Bourgain [6]. For some partial results and a further discussion, see [78].

**4.4. No-three-in-a-line.** Let  $p$  be an odd prime, and suppose that  $A \subset \text{PG}(2, p)$  is a set containing no three distinct points in a line.<sup>9</sup> (Here,  $\text{PG}(2, p)$  is the 2-dimensional projective space over  $\mathbb{F}_p$ , thus  $|\text{PG}(2, p)| = p^2 + p + 1$ ). It is very easy to see that  $|A| \leq p + 2$  and an exercise to show that  $|A| \leq p + 1$ . Equality occurs when  $A$  is a conic. Remarkably, a celebrated result of Segre [90] shows that in fact equality occurs *only* when  $A$  is a conic.

The stability question was resolved by Voloch [98], building upon remarkable work of Segre. Voloch shows that any  $A$  with no three-in-a-line and  $|A| \geq \frac{44}{45}p$  is contained in a conic. This argument is quite deep, depending on an application of the polynomial method [95] as well as bounds of Stöhr and Voloch [91] about counting points on high degree curves.

The robustness question, that is to say the classification of those  $A$  with  $|A| \geq \frac{1}{K}p$ , is very interesting. There are examples coming from cubic curves, such as  $A = \{(x : x^3 : 1) : 0 < x < p/3\}$ . So far as I am aware there is no example in the literature to contradict the possibility that all sets  $A \subset \text{PG}(2, p)$  with no-three-on-a-line and  $|A| \geq \frac{1}{K}p$  have all but  $o(p)$  of their points lying on a curve of degree at most 3. So far as I am aware no-one has explicitly conjectured this either, so perhaps I shall take this opportunity to do so.

There is a superficial link to a notorious problem of Dudeney [23] about whether there is a set  $A$  of  $2N$  points on the grid  $[N] \times [N]$  with no three in a line. There are many fewer colinear triples in  $[N] \times [N]$  than in  $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$  for  $p \sim N$ , however, so the study of sets  $A$  such as this is likely to be even harder than the problem discussed above. Nonetheless, the best-known examples (with  $|A| \sim 3N/2$ , see [57]) are given by very algebraic constructions. It seems likely that the answer to Dudeney’s question is negative.

**4.5. Sidon sets.** Suppose that  $A \subset [N]$  is a set with the property that all pairwise sums  $x + y$  with  $x, y \in A$  are distinct, apart from the obvious coincidences  $x + y = y + x$ . Such a set  $A$  is called a Sidon set. It is very easy to see that  $|A| \ll \sqrt{N}$ , and with more care (an argument of Erdős and Turán) one may show that  $|A| \leq (1 + o(1))\sqrt{N}$ . There are examples

---

<sup>9</sup>Such sets are called “arcs” in the literature, which is extremely extensive.

of Sidon sets  $A$  with  $|A| = (1 - o(1))\sqrt{N}$ , all constructed in a highly algebraic manner using finite fields. There are different variants due to Bose, Ruzsa and Singer. It is possible that the stability question (that is, the classification of those Sidon set  $A$  with  $|A| = (1 - o(1))\sqrt{N}$ ) has a satisfactory answer, but there is no obvious guess, based on the known examples, as to what it might be. The robustness question, that is to say the classification of those  $A$  with  $|A| \geq \frac{1}{K}\sqrt{N}$ , is of course even more difficult. A discussion of it was had on the blog of Tim Gowers [38]. In commenting on that discussion, Terence Tao raised the possibility that an answer to this question could lead to progress on a famous and old problem of Erdős, namely to determine if there is an additive basis  $\mathcal{A}$  of the natural numbers of order 2 (i.e.  $\mathcal{A} + \mathcal{A} = \mathbb{N}$ ) with an absolute bound on the number of representations of  $x$  as a sum of two elements of  $\mathcal{A}$ .

**4.6. Acknowledgements.** The author is supported by ERC Starting Grant number 274938 *Approximate algebraic structure and applications*. I would like to thank Sean Eberhard, Bryna Kra, Freddie Manners, Peter Sarnak and Terence Tao for comments on a draft of this article. I wish to thank the last of these for our extensive collaboration over the last decade, which has so far led to 30 joint papers. I also thank my other coauthors, three of whom are speaking at this congress. It is only because of these mathematicians that I have the opportunity to present these topics at the 2014 ICM.

## References

- [1] N. Alon, T. Kaufman, M. Krivelevich, S. Litsyn and D. Ron, *Testing low-degree polynomials over  $\text{GF}(2)$* , RANDOM-APPROX 2003, 188–199.
- [2] A. Balog and E. Szemerédi, *A statistical theorem of set addition*, Combinatorica **14** (1994), 263–268.
- [3] B. Barak, Z. Dvir, A. Yehudayoff and A. Wigderson, *Rank bounds for design matrices with applications to combinatorial geometry and locally correctable codes*, In Proceedings of the 43rd annual ACM symposium on Theory of computing, STOC '11, 519–528.
- [4] V. Bergelson, B. Host and B. Kra, *Multiple recurrence and nilsequences*, (with an appendix by I. Ruzsa), Invent. Math. **160**, no. 2 (2005), 261–303.
- [5] Y. Bilu, *Structure of sets with small sumset*, Structure Theory of Set Addition, Astérisque **258** (1999), 77–108.
- [6] J. Bourgain, *Estimates related to sumfree subsets of sets of integers*, Israel J. Math. **97** (1997), 71–92.
- [7] J. Bourgain and A. Gamburd, *Uniform expansion bounds for Cayley graphs of  $\text{SL}_2(\mathbb{F}_p)$* , Ann. Math. **167** (2008), no. 2, 625–642.
- [8] J. Bourgain, A. Gamburd and P. Sarnak, *Affine linear sieve, expanders and sum-product*, Invent. Math. **179** (2010), no. 3, 559–644.
- [9] J. Bourgain, A. Glibichuk and S. Konyagin, *Estimates for the number of sums and products and for exponential sums in fields of prime order*, J. London Math. Soc. **73** (2006), no. 2, 380–398.

- [10] J. Bourgain, N. Katz, and T. C. Tao. *A sum-product estimate in finite fields, and applications*, *Geom. Funct. Anal. (GAFA)* **14** (2004), no. 1, 27–57.
- [11] E. Breuillard, B. J. Green, and T. C. Tao, *Approximate subgroups of linear groups*, *Geom. Funct. Anal. (GAFA)* **21** (2011), no. 4, 774–819.
- [12] ———, *Suzuki groups as expanders*, *Groups Geom. Dyn.* **5** (2011), no. 2, 281–299.
- [13] ———, *The structure of approximate groups*, *Publ. Math. IHES* **116** (2012), no. 1, 115–221.
- [14] ———, *Small Doubling in Groups*, *Erdős Centennial, Bolyai Society Mathematical Studies* **25** (2013), 129–151.
- [15] E. Breuillard, B. J. Green, R. Guralnick and T. C. Tao, *Expansion in Finite Simple Groups of Lie Type*, to appear in *J. European Math. Soc.*
- [16] O. A. Camarena and B. Szegedy, *Nilspaces, nilmanifolds and their automorphisms*, 2010 preprint, <http://arxiv.org/abs/1009.3825>.
- [17] M.-C. Chang, *A polynomial bound in Freiman’s theorem*, *Duke Math. J.* **113** (2002), no. 3, 399–419.
- [18] ———, *Product theorems in  $SL_2$  and  $SL_3$* , *J. Inst. Math. Jussieu* **7** (2008), no. 1, 1–25.
- [19] G. Cooperman, *Towards a practical, theoretically sound algorithm for random generation in finite groups*, unpublished, arXiv:math/0205203.
- [20] J. P. Conze and E. Lesigne, *Sur un théorème ergodique pour des mesures diagonales*, *C. R. Acad. Sci. Paris* **306** (1988), 491–493.
- [21] E. Croot and O. Sisask, *A probabilistic technique for finding almost-periods of convolutions*, *Geom. Funct. Anal. (GAFA)* **20** (2010), no. 6, 1367–1396.
- [22] J. D. Dixon, *Generating random elements in finite groups*, *Electronic J. Comb.* **15** (2008), R94.
- [23] H. E. Dudeney, *Amusements in Mathematics*, Nelson, Edinburgh, 1917.
- [24] T. Eisner and T. C. Tao, *Large values of the Gowers-Host-Kra seminorms*, *J. Anal. Math.* **117** (2012), 133–186.
- [25] G. Elekes and Z. Király, *On the combinatorics of projective mappings*, *J. Algebraic Combin.* **14** (2001), no. 3, 183–197.
- [26] A. Figalli and D. Jerison, *Quantitative stability for sumsets in  $\mathbb{R}^n$* , 2013 preprint.
- [27] ———, *Quantitative stability for the Brunn-Minkowski inequality*, 2013 preprint.
- [28] J. J. F. Fournier, *Sharpness in Young’s inequality for convolution*, *Pacific J. Math.* **72** (1977), no. 2, 383–397.
- [29] G. A. Freiman, *Groups and the inverse problems of additive number theory*, In *Number-theoretic studies in the Markov spectrum and in the structural theory of set addition (Russian)*, 175–183, Kalinin. Gos. Univ., Moscow, 1973.
- [30] ———, *Foundations of a structural theory of set addition*, American Mathematical Society, Providence, R. I., 1973, Translated from the Russian, *Translations of Mathematical Monographs*, Vol. 37.
- [31] H. Furstenberg, *Recurrence in ergodic theory and combinatorial number theory*, Princeton University Press 1981.

- [32] ———, *Nonconventional ergodic averages*, The legacy of John von Neumann (Hempstead, NY, 1988), Proc. Sympos. Pure. Math. **50** (1990), 43–56, AMS (Providence, RI).
- [33] ———, *From the Erdős-Turán conjecture to ergodic theory – the contribution of combinatorial number theory to dynamics*, in Paul Erdős and his mathematics, I (Budapest, 1999), 261–277, Bolyai Soc. Math. Stud. **11**, János Bolyai Math. Soc., Budapest, 2002.
- [34] 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 Convergence in ergodic theory and probability (Columbus, OH, 1993), 193–227, Ohio State Univ. Math. Res. Inst. Publ., **5**, de Gruyter, Berlin, 1996.
- [35] W. T. Gowers, *A new proof of Szemerédi’s theorem for progressions of length four*, Geom. Funct. Anal. (GAFA) **8** (1998), no. 3, 529–551.
- [36] ———, *A new proof of Szemerédi’s theorem*, Geom. Funct. Anal. (GAFA) **11** (2001), no. 3, 465–588.
- [37] ———, *Quasirandom groups*, Combin. Probab. Comput. **17** (2008), no. 3, 363–387.
- [38] ———, *What are dense Sidon subsets of  $\{1, 2, \dots, n\}$  like?*, blog post. Available at [gowers.wordpress.com/2012/07/13/what-are-dense-sidon-subsets-of-12-n-like/](http://gowers.wordpress.com/2012/07/13/what-are-dense-sidon-subsets-of-12-n-like/).
- [39] B. J. Green, *Notes on the Polynomial Freiman-Ruzsa Conjecture*, unpublished. Available at [people.maths.ox.ac.uk/greenbj/papers/PFR.pdf](http://people.maths.ox.ac.uk/greenbj/papers/PFR.pdf).
- [40] ———, *Generalising the Hardy-Littlewood method for primes*, International Congress of Mathematicians. Vol. II, 373–399, Eur. Math. Soc., Zürich, 2006.
- [41] ———, *Approximate groups and their applications: work of Bourgain, Gamburd, Helfgott and Sarnak*, 25 pages, Current Events Bulletin of the AMS, 2010.
- [42] ———, *Barbados lecture notes*, 2010, Transcript available at <http://www.cs.mcgill.ca/denis/additive-lectures-v2.pdf>.
- [43] B. J. Green and A. J. Harper, *Inverse questions for the large sieve*, to appear in Geom. Funct. Anal. (GAFA).
- [44] B. J. Green and I. Z. Ruzsa, *Freiman’s theorem in an arbitrary abelian group*, J. Lond. Math. Soc. **75** (2007), no. 2, 163–175.
- [45] B. J. Green and T. C. Tao, *An inverse theorem for the Gowers  $U^3$ -norm, with applications*, Proc. Edinburgh Math. Soc. **51**, no. 1, 71–153.
- [46] ———, *Compressions, convex geometry and the Freiman-Bilu theorem*, Quart. J. Math (Oxford) **57** (2006), no. 4, 495–504.
- [47] ———, *Linear equations in primes*, Ann. Math **171** (2010), no. 3, 1753–1850.
- [48] ———, *An equivalence between inverse sumset theorems and inverse conjectures for the  $U^3$ -norm*, Math. Proc. Camb. Phil. Soc. **149** (2010), no. 1, 1–19.
- [49] ———, *The quantitative behaviour of polynomial orbits on nilmanifolds*, Ann. Math. **175** (2012), no. 2, 465–540.
- [50] ———, *The Möbius function is strongly orthogonal to nilsequences*, Ann. Math. **175** (2012), no. 2, 541–566.
- [51] ———, *On sets defining few ordinary lines*, Disc. Comp. Geom. **50** (2013), no. 2,

409–468.

- [52] B. J. Green, T. C. Tao and T. Ziegler, *An inverse theorem for the Gowers  $U^4[N]$  norm*, Glasgow Math. J. **53** (2011), 1–50.
- [53] B. J. Green, T. C. Tao and T. Ziegler, *An inverse theorem for the Gowers  $U^{s+1}[N]$ -norm*, announcement, Electronic Research Announcements **18** (2011), 69–90.
- [54] B. J. Green, T. C. Tao and T. Ziegler, *An inverse theorem for the Gowers  $U^{s+1}[N]$ -norm*, Ann. Math **176** (2012), no. 2, 1231–1372.
- [55] L. W. Green, *Spectra of nilflows*, Bull. Amer. Math. Soc. **67** (1961) 414–415.
- [56] M. Gromov, *Groups of polynomial growth and expanding maps*, Publ. Math. IHES **53** (1981), 53–73.
- [57] R. R. Hall, T. H. Jackson, A. Sudbery and K. Wild, *Some advances in the no-three-in-line problem*, J. Combinatorial Theory Ser. A **18** (1975), 336–341.
- [58] G. H. Hardy and J. E. Littlewood, *Some new properties of Fourier constants*, Math. Ann. **97** (1927), 159–209.
- [59] H. A. Helfgott, *Growth and generation in  $SL_2(\mathbb{Z}/p\mathbb{Z})$* , Ann. Math. **167** (2008), 601–623.
- [60] ———, *Growth in  $SL_3(\mathbb{Z}/p\mathbb{Z})$* , J. Eur. Math. Soc. **13** (2011), no. 3, 761–851.
- [61] H. A. Helfgott and A. Venkatesh, *How small must ill-distributed sets be?*, in Analytic Number Theory: Essays in honour of Klaus Roth, 224–234, Cambridge Univ. Press, Cambridge, 2009.
- [62] S. Hoory, N. Linial, and A. Wigderson, *Expander graphs and their applications*, Bull. Amer. Math. Soc. (N.S.) **43** (2006), no. 4, 439–561.
- [63] B. Host and B. Kra, *Nonconventional ergodic averages and nilmanifolds*, Ann. Math. **161** (2005), no. 1, 397–488.
- [64] B. Host and B. Kra, *Analysis of two step nilsequences*, Annales de l’institut Fourier **58** (2008), no. 5, 1407–1453.
- [65] E. Hrushovski, *Stable group theory and approximate subgroups*, J. Amer. Math. Soc. **25** (2012), no. 1, 189–243.
- [66] M. Kassabov, *Symmetric groups and expander graphs*, Invent. Math. **170** (2007), no. 2, 327–354.
- [67] M. Kassabov, A. Lubotzky and N. Nikolov, *Finite simple groups as expanders*, Proc. Nat. Acad. Sci. **103** (2006), no. 16, 6116–6119.
- [68] B. Kleiner, *A new proof of Gromov’s theorem on groups of polynomial growth*, J. Amer. Math. Soc. **23** (2010), no. 3, 815–829.
- [69] S. V. Konyagin, *On the Littlewood problem*, Izv. Akad. Nauk SSSR Ser. Mat. **45** (1981), no. 2, 243–265.
- [70] D. Kotschick, *What is a quasimorphism?*, Notices Amer. Math. Soc. **51** (2004), no. 2, 208–209.
- [71] B. Kra, *From combinatorics to ergodic theory and back again*, Proceedings of International Congress of Mathematicians, Madrid 2006, volume III, 57–76.
- [72] M. J. Larsen and R. Pink, *Finite subgroups of algebraic groups*, J. Amer. Math. Soc. **24** (2011), no. 4, 1105–1158.

- [73] A. Leibman, *Polynomial sequences in groups*, Journal of Algebra **201** (1998), 189–206.
- [74] ———, *Pointwise convergence of ergodic averages for polynomial sequences of translations on a nilmanifold*, Ergodic Th. Dyn. Systems **25** (2005), no. 1, 201–213.
- [75] J. E. Littlewood, *Some problems in real and complex analysis*, Heath Mathematical Monographs, D. C. Heath and co., Lexington, MA 1968.
- [76] S. Lovett, *Equivalence of polynomial conjectures in additive combinatorics*, Combinatorica **32** (2012), no. 5, 607–618.
- [77] O. C. McGehee, L. Pigno and B. Smith, *Hardy’s inequality and the  $L^1$ -norm of exponential sums*, Ann. Math. **113** (1981), no. 3, 613–618.
- [78] G. Petridis, *The  $L^1$ -norm of exponential sums in  $\mathbb{Z}^d$* , Math. Proc. Camb. Phil. Soc. **154** (2013), no. 3, 381–392.
- [79] L. Pyber and E. Szabó, *Growth in finite simple groups of Lie type of bounded rank*, preprint (2010), arXiv:1005.1858.
- [80] I. Z. Ruzsa, *Sums of finite sets*, Number theory (New York, 1991–1995), 281–293. Springer, New York, 1996.
- [81] ———, *An analog of Freiman’s theorem in groups*, in Structure Theorem of Set Addition, Astérisque **258** (1999) 323–326.
- [82] A. Salehi-Golsefidy and P. Sarnak, *Affine sieve*, J. Amer. Math. Soc. **26** (2013), no. 4, 1085–1105.
- [83] A. Samorodnitsky, *Low-degree tests at large distances*, STOC 2007.
- [84] T. Sanders, *On a nonabelian Balog-Szemerédi-type lemma*, J. Aust. Math. Soc. **89** (2010), no. 1, 127–132.
- [85] ———, *On the Bogolyubov-Ruzsa lemma*, Anal. PDE **5** (2012), no. 3, 627–655.
- [86] ———, *The structure theory of set addition revisited*, Bull. Amer. Math. Soc. (N.S.) **50** (2013), no. 1, 93–127.
- [87] P. Sarnak, *Equidistribution and primes*, available at <http://web.math.princeton.edu/sarnak/EquidPrimes.pdf>.
- [88] P. Sarnak and X. X. Xue, *Bounds for multiplicities of automorphic representations*, Duke Math. J. **64** (1991), no. 1, 207–227.
- [89] T. Schoen, *Near optimal bounds in Freiman’s theorem*, Duke Math. J. **158** (2011), no. 1, 1–12.
- [90] B. Segre, *Ovals in a finite projective plane*, Canadian J. Math. **7** (1955), 414–416.
- [91] K.-O.- Stöhr and J. F. Voloch, *Weierstrass points and curves over finite fields*, Proc. London Math. Soc. (3) **52** (1986), no. 1, 1–19.
- [92] J. Sylvester, Mathematical Question 2571, Educational Times, February 1868.
- [93] B. Szegedy, *On higher-order Fourier analysis*, 2012 preprint, <http://arxiv.org/abs/1203.2260>.
- [94] T. C. Tao, *Product set estimates for non-commutative groups*, Combinatorica **28** (2008), no. 5, 547–594.
- [95] ———, *Algebraic combinatorial geometry: the polynomial method in arithmetic combinatorics, incidence combinatorics, and number theory*, preprint 2013, <http://>

- arxiv.org/abs/1310.6482.
- [96] T. C. Tao and V. H. Vu, *Additive Combinatorics*, Cambridge Stud. Adv. Math. **105** (2006).
- [97] P. P. Varjú, *Expansion in  $SL_d(O_K/I)$ ,  $I$  square-free*, J. Eur. Math. Soc. **14** (2012), no. 1, 273–305.
- [98] J. F. Voloch, *Arcs in projective planes over prime fields*, J. Geom. **38** (1990), no. 1–2, 198–200.
- [99] M. N. Walsh, *The inverse sieve problem in high dimensions*, Duke Math. J. **161** (2012), no. 10, 2001–2022.
- [100] ———, *The algebraicity of ill-distributed sets*, preprint. Available at <http://arxiv.org/abs/1307.0259>.
- [101] W. H. Young, *On the multiplication of successions of Fourier constants*, Proc. Roy. Soc. London Ser. A **87** (1912), 331–339.
- [102] T. Ziegler, *Universal Characteristic Factors and Furstenberg Averages*, J. Amer. Math. Soc. **20** (2007), 53–97.

Mathematical Institute, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford OX2 6GG

E-mail: [ben.green@maths.ox.ac.uk](mailto:ben.green@maths.ox.ac.uk)



# Mori geometry meets Cartan geometry: Varieties of minimal rational tangents

Jun-Muk Hwang

**Abstract.** We give an introduction to the theory of varieties of minimal rational tangents, emphasizing its aspect as a fusion of algebraic geometry and differential geometry, more specifically, a fusion of Mori geometry of minimal rational curves and Cartan geometry of cone structures.

**Mathematics Subject Classification (2010).** Primary 14J40, 53B99; Secondary 14J45.

**Keywords.** Varieties of minimal rational tangents, uniruled projective manifolds, Cartan geometry, G-structures.

## 1. Introduction: a brief prehistory

Lines have been champion figures in classical geometry. Together with circles, they dominate the entire geometric contents of Euclid. Their dominance is no less strong in projective geometry. Classical projective geometry is full of fascinating results about intricate combinations of lines. As geometry entered the modern era, lines evolved into objects of greater flexibility and generality while retaining all the beauty and brilliance of classical lines. As Euclidean geometry developed into Riemannian geometry, for example, lines were replaced by geodesics which then inherited all the glory of Euclidean lines.

In the transition from classical projective geometry to complex projective geometry, real lines have been replaced by complex lines. Lines over complex numbers have all the power of lines in classical projective geometry and even more: results of greater elegance and harmony are obtained over complex numbers. A large number of results on lines and their interactions with other varieties have been obtained in complex projective geometry, their dazzling beauty no less impressive than that of classical geometry. But as complex projective geometry develops further into complex geometry and abstract algebraic geometry, which emphasize intrinsic properties of complex manifolds and abstract varieties, the notion of lines in projective space seems to be too limited for it to keep its leading role.

Firstly, to be useful in intrinsic geometry of projective varieties in projective space, lines should lie on the projective varieties. But most projective varieties do not contain lines. Even when a projective manifold contains lines, the locus of lines is, often, small and then such a locus is usually regarded as an exceptional part. Of course, there are many important varieties that are covered by lines, but they belong to a limited class from the general perspective of classification theory of varieties. In short,

*(\*) the class of projective manifolds covered by lines seems to be too special from the perspective of the general theory of complex manifolds or algebraic varieties.*

Secondly, many of the methods employed to use lines on varieties in projective space depend on the extrinsic geometry of ambient projective space. They do not truly belong to intrinsic geometry of the varieties. Such geometric arguments are undoubtedly useful in fathoming deeper geometric properties of varieties which are described explicitly, at least to some extent. But can such methods yield results on *a priori* unknown varieties, defined abstractly by intrinsic conditions? In short,

*(\*\*) tools employed in line geometry are not intrinsic enough to handle intrinsic problems on abstractly described varieties.*

These concerns show that lines in projective space have a rather limited role in the modern development of complex algebraic geometry. Is there a more general and more powerful notion in complex algebraic geometry that can replace the role of lines, as geodesics do in Riemannian geometry? No serious candidate had emerged until Mori's groundbreaking work [37].

In the celebrated paper [37], Mori shows that a large class of projective manifolds, including all Fano manifolds, are covered by certain intrinsically defined rational curves that behave like lines in many respects. Let us call these rational curves 'minimal rational curves'. If a projective manifold embedded in projective space is covered by lines, these lines are minimal rational curves of the projective manifold, so the notion of minimal rational curves can be viewed as an intrinsic generalization of lines.

The class of projective manifolds covered by minimal rational curves are called uniruled projective manifolds. Generalizing Mori's result, Miyaoka and Mori have proved in [32] that a projective manifold is uniruled if its anti-canonical bundle satisfies a certain positivity condition. This implies that uniruled projective manifolds form a large class of algebraic varieties. Furthermore, the minimal model program, a modern structure theory of higher-dimensional algebraic varieties, predicts that uniruled projective manifolds are precisely those projective manifolds that do not admit minimal models. Thus

*projective manifolds covered by minimal rational curves form a distinguished class of manifolds, worthy of independent study from the view-point of classification theory of general projective varieties, and at the same time, large enough to contain examples of great diversity.*

This overcomes the limitation (\*) of the class of projective manifolds covered by lines.

Furthermore, Mori's work exhibits how to use minimal rational curves in an intrinsic way to obtain geometric information on uniruled projective manifolds. The main tool here is the deformation theory of curves, a machinery of modern complex algebraic geometry—somewhat reminiscent of the use of variational calculus in the local study of geodesics in Riemannian geometry. An example is the property that a minimal rational curve cannot be deformed when two distinct points on the curve are fixed. This result generalizes the fundamental postulate of classical geometry that "two points determine one line". The important point is that such a classical property of lines can be recovered by modern deformation theory in an abstract setting.

*Deformation theory of rational curves is a powerful technique applicable to conceptual problems on varieties defined in abstract intrinsic terms.*

In [37], Mori, in fact, has resolved one of the toughest problems of this kind, the Hartshorne conjecture, characterizing projective space by the positivity of the tangent bundle. The methods employed in the theory of minimal rational curves are certainly free of the concern (\*\*) on the tools of line geometry.

These considerations indicate that minimal rational curves can serve as the natural generalization of lines, overcoming the limitations of lines, while inheriting their powerful and elegant features.

Our main interest is the geometry of minimal rational curves in uniruled projective manifolds. As in Mori's work, we would like to see how minimal rational curves can be used to control the intrinsic geometry of uniruled projective manifolds. One guiding problem is the following question on recognizing a given uniruled projective manifolds by minimal rational curves.

**Problem 1.1.** Let  $S$  be a (well-known) uniruled projective manifold. Given another uniruled projective manifold  $X$ , what properties of minimal rational curves on  $X$  guarantee that  $X$  is biregular (i.e. isomorphic as abstract algebraic varieties) to  $S$ ?

Here, the setting of the problem is algebraic geometry and by properties of minimal rational curves, we mean algebro-geometric properties. When  $S$  is projective space, a version of this problem is precisely what Mori solved in [37]. The initial goal of [37] was to prove the Hartshorne conjecture, which characterizes projective space by certain positivity property of the tangent bundle. After showing that the projective manifold in question is uniruled, Mori used the tangent directions of minimal rational curves to finish the proof. This part of Mori's proof has been greatly strengthened by the later work of Cho-Miyaoka-Shepherd-Barron [4], which says roughly the following (see Theorem 3.16 for a precise statement).

**Theorem 1.2.** *Suppose for a general point  $x$  on a uniruled projective manifold  $X$  and a general tangent direction  $\alpha \in \mathbb{P}T_x(X)$ , there exists a minimal rational curve through  $x$  tangent to  $\alpha$ . Then  $X$  is projective space.*

This is a very satisfactory answer to Problem 1.1 when  $S$  is projective space. It includes, as special cases, many previously known characterizations of projective space. One may wonder why the condition on minimal rational curves here is formulated in terms of their tangent directions, not in terms of some other properties of minimal rational curves. The essential reason is because the main technical tool to handle minimal rational curves is the deformation theory of curves, as mentioned before. The tangential information of curves is essential in deformation theory. For this reason, it is natural and also useful to give conditions in terms of tangent directions of minimal rational curves.

What about other uniruled projective manifolds? When  $S$  is different from projective space, Theorem 1.2 says that minimal rational curves on  $S$  exist only in some distinguished directions. Thus in the setting of Problem 1.1, it is natural to consider

*the subvariety  $\mathcal{C}_s \subset \mathbb{P}T_s(S)$  consisting of the directions of minimal rational curves through  $s \in S$  and the corresponding subvariety  $\mathcal{C}_x \subset \mathbb{P}T_x(X)$ .*

When  $S$  is projective space, we have  $\mathcal{C}_s = \mathbb{P}T_s(S)$  for any  $s \in S$ . Theorem 1.2 says that a uniruled projective manifold  $X$  is projective space if and only if  $\mathcal{C}_x = \mathbb{P}T_x(X)$  at some point  $x \in X$ . In other words, if a uniruled projective manifold has the same type of  $\mathcal{C}_x$  as projective space, then it is projective space. Based on this observation, we can refine our guiding Problem 1.1 as follows.

**Problem 1.3.** Let  $S$  be a (well-known) uniruled projective manifold. Given another uniruled projective manifold  $X$ , what properties of  $\mathcal{C}_x \subset \mathbb{P}T_x(X)$  for general points  $x \in X$  guarantee that  $X$  is biregular to  $S$ ?

Comparing this with Theorem 1.2, one may wonder why we are asking for information on  $\mathcal{C}_x$  for general points  $x \in X$ , instead of a single point  $x \in X$  as in Theorem 1.2. This is because the information at one point  $x \in X$  seems to be too weak to characterize  $X$  when  $\mathcal{C}_x \neq \mathbb{P}T_x(X)$ . The equality  $\mathcal{C}_x = \mathbb{P}T_x(X)$  implies that minimal rational curves through one point  $x$  cover the whole of  $X$ . This is why in Theorem 1.2 the information at one point is sufficient to control the whole of  $X$ . If  $\mathcal{C}_x \neq \mathbb{P}T_x(X)$ , minimal rational curves through one point  $x$  cover only small part of  $X$ . Besides, the subvariety  $\mathcal{C}_s \subset \mathbb{P}T_s(S)$  may change as the point  $s \in S$  varies and so the expected condition is not just on  $\mathcal{C}_x$  for a single  $x$ , but on the family

$$\{\mathcal{C}_x \subset \mathbb{P}T_x(X), \text{ general } x \in X\}.$$

This is why we are asking for the data  $\mathcal{C}_x$  for all general  $x \in X$ .

Now as in Problem 1.1, the properties of  $\mathcal{C}_x$  that we are looking for in Problem 1.3 are algebro-geometric properties. In algebraic geometry, however, to use properties of such a family of varieties to control the whole of  $X$ , we usually need to have good information not only on general members of the family, but also on the potential degeneration of the family. Thus it may look more reasonable to require, in Problem 1.3, some additional properties on the behavior of the family  $\mathcal{C}_x$  under degeneration. But such additional conditions would diminish the true interest of Problem 1.3. This is because in the context of intrinsic geometry of uniruled manifolds, the properties of  $\mathcal{C}_x$  we are looking for should be checkable by deformation theory of curves. Deformation theory of rational curves works well at general points of nonsingular varieties, but not so at special points. Thus it is important to find conditions for  $\mathcal{C}_x$  only for general  $x \in X$  in Problem 1.3. But then controlling the whole of  $X$  using the algebraic behavior of  $\{\mathcal{C}_x \subset \mathbb{P}T_x(X), \text{ general } x \in X\}$  becomes a serious issue.

This was exactly the issue puzzling me when I first encountered a version of Problem 1.3 about twenty years ago. At that time, I was working on the deformation rigidity of Hermitian symmetric spaces in the setting of algebraic geometry. I refer the reader to [11] for the details on this rigidity problem. Here it suffices to say that the deformation rigidity of Hermitian symmetric spaces was a question originated from Kodaira-Spencer's work in 1950's and the question itself did not involve rational curves. I was trying to attack this question employing Mori's approach of minimal rational curves, which naturally led to a version of Problem 1.3 when  $S$  is an irreducible Hermitian symmetric space. In the setting of this rigidity question, I could derive a certain amount of algebro-geometric information on  $\mathcal{C}_x$  for general  $x \in X$ , but I was unable to figure out how to proceed from there, essentially because of the above difficulty, that it is hard to control the whole of  $X$  by algebro-geometric information on  $\mathcal{C}_x$  for general  $x \in X$ .

There was one hope. A few years earlier, Ngaiming Mok had overcome an obstacle of a similar kind in [33]. In that work, Mok solved what is called the generalized Frankel conjecture, which asks for a characterization of Hermitian symmetric spaces among Kähler manifolds in terms of a curvature condition. The Frankel conjecture itself is the Kähler version of the Hartshorne conjecture and was settled by Siu-Yau [40] around the time Mori solved the Hartshorne conjecture. Since the method used by Siu and Yau was rather restrictive, Mok naturally took the approach of Mori and encountered a situation similar to Problem 1.3. Now in his situation, there is a Riemannian metric on  $X$  and Mok could relate

$\mathcal{C}_x$  to a suitably deformed Riemannian metric. This enables him to show that  $X$  is Hermitian symmetric space using Berger's work on Riemannian holonomy. Roughly speaking, in [33]

*the difficulty in Problem 1.3 was overcome by relating  $\mathcal{C}_x$  to a Riemannian structure.*

This shows that differential geometry can be a recourse for Problem 1.3 when  $S$  is a Hermitian symmetric space. Indeed, compared with tools in algebraic geometry, methods of differential geometry tend to be more effective when the available data are only at general points of a manifold. Motivated by this, I tried to imitate Mok's argument in the setting of the deformation rigidity problem. However, the nature of the deformation rigidity problem is purely algebro-geometric and it is very hard to relate it to Riemannian structures. As a matter of fact, there had already been some unsuccessful attempts in 1960's to use Riemannian structures for the deformation rigidity question.

This was precisely the problem I was agonizing over when I attended my first ICM: Zürich 1994. Having come to the congress just to have fun listening to the new developments in mathematics, I found that Mok was there as a speaker and managed to have a chat with him. When I told him about the above difficulty in applying the approach of [33] to the deformation rigidity problem, he gave me an enlightening comment: besides the Riemannian metric, there is another differential geometric structure, a certain holomorphic G-structure, which can be used to characterize a Hermitian symmetric space. His suggestion was that one might be able to construct these G-structures using the information on  $\mathcal{C}_x$  for general  $x \in X$  and from this to recover Hermitian symmetric spaces. This suggestion looked promising because algebro-geometric data are closer to holomorphic structures than Riemannian structures.

Soon after the congress, I started looking into G-structures. I realized that there is a far-reaching generalization of Riemannian structures by Elie Cartan and the G-structures modeled on Hermitian symmetric spaces are special examples of Cartan's general theory of geometric structures. To recover these G-structures, it was necessary to investigate the geometry of  $\mathcal{C}_x$ 's in depth in the setting of the deformation rigidity problem. I had subsequent communications with Mok, and we started working together on this problem. Our collaboration was successful, leading to a solution of the deformation rigidity problem in [18]. But the most exciting point in our work was not the deformation rigidity itself. As mentioned, the essential part of [18] is to construct on  $X$  the G-structures modeled on Hermitian symmetric spaces. It turns out that a crucial point of this construction lies in a study of the behavior of  $\mathcal{C}_x$ 's not just as a family of projective algebraic varieties, but as data imposed on the tangent bundle of an open subset of  $X$ . In other words, we had to treat these  $\mathcal{C}_x$ 's as if

*the union of  $\mathcal{C}_x$ 's for general  $x \in X$  is a differential geometric structure.*

And why not? Such a family of subvarieties in  $\mathbb{P}T_x(X)$  is a legitimate example of Cartan's general geometric structures! So what happened can be summarized as follows. Initially we had been trying to relate  $\mathcal{C}_x$ 's to some differential geometric structures. These differential geometric structures were Riemannian structures in [33] and then G-structures in [18]. But actually, they have been there all along, namely,  $\mathcal{C}_x$ 's themselves!

Now once we accept  $\mathcal{C}_x$ 's as a differential geometric structure, there is no need to restrict ourselves to Hermitian symmetric spaces. This geometric structure exists for any uniruled projective manifold  $S$  and its minimal rational curves! This was an epiphany for me. We realized that the variety  $\mathcal{C}_x$  deserves a name of its own and endowed it with the appellation, somewhat uncharming, 'variety of minimal rational tangents'.

Realizing the varieties of minimal rational tangents as geometric structures opens up an approach to Problem 1.3 via Cartan geometry. In fact, Mok and I were able to show in [20] (see Theorem 4.6 for a precise statement)

**Theorem 1.4.** *Assume that  $S$  is a fixed uniruled projective manifold with  $b_2(S) = 1$  and  $\mathcal{C}_s$  at a general point  $s \in S$  is a smooth irreducible variety of positive dimension. If  $X$  is a uniruled projective manifold with  $b_2(X) = 1$  and the differential geometric structures defined by  $\mathcal{C}_s$ 's and  $\mathcal{C}_x$ 's are locally equivalent in the sense of Cartan, then  $S$  and  $X$  are biregular.*

This means that for a large and interesting class of uniruled projective manifolds, Problem 1.3 can be solved by studying the Cartan geometry of the structures defined by  $\mathcal{C}_x$ 's. By Theorem 1.4, the essence of Problem 1.3 has become

*searching for algebro-geometric properties of varieties of minimal rational tangents which make it possible to control the Cartan geometry of the geometric structures defined by them.*

As we will see in Section 4, this search has been successful in a number of cases and Problem 1.3 has been answered for some uniruled projective manifolds, including irreducible Hermitian symmetric spaces.

Since [18], the theory of varieties of minimal rational tangents has seen exciting developments and has found a wide range of applications in algebraic geometry. For surveys on these developments and applications, we refer the reader to [9, 19, 27, 36]. The purpose of this article is to give an introduction to one special aspect of the theory, the development centered around Problem 1.3. This is a special aspect, because many results on varieties of minimal rational tangents and their applications are not directly related to it. Yet, this is the most fascinating aspect: it offers an area for a fusion of algebraic geometry and differential geometry, more specifically, a fusion of Mori geometry of minimal rational curves and Cartan geometry of cone structures. We will stick to the core of this aspect and will not go into the diverse issues arising from it. Those interested in further directions of explorations may find my MSRI article [13] useful.

**Conventions.** We will work over the complex numbers and all our objects are holomorphic. Open sets refer to the Euclidean topology, unless otherwise stated. All manifolds are connected. A projective manifold is a smooth irreducible projective variety. A variety is a complex analytic set which is not necessarily irreducible, but has finitely many irreducible components. A general point of a manifold or an irreducible variety means a point in a dense open subset.

## 2. Cartan geometry: Cone structures

*A priori*, this section is about local differential geometry and has nothing to do with rational curves. We will introduce a class of geometric structures, cone structures, and some related notions. In a simpler form, cone structures have already appeared in twistor theory (see [31]), but as they are not widely known, I will try to give a detailed introduction.

**Definition 2.1.** For a complex manifold  $M$ , let  $\pi : \mathbb{P}T(M) \rightarrow M$  be the projectivized tangent bundle. A *smooth cone structure* on  $M$  is a closed nonsingular subvariety  $\mathcal{C} \subset$

$\mathbb{P}T(M)$  such that all components of  $\mathcal{C}$  have the same dimension and the restriction  $\varpi := \pi|_{\mathcal{C}}$  is a submersion.

We may restrict our discussion to smooth cone structures. Understanding the geometry of smooth cone structures is already challenging and lots of examples of smooth cone structures remain uninvestigated. To have a satisfactory general theory, however, we need to allow certain singularity in  $\mathcal{C}$ . This necessitates the following somewhat technical definition. (Readers not familiar with singularities may skip this definition and just stick to Definition 2.1, regarding  $\nu : \tilde{\mathcal{C}} \rightarrow \mathcal{C}$  as an identity map in the subsequent discussion.)

**Definition 2.2.** A cone structure on a complex manifold  $M$  is a closed subvariety  $\mathcal{C} \subset \mathbb{P}T(M)$  the normalization  $\nu : \tilde{\mathcal{C}} \rightarrow \mathcal{C}$  of which satisfies the following conditions.

- (1) All components of  $\tilde{\mathcal{C}}$  are smooth and have the same dimension.
- (2) The composition  $\varpi := \pi \circ \nu : \tilde{\mathcal{C}} \rightarrow M$  is a submersion. In particular, the relative tangent bundle  $T^\varpi \subset T(\tilde{\mathcal{C}})$  is a vector subbundle.
- (3) There is a vector subbundle  $\mathcal{T} \subset T(\tilde{\mathcal{C}})$  with  $T^\varpi \subset \mathcal{T}$  and  $\text{rank}(\mathcal{T}) = \text{rank}(T^\varpi) + 1$ , such that for any  $\alpha \in \tilde{\mathcal{C}}$  and any  $v \in \mathcal{T}_\alpha \setminus T^\varpi_\alpha$ , the nonzero vector  $d\varpi_\alpha(v) \in T_{\varpi(\alpha)}(M)$  satisfies

$$[d\varpi_\alpha(v)] = \nu(\alpha) \text{ as elements of } \mathbb{P}T_{\varpi(\alpha)}(M).$$

The conditions (1) and (2) say that  $\mathcal{C}$  is allowed to be singular but it becomes smooth after normalization and the natural projection to  $M$  becomes a submersion. Note that on  $\mathbb{P}T(M)$ , we have the tautological line bundle  $\xi \subset \pi^*T(M)$ . The condition (3) says that the quotient line bundle  $\mathcal{T}/T^\varpi$  is naturally isomorphic to  $\nu^*\xi$ . Another useful interpretation of the condition (3) is in terms of the following

**Definition 2.3.** Given a cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$ , let  $\text{Sm}(\mathcal{C}) \subset \mathcal{C}$  be the maximal dense open subset such that

$$\pi|_{\text{Sm}(\mathcal{C})} : \text{Sm}(\mathcal{C}) \rightarrow \pi(\text{Sm}(\mathcal{C}))$$

is a submersion.

Denote by  $\mathcal{T}^{\mathbb{P}T(M)} \subset T(\mathbb{P}T(M))$  the inverse image of the tautological bundle  $\xi \subset \pi^*T(M)$  under  $d\pi : T(\mathbb{P}T(M)) \rightarrow \pi^*T(M)$ . Then the condition (3) means that the vector bundle  $\mathcal{T}^{\mathbb{P}T(M)} \cap T(\text{Sm}(\mathcal{C}))$  on  $\text{Sm}(\mathcal{C})$ , after pulling back to  $\tilde{\mathcal{C}}$  by  $\nu$ , extends to a vector subbundle of  $T(\tilde{\mathcal{C}})$ . From this interpretation of (3), it is easy to see that

**Proposition 2.4.** A cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$  is a smooth cone structure if and only if  $\mathcal{C}$  is normal, i.e., the normalization  $\nu : \tilde{\mathcal{C}} \rightarrow \mathcal{C}$  is biholomorphic.

All three conditions (1)-(3) for cone structures are of local nature on  $M$ . This implies

**Proposition 2.5.** Given a cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$  and a connected open subset  $U \subset M$ , the restriction

$$\mathcal{C}|_U := \mathcal{C} \cap \mathbb{P}T(U) \subset \mathbb{P}T(U)$$

is a cone structure on the complex manifold  $U$ .

By Proposition 2.5, we can view a cone structure as a geometric structure on  $M$ . We are interested in Cartan geometry of cone structures. In particular, isomorphisms in cone structures are given by the following

**Definition 2.6.** A cone structures  $\mathcal{C} \subset \mathbb{P}T(M)$  on a complex manifold  $M$  is *equivalent* to a cone structure  $\mathcal{C}' \subset \mathbb{P}T(M')$  on a complex manifold  $M'$  if there exists a biholomorphic map  $\varphi : M \rightarrow M'$  such that the projective bundle isomorphism  $\mathbb{P}d\varphi : \mathbb{P}T(M) \rightarrow \mathbb{P}T(M')$  induced by the differential  $d\varphi : T(M) \rightarrow T(M')$  of  $\varphi$  satisfies  $\mathbb{P}d\varphi(\mathcal{C}) = \mathcal{C}'$ .

It is convenient to have a localized version of this:

**Definition 2.7.** For a cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$  (resp.  $\mathcal{C}' \subset \mathbb{P}T(M')$ ) and a point  $x \in M$  (resp.  $x' \in M'$ ), we say that  $\mathcal{C}$  at  $x$  is *equivalent* to  $\mathcal{C}'$  at  $x'$  if there exists a neighborhood  $U \subset M$  of  $x$  and a neighborhood  $U' \subset M'$  of  $x'$  such that the restriction  $\mathcal{C}|_U$  is equivalent to  $\mathcal{C}'|_{U'}$  as cone structures. We say that  $\mathcal{C}$  is *locally equivalent* to  $\mathcal{C}'$  if there are points  $x \in M$  and  $x' \in M'$  such that  $\mathcal{C}$  at  $x$  is equivalent to  $\mathcal{C}'$  at  $x'$ .

Let us give one simple example of a cone structure. Let  $V$  be a vector space and  $Z \subset \mathbb{P}V$  be a projective variety all components of which have the same dimension such that the normalization  $\tilde{Z}$  is nonsingular. Via the canonical isomorphism  $T(V) = V \times V$ , the projectivized tangent bundle  $\mathbb{P}T(V) = V \times \mathbb{P}V$  contains the subvariety  $\mathcal{C} := V \times Z \subset \mathbb{P}T(V)$ . This is a cone structure. Indeed the normalization  $\tilde{\mathcal{C}}$  is just  $V \times \tilde{Z}$  which is smooth and  $\varpi : \tilde{\mathcal{C}} \rightarrow V$  is just the projection  $V \times \tilde{Z} \rightarrow V$  which is a submersion, verifying the conditions (1) and (2) of Definition 2.2. The tautological line bundle of  $Z \subset \mathbb{P}V$  induces a line bundle  $\chi$  in  $T(V \times \tilde{Z})$  via the normalization morphism  $\tilde{Z} \rightarrow Z$  and the subbundle  $\mathcal{T} = T^\varpi + \chi$  of  $T(\tilde{\mathcal{C}})$  satisfies the condition (3).

**Definition 2.8.** The cone structure  $V \times Z \subset \mathbb{P}T(V)$  on  $V$  defined above is called the *flat cone structure with a fiber*  $Z \subset \mathbb{P}V$ . We will denote it by  $\text{Flat}_V^Z \subset \mathbb{P}T(V)$ . A cone structure on a complex manifold  $M$  is *locally flat* if it is locally equivalent to  $\text{Flat}_V^Z$  for some  $Z \subset \mathbb{P}V$  with  $\dim V = \dim M$ .

**Definition 2.9.** Let  $Z \subset \mathbb{P}V$  be a projective variety. A cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$  is *Z-isotrivial* if for a general  $x \in M$ , the fiber

$$\mathcal{C}_x = \mathcal{C} \cap \mathbb{P}T_x(M) \subset \mathbb{P}T_x(M)$$

is isomorphic to  $Z \subset \mathbb{P}V$  as a projective variety, i.e., a suitable linear isomorphism  $T_x(M) \rightarrow V$  sends  $\mathcal{C}_x$  to  $Z$ . A cone structure is *isotrivial* if it is  $Z$ -isotrivial for some  $Z$ .

A locally flat cone structure is isotrivial. But an isotrivial cone structure needs not be locally flat. Some isotrivial smooth cone structures are very familiar objects in differential geometry. When  $Z \subset \mathbb{P}V$  is a linear subspace of dimension  $p$ , a  $Z$ -isotrivial cone structure on  $M$  is just a Pfaffian system of rank  $p + 1$  on  $M$ . It is locally flat if and only if the Pfaffian system is involutive, i.e., it comes from a foliation. When  $Z \subset \mathbb{P}V$  is a nonsingular quadric hypersurface, a  $Z$ -isotrivial cone structure is a conformal structure on  $M$ . It is locally flat if and only if it is locally conformally flat. A natural generalization of the conformal structure is the cone structure modeled on an irreducible Hermitian symmetric space  $S = G/P$ . The isotropy action of  $P$  on the tangent space  $T_o(S)$  at the base point  $o \in S$  has a unique closed orbit  $\mathcal{C}_o \subset \mathbb{P}T_o(S)$ . A  $Z$ -isotrivial cone structure where  $Z \subset \mathbb{P}V$  is isomorphic to  $\mathcal{C}_o \subset \mathbb{P}T_o(S)$  is called an *almost S-structure*. A conformal structure is exactly an almost  $S$ -structure where  $S$  is a nonsingular quadric hypersurface, equivalently, an irreducible Hermitian symmetric space of type IV. The natural almost  $S$ -structure  $\mathcal{C} \subset \mathbb{P}T(S)$  given by the translate of  $\mathcal{C}_o$  by  $G$ -action is locally flat, which can be seen by Harish-Chandra coordinates

of irreducible Hermitian symmetric spaces (see Section (1.2) in [35] for a presentation in terms of explicit coordinates for Grassmannians). The  $G$ -structure on an irreducible Hermitian symmetric space  $S$  referred to in Section 1 is essentially equal to the cone structure  $\mathcal{C} \subset \mathbb{P}T(S)$ .

How do we check the local equivalence of two cone structures? A general method of checking equivalence of geometric structures has been formulated by Elie Cartan [2]. The fundamental apparatus in Cartan’s method is a coframe.

**Definition 2.10.** Let  $V$  be a vector space and let  $M$  be a complex manifold with  $\dim V = \dim M$ . A *coframe* on  $M$  is a trivialization  $\omega : T(M) \rightarrow M \times V$ , equivalently, a  $V$ -valued 1-form on  $M$  such that  $\omega_x : T_x(M) \rightarrow V$  is an isomorphism for each  $x \in M$ . We will denote by  $\mathbb{P}\omega : \mathbb{P}T(M) \rightarrow M \times \mathbb{P}V$  the trivialization of the projectivized tangent bundle induced by  $\omega$ . Given a coframe, there exists a  $\text{Hom}(\wedge^2 V, V)$ -valued function  $\sigma^\omega$  on  $M$ , called the *structure function* of  $\omega$ , such that

$$d\omega = \sigma^\omega(\omega \wedge \omega).$$

A coframe is *closed* if  $d\omega = 0$ , i.e., the structure function  $\sigma^\omega$  is identically zero. A coframe is *conformally closed* if there exists a holomorphic function  $f$  on an open subset  $U \subset M$  such that  $f\omega$  is closed on  $U$ .

The following is a simple consequence of the Poincaré lemma (see Theorem 3.4 in [12]).

**Proposition 2.11.** *Let  $V^\vee \subset \text{Hom}(\wedge^2 V, V)$  be the natural inclusion of the dual space of  $V$  given by contracting with one factor. When  $\dim M \geq 3$ , a coframe  $\omega$  is conformally closed if and only if  $\sigma^\omega$  takes values in  $V^\vee$ .*

Although Cartan’s method is applicable to the equivalence problem for arbitrary cone structures, its actual implementation can be challenging, depending on the type of the cone structure. For isotrivial cone structures, however, this becomes simple:

**Definition 2.12.** Let  $\mathcal{C} \subset \mathbb{P}T(M)$  be a  $Z$ -isotrivial cone structure for a projective variety  $Z \subset \mathbb{P}V$ . A coframe  $\omega : T(M) \rightarrow M \times V$  is *adapted* to the cone structure if  $\mathbb{P}\omega(\mathcal{C}) = Z$ .

**Proposition 2.13.** *An isotrivial cone structure is locally flat if and only if after restricting to an open subset, it admits a conformally closed adapted coframe.*

Since an isotrivial cone structure always admits an adapted coframe, we can use Proposition 2.11 and Proposition 2.13 to check the local flatness of an isotrivial cone structure. One difficulty here is that there may be several different adapted coframes, so we need to choose the right one. Different choices of adapted coframes are related by the linear automorphism group of the fiber. Let us elaborate this point.

For a projective variety  $Z \subset \mathbb{P}V$ , let  $\widehat{Z} \subset V$  be its homogeneous cone. Denote by  $\text{Aut}(\widehat{Z}) \subset \text{GL}(V)$  the linear automorphism group of  $\widehat{Z}$  and by  $\text{aut}(\widehat{Z}) \subset \mathfrak{gl}(V)$  its Lie algebra. Since  $\widehat{Z} \subset V$  is a cone, the Lie algebra  $\text{aut}(\widehat{Z})$  always contains the scalars  $\mathbb{C}$ . When  $\text{aut}(\widehat{Z}) = \mathbb{C}$ , a  $Z$ -isotrivial cone structure has a unique adapted coframe up to multiplication by functions. Consequently, the method of Proposition 2.13 essentially determines the local flatness of  $Z$ -isotrivial cone structures when  $\text{aut}(\widehat{Z}) = \mathbb{C}$ .

When  $\text{aut}(\widehat{Z}) \neq \mathbb{C}$ , however, compositions with  $\text{Aut}(\widehat{Z})$ -valued functions give rise to many different choices of adapted coframes for a  $Z$ -isotrivial cone structure. In this case,

Proposition 2.13 is not decisive and we have to consider the problem of choosing the right coframe. This leads to the equivalence problem for G-structures where G corresponds to the group  $\text{Aut}(\widehat{Z}) \subset \text{GL}(V)$ . The general theory of G-structures has been developed by many mathematicians. In particular, for the G-structures modeled on Hermitian symmetric spaces, [5] and [39] provide a calculable criterion for local flatness in terms of the vanishing of certain ‘curvature tensors’, which are more elaborate version of the structure functions  $\sigma^\omega$ .

It turns out that the cone structures we are interested in are equipped with some additional structures.

**Definition 2.14.** Let  $\mathcal{C} \subset \mathbb{P}T(M)$  be a cone structure. From the condition (3) in Definition 2.2, we have an exact sequence of vector bundles on  $\widetilde{\mathcal{C}}$

$$0 \rightarrow T^\omega \rightarrow \mathcal{T} \rightarrow \nu^*\xi \rightarrow 0.$$

A line subbundle  $\mathcal{F} \subset \mathcal{T}$  is called a *connection* of the cone structure if  $\mathcal{F}$  splits this exact sequence. Thus a connection exists if and only if this exact sequence splits.

All the cone structures we are to meet have certain canonically defined connections. These connections will have some special properties.

**Definition 2.15.** In Definition 2.14,  $\mathcal{F}$  is a *characteristic connection* if  $[\mathcal{F}, [\mathcal{F}, \mathcal{T}]] \subset [\mathcal{F}, \mathcal{T}]$  at general points of  $\widetilde{\mathcal{C}}$ . The inclusion means that for any local section  $f$  of  $\mathcal{F}$  and any local section  $v$  of  $T^\omega$  regarded as local vector fields in some open subset of  $\widetilde{\mathcal{C}}$ , the Lie bracket  $[f, [f, v]]$  is a local section of  $[\mathcal{F}, \mathcal{T}]$ .

The most important property of a characteristic connection is its uniqueness for a large class of cone structures. This condition is formulated in terms of the Gauss map and the projective second fundamental form. Let us recall the definition.

**Definition 2.16.** Let  $Z \subset \mathbb{P}V$  be an irreducible projective variety of dimension  $p$ . The *Gauss map* of  $Z$  is the morphism  $\gamma : \text{Sm}(Z) \rightarrow \text{Gr}(p + 1, V)$  defined on the smooth locus of  $Z$  by associating to a smooth point  $\alpha$  of  $Z$  the affine tangent space  $T_\alpha(\widehat{Z}) \subset V$ , the tangent space of the homogeneous cone  $\widehat{Z} \subset V$  along  $\alpha$ . We say that the Gauss map of  $Z$  is *nondegenerate* if  $\gamma$  is generically finite over its image. Let  $\alpha \in \text{Sm}(Z)$  be a smooth point of  $Z$  and let  $N_{Z,\alpha}$  be the normal space of  $Z$  inside  $\mathbb{P}V$  at  $\alpha$ . The differential of  $\gamma$  defines a homomorphism

$$\text{II}_{Z,\alpha} : \text{Sym}^2 T_\alpha(Z) \rightarrow N_{Z,\alpha},$$

called the *projective second fundamental form*. We say that  $\text{II}_{Z,\alpha}$  is *nondegenerate* if its null space

$$\text{Null}_{\text{II}_{Z,\alpha}} = \{v \in T_\alpha(Z), \text{II}_{N_{Z,\alpha}}(v, u) = 0 \text{ for all } u \in T_\alpha(Z)\}$$

is zero. Then the Gauss map of  $Z$  is nondegenerate if the projective second fundamental form of  $Z$  is nondegenerate at a general point of  $Z$ . It is well-known (e.g. Theorem 3.4.2 in [25]) that if an irreducible projective variety  $Z$  is smooth and not a linear subspace of  $\mathbb{P}V$ , then the Gauss map of  $Z$  is nondegenerate, or equivalently, the projective second fundamental form  $\text{II}_{Z,\alpha}$  is nondegenerate at a general point  $\alpha \in Z$ .

The uniqueness result for a characteristic connection is the following result from [21].

**Theorem 2.17.** *Let  $\mathcal{C} \subset \mathbb{P}T(M)$  be a cone structure such that all components of the fiber  $\mathcal{C}_x$  for a general  $x \in M$  have nondegenerate Gauss maps. Then  $\mathcal{C}$  has at most one characteristic connection.*

It is easy to see that the flat cone structure  $\text{Flat}_V^Z$  in Definition 2.8 has a characteristic connection given by the intersection of  $\mathcal{T}$  with the fibers of the projection map  $V \times \tilde{Z} \rightarrow \tilde{Z}$ . Cone structures admitting characteristic connections have certain amount of flatness, although this is not easy to explicate. One manifestation is the following proposition (see Theorem 6.2 in [13] for a proof). Although it is stated here without any regard to minimal rational curves, a version of this proposition is first discovered in [18] for varieties of minimal rational tangents and has been the key revelation on the significance of the differential geometric interpretation of the varieties of minimal rational tangents, as mentioned in Section 1.

**Theorem 2.18.** *Given a cone structure  $\mathcal{C} \subset \mathbb{P}T(M)$  admitting a characteristic connection, denote by  $\text{Pf}(\mathcal{C})$  the Pfaffian system defined on a dense open subset of  $M$  by the linear span of the homogeneous cone  $\widehat{\mathcal{C}} \subset T(M)$ . Then for any  $\alpha \in \text{Sm}(\widehat{\mathcal{C}})$  and  $\beta \in T_\alpha(\widehat{\mathcal{C}}) \cap T_\alpha^\pi$ , and any local sections  $\vec{\alpha}$  and  $\vec{\beta}$  of  $\text{Pf}(\mathcal{C})$  extending  $\alpha$  and  $\beta$ , the Lie bracket  $[\vec{\alpha}, \vec{\beta}]$  belongs to  $\text{Pf}(\mathcal{C})$  at the point  $\pi(\alpha)$ .*

When  $\mathcal{C} \subset \mathbb{P}T(M)$  itself is a Pfaffian system, i.e., when  $\text{Pf}(\mathcal{C}) = \mathcal{C}$ , Proposition 2.18 says that the existence of a characteristic connection on  $\mathcal{C}$  implies that  $\mathcal{C}$  is involutive. This is an example of the statement that characteristic connections contain certain amount of flatness. Another example of this phenomena is the next result from [12]:

**Theorem 2.19.** *Let  $Z \subset \mathbb{P}V$  be a smooth hypersurface of degree  $\geq 4$ . Let  $\mathcal{C} \subset \mathbb{P}T(M)$  be a  $Z$ -isotrivial cone structure. If  $\mathcal{C}$  has a characteristic connection, then it is locally flat.*

The above results show that the existence of characteristic connections impose severe restrictions on isotrivial cone structures. We expect similar restrictions on non-isotrivial cone structures, although no specific results are known.

Being a characteristic connection is a local property of a connection  $\mathcal{F}$  on  $\tilde{\mathcal{C}}$ : the condition in Definition 2.15 is to be checked on an open subset of  $\tilde{\mathcal{C}}$ . The connections we are interested in have another important property which is of a global nature. To introduce this property, we note that there is a natural vector subbundle  $\mathcal{P} \subset T(\text{Sm}(\mathcal{C}))$  defined as follows. At a smooth point  $\alpha \in \mathcal{C}$  with  $x = \pi(\alpha)$ , we have the differential  $d\varpi_\alpha : T_\alpha(\mathcal{C}) \rightarrow T_x(M)$  of the projection  $\varpi = \pi|_{\text{Sm}(\mathcal{C})} : \text{Sm}(\mathcal{C}) \rightarrow M$ . Since the fiber  $\mathcal{C}_x \subset \mathbb{P}T_x(M)$  of  $\pi|_{\mathcal{C}}$  is smooth at  $\alpha$  by Definition 2.2 (2), we have the affine tangent space  $T_\alpha(\widehat{\mathcal{C}}_x) \subset T_x(M)$ . Define  $\mathcal{P}_\alpha \subset T_\alpha(\mathcal{C})$  by

$$\mathcal{P}_\alpha := d\varpi_\alpha^{-1}(T_\alpha(\widehat{\mathcal{C}}_x)).$$

This defines a vector bundle  $\mathcal{P}$  on  $\text{Sm}(\mathcal{C})$ .

**Definition 2.20.** View  $T^\varpi \subset \mathcal{T}$  on  $\tilde{\mathcal{C}}$  and a connection  $\mathcal{F} \subset \mathcal{T}$  as vector bundles on  $\text{Sm}(\mathcal{C})$  via the normalization morphism  $\nu : \tilde{\mathcal{C}} \rightarrow \mathcal{C}$  which is an isomorphism over  $\text{Sm}(\mathcal{C})$ . Then we have  $\mathcal{T} \subset \mathcal{P} \subset T(\text{Sm}(\mathcal{C}))$ . A connection  $\mathcal{F} \subset \mathcal{T}$  is  $\mathcal{P}$ -splitting if there exists a vector subbundle  $\mathcal{W} \subset \mathcal{P}$  on  $\text{Sm}(\mathcal{C})$  that splits

$$0 \rightarrow T^\varpi \rightarrow \mathcal{P} \rightarrow \mathcal{P}/T^\varpi \rightarrow 0$$

such that  $\mathcal{W} \cap \mathcal{T} = \mathcal{F}$  on  $\text{Sm}(\mathcal{C})$ .

The connections we are interested in are  $\mathcal{P}$ -splitting. The significance of the  $\mathcal{P}$ -splitting property has been noticed only recently in [14], and many of its implications are yet to be discovered. It is used in [14] in the following way.

**Theorem 2.21.** *Let  $\mathcal{C} \subset \mathbb{P}T(M)$  be a smooth cone structure of codimension 1. In other words,  $\mathcal{C}$  is a smooth hypersurface in  $\mathbb{P}T(M)$  such that  $\varpi : \mathcal{C} \rightarrow M$  is a submersion. If  $\mathcal{C}$  has a  $\mathcal{P}$ -splitting connection and  $\dim M \geq 4$ , then it is isotrivial.*

In particular, if the degree of the fiber  $\mathcal{C}_x$  in Theorem 2.21 is at least 4 and the  $\mathcal{P}$ -splitting connection is also a characteristic connection, then  $\mathcal{C}$  is locally flat by Theorem 2.19. Actually, the requirement in Theorem 2.19 that the degree  $d$  of the hypersurface is at least 4 can be weakened to  $d \geq 3$  if the connection is  $\mathcal{P}$ -splitting. Thus at least for smooth cone structures of codimension 1, the existence of a  $\mathcal{P}$ -splitting connection has a significant consequence. This is to be contrasted with cone structures of codimension 1 that are not smooth. There are examples discovered in [3] of cone structures of codimension 1 that have  $\mathcal{P}$ -splitting characteristic connections but are not isotrivial.

Cartan geometry of cone structures with  $\mathcal{P}$ -splitting characteristic connections is our central interest from the differential geometric side. As we will see in Section 4, there are lots of examples of cone structures with  $\mathcal{P}$ -splitting characteristic connections. Properties of such structures are intricately related to the projective geometry of the fibers  $\mathcal{C}_x$ . Thus this Cartan geometry has an inseparable link with projective algebraic geometry. This is analogous to the fact that Cartan geometry of  $G$ -structures has an intimate link with representation theory of Lie groups. For example, the proofs of Theorem 2.19 and Theorem 2.21 use cohomological properties of smooth hypersurfaces in projective space. The number of results in this direction is still very small and the investigation of cone structures with  $\mathcal{P}$ -splitting characteristic connections is a wide open area.

Let us close this section with one remark. There is an additional property, which could be called the *admissibility* of a connection, that holds for all connections we are interested in. This property arises from Bernstein-Gindikin's admissibility condition in integral geometry [1]. Significant consequences of admissibility have not yet been found in connection with the topic of this article, which is the reason I have skipped discussing this property. However, this additional condition may lead to interesting discoveries in the future.

### 3. Mori geometry: minimal rational curves

Our major interest in algebraic geometry is in uniruled projective manifolds, i.e., projective manifolds covered by rational curves. Recall that a rational curve  $C$  on a projective manifold  $X$  is a curve  $C \subset X$  with normalization  $\nu_C : \mathbb{P}^1 \rightarrow C$  by  $\mathbb{P}^1$ . The set  $\text{RatCurves}(X)$  of all rational curves on  $X$  can be given a scheme structure and its normalization is denoted by  $\text{RatCurves}^n(X)$ . Each irreducible component  $\mathcal{K}$  of  $\text{RatCurves}^n(X)$  is a quasi-projective variety equipped with the universal  $\mathbb{P}^1$ -bundle  $\rho_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow \mathcal{K}$  and the associated cycle morphism  $\mu_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow X$ . This means that for each  $z \in \mathcal{K}$ , the corresponding rational curve  $C \subset X$  is given by  $\mu_{\mathcal{K}}(\rho_{\mathcal{K}}^{-1}(z))$  and the morphism

$$\nu_C := \mu_{\mathcal{K}}|_{\rho_{\mathcal{K}}^{-1}(z)} : \mathbb{P}^1 \rightarrow C = \mu_{\mathcal{K}}(\rho_{\mathcal{K}}^{-1}(z))$$

is the normalization of  $C$ . For a rigorous presentation of this foundational material, we refer the reader to [28].

Now I am going to introduce a number of terms related to uniruled manifolds and rational curves. I should warn the reader that most of these are *not standard*: they appear under different names in the literature. As is the case in any growing area of mathematics, the technical terms have not yet been completely standardized. I believe the terms introduced below are shorter and more intuitive than some of the ones in use (including some in my own papers) for nonexperts to remember their meaning. You may regard the definitions below as nicknames we will use in this article. To start with, we can give a precise definition of a uniruled projective manifold in the following form.

**Definition 3.1.** An irreducible component  $\mathcal{K}$  of  $\text{RatCurves}^n(X)$  is called a *uniruling* on  $X$  if the cycle morphism  $\mu_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow X$  is dominant. A projective manifold  $X$  is *uniruled* if it has a uniruling. For a line bundle  $L$  on  $X$ , we will denote by  $\deg_L(\mathcal{K})$  the  $L$ -degree of a member of  $\mathcal{K}$ .

A fundamental tool in the study of unirulings on  $X$  is the deformation theory of rational curves on  $X$ , or equivalently, the deformation theory of morphisms  $\mathbb{P}^1 \rightarrow X$ . By the classical Kodaira theory, given a rational curve  $C \subset X$ , the first-order deformation of the normalization morphism  $\nu_C : \mathbb{P}^1 \rightarrow X$  regarded as a map to  $X$  is controlled by the pull-back  $\nu_C^*T(X)$  of the tangent bundle of  $X$ . In this regard, the following definition is fundamental.

**Definition 3.2.** A rational curve  $C \subset X$  is *free* if  $\nu_C^*T(X)$  is semi-positive, i.e., of the form  $\mathcal{O}(a_1) \oplus \cdots \oplus \mathcal{O}(a_n)$ ,  $n = \dim X$ , with  $a_i \geq 0$  for all  $i$ .

Free rational curves have a nice deformation theory because  $H^1(\mathbb{P}^1, \nu_C^*T(X)) = 0$  by the semi-positivity of  $\nu_C^*T(X)$ . This cohomology group contains the obstruction to realizing deformations of  $\nu_C$  from its infinitesimal deformations in  $H^0(\mathbb{P}^1, \nu_C^*T(X))$ . Thus the vanishing implies the following.

**Theorem 3.3.** Let  $\mathcal{K}$  be an irreducible component of  $\text{RatCurves}^n(X)$ . Denote by  $\mathcal{K}^{\text{free}} \subset \mathcal{K}$  the parameter space of members of  $\mathcal{K}$  that are free. Then  $\mathcal{K}$  is a uniruling if and only if  $\mathcal{K}^{\text{free}}$  is nonempty. In this case,  $\mathcal{K}^{\text{free}}$  is a Zariski open subset of the smooth locus of  $\mathcal{K}$ .

Given a uniruling  $\mathcal{K}$  on  $X$  and a point  $x \in X$ , let  $\mathcal{K}_x$  be the normalization of the subvariety of  $\mathcal{K}$  parametrizing members of  $\mathcal{K}$  passing through  $x$ . When  $x$  is a general point of  $X$ , the structure of  $\mathcal{K}_x$  is particularly nice:

**Theorem 3.4.** For a uniruling  $\mathcal{K}$  on a projective manifold  $X$  and a general point  $x \in X$ , all members of  $\mathcal{K}_x$  belongs to  $\mathcal{K}^{\text{free}}$ . Furthermore, the variety  $\mathcal{K}_x$  is a finite union of smooth quasi-projective varieties of dimension  $\deg_{K_X^{-1}}(\mathcal{K}) - 2$ .

Both Theorem 3.3 and Theorem 3.4 must have been known before [37], although their significance has not been fully recognized until Mori's work. Now we are ready to introduce minimal rational curves.

**Definition 3.5.** A uniruling  $\mathcal{K}$  on a projective manifold  $X$  is *unbreakable* if  $\mathcal{K}_x$  is projective for a general  $x \in X$ . In other words,  $\mathcal{K}$  is an unbreakable uniruling if a general fiber of the cycle morphism  $\mu_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow X$  is nonempty and complete. Members of an unbreakable uniruling on  $X$  will be called *minimal rational curves* on  $X$ .

Unbreakable unirulings exist on any uniruled projective manifold. To see this, we need the following notion.

**Definition 3.6.** Let  $L$  be an ample line bundle on a projective manifold  $X$ . A uniruling  $\mathcal{K}$  is a *minimal with respect to  $L$* , if  $\deg_L(\mathcal{K})$  is minimal among all unirulings of  $X$ . A uniruling is a *minimal uniruling* if it is minimal with respect to some ample line bundle. Minimal unirulings exist on any uniruled projective manifold and they are unbreakable.

It is essential to understand the geometric idea behind the unbreakability of minimal unirulings. Suppose for a uniruling  $\mathcal{K}$ , which is minimal with respect to an ample line bundle  $L$ , the variety  $\mathcal{K}_x$  is not projective for a general point  $x \in X$ . Then the members of  $\mathcal{K}_x$  degenerate to reducible curves all components of which are rational curves of smaller  $L$ -degree than the members of  $\mathcal{K}$  and some components of which pass through  $x$ . Collecting those components passing through  $x$  as  $x$  varies over the general points of  $X$  gives rise to another uniruling  $\mathcal{K}'$  satisfying  $\deg_L(\mathcal{K}') < \deg_L(\mathcal{K})$ , a contradiction to the minimality of  $\deg_L(\mathcal{K})$ . This argument gives an intuitive picture behind the definition of an unbreakable uniruling: if a uniruling is not unbreakable, its members can be broken into members of another uniruling. More figuratively speaking, *if a uniruling is not unbreakable, it can be broken into a smaller uniruling.*

It is worth introducing a special class of minimal unirulings, which are particularly interesting from the viewpoint of (extrinsic) projective algebraic geometry:

**Definition 3.7.** Let  $L$  be an ample line bundle on a projective manifold  $X$  that is base-point free. A uniruling  $\mathcal{K}$  on  $X$  is a *uniruling by lines* if  $\deg_L(\mathcal{K}) = 1$ . Geometrically, this means that there is a morphism  $j : X \rightarrow \mathbb{P}^N$  which is finite over  $j(X)$  and sends members of  $\mathcal{K}$  to lines in  $\mathbb{P}^N$ .

Most of the classical examples of unbreakable unirulings are unirulings by lines and the morphism  $j$  is often an embedding. For example, smooth complete intersections of low degree in  $\mathbb{P}^N$  are covered by lines and so are Grassmannians under the Plücker embedding. But there are many examples of unirulings by lines where  $j$  is not an embedding. Also there are many minimal unirulings which are not unirulings by lines: hypersurfaces of degree  $n$  in  $\mathbb{P}^n$  have minimal unirulings by conics. Also there are many unbreakable unirulings that are not minimal. Trivial examples can be constructed on a product  $X = X_1 \times X_2$  of two uniruled manifolds. There are more interesting examples in [3] and [30] where these non-minimal unbreakable unirulings rather than minimal unirulings play crucial roles. We refer the reader to [24] and [28] for many examples of unbreakable unirulings.

All these examples illustrate that unbreakable unirulings and minimal rational curves, which exist on any uniruled projective manifolds, are genuine extensions of the classical notion of unirulings by lines:

$$\{\text{unirulings by lines}\} \subset \{\text{minimal unirulings}\} \subset \{\text{unbreakable unirulings}\}.$$

The important point is that this extension retains many geometric properties of unirulings by lines. The most fundamental example is the following. Let  $x \neq y$  be two distinct points on a projective manifold  $X$ . If  $\mathcal{K}$  is a uniruling by lines on  $X$ , we know that there exists at most one member of  $\mathcal{K}$  through  $x$  and  $y$ . Mori shows that a weaker version of this property continues to hold for unbreakable unirulings:

**Theorem 3.8.** *Let  $\mathcal{K}$  be an unbreakable uniruling. Then for a general point  $x \in X$  and any other point  $y \in X$ , there does not exist a positive-dimensional family of members of  $\mathcal{K}$  that pass through both  $x$  and  $y$ . In particular,  $\dim \mathcal{K}_x \leq \dim X - 1$ .*

Theorem 3.8 is proved by what is called the ‘bend-and-break’ argument. Geometrically, it says that any 1-dimensional family of rational curves which share two distinct points in common must degenerate into a reducible curve. This is the most important geometric property of an unbreakable uniruling. Combined with Theorem 3.4, the bound on  $\dim \mathcal{K}_x$  in Theorem 3.8 implies that  $\deg_{\mathcal{K}_X^{-1}}(\mathcal{K}) \leq \dim X + 1$  for any unbreakable uniruling. In other words, the  $\mathcal{K}_X^{-1}$ -degrees of minimal rational curves are bounded by  $\dim X + 1$ . The fact that all uniruled projective manifolds are covered by rational curves of small degree is of fundamental significance and has eventually developed into the boundedness of Fano manifolds, for which we refer the reader to Chapter V of [28].

An infinitesimal version of Theorem 3.8 is important for us. A key notion here is the following

**Definition 3.9.** A rational curve  $C \subset X$  is *unbending* if under the normalization  $\nu_C : \mathbb{P}^1 \rightarrow C \subset X$ , the vector bundle  $\nu_C^*T(X)$  has the form

$$\nu_C^*T(X) \cong \mathcal{O}(2) \oplus \mathcal{O}(1)^p \oplus \mathcal{O}^{n-1-p}$$

for some integer  $p$  satisfying  $0 \leq p \leq n - 1$ , where  $n = \dim X$ .

What is the rationale behind the name ‘unbending’? Note that a free rational curve  $C \subset X$  is unbending if and only if for any two distinct points  $x \neq y \in \mathbb{P}^1$  and their maximal ideals  $\mathfrak{m}_x$  and  $\mathfrak{m}_y$  in  $\mathcal{O}_{\mathbb{P}^1}$ ,

$$H^0(\mathbb{P}^1, \nu_C^*T(X) \otimes \mathfrak{m}_x \otimes \mathfrak{m}_y) = H^0(\mathbb{P}^1, T(\mathbb{P}^1) \otimes \mathfrak{m}_x \otimes \mathfrak{m}_y).$$

In the standard deformation theory, this means that  $C$  does not have infinitesimal deformation fixing two distinct points. Figuratively, we can say that ‘an unbending curve cannot be bent infinitesimally’. Comparing this with Theorem 3.8, we wonder whether members of an unbreakable uniruling are unbending. This is indeed the case for general members:

**Theorem 3.10.** *A general member of an unbreakable uniruling is unbending.*

Theorem 3.10 is not a direct consequence of Theorem 3.8 because the notion of an unbending rational curve gives information on infinitesimal deformation only, while Theorem 3.8 is concerned with an actually realized deformation. Theorem 3.10 enables us to control the behavior of the tangent directions of members of an unbreakable uniruling. To study this behavior systematically, it is convenient to introduce the tangent map.

**Definition 3.11.** For any uniruling  $\mathcal{K}$  on a projective manifold  $X$  and a point  $x \in X$ , the rational map  $\tau_x : \mathcal{K}_x \dashrightarrow \mathbb{P}T_x(X)$  sending a member of  $\mathcal{K}_x$  that is smooth at  $x$  to its tangent direction is called the *tangent map* at  $x$ .

If  $C$  is an unbending member of  $\mathcal{K}_x$ , the tangent map can be extended to  $[C] \in \mathcal{K}_x$ , even when  $C$  is singular at  $x$ , because the differential  $d\nu_C : T(\mathbb{P}^1) \rightarrow \nu_C^*T(X)$  is injective. In fact, a stronger result holds.

**Theorem 3.12.** *In Definition 3.11, if  $C$  is an unbending member of  $\mathcal{K}_x$ , the tangent map  $\tau_x$  is well-defined and immersive at  $[C] \in \mathcal{K}_x$ .*

In particular, Theorem 3.10 implies that for an unbreakable uniruling  $\mathcal{K}$  and a general point  $x \in X$ , the tangent map  $\tau_x$  is generically finite over its image.

It is worth comparing Theorem 3.10 with Theorem 3.3. Theorem 3.3 suggests that the freeness of rational curves is an ‘individualized’ version of the notion of a uniruling. In a similar way, Theorem 3.10 suggests that the unbending property of rational curves is an ‘individualized’ version of the notion of an unbreakable uniruling. The correspondence in the latter case, however, is less exact: general members of an unbreakable uniruling are unbending, but a uniruling whose general members are unbending is not necessarily unbreakable.

There is another important difference between the two correspondences. When  $\mathcal{K}$  is a uniruling, we know that every member of  $\mathcal{K}_x$  for a general  $x \in X$  is free by Theorem 3.4. Is it true that for an unbreakable uniruling  $\mathcal{K}$ , every member of  $\mathcal{K}_x$  for a general  $x \in X$  is unbending? Using Theorem 3.12, we can formulate this question as follows:

**Question 3.13.** *For an unbreakable uniruling  $\mathcal{K}$  on a projective manifold  $X$ , is the tangent map  $\tau_x : \mathcal{K}_x \dashrightarrow \mathbb{P}T_x(X)$  at a general point  $x \in X$  extendable to an immersion  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$ ?*

It is easy to see that if  $\mathcal{K}$  is a uniruling by lines, then the answer is yes:  $\tau_x$  is an embedding. This corresponds to the classical property of lines that they are determined by their tangent directions at a given point. Encouraged by this special case, it has been expected that the answer to Question 3.13 is affirmative for all unbreakable unirulings, or at least for all minimal unirulings. Recently, however, counterexamples have been discovered: an unbreakable one in [3] and then a minimal one in [17].

Although not all members of  $\mathcal{K}_x$  are as nice as we would wish them to be, they are still considerably well behaved, as the following result of Kebekus shows. In [26], an in-depth analysis of singularities of members of  $\mathcal{K}_x$  has been carried out. Among other things, Kebekus has shown

**Theorem 3.14.** *For an unbreakable uniruling  $\mathcal{K}$  and a general point  $x \in X$ , all members of  $\mathcal{K}_x$  are immersed at the point corresponding to  $x$ .*

To elaborate, a member of  $\mathcal{K}_x$  is given by a morphism  $\nu_C : \mathbb{P}^1 \rightarrow C \subset X$  with a point  $o \in \mathbb{P}^1$  satisfying  $\nu_C(o) = x$ . Theorem 3.14 says that  $(d\nu_C)_o : T_o(\mathbb{P}^1) \rightarrow T_x(X)$  is injective. Using this, Kebekus has shown the following important result.

**Theorem 3.15.** *In the setting of Theorem 3.14, the tangent morphism  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$  can be defined by assigning to each member  $C$  of  $\mathcal{K}_x$  its tangent direction*

$$\mathbb{P}(d\nu_C(T_o(\mathbb{P}^1))) \in \mathbb{P}T_x(X).$$

*This morphism  $\tau_x$  is finite over its image.*

Kebekus’s study of singularities of members of  $\mathcal{K}_x$ , including Theorem 3.14, plays an important role in the proof of Theorem 1.2 due to Cho, Miyaoka and Shepherd-Barron [4]. In terms of Theorem 3.15, we can state it as follows.

**Theorem 3.16.** *Let  $X$  be a uniruled projective manifold and let  $\mathcal{K}$  be an unbreakable uniruling. Suppose for a general point  $x \in X$ , the tangent morphism  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$  is dominant. Then  $X$  is projective space and  $\mathcal{K}$  is the space of lines.*

By Theorem 3.15, the morphism  $\tau_x$  in Theorem 3.16 is a finite covering of projective space. The essential point in the proof of Theorem 3.16 is to prove that  $\tau_x$  is a birational morphism, thus an isomorphism. In fact, it is fairly easy to show that  $X$  is projective space if  $\tau_x$  is an isomorphism. In this sense, the following result in [21] is a generalization of Theorem 3.16.

**Theorem 3.17.** *Let  $X$  be a uniruled projective manifold and let  $\mathcal{K}$  be an unbreakable uniruling. For a general point  $x \in X$ , the tangent morphism  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$  is birational over its image.*

Combining Theorem 3.15 and Theorem 3.17, we see that  $\tau_x$  is the normalization of its image in  $\mathbb{P}T_x(X)$ .

The collection of tangent morphisms  $\{\tau_x, \text{ general } x \in X\}$  can be assembled into a single map. Recall that we have the universal  $\mathbb{P}^1$ -bundle  $\rho_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow \mathcal{K}$  and the cycle morphism  $\mu_{\mathcal{K}} : \text{Univ}_{\mathcal{K}} \rightarrow X$ . When  $\mathcal{K}$  is unbreakable, the variety  $\mathcal{K}_x$  for a general  $x \in X$  can be identified with  $\mu_{\mathcal{K}}^{-1}(x) \subset \text{Univ}_{\mathcal{K}}$  and we have a rational map  $\tau : \text{Univ}_{\mathcal{K}} \dashrightarrow \mathbb{P}T(X)$  with a commuting diagram

$$\begin{array}{ccc} \text{Univ}_{\mathcal{K}} & \xrightarrow{\tau} & \mathbb{P}T(X) \\ \mu_{\mathcal{K}} \downarrow & & \downarrow \pi \\ X & = & X \end{array}$$

such that the tangent morphism  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$  is just the fiber of this diagram at a general point  $x \in X$ . In terms of  $\tau$ , we can summarize Theorem 3.4, Theorem 3.14 and Theorem 3.17 as follows.

**Theorem 3.18.** *Let  $\mathcal{K}$  be an unbreakable uniruling on a projective manifold  $X$  and let  $\tau : \text{Univ}_{\mathcal{K}} \dashrightarrow \mathbb{P}T(X)$  be the tangent map. Then there exists a Zariski open subset  $X_o \subset X$  satisfying the following properties.*

- (i) *The dense open subset  $\text{Univ}_{\mathcal{K}}^o := \mu_{\mathcal{K}}^{-1}(X_o)$  of  $\text{Univ}_{\mathcal{K}}$  is a smooth quasi-projective variety.*
- (ii) *The morphism  $\mu_{\mathcal{K}}$  restricted to  $\text{Univ}_{\mathcal{K}}^o$  is a submersion over  $X_o$ .*
- (iii) *The restriction of  $\tau$  gives a morphism  $\tau^o : \text{Univ}_{\mathcal{K}}^o \rightarrow \mathbb{P}T(X_o)$  which is normalization of its image.*
- (iv) *The two vector subbundles  $T^{\mu_{\mathcal{K}}}$  and  $T^{\rho_{\mathcal{K}}}$  of  $T(\text{Univ}_{\mathcal{K}}^o)$  are transversal.*

In fact, (i) and (ii) follow from Theorem 3.4, (iii) follows from Theorem 3.17 and finally (iv) follows from Theorem 3.14.

In this section, we have collected some of the key results on minimal rational curves that are needed for the next section. There are many other results on minimal rational curves which we have omitted, for which we refer the reader to [28] and the survey papers cited in Section 1.

### 4. From Mori to Cartan: VMRT-structures

In Section 3, we have seen that starting from a uniruled projective manifold  $X$  by choosing an unbreakable uniruling  $\mathcal{K}$  and a general point  $x \in X$ , we obtain  $\mathcal{K}_x$ , a finite union of projective manifolds. Since  $\dim \mathcal{K}_x < \dim X$  by Theorem 3.8, the variety  $\mathcal{K}_x$  is likely to be simpler than  $X$ . This opens up the possibility of using  $\mathcal{K}_x$  to study the geometry of  $X$ . Indeed, at least for unirulings by lines, the variety  $\mathcal{K}_x$  has been used in this way in many classical constructions. Now Theorem 3.15 and Theorem 3.17 show a bigger advantage of this  $\mathcal{K}_x$ : it is provided with a morphism  $\tau_x : \mathcal{K}_x \rightarrow \mathbb{P}T_x(X)$  which is almost an embedding,

a normalization of its image. That is, starting from the intrinsic information of an unbreakable uniruling, we obtain a natural projective subvariety  $\text{Im}(\tau_x) \subset \mathbb{P}T_x(X)$ . The extrinsic projective geometry of  $\text{Im}(\tau_x)$  yields intrinsic information on  $X$  and  $\mathcal{K}$ . Since  $\tau_x$  is just the normalization of its image, it seems more advantageous to look at  $\text{Im}(\tau_x)$  rather than  $\mathcal{K}_x$ . This motivates the following definition.

**Definition 4.1.** In the above setting, the image  $\text{Im}(\tau_x)$  is denoted by  $\mathcal{C}_x \subset \mathbb{P}T_x(X)$  and called the *variety of minimal rational tangents* (abbr. VMRT) at  $x$  associated to  $\mathcal{K}$ .

This shift of attention from  $\mathcal{K}_x$  to  $\mathcal{C}_x$  connects Mori geometry to Cartan geometry:

**Definition 4.2.** In the setting of Theorem 3.18, put  $\tilde{\mathcal{C}} := \text{Univ}_{\mathcal{K}}^o$  and  $\mathcal{C} := \tau(\text{Univ}_{\mathcal{K}}^o)$ . Then  $\nu = \tau^o : \tilde{\mathcal{C}} \rightarrow \mathcal{C}$  is a normalization morphism. Putting  $\mathcal{T} := T^{\rho\kappa} \oplus T^{\mu\kappa}$  on  $\tilde{\mathcal{C}}$ , Theorem 3.18 (iv) says that  $\mathcal{T}$  is a vector subbundle of  $T(\tilde{\mathcal{C}})$ . It satisfies the condition (3) of Definition 2.2, so  $\mathcal{C} \subset \mathbb{P}T(X_o)$  is a cone structure on the complex manifold  $X_o$ . This cone structure is called the *VMRT-structure* of the unbreakable uniruling  $\mathcal{K}$ . Moreover  $T^{\rho\kappa}$  gives a connection  $\mathcal{F}$  on this cone structure, called the *tautological connection* on  $\mathcal{C}$ .

The following is proved in [14] and [21].

**Theorem 4.3.** *The tautological connection  $\mathcal{F}$  on the VMRT-structure in Definition 4.2 is a  $\mathcal{P}$ -splitting characteristic connection.*

It follows that a choice of an unbreakable uniruling on a uniruled projective manifold gives rise to a cone structure with a natural  $\mathcal{P}$ -splitting characteristic connection: the VMRT-structure with the tautological connection. This provides diverse examples of such cone structures. How much diversity? We will see shortly that this transition from algebraic geometry to differential geometry

$$(X, \mathcal{K}) \Rightarrow (\mathcal{C} \subset \mathbb{P}T(X_o), \mathcal{F})$$

is injective, under some topological assumptions. Thus one may say that such cone structures are almost as diverse as all uniruled projective manifolds and all unbreakable unirulings on them. Referring the reader to [9] and [19] for many interesting examples of VMRT-structures, let us just note three salient features of this diversity.

Firstly, for any irreducible smooth projective variety  $Z \subset \mathbb{P}V$ , there exists a VMRT-structure which is  $Z$ -isotrivial and locally flat (Example 1.7 in [12]). In fact, writing  $W = V \oplus \mathbb{C}$ , we can regard  $Z \subset \mathbb{P}V \subset \mathbb{P}W$  as a smooth subvariety of  $\mathbb{P}W$  contained in the hyperplane  $\mathbb{P}V$ . Then the blow-up  $X := \text{Bl}_Z(\mathbb{P}W)$  has an unbreakable uniruling whose general members are proper transforms of lines in  $\mathbb{P}W$  intersecting  $Z$  at one point. Then for  $x \in X$  over a point on  $\mathbb{P}W \setminus \mathbb{P}V$ , the VMRT  $\mathcal{C}_x \subset \mathbb{P}T_x(X)$  is isomorphic to  $Z \subset \mathbb{P}V$  as projective varieties. It is easy to see that this VMRT-structure is locally flat.

Secondly, there are isotrivial VMRT-structures which are not locally flat. A simple example of this type is provided by a homogeneous contact manifold  $X$  with second Betti number 1, different from projective space. This is a homogeneous projective manifold  $X = G/P$  equipped with a  $G$ -invariant holomorphic contact structure  $\mathcal{D} \subset T(X)$ , i.e., a Pfaffian system of corank 1 which is maximally non-integrable. It has a uniruling by lines, the VMRT-structure  $\mathcal{C} \subset \mathbb{P}T(X)$  of which is  $G$ -invariant and hence it is isotrivial. Furthermore, at each  $x \in X$ , the VMRT  $\mathcal{C}_x$  is contained in  $\mathbb{P}\mathcal{D}_x$  and, in fact, spans  $\mathbb{P}\mathcal{D}_x$  (see [8]). This implies that  $\mathcal{C}$  is not locally flat, because if it were, then  $\mathcal{D}$  would be integrable.

Thirdly, there are many examples of non-isotrivial VMRT-structures. Actually, most of the VMRT-structures on projective manifolds with second Betti number 1 are expected to be non-isotrivial. A most transparent example is given by the moduli variety  $SU_C(r, d)$  of stable bundles on a curve  $C$  of genus greater than 3 with rank  $r$  and fixed determinant of degree  $r$ . When  $(r, d) = 1$ , this is a projective manifold with second Betti number 1. There is a minimal uniruling given by ‘Hecke curves’, certain families of stable bundles on  $C$  arising from a geometric version of the Hecke correspondence introduced in [38]. Its VMRT’s are iterated projective bundles on  $C$  constructed from the universal bundle on  $C \times SU_C(r, d)$ , which shows that the VMRT-structure is not isotrivial (see [10] and [23] for details).

These three points illustrate the diversity of VMRT-structures. But even on these specific points, our understanding of this diversity is very limited.

Regarding the examples of locally flat VMRT-structures, our construction does not work if  $Z$  is reducible. In fact, if  $\mathcal{K}_x$  is reducible, there is a global constraint on the VMRT-structure coming from the irreducibility of  $\mathcal{K}$ . For example, we cannot have  $\mathcal{C}_x$  consisting of two components only one of which is linear. In fact, if this were to happen, then collecting the linear components for general  $x \in X$ , we would get an irreducible component of  $\mathcal{C}$  different from  $\mathcal{C}$ , a contradiction to the irreducibility of  $\mathcal{C}$ . Also the construction does not work if  $Z$  is singular with smooth normalization. Which singular varieties can be realized as the VMRT of an unbreakable uniruling is a completely open question.

As for the examples of non-isotrivial VMRT-structures, although we expect that ‘most’ VMRT-structures are non-isotrivial, this has been verified only in a small number of cases because checking the non-isotriviality can be tricky even when  $X$  is a simple well-known variety. We do not have a general method for doing this. An approach to prove that a certain type of VMRT-structures are not isotrivial has been proposed in [12], but this has not been implemented in concrete examples. So far the non-isotriviality is known only for a handful of examples, and only by explicit direct description of the VMRT’s. Besides  $SU_C(r, d)$  explained above, there are only two more types of examples we know of. In [29], Landsberg and Robles show that the uniruling by lines on a general smooth hypersurface of degree  $d, 3 \leq d \leq n$ , in  $\mathbb{P}^{n+1}$  has non-isotrivial VMRT-structure. Another case is in [16] which shows the non-isotriviality of the VMRT-structure for the uniruling by lines on a double cover of  $\mathbb{P}^n$  branched along a general smooth hypersurface of degree  $2m, 2 \leq m \leq n - 1$ .

Now that we have many cone structures with  $\mathcal{P}$ -splitting characteristic connections arising from unbreakable unirulings on projective manifolds, it is natural to study the local equivalence problem for these structures in the sense of Cartan, which we can formulate as follows.

**Definition 4.4.** Let  $X^1$  (resp.  $X^2$ ) be a uniruled projective manifold with an unbreakable uniruling  $\mathcal{K}^1$  (resp.  $\mathcal{K}^2$ ). Let  $\mathcal{C}^1 \subset \mathbb{P}T(X^1_o)$  (resp.  $\mathcal{C}^2 \subset \mathbb{P}T(X^2_o)$ ) be the corresponding VMRT-structure with the tautological connection  $\mathcal{F}^1$  (resp  $\mathcal{F}^2$ ). Suppose that there exists a connected Euclidean open subset  $U^1 \subset X^1_o$  (resp.  $U^2 \subset X^2_o$ ) and a biholomorphic map  $\varphi : U^1 \rightarrow U^2$  such that  $\mathbb{P}d\varphi : \mathbb{P}T(U^1) \rightarrow \mathbb{P}T(U^2)$  sends  $\mathcal{C}^1|_{U^1}$  to  $\mathcal{C}^2|_{U^2}$  and  $\mathcal{F}^1|_{U^1}$  to  $\mathcal{F}^2|_{U^2}$ . Then we say that the VMRT-structures  $\mathcal{C}^1$  and  $\mathcal{C}^2$  are *locally equivalent* and  $\varphi : U^1 \rightarrow U^2$  is a *local equivalence map* of the two VMRT-structures.

This is of course a natural definition from the viewpoint of Cartan geometry. But does it have significant implications in algebraic geometry? The following result from [20] shows that this is indeed so, under some topological assumptions.

**Theorem 4.5.** *In the setting of Definition 4.4, assume that  $b_2(X^1) = 1$  and  $\dim \mathcal{K}_x^1 > 0$  for a general  $x \in X_1$ , or equivalently,  $\deg_{K_{X^1}}(\mathcal{K}^1) \geq 3$ . Then an equivalence map  $\varphi$  can be extended to a rational map  $\Phi : X^1 \dashrightarrow X^2$ , i.e.,  $\varphi = \Phi|_{U^1}$ . If furthermore,  $b_2(X^2) = 1$ , then  $\Phi$  is biregular.*

In other words, within the class of uniruled projective manifolds of second Betti number 1 and unirulings of anti-canonical degree at least 3, the correspondence

$$(X, \mathcal{K}) \Rightarrow (\mathcal{C} \subset \mathbb{P}T(X_o), \mathcal{F})$$

is injective. The requirement in Theorem 4.5 that the projective manifolds have second Betti number 1 is necessary. Indeed, it is easy to construct examples of  $X^1$  with  $b_2(X^1) > 1$  where an analogue of Theorem 4.5 fails. However, this condition is not a serious handicap. In fact, projective manifolds with second Betti number 1 form a large class of projective varieties where general structure theories of higher dimensional algebraic geometry, like the minimal model program, give little direct information. That VMRT-theory is effective for uniruled projective manifolds with second Betti number 1 means that it could complement these general structure theories.

The proof of Theorem 4.5 is by analytic continuation of the map  $\varphi$  along members of  $\mathcal{K}^1$ . This analytic continuation corresponds to the construction of the developing map in Cartan geometry. Here  $X^2$  is regarded as the model of the geometric structure and the map  $\varphi$  is developed to  $\Phi$ . The topological condition  $b_2(X^1) = 1$  is used to guarantee that the analytic continuation can be carried out to cover the whole of  $X^1$ . The potential multi-valuedness of the analytic continuation is taken care of by the condition  $\dim \mathcal{K}_x^1 > 0$ .

From Definition 4.4, the condition required for the map  $\varphi$  in Theorem 4.5 is that  $\mathbb{P}d\varphi$  preserves the cone structure and also the characteristic connection. These are differential geometric conditions. To be useful in algebraic geometry, we need a way to find algebro-geometric conditions to guarantee them. In most interesting cases, the requirement of preserving the connection can be replaced by algebro-geometric condition in the following form:

**Theorem 4.6.** *Let  $X^1$  (resp.  $X^2$ ) be a uniruled projective manifold with  $b_2(X^1) = 1$  (resp.  $b_2(X^2) = 1$ ). Let  $\mathcal{K}^1$  (resp.  $\mathcal{K}^2$ ) be an unbreakable uniruling on  $X^1$  (resp.  $X^2$ ). Let  $\mathcal{C}^1 \subset \mathbb{P}T(X_o^1)$  (resp.  $\mathcal{C}^2 \subset \mathbb{P}T(X_o^2)$ ) be the corresponding VMRT-structure. Suppose that there exists a connected Euclidean open subset  $U^1 \subset X_o^1$  (resp.  $U^2 \subset X_o^2$ ) and a biholomorphic map  $\varphi : U^1 \rightarrow U^2$  such that  $\mathbb{P}d\varphi : \mathbb{P}T(U^1) \rightarrow \mathbb{P}T(U^2)$  sends  $\mathcal{C}^1|_{U^1}$  to  $\mathcal{C}^2|_{U^2}$ . If the VMRT  $\mathcal{C}_x^1$  at a general point  $x \in U^1$  has positive dimension and nondegenerate Gauss map, then there exists a biregular morphism  $\Phi : X^1 \rightarrow X^2$  such that  $\Phi|_{U^1} = \varphi$ .*

This is a simple combination of Theorem 2.17 and Theorem 4.5. Note that one Cartanian condition in Theorem 4.5, i.e., that  $\varphi$  preserves the connection  $\mathcal{F}$ , has been replaced by an algebro-geometric condition on the fiber  $\mathcal{C}_x$ , the nondegeneracy of Gauss map. The latter condition holds for the VMRT of a large class of uniruled projective manifolds and unbreakable unirulings, so we may concentrate on VMRT-structures satisfying Theorem 4.6. Thus our main problem becomes finding algebro-geometric conditions for the local equivalence of the cone structures. In this regard, the most central problem connecting Mori geometry and Cartan geometry is the following:

**Problem 4.7.** Let  $X^1$  (resp.  $X^2$ ) be a uniruled projective manifold with an unbreakable uniruling  $\mathcal{K}^1$  (resp.  $\mathcal{K}^2$ ). Let  $\mathcal{C}^1 \subset \mathbb{P}T(X_o^1)$  (resp.  $\mathcal{C}^2 \subset \mathbb{P}T(X_o^2)$ ) be the corresponding

VMRT-structure with the tautological connection  $\mathcal{F}^1$  (resp  $\mathcal{F}^2$ ). Suppose that there exists a connected Euclidean open subset  $U^1 \subset X^1_o$  (resp.  $U^2 \subset X^2_o$ ) and a biholomorphic map  $\varphi : U^1 \rightarrow U^2$  such that  $\mathcal{C}^1|_{U^1}$  and  $\mathcal{C}^2|_{U^2}$  are isomorphic as families of projective varieties. More precisely, suppose that we have a biholomorphic map  $\psi : \mathbb{P}T(U^1) \rightarrow \mathbb{P}T(U^2)$  satisfying  $\psi(\mathcal{C}^1|_{U^1}) = \mathcal{C}^2|_{U^2}$  with a commuting diagram

$$\begin{array}{ccc} \mathbb{P}T(U^1) & \xrightarrow{\psi} & \mathbb{P}T(U^2) \\ \downarrow & & \downarrow \\ U^1 & \xrightarrow{\varphi} & U^2 \end{array} .$$

Does this imply that the two cone structures  $\mathcal{C}^1|_{U^1}$  and  $\mathcal{C}^2|_{U^2}$  are locally equivalent? In other words, can we choose  $\varphi$  such that  $\psi = \mathbb{P}d\varphi$ ?

What makes Problem 4.7 exciting and challenging is that it needs insights and techniques for the fusion of the algebraic geometry of the projective variety  $\mathcal{C}_x \subset \mathbb{P}T_x(X)$  and the differential geometry of the cone structure  $\mathcal{C}$  via the connection  $\mathcal{F}$ . At present, no plausible uniform method to handle Problem 4.7 seems conceivable and, depending on the type of the family of varieties  $\mathcal{C}_x$ 's, different tools are required. A practicable approach is to try the cases where the projective geometry of the family  $\mathcal{C}_x$ 's is simple and the differential geometric machinery of the cone structure is available. A most reasonable candidate is the case when the VMRT-structure is  $Z$ -isotrivial for a projective variety  $Z \subset \mathbb{P}V$  of simple type.

Even for  $Z$ -isotrivial cases, Problem 4.7 is highly challenging. The answer is not always affirmative. There are examples of two uniruled projective manifolds with  $Z$ -isotrivial VMRT-structures for the same smooth irreducible  $Z \subset \mathbb{P}V$ , which are not locally equivalent. Since we have seen that a locally flat  $Z$ -isotrivial VMRT-structure exists for any given  $Z \subset \mathbb{P}V$ , it suffices to give examples of  $Z$ -isotrivial VMRT-structures that are not locally flat. We have already mentioned that homogeneous contact manifolds provide such examples where  $Z \subset \mathbb{P}V$  is contained in a hyperplane in  $\mathbb{P}V$ . There are also examples where  $Z$  spans  $\mathbb{P}V$ . A simple example is provided by symplectic Grassmannians. Given a vector space  $W$  with a symplectic form  $\omega$ , let us denote by  $\text{Gr}_\omega(k, W)$  the set of  $k$ -dimensional subspaces of  $W$  isotropic with respect to  $\omega$ . This projective manifold  $\text{Gr}_\omega(k, W)$  has a uniruling by lines. The associated VMRT-structure is  $Z$ -isotrivial for some  $Z$  because  $\text{Gr}_\omega(k, W)$  is a homogeneous space. It turns out that  $Z \subset \mathbb{P}V$  is nondegenerate. If  $2k < \dim W$ , the automorphism group  $\text{Aut}(\widehat{Z})$  acts on  $Z$  with two orbits and the unique closed orbit  $Z' \subset Z$  is degenerate. The corresponding subvariety  $\mathcal{C}' \subset \mathcal{C}$  of the VMRT-structure spans a Pfaffian system  $\mathcal{D} \subset T(\text{Gr}_\omega(k, W))$  which is not integrable. This non-integrability implies that the VMRT-structure is not locally flat (see [22] for details).

This example shows that the answer to Problem 4.7 for  $Z$ -isotrivial cases would depend on the variety  $Z$ . This raises the following purely local problem in Cartan geometry:

**Problem 4.8.** Find algebro-geometric conditions on a smooth irreducible  $Z \subset \mathbb{P}V$  which guarantee that every  $Z$ -isotrivial cone structure with a  $\mathcal{P}$ -splitting characteristic connection is locally flat.

As we have already mentioned in Section 2, there are two types of  $Z$  where the answer is known. When  $Z$  is a linear subspace of  $\mathbb{P}V$ , we have seen that the corresponding Pfaffian system is involutive, thus the cone structure is locally flat. Another case is when  $Z \subset \mathbb{P}V$  is a smooth hypersurface of degree  $\geq 3$  by Theorem 2.19 together with the comment

after Theorem 2.21. At the moment, these are the only cases where Problem 4.8 has been answered, even though it is reasonable to expect that it is locally flat for most choices of  $Z \subset \mathbb{P}V$ .

One important case of  $Z$ -isotrivial VMRT-structures where an affirmative answer is obtained for Problem 4.7 is when  $Z$  is the VMRT of an irreducible Hermitian symmetric space. This is not done through the local approach of Problem 4.8. Instead, a mixture of local and global methods has settled the question successfully. As mentioned in Section 2, differential geometric machinery of such structures has been developed by differential geometers and there are natural curvature tensors whose vanishing implies the local flatness. In [34], Mok shows that if a VMRT-structure is  $Z$ -isotrivial where  $Z$  is the VMRT of an irreducible Hermitian symmetric space, then the  $Z$ -isotrivial cone structure can be extended to a neighborhood of a general member of the unbreakable uniruling. Then the unbending property of the rational curve can be used to show the vanishing of the curvature tensors. This way, Mok proves

**Theorem 4.9.** *Let  $X$  be a uniruled projective manifold with second Betti number 1. Suppose there exists an unbreakable uniruling whose VMRT-structure is  $Z$ -isotrivial where  $Z$  is the VMRT of an irreducible Hermitian symmetric space  $G/P$ . Then  $X$  is biregular to  $G/P$ .*

This resolves Problem 1.3 for irreducible Hermitian symmetric spaces. Mok's method has been extended to any rational homogeneous space  $G/P$  where  $P$  is a maximal parabolic subgroup of a complex simple Lie group  $G$ , associated to a long root of  $G$  in [6] and [34]. Whether this remains true when  $P$  is associated to a short root of  $G$ , such as symplectic Grassmannians, is a most tantalizing problem in this direction. Also, it would be desirable to recover Theorem 4.9 by answering Problem 4.8 for the corresponding  $Z$ .

Combining Theorem 4.9 with the affirmative answer to Problem 4.8 when  $Z$  is a smooth hypersurface of degree  $\geq 3$ , we have the following result in [14].

**Theorem 4.10.** *Let  $X$  be a uniruled projective manifold of dimension  $\geq 4$  with second Betti number 1. Assume that there is an unbreakable uniruling such that the VMRT at a general point is a smooth hypersurface. Then  $X$  is a quadric hypersurface.*

These few examples are all the results we currently have for Problem 4.7 for  $Z$ -isotrivial cases. They correspond to the simplest type of projective varieties, hypersurfaces and some homogeneous varieties. Any other projective variety  $Z \subset \mathbb{P}V$  seems to be a new challenge. Even less -essentially no concrete result- is known for non-isotrivial cases of Problem 4.7.

There is a natural extension of the equivalence problem for VMRT-structures to the setting of submanifold geometry. In the investigation of any type of geometric structures, it is natural to study submanifolds inheriting such structures. From this perspective, among submanifolds in a uniruled projective manifold  $X$ , of particular importance are uniruled submanifolds with VMRT-structures which are compatible with the VMRT-structure of the ambient manifold. The study of such submanifolds from an intrinsic viewpoint, which has been initiated in [35], is based on the fact that the analytic continuation Theorem 4.5 can be easily extended to submanifold geometry in the following sense.

**Theorem 4.11.** *Let  $X^1$  (resp.  $X^2$ ) be a uniruled projective manifold with an unbreakable uniruling  $\mathcal{K}^1$  (resp.  $\mathcal{K}^2$ ). Let  $\mathcal{C}^1 \subset \mathbb{P}T(X^1_o)$  (resp.  $\mathcal{C}^2 \subset \mathbb{P}T(X^2_o)$ ) be the corresponding VMRT-structure with the tautological connection  $\mathcal{F}^1$  (resp.  $\mathcal{F}^2$ ). Suppose that there exists a connected Euclidean open subset  $U^1 \subset X^1_o$  (resp.  $U^2 \subset X^2_o$ ) and an embedding  $\varphi : U^1 \rightarrow$*

$U^2$  such that  $\mathbb{P}d\varphi : \mathbb{P}T(U^1) \rightarrow \mathbb{P}T(U^2)$  sends  $\mathcal{C}^1|_{U^1}$  into  $\mathcal{C}^2|_{U^2}$  and  $\mathcal{F}^1$  into  $\mathcal{F}^2$ . Assume that  $b_2(X^1) = 1$  and  $\dim \mathcal{K}_x^1 > 0$  for a general  $x \in X_1$ . Then  $\varphi$  can be extended to a rational map  $\Phi : X^1 \dashrightarrow X^2$ , i.e.,  $\varphi = \Phi|_{U^1}$ .

The first task is to replace the preservation of connections by algebro-geometric conditions, just as Theorem 4.6 is an improvement over Theorem 4.5. This is done in [7] and the following analogue of Theorem 4.6 in submanifold geometry is obtained.

**Theorem 4.12.** *Let  $X^1$  (resp.  $X^2$ ) be a uniruled projective manifold with a VMRT-structure  $\mathcal{C}^1 \subset \mathbb{P}T(X^1_o)$  (resp.  $\mathcal{C}^2 \subset \mathbb{P}T(X^2_o)$ ). Assume that  $b_2(X^1) = 1$  and  $\dim \mathcal{C}_x^1 > 0$  for  $x \in X^1_o$ . Suppose that there exists a connected Euclidean open subset  $U^1 \subset X^1_o$  (resp.  $U^2 \subset X^2_o$ ) and an embedding  $\varphi : U^1 \rightarrow U^2$  satisfying the following two conditions.*

- (i)  $\mathbb{P}d\varphi(\mathcal{C}^1|_{U^1}) \subset \mathcal{C}^2|_{\varphi(U^1)}$  and
- (ii) for a general point  $z \in U^1$  with  $x = \varphi(z)$  and a general smooth point  $\beta \in \mathcal{C}_z^1$ , the image  $\alpha := \mathbb{P}d\varphi(\beta)$  is a smooth point of  $\mathcal{C}_x^2$  and

$$\{v \in T_\alpha(\mathcal{C}_x^2), \Pi_{\mathcal{C}_x^2, \alpha}(v, u) = 0 \text{ for all } u \in d(\mathbb{P}d\varphi)(T_\beta(\mathcal{C}_z^1))\} = 0.$$

Then there exists a rational map  $\Phi : X^1 \dashrightarrow X^2$  such that  $\Phi|_{U^1} = \varphi$ .

To be precise, the statement in [7] is slightly weaker, although their argument essentially proves Theorem 4.12. The full statement of Theorem 4.12 is given in [15] with a simplified proof.

The condition (ii) in Theorem 4.12 is a natural extension of the Gauss map condition in Theorem 4.6. Note that the Cartanian condition in Theorem 4.11 on the characteristic connection is replaced by the algebro-geometric condition (ii).

Theorem 4.11 goes beyond Theorem 4.6 even when  $\dim X^1 = \dim X^2$  because  $\dim \mathcal{K}^1$  can be strictly smaller than  $\dim \mathcal{K}^2$ . For example, it can describe a finite morphism  $X^1 \rightarrow X^2$  which sends members of  $\mathcal{K}^1$  to members of  $\mathcal{K}^2$ .

One immediate question is whether the map  $\Phi$  can be extended to a morphism  $\Phi : X^1 \rightarrow X^2$  when  $b_2(X^2) = 1$ , as in Theorem 4.6. In [7], an affirmative answer is given when  $X^1$  is modeled on a special class of Schubert submanifolds in rational homogeneous spaces  $X^2 = G/P$ .

At present, submanifold theory in the interaction of Mori geometry and Cartan geometry is in an incipient stage. Many natural questions can be raised, but the most fundamental one is an analogue of Problem 4.7 in the submanifold setting. This would lead to a deeper aspect of Cartan geometry. Interesting applications to algebraic geometry are yet to come.

In conclusion, VMRT-structures on uniruled projective manifolds provide us with a large number of examples in great diversity of cone structures admitting  $\mathcal{P}$ -splitting characteristic connections. Cartan geometry of these structures has been understood for only a few special cases and a wide range of examples and problems remain to be explored. The small number of results we have seen so far have already found interesting applications in algebraic geometry. Further development will undoubtedly bring more exciting applications. This will be a fertile ground for interactions between differential geometry and algebraic geometry.

**Acknowledgment.** Supported by National Researcher Program 2010-0020413 of NRF. I am very grateful to Ngaiming Mok and Richard Weiss for valuable comments and helpful suggestions.

## References

- [1] Bernstein, J. and Gindikin, S., *Notes on integral geometry for manifolds of curves*, Amer. Math. Soc. Transl. Ser. (2) **210** (2003), 57–80.
- [2] Cartan, E., *Les groupes de transformations continus, infinis, simples*, Ann. Sci. Ecole Norm. Sup. **26** (1909), 93–161.
- [3] Casagrande, C. and Druel, S., *Locally unsplit families of rational curves of large anti-canonical degree on Fano manifolds*, preprint, 2012.
- [4] Cho, K., Miyaoka, Y., and Shepherd-Barron, N., *Characterizations of projective space and applications to complex symplectic manifolds*, Adv. Stud. Pure Math. **35** (2002), 1–88.
- [5] Guillemin, V., *The integrability problem for G-structures*, Trans. Amer. Math. Soc. **116** (1965), 544–560.
- [6] Hong, J. and Hwang, J.-M., *Characterization of the rational homogeneous space associated to a long simple root by its variety of minimal rational tangents*, Adv. Stud. Pure Math. **50** (2008), 217–236.
- [7] Hong, J. and Mok, N., *Analytic continuation of holomorphic maps respecting varieties of minimal rational tangents and applications to rational homogeneous manifolds*, J. Diff. Geom. **86** (2010), 539–567.
- [8] Hwang, J.-M., *Rigidity of homogeneous contact manifolds under Fano deformation*, J. reine angew. Math. **486** (1997), 153–163.
- [9] ———, *Geometry of minimal rational curves on Fano manifolds*, in School on Vanishing Theorems and Effective Results in Algebraic Geometry (Trieste, 2000), 335–393, ICTP Lect. Notes 6, Abdus Salam Int. Cent. Theoret. Phys., Trieste, 2001.
- [10] ———, *Hecke curves on the moduli space of vector bundles over an algebraic curve*, in Algebraic Geometry in East Asia (Kyoto, August 3-10, 2001), World Scientific, 2003, pp. 155–164.
- [11] ———, *Rigidity of rational homogeneous spaces*, in Proceedings of ICM 2006 Madrid, volume II, European Mathematical Society, 2006, pp. 613–626.
- [12] ———, *Equivalence problem for minimal rational curves with isotrivial varieties of minimal rational tangents*, Ann. scient. Ec. Norm. Sup. **43** (2010), 607–620.
- [13] ———, *Geometry of varieties of minimal rational tangents*, in Current Developments in Algebraic Geometry, 197–222, MSRI Publ. 59, Cambridge University Press, New York, 2012.
- [14] ———, *Varieties of minimal rational tangents of codimension 1*, Ann. scient. Ec. Norm. Sup. **46** (2013), 629–649.
- [15] ———, *Cartan-Fubini type extension of holomorphic maps respecting varieties of minimal rational tangents*, preprint, 2014.
- [16] Hwang, J.-M. and Kim, H., *Varieties of minimal rational tangents on double covers of projective space*, Math. Zeit. **275** (2013), 109–125.
- [17] ———, *Varieties of minimal rational tangents on Veronese double cones*, preprint, 2013.
- [18] Hwang, J.-M. and Mok, N., *Rigidity of irreducible Hermitian symmetric spaces of the*

- compact type under Kähler deformation*, Invent. math. **131** (1998), 393–418.
- [19] ———, *Varieties of minimal rational tangents on uniruled projective manifolds*, in Several complex variables (Berkeley, CA, 1995–1996), 351–389, MSRI Publ. 37, Cambridge Univ. Press, Cambridge, 1999.
- [20] ———, *Cartan-Fubini type extension of holomorphic maps for Fano manifolds of Picard number 1*, Journal Math. Pures Appl. **80** (2001), 563–575.
- [21] ———, *Birationality of the tangent map for minimal rational curves*, Asian J. Math. **8** (2004), 51–63.
- [22] ———, *Prolongations of infinitesimal linear automorphisms of projective varieties and rigidity of rational homogeneous spaces of Picard number 1 under Kähler deformation*, Invent. math. **160** (2005), 591–645.
- [23] Hwang, J.-M. and Ramanan, S., *Hecke curves and Hitchin discriminant*, Ann. scient. Ec. Norm. Sup. **37** (2004), 801–817.
- [24] Iskovskikh, V. A. and Prokhorov, Yu. G., *Fano varieties (Algebraic geometry, V)*. Encyclopaedia Math. Sci. 47, Springer, Berlin, 1999.
- [25] Ivey, T. A. and Landsberg, J. M., *Cartan for beginners*. Graduate Studies in Math. 61, American Math. Soc., Providence, RI, 2003.
- [26] Kebekus, S., *Families of singular rational curves*, J. Algebraic Geom. **11** (2002), 245–256.
- [27] Kebekus, S. and Sola Conde, L., *Existence of rational curves on algebraic varieties, minimal rational tangents, and applications*, in Global aspects of complex geometry, Springer, Berlin, 2006, pp. 359–416.
- [28] Kollár, J., *Rational curves on algebraic varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete, 3 Folge, Band 32, Springer Verlag, 1996.
- [29] Landsberg, J. M. and Robles, C., *Lines and osculating lines of hypersurfaces*, J. Lond. Math. Soc. (2) **82** (2010), 733–746.
- [30] Lau, C.-H., *Holomorphic maps from rational homogeneous spaces onto projective manifolds*, J. Algebraic Geom. **18** (2009), 223–256.
- [31] Manin, Y. I., *Gauge Field Theory and Complex Geometry*, Grundlehren der mathematischen Wissenschaften 289, 2nd ed. Springer Verlag, 1997.
- [32] Miyaoka, Y. and Mori, S., *A numerical criterion for uniruledness*, Ann. of Math. **124** (1986), 65–69.
- [33] Mok, N., *The uniformization theorem for compact Kähler manifolds of nonnegative holomorphic bisectional curvature*, J. Differential Geom. **27** (1988), 179–214.
- [34] ———, *Recognizing certain rational homogeneous manifolds of Picard number 1 from their varieties of minimal rational tangents*, in Third International Congress of Chinese Mathematicians, 41–61, AMS/IP Stud. Adv. Math. 42, Amer. Math. Soc., Providence, RI, 2008.
- [35] ———, *Characterization of standard embeddings between complex Grassmannians by means of varieties of minimal rational tangents*, Sci. China Ser. A **51** (2008), 660–684.
- [36] ———, *Geometric structures on uniruled projective manifolds defined by their vari-*

- eties of minimal rational tangents*, *Asterisque* **322** (2008), 151–205.
- [37] Mori, S., *Projective manifolds with ample tangent bundles*, *Ann. of Math.* **110** (1979), 593–606.
- [38] Narasimhan, M. S. and Ramanan, S., *Geometry of Hecke cycles*, in I. C. P. Ramanujam-a tribute, 291–345, *Tata Inst. Fund. Res. Studies in Math.* 8, Springer, Berlin-New York, 1978.
- [39] Ochiai, T., *Geometry associated with semisimple flat homogeneous spaces*, *Trans. Amer. Math. Soc.* **152** (1970), 159–193.
- [40] Siu, Y.-T. and Yau, S.-T., *Compact Kähler manifolds of positive bisectional curvature*, *Invent. Math.* **59** (1980), 189–204.

Korea Institute for Advanced Study, Hoegiro 85, Seoul, 130-722, Korea

E-mail: [jmhwang@kias.re.kr](mailto:jmhwang@kias.re.kr)

# The structure of algebraic varieties

János Kollár

**Abstract.** The aim of this address is to give an overview of the main questions and results of the structure theory of higher dimensional algebraic varieties.

**Mathematics Subject Classification (2010).** 14E30, 14B05, 14D20.

**Keywords.** Algebraic variety, Mori program, moduli questions.

## 1. Early history: Euler, Abel, Jacobi, Riemann

Our story, like many others in mathematics, can be traced back at least to Euler who studied elliptic integrals of the form

$$\int \frac{dx}{\sqrt{x^3 + ax^2 + bx + c}}.$$

The study of integrals of algebraic functions was further developed by Abel and Jacobi. From our point of view the next major step was taken by Riemann. Instead of dealing with a multi-valued function like  $\sqrt{x^3 + ax^2 + bx + c}$ , Riemann looks at the complex algebraic curve

$$C := \{(x, y) : y^2 = x^3 + ax^2 + bx + c\} \subset \mathbb{C}^2.$$

Then the above integral becomes

$$\int_{\Gamma} \frac{dx}{y}$$

for some path  $\Gamma$  on the algebraic curve  $C$ . More generally, a polynomial  $g(x, y)$  implicitly defines  $y := y(x)$  as a multi-valued function of  $x$  and for any meromorphic function  $h(u, v)$ , the multi-valued integral

$$\int h(x, y(x)) dx$$

becomes a single valued integral

$$\int_{\Gamma} h(x, y) dx$$

for some path  $\Gamma$  on the algebraic curve  $C(g) := (g(x, y) = 0) \subset \mathbb{C}^2$ . Substitutions that transform one integral associated to a polynomial  $g_1$  into another integral associated to a  $g_2$  can be now seen as algebraic maps between the curves  $C(g_1)$  and  $C(g_2)$ .

Riemann also went further. As a simple example, consider the curve  $C$  defined by  $(y^2 = x^3 + x^2)$  and notice that  $(t^3 - t)^2 \equiv (t^2 - 1)^3 + (t^2 - 1)^2$ . Thus the substitution  $x = t^2 - 1$ ,  $y = t^3 - t$  (with inverse  $t = y/x$ ) allows us to transform any integral

$$\int h(x, \sqrt{x^3 + x^2}) dx \quad \text{into} \quad \int h(t^2 - 1, t^3 - t) \cdot 2t dt.$$

To put it somewhat differently, the map

$$t \mapsto (x = t^2 - 1, y = t^3 - t) \quad \text{and its inverse} \quad (x, y) \mapsto t = y/x$$

establish an isomorphism

$$\left\{ \begin{array}{l} \text{meromorphic functions} \\ \text{on the curve } (y^2 = x^3 + x^2) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} \text{meromorphic functions} \\ \text{on the complex plane } \mathbb{C} \end{array} \right\}.$$

It is best to work with meromorphic functions on  $\mathbb{C}$  that are also meromorphic at infinity; these live naturally on the Riemann sphere  $\mathbb{CP}^1$ . We can now state Riemann's fundamental theorem as follows.

**Theorem 1.1** (Riemann, 1851). *For every algebraic curve  $C \subset \mathbb{C}^2$  there is a unique, compact Riemann surface  $S$  and a meromorphic map  $\phi : S \dashrightarrow C$  with meromorphic inverse  $\phi^{-1} : C \dashrightarrow S$  such that*

$$f_C \mapsto f_S := f_C \circ \phi \quad \text{and} \quad f_S \mapsto f_C := f_S \circ \phi^{-1}$$

*establish an isomorphism between the meromorphic function theory on  $C$  and the meromorphic function theory on  $S$ .*

## 2. Main questions, informally

We can now give an initial formulation of the two main problems that we consider; the precise versions are stated in Sections 6 and 10. The first is a direct higher-dimensional analog of the results of Riemann. (See Section 3 for basic definitions.)

**Main Question 2.1.** *Given an algebraic variety  $X$ , is there another algebraic variety  $X^m$  such that*

1. *the meromorphic function theories of  $X$  and of  $X^m$  are isomorphic and*
2. *the geometry of  $X^m$  is the "simplest" possible?*

Riemann's theorem says that, in dimension 1, "simplest" should mean smooth and compact, but in higher dimensions smoothness is not the right notion. One of the hardest aspects of the theory was to understand what the correct concept of "simplest" should be.

So far we have dealt with individual algebraic varieties. A salient feature of algebraic geometry is that by continuously varying the coefficients of the defining polynomials we get continuously varying families of algebraic varieties. We can thus study how to transform a family  $\{X_t : t \in T\}$  of varieties into its "simplest" form. A tempting idea is to take the "simplest" forms  $\{X_t^m : t \in T\}$  obtained previously. Unfortunately, this fails already in

dimension 1. Starting with a family of curves  $\{C_t : t \in T\}$ , the corresponding Riemann surfaces  $\{S_t : t \in T\}$  form a continuously varying family over a dense open subset  $T^0 \subset T$  but not everywhere.

For curves the correct answer was found by Deligne and Mumford in 1969. We use the guidance provided by this 1-dimensional case and the answer to the first Main Question to answer the second.

**Main Question 2.2.** *What are the “simplest” families of algebraic varieties? How can one transform an arbitrary family into one of the “simplest” families?*

### 3. What are algebraic varieties?

Here we quickly recall the basic concepts and definitions that we use. For general introductory texts, see [30, 66, 73].

An *affine algebraic set* in  $\mathbb{C}^N$  is the common zero-set of some polynomials

$$\begin{aligned} X^{\text{aff}} &= X^{\text{aff}}(f_1, \dots, f_r) \\ &= \{(x_1, \dots, x_N) : f_1(x_1, \dots, x_N) = \dots = f_r(x_1, \dots, x_N) = 0\} \subset \mathbb{C}^N. \end{aligned}$$

It is especially easy to visualize *hypersurfaces*  $X(f) \subset \mathbb{C}^N$  defined by 1 equation. Usually we count complex dimensions, thus  $\dim \mathbb{C}^N = N$  and  $\dim X$  is one half of the usual topological dimension of  $X$ . In low dimensions we talk about *curves*, *surfaces*, *3-folds*. Thus, somewhat confusingly, an algebraic curve is a (possibly singular) Riemann surface.

An affine algebraic set  $X$  is called *irreducible* if it can not be written as a union of two algebraic sets in a nontrivial way. Such sets are called *affine algebraic varieties*. Every algebraic set  $X$  is a finite union of algebraic varieties  $X = \cup_i X_i$  such that  $X_i \not\subset X_j$  for  $i \neq j$ . Such a decomposition is unique, up to permuting the indices. Thus from now on we are interested mainly in algebraic varieties.

For example, the irreducible components of a hypersurface  $X(f)$  correspond to the irreducible factors of  $f$ , thus  $X(f)$  is irreducible iff  $f$  is a power of an irreducible polynomial.

An affine algebraic set  $X^{\text{aff}}$  is compact iff it is 0-dimensional, thus it is almost always better to work with the closure of  $X^{\text{aff}}$  in the complex projective space

$$X := X^{\text{proj}} \subset \mathbb{C}\mathbb{P}^N.$$

Thus we get *projective algebraic sets* and *projective varieties*. Finally, a *quasi-projective variety* is an open subset  $U$  of a projective variety  $X$  whose complement  $X \setminus U$  is a projective algebraic set. Note that  $U$  is a “very large” subset of  $X$ , in particular  $U$  is dense in  $X$ . This is a key feature of algebraic geometry: all open subsets are “very large.”

On a complex projective space  $\mathbb{C}\mathbb{P}^N$  the homogeneous coordinates  $(x_0 : \dots : x_N)$  are defined only up to multiplication by a scalar. Thus one can not evaluate a polynomial

$$p(x_0, \dots, x_N) \in \mathbb{C}[x_0, \dots, x_N],$$

at a point of  $\mathbb{C}\mathbb{P}^N$ . However, if  $p$  is homogeneous of degree  $d$  then

$$p(\lambda x_0, \dots, \lambda x_N) = \lambda^d p(x_0, \dots, x_N).$$

Thus the zero set of  $p$  is well-defined and a quotient of two homogeneous polynomials of the same degree

$$f(x_0, \dots, x_N) = \frac{p_1(x_0, \dots, x_N)}{p_2(x_0, \dots, x_N)}$$

is also well-defined (except where  $p_2$  vanishes). These are the *rational functions* on  $\mathbb{CP}^N$ . By restriction, we get rational functions on any projective variety  $X \subset \mathbb{CP}^N$ .

At first sight these seem downright antiquated definitions; a modern theory ought to be local. That is, one should consider varieties that are locally defined by analytic functions and work with meromorphic functions on them. However, we know that every meromorphic function on  $\mathbb{CP}^1$  is rational and the same holds in all dimensions.

**Theorem 3.1** (Chow, 1949; Serre, 1956). *Let  $M \subset \mathbb{CP}^N$  be a closed subset that can be locally given as the common zero set of analytic functions. Then*

- (1)  *$M$  is algebraic, that is, it can be globally given as the common zero set of homogeneous polynomials and*
- (2) *every meromorphic function  $f$  on  $M$  is algebraic, that is,  $f$  can be globally given as the quotient of two homogeneous polynomials of the same degree.*

Now we come to a key feature of algebraic geometry. There are two competing notions of “map” and two competing notions of “isomorphism.”

**Definition 3.2** (Map and morphism). Let  $X \subset \mathbb{CP}^N$  be an algebraic variety and  $f_0, \dots, f_M$  nonzero rational functions on  $X$ . They define a *map* (or rational map)

$$\mathbf{f} : X \dashrightarrow \mathbb{CP}^M \quad \text{given by} \quad p \mapsto (f_0(p) : \dots : f_M(p)) \in \mathbb{CP}^M.$$

To start with,  $\mathbf{f}$  is only defined at a point  $p$  if none of the  $f_i$  has a pole at  $p$  and not all of the  $f_i$  vanish at  $p$ . However, since the projective coordinates are defined only up to a scalar multiple,  $(gf_0, \dots, gf_M)$  define the same map for any rational function  $g$ , thus it can happen that  $\mathbf{f}$  is everywhere defined. In this case it is called a *morphism*. A map is denoted by  $\dashrightarrow$  and a morphism by  $\rightarrow$ .

For example, projecting  $\mathbb{CP}^2$  from the origin  $(0:0:1)$  to the line at infinity is given by

$$\pi : (x:y:z) \mapsto \left(\frac{x}{z} : \frac{y}{z}\right) = \left(\frac{x}{y} : 1\right) = \left(1 : \frac{y}{x}\right).$$

Thus  $\pi$  is defined everywhere except at  $(0:0:1)$ .

**Definition 3.3** (Isomorphism). Two quasi-projective varieties  $X \subset \mathbb{CP}^N$  and  $Y \subset \mathbb{CP}^M$  are *isomorphic* if there are morphisms

$$f : X \rightarrow Y \quad \text{and} \quad g : Y \rightarrow X$$

that are inverses of each other. Isomorphism is denoted by  $X \cong Y$ .

We will think of isomorphic varieties as being essentially the same. Using maps instead of morphisms in the above definition yields the notion of birational equivalence. This notion is unique to algebraic geometry; it has no known analog in topology or differential geometry.

**Definition 3.4** (Birational equivalence). Two quasi-projective varieties  $X \subset \mathbb{C}\mathbb{P}^N$  and  $Y \subset \mathbb{C}\mathbb{P}^M$  are *birational* (in old terminology, birationally isomorphic) if there are rational maps

$$f : X \dashrightarrow Y \quad \text{and} \quad g : Y \dashrightarrow X$$

such that the following equivalent conditions hold.

1.  $\phi_Y \mapsto \phi_X := \phi_Y \circ f$  and  $\phi_X \mapsto \phi_Y := \phi_X \circ g$  establish an isomorphism between the meromorphic (=rational) function theory on  $X$  and the meromorphic (=rational) function theory on  $Y$ .
2. There are algebraic subsets  $Z \subsetneq X$  and  $W \subsetneq Y$  such that  $(X \setminus Z) \cong (Y \setminus W)$ .

As an example, consider the affine surface  $S := (x^2 + y^2 = z^3) \subset \mathbb{C}^3$ . It is birational to  $\mathbb{C}_{uv}^2$  as shown by the rational maps

$$f : (x, y, z) \dashrightarrow \left(\frac{x}{z}, \frac{y}{z}\right) \quad \text{and} \quad g : (u, v) \rightarrow (u(u^2 + v^2), v(u^2 + v^2), u^2 + v^2).$$

Here  $f$  is not defined if  $z = 0$  while  $g$  is everywhere defined but it maps the pair of lines  $(u = \pm iv)$  to the origin  $(0, 0, 0)$ . Thus

$$S \setminus (z = 0) \cong \mathbb{C}^2 \setminus (u^2 + v^2 = 0) \quad \text{but} \quad S \not\cong \mathbb{C}^2.$$

**Basic rule of thumb 3.5.** Let  $X, Y$  be algebraic varieties that are birational to each other. Many questions of algebraic geometry about  $X$  can be answered by

- first studying the same question on  $Y$  and then
- studying a similar question involving the lower dimensional algebraic sets  $Z$  and  $W$  as in (3.4.2).

The aim of the Minimal Model Program is to exploit this in two steps.

- Given a question and a variety  $X$ , find a variety  $Y$  that is birational to  $X$  such that the geometry of  $Y$  is “best adapted” to studying the particular question. This is a variant of the first Main Question.
- Set up the appropriate dimension induction to deal with the exceptional sets  $Z \subset X$  and  $W \subset Y$ .

*Important aside.* More generally, if we decompose an algebraic variety into disjoint locally closed pieces, then the collection of the pieces carries a lot of information about the variety. I would like to stress that this is a rather noteworthy fact about algebraic geometry. For instance, if we decompose a simplicial complex into its simplices, then usually the only information we retain is the dimension and the Euler characteristic. By contrast, all the homology groups of a smooth, projective algebraic variety can be recovered from the pieces. This is a key consequence of Hodge Theory, as formulated by Deligne, and is a starting point of Grothendieck’s theory of motives.

### 4. Classical results

After the study of algebraic curves, two main avenues of investigations were pursued. One direction focused on the local study of varieties with a main aim of resolving them completely. The other direction aimed to understand the global structure of algebraic surfaces. These are both still very active research areas. We recall a few of the main results that are relevant for the general theory. For detailed treatments and for references see [6, 45].

**Resolution of singularities.** Riemann's theorem says that every singular algebraic curve  $C$  is birational to a smooth, compact curve (or Riemann surface). The first steps toward answering the Main Questions in higher dimensions focused on this problem: *Is every algebraic variety birational to a smooth, projective variety?*

**Definition 4.1.** A variety  $X \subset \mathbb{C}^N$  is smooth and has dimension  $d$  at a point  $p \in X$  iff the following equivalent conditions hold.

1.  $X \subset \mathbb{C}^N \cong \mathbb{R}^{2N}$  is a  $C^\infty$ -submanifold of (real) dimension  $2d$  near  $p$ .
2. One can choose coordinates  $z_1, \dots, z_N$  and equations  $f_1, \dots, f_{N-d}$  of  $X$  such that  $(f_1 = \dots = f_{N-d} = 0)$  coincides with  $X$  near  $p$  and the Jacobian matrix  $(\partial f_i / \partial z_j : 1 \leq i, j \leq N - d)$  is invertible at  $p$ .
3. There are holomorphic functions  $\phi_i = \phi_i(w_1, \dots, w_d)$  defined near the origin and constants  $c_i$  such that

$$(w_1, \dots, w_d) \mapsto (\phi_1(\mathbf{w}), \dots, \phi_{N-d}(\mathbf{w}), w_1 + c_1, \dots, w_d + c_d)$$

maps a small ball  $\mathbf{0} \in \mathbb{B}^d(\epsilon) \subset \mathbb{C}^d$  onto a neighborhood of  $p \in X$ .

In the latter case we view  $(w_1, \dots, w_d)$  as *local analytic coordinates* on  $X$  near  $p$ . (It is an ever present technical problem that there is no good notion of local algebraic coordinates. Open algebraic neighborhoods are too large to admit a single-valued coordinate system.)

On an algebraic variety  $X$  the set of singular points turns out to be an algebraic subset, denoted by  $\text{Sing } X \subset X$ . For every variety  $X$ , a generalization of Riemann's method (1.1) produces a new variety  $X^n \rightarrow X$ , called the *normalization* of  $X$ , such that  $\text{Sing}(X^n)$  has codimension  $\geq 2$  in  $X^n$ . Thus, in higher dimensions, one usually works with *normal varieties* whose singular set has codimension  $\geq 2$ .

To make the singular set even smaller, or to get rid of it completely, turned out to be very difficult. The final result was established by Hironaka in 1964.

**Theorem 4.2** (Resolution of singularities). *For every algebraic variety  $X$ , there are (very many) smooth, projective varieties  $X^{\text{sm}}$  birational to  $X$ .*

*If  $X$  is projective, one can arrange to have a morphism  $f : X^{\text{sm}} \rightarrow X$  that is an isomorphism over  $X \setminus \text{Sing } X$ .*

**Algebraic surfaces.** By resolution of singularities, any projective surface  $S$  is birational to a smooth projective surface  $S^{\text{sm}}$ , but, in contrast with the theory of curves, there are many such smooth projective surfaces  $S^{\text{sm}}$ . We can thus reformulate the first Main Question: Is there a "simplest" one among all smooth projective surfaces birational to  $S$ ?

To answer this question, first we study how to make a smooth projective surface more "complicated."

**Definition 4.3** (Blowing-up). Let  $S$  be a smooth algebraic surface and  $p \in S$  a point. *Blowing-up* is an operation that creates a new smooth surface  $B_p S$  by removing  $p$  and replacing it with a  $\mathbb{C}\mathbb{P}^1$  corresponding to all the tangent directions of  $S$  at  $p$ . Collapsing the new  $\mathbb{C}\mathbb{P}^1$  to a point gives a morphism  $\pi : B_p S \rightarrow S$ .

In local coordinates, it can be described as follows. Start with the unit ball  $\mathbb{B}_{xy}^2 \subset \mathbb{C}_{xy}^2$  and  $\mathbb{C}\mathbb{P}_{st}^1$  where the subscripts name the coordinates. Set

$$B_0 \mathbb{B}_{xy}^2 := (xt - ys = 0) \subset \mathbb{B}_{xy}^2 \times \mathbb{C}\mathbb{P}_{st}^1.$$

Let  $\pi : B_0\mathbb{B}_{xy}^2 \rightarrow \mathbb{B}_{xy}^2$  denote the coordinate projection. If  $(x, y) \neq (0, 0)$  then  $\pi^{-1}(x, y)$  is the single point  $(x, y) \times (s:t)$ . However, if  $x = y = 0$  then  $(s:t)$  can be arbitrary, thus  $\pi^{-1}(0, 0) \cong \mathbb{C}\mathbb{P}_{st}^1$ .

Note that  $s/t = x/y$  is the natural coordinate on  $\mathbb{C}\mathbb{P}_{st}^1$ , thus blowing up is akin to switching to polar coordinates since the polar angle  $\theta$  equals  $\tan^{-1}(x/y)$ .

One can blow up any number of points of  $S$  and then repeat by blowing up some of the new points of  $B_p S$ . Thus blowing up is a cheap way to get infinitely many new smooth surfaces out of one.

**Definition 4.4.** A smooth projective surface is called *minimal* if it can not be obtained from another smooth, projective surface by blowing up.

This notion allows us to get a very good analog of Riemann’s theorem 1.1.

**Theorem 4.5** (Enriques, 1914 ; Kodaira, 1966). *For every projective, algebraic surface  $S$ , exactly one of the following holds.*

1. (Minimal model) *There is a unique, minimal surface  $S^m$  birational to  $S$ .*
2.  *$S$  is birational to  $C \times \mathbb{C}\mathbb{P}^1$  for a unique, smooth, projective curve  $C$ .*

**4.6** (Du Val singularities). It was also gradually understood that instead of working with the minimal model  $S^m$ , it is sometimes better to use a slightly singular *canonical model*  $S^{\text{can}}$ . The resulting singularities were first classified by Du Val in 1934; the list is quite short, ranging from the simplest  $(x^2 + y^2 + z^2 = 0)$  to the most complicated  $(x^2 + y^3 + z^5 = 0)$ . They are also called *rational double points*.

Their importance was not generally recognized until the 1960’s when they were rediscovered from many different points of view; see [18] for a survey.

## 5. The first Chern class and the Ricci curvature

The first Chern class, which is closely related to the Ricci curvature, carries much of the important information about the structure of a variety. We follow the differential geometry sign conventions; algebraic geometers usually work with the *canonical class*, which is (a slight refinement of) the negative of the first Chern class.

**5.1** (Complex volume forms). A measure on  $\mathbb{R}^n$  can be identified with an  $n$ -form

$$s(x_1, \dots, x_n) \cdot dx_1 \wedge \dots \wedge dx_n.$$

Thus a measure on a real manifold  $M$  is an  $n$ -form that in local coordinates can be written as above.

Similarly, on a smooth variety  $X$  of dimension  $n$  a *complex volume form* is an  $n$ -form  $\omega$  that in local holomorphic coordinates can be written as

$$h(z_1, \dots, z_n) \cdot dz_1 \wedge \dots \wedge dz_n.$$

Thus a complex volume form  $\omega$  gives a real volume form  $\left(\frac{\sqrt{-1}}{2}\right)^n \omega \wedge \bar{\omega}$  where the constant comes from the formula

$$dz \wedge d\bar{z} = (dx + \sqrt{-1}dy) \wedge (dx - \sqrt{-1}dy) = -2\sqrt{-1} dx \wedge dy.$$

(There is usually an additional  $\pm$ , depending on one’s orientation conventions.)

From the point of view of differential geometry, one would like to use  $C^\infty$  complex volume forms, that is, the  $h(z_1, \dots, z_n)$  should be nowhere zero  $C^\infty$ -functions. Algebraic geometry, however, prefers meromorphic volume forms where the  $h(z_1, \dots, z_n)$  are meromorphic functions. (See (9.2.1) for some explicit examples.) Thus the ideal situation is when a complex volume form is given by nowhere zero holomorphic functions  $h(z_1, \dots, z_n)$ . This is possible only for Calabi–Yau varieties; they form a very special but important subclass (6.2).

Thus in general we try to understand how to connect  $C^\infty$  and meromorphic volume forms.

On the differential geometry side the key notion is the curvature which defines the Chern form.

**Definition 5.2** (Chern form and Chern class). Let  $\omega$  be a  $C^\infty$  complex volume form. The *first Chern form* or *Ricci curvature form* of  $(X, \omega)$  is the 2-form

$$\tilde{c}_1(X, \omega) := \frac{\sqrt{-1}}{\pi} \partial \bar{\partial} \log |h(z_1, \dots, z_n)| = \frac{\sqrt{-1}}{\pi} \sum_{ij} \frac{\partial^2 \log |h(\mathbf{z})|}{\partial z_i \partial \bar{z}_j} dz_i \wedge d\bar{z}_j.$$

As a 2-form, this depends on the choice of the volume form  $\omega$ , but it gives a well defined De Rham cohomology class  $c_1^{\mathbb{R}}(X) \in H_{DR}^2(X, \mathbb{R})$  which actually lifts to an integral cohomology class

$$c_1(X) \in H^2(X, \mathbb{Z}),$$

called the *first Chern class* of  $X$ .

**Definition 5.3** (Algebraic degree). Let  $X$  be a smooth, projective variety and  $C \subset X$  an algebraic curve. It is not hard to see that there is always a meromorphic volume form  $\omega_m$  that is defined and nonzero at all but finitely many points of  $C$ . We define the *degree* of  $\omega_m$  on  $C$  as

$$\deg_C \omega_m := \#(\text{zeros of } \omega_m \text{ on } C) - \#(\text{poles of } \omega_m \text{ on } C),$$

where both zeros and poles are counted with multiplicities.

The Chern form and the algebraic degree are connected by the Gauss–Bonnet theorem.

**Theorem 5.4.** *Let  $X$  be a smooth, projective variety. Let  $\omega_r$  be a  $C^\infty$  complex volume form and  $\omega_m$  a meromorphic volume form. Then, for every algebraic curve  $C \subset X$*

$$\int_C c_1(X) = \int_C \tilde{c}_1(X, \omega_r) = -\deg_C \omega_m. \quad (5.4.1)$$

(The minus sign comes from the happenstance that differential geometers prefer to work with the tangent bundle while the volume forms use the (determinant of the) cotangent bundle.)

**Positivity/negativity and complex differential geometry.** In differential geometry it is especially nice to work with metrics whose curvature is everywhere positive (or everywhere zero or everywhere negative) but these rarely exist. A usual weakening is to work with Kähler metrics that satisfy the *Einstein condition*: the Ricci curvature should be a constant multiple of the metric; see [64, Chap.19] for definitions and an introduction.

If this *Einstein constant* is positive, then in (5.4.1) we integrate an everywhere positive form. Thus  $\int_C c_1(X)$  is positive for every curve  $C$ . We hope that in this case there are meromorphic volume forms with poles (but no zeros).

Similarly, if the Einstein constant is negative, then in (5.4.1) we integrate an everywhere negative form. Thus  $\int_C c_1(X)$  is negative for every curve  $C$ . We hope that in this case there are holomorphic volume forms (usually with zeros).

Algebraic geometry can be used to understand the numbers  $\deg_C \omega_m$ , hence the values of the integrals  $\int_C c_1(X)$ . It is a very difficult task to use the positivity/negativity of the integrals  $\int_C c_1(X)$  to obtain a Kähler metric with positive/negative Einstein constant.

For smooth varieties Aubin and Yau proved existence in 1977 when  $\int_C c_1(X)$  is always negative or when  $\int_C c_1(X)$  is always zero. The singular case is treated in [7, 20]. The positive curvature case is more subtle; a complete answer is not yet known.

While our approach to the structure of varieties is guided by these curvature considerations, in algebraic geometry we can understand only the algebraic degree of the first Chern class. Thus we look at the functional

$$C \mapsto \int_C c_1(X)$$

and focus on those varieties where this is everywhere negative (or everywhere zero or everywhere positive).

The Main Conjecture then asserts that every variety can be built up from these special varieties in a rather clear process.

## 6. The main conjecture

On a typical variety  $X$ , the Chern class  $c_1(X)$  is positive on some curves and negative on others, in a rather unpredictable way. Using the first Chern class and the theory of algebraic surfaces as our guide, we focus on three basic “especially simple” types of smooth, projective varieties. These are the “building blocks” of all algebraic varieties.

**6.1 (Negatively curved).** These are the varieties where  $\int_C c_1(X)$  is negative for every curve  $C \subset X$ . This is the largest class of the three.

**6.2 (Flat or Calabi–Yau).** Here  $\int_C c_1(X)$  is zero for every curve  $C \subset X$ . They play an especially important role in string theory and mirror symmetry; see [32, 77] for introductions.

**6.3 (Positively curved or Fano).** Here  $\int_C c_1(X)$  is positive for every curve. There are few of these varieties in each dimension, but they occur especially frequently in applications.

A simple set of examples to keep in mind is the following. A smooth hypersurface  $X_d \subset \mathbb{C}P^n$  of degree  $d$  is negatively curved if  $d > n + 1$ , flat if  $d = n + 1$  and positively curved if  $d < n + 1$ .

A variety in any of these 3 classes is considered “simplest,” but we do not yet have enough “simplest” varieties for answering the first Main Question. For example, taking products of these we get examples where  $c_1(X)$  has different signs on different curves. Two of these possible “mixed types” are relevant for us.

Consider a product  $X := N \times F$  of a negatively curved and of a flat variety. It is clear that  $\int_C c_1(X) \leq 0$  for every curve  $C \subset X$  and  $\int_C c_1(X) = 0$  only if  $C$  lies in a fiber of the first projection  $N \times F \rightarrow N$ . This observation leads to the 4th class.

**6.4 (Semi-negatively curved or Kodaira–Iitaka type).** Here  $\int_C c_1(X) \leq 0$  for every curve  $C \subset X$  and there is a unique morphism  $I_X : X \rightarrow I(X)$  such that  $\int_C c_1(X) = 0$  iff  $C$  is contained in a fiber of  $I_X$ .

This includes the classes 6.1–6.2:  $I_X$  is an isomorphism for negatively curved varieties and a constant map in the flat case.

In the intermediate cases, when  $0 < \dim I(X) < \dim X$ , almost all fibers of  $I_X$  are Calabi–Yau varieties. Thus one can view these as families of (lower dimensional) Calabi–Yau varieties parametrized by the (lower dimensional) variety  $I(X)$ . If we understand families of (lower dimensional) varieties well enough, we understand  $X$ . (This is one of the reasons we are interested in the second Main Question.) Furthermore, in these cases  $I(X)$  is negatively curved in a “suitable sense,” though we do not yet have a final agreed-upon definition of what this means.

Next consider a product  $X := N \times P$  of a negatively curved and of a positively curved variety. If a curve  $C$  lies in a fiber of the first projection then  $\int_C c_1(X) > 0$ , but there are many other such curves. Nonetheless, the first projection is uniquely determined by  $X$  and this leads to the definition of the 5th class.

**6.5 (Positive fiber type).** I really would like to say that in these cases there is a unique morphism  $m_X : X \rightarrow M(X)$  such that  $M(X)$  is semi-negatively curved and  $c_1(X)$  is positive on all the fibers. (To avoid trivial cases, we also assume that  $\dim M(X) < \dim X$ .) This, unfortunately, still does not give enough “simplest” varieties for the first Main Question. It took quite some time to arrive at the correct definition, to be discussed in Section 7.

We can now state a precise version of the first Main Question.

**Main Conjecture 6.6.** *Every algebraic variety  $X$  is birational to a variety  $X^m$  that is either of type (6.4) or of type (6.5).*

**Complement.**  $X^m$  – especially in case (6.4) – is called a *minimal model* of  $X$ .

In the semi-negatively curved case  $I(X^m)$  is unique but  $X^m$  itself is not. However, it is quite well understood how the different  $X^m$  are related to each other. (This is the story of *flops*, see [27, 37].) By contrast, in case (6.5) it is very hard to determine when two such varieties  $X_1^m$  and  $X_2^m$  are birational.

**Caveat.** While the Main Conjecture is expected to be true, in general one has to allow *terminal* singularities – to be defined in (9.3) – on  $X^m$ .

This was a rather difficult point historically since over a century of experience suggested that singularities should be avoided. For surfaces terminal = smooth, thus the issue of singularities did not come up in Theorem 4.5.

By now the correct classes of singularities have been established and, for many questions we consider, they do not seem to cause any problems. We describe these singularities in Section 9.

**6.7 (Traditional names).** A variety  $X$  is said to be of *general type* if  $\dim I(X^m) = \dim X$ . In this case  $X \dashrightarrow I(X^m)$  is birational and  $I(X) := I(X^m)$  is called the *canonical model* of  $X$ ; it has canonical singularities (9.4). We see in Section 10 that the second Main Question has a good answer for families of canonical models.

The *Kodaira dimension* of a variety  $X$  is the dimension of  $I(X^m)$ .

The Kodaira dimension is defined to be  $-\infty$  for the class (6.5).

The Main Conjecture is usually broken down into two parts that are, in principle, independent of each other. The first part separates the classes 6.4 and 6.5 from each other and the second part provides the structural description in case 6.4. These forms first appear in Reid’s paper [71, Sec.4].

**6.7.1 Minimal Model Conjecture.** Every algebraic variety  $X$  is birational to a variety  $X^m$  such that either  $c_1(X^m)$  is semi-negative or there is a morphism to a lower dimensional variety  $\pi : X^m \rightarrow S$  such that  $\int_C c_1(X^m) > 0$  if  $C$  is contained in a fiber of  $\pi$ . (In the second case the map  $\pi$  need not be unique and it does not give the best structural description.)

**6.7.2 Abundance Conjecture.** If  $c_1(Y)$  is semi-negative then there is a unique morphism  $I_Y : Y \rightarrow I(Y)$  such that  $\int_C c_1(Y) = 0$  iff  $C$  is contained in a fiber of  $I_Y$ .

## 7. Rationally connected varieties

Before we consider minimal models, we describe the structure we expect for varieties in the 5th class (6.5). An introduction aimed at non-specialists is given in [43]. More detailed accounts are in [5, 39].

Clebsch and Max Noether noticed around 1860–1870 that, when the numerical invariants suggest that a surface could be birational to  $\mathbb{C}P^2$ , then it is. The final result along these lines was established by Castelnuovo in 1896.

Analogous questions in higher dimension turned out to be much harder. Fano classified smooth positively curved 3–folds around 1930. (He missed some cases though, so did subsequent “complete” lists produced in the 1970’s and then in the 1980’s. The (hopefully) final list was not established until 2003.) This is, however, one area where the singularities do matter; we still do not know all positively curved 3–folds with terminal singularities.

It appears that instead of global descriptions we should focus on *rational curves* in a variety; these are the images of morphisms  $\phi : \mathbb{C}P^1 \rightarrow X$ . For a projective variety  $X$ , the following dichotomy is quite easy to establish.

- i) either the rational curves cover a subset of  $X$  which is *meager* (that is, a countable union of nowhere dense closed subsets)
- ii) or the rational curves cover all of  $X$ .

These two cases correspond to the alternatives in the Main Conjecture. That is, if  $X$  is birational to a semi-negatively curved variety then rational curves cover a meager subset and, conjecturally, the converse also holds.

The correct approach to the best structural description of the 5th class 6.5 was not discovered until 1992 (Kollár–Miyaoka–Mori [50]). The key observation is that we should even change the class 6.3. Instead of a curvature description, we should focus on rational curves contained in a variety.

**Definition 7.1.** A projective variety  $X$  is called *rationally connected* if, for any number of points  $x_1, \dots, x_r \in X$ , there is a morphism  $\phi : \mathbb{C}P^1 \rightarrow X$  whose image passes through  $x_1, \dots, x_r$ .

I claim that rationally connected varieties constitute the “correct” birational version of being positively curved. This is not a precise mathematical assertion since not every ratio-

nally connected variety is birational to a positively curved variety, not even when singularities are allowed. Rather, the assertion is that any answer to the first Main Question needs to work with rational connectedness instead of positivity of curvature.

**7.2** (Supporting evidence). It is easy to see that  $\mathbb{C}\mathbb{P}^n$  is rationally connected. More generally, every positively curved variety is rationally connected (Nadel [67], Campana [10], Kollár–Miyaoka–Mori [49], Zhang [80]).

Being rationally connected is invariant under smooth deformations and birational maps [49].

Rationally connected varieties share key arithmetic properties of rational varieties over  $p$ -adic fields (Kollár [42]), finite fields (Kollár–Szabó [54], Esnault [19]) and function fields of curves (Graber–Harris–Starr [22], de Jong–Starr [16]).

The loop space of a rationally connected variety is also rationally connected (Lempert–Szabó [59]).

The notion of rational connectedness allows us to give the correct description of the class 6.5. A weaker variant is proved in [50]; the form below combines this with [22].

**Theorem 7.3.** *Let  $X$  be a variety that is covered by rational curves. Then there is a unique (up to birational equivalence) map  $m_X : X \dashrightarrow M(X)$  such that*

- (1) *almost all fibers of  $m_X$  are rationally connected and*
- (2) *rational curves cover only a meager subset of  $M(X)$ .*

There are two main open geometric problems about rationally connected varieties. The first concerns a topological characterization. In its naive form the question asks: What can we tell about a variety from its underlying topological space? It seems that the answer is: not much. However, the underlying topological space of a smooth variety carries a natural symplectic structure and this seems to incorporate much more information.

**Conjecture 7.4** ([41, Conj.4.2.7]). *Being rationally connected is a property of the underlying symplectic structure.*

For partial results see [41, 78].

The other problem asks if we could strengthen the definition of rationally connected varieties. Note that  $\mathbb{C}\mathbb{P}^n$  contains not just many rational curves but also many higher dimensional rational subvarieties (hyperplanes, hyperquadrics, ...). Maybe this is also a general property of rationally connected varieties? As far as I know, 3-dimensional rationally connected varieties always contain rational surfaces. I believe, however, that this is not the case in higher dimension.

**Conjecture 7.5** ([43, Prob.56]). *Many rationally connected varieties do not contain any rational surface.*

## 8. Minimal models

This is a short history of *Mori's program*, also called the *Minimal Model Program* and frequently abbreviated as *MMP*. For general introductions see [12, 52] or the technically more detailed [13, 25, 33].

**8.1 (Iitaka’s program, 1970–85).** This approach predates the Main Conjecture. At the beginning it was not even suspected that the Main Conjecture could be true, in fact, lacking the right class of singularities, it was assumed that the Main Conjecture would fail for most varieties. Thus the aim of Iitaka’s program was to sort varieties into 5 broad types that (as we now know) exactly correspond to the ones in (6.1–6.5). The main contributors were, in rough historical order, Iitaka, Ueno, Fujita, Kawamata, Viehweg and Kollár; see [62, 74] for surveys.

**8.2 (Canonical and terminal singularities, Reid 1980–83).** Reid was studying higher dimensional analogs of Du Val singularities of surfaces (4.6); obtaining rather complete descriptions in dimension 3. It was quite important that when Mori’s program lead to singularities, the relevant classes were already there and were known to be well behaved. An especially readable account is [72].

**8.3 (The birth of Mori’s program, 1981–88).** Mori’s groundbreaking paper [61] introduces 3 new ideas.

If  $c_1(X)$  is not semi-negative then, by definition,  $c_1(X)$  is positive on some curve  $C \subset X$ . Mori first proves that there is such a rational curve; that is, there is a morphism  $\phi : \mathbb{C}P^1 \rightarrow X$  such that  $c_1(X)$  is positive on its image. It is quite remarkable that the proof goes through algebraic geometry over finite fields. To this day there is no proof known that avoids this; in particular this step is not yet known for complex manifolds that are not algebraic.

Second, he identifies the “most positive” such maps  $\phi : \mathbb{C}P^1 \rightarrow X$ ; this is called *extremal ray theory*.

Third, in dimension 3 he gives a complete description of all extremal rays and the resulting map  $X \rightarrow X_1$  that removes the “most positive” part of  $X$ .

The program now seems clear (at least in dimension 3). Repeat the procedure for  $X_1$  and prove that after finitely many steps we end up with  $X \rightarrow X_1 \rightarrow \dots \rightarrow X_r$  such that  $c_1(X_r)$  is semi-negative. This is called *Mori’s program* or *Minimal Model program*.

There are two, rather formidable, problems. In many cases the new variety  $X_1$  is smooth but sometimes it is singular. Luckily, these singularities have been studied by Reid, at least in dimension 3. Still, it is necessary to establish the above 3 steps for singular varieties. This was accomplished rather rapidly by Kawamata, Reid, Shokurov and Kollár. The program was first written down in [71, Sec.4].

The more serious problem is that in some cases taking the contraction  $X_i \rightarrow X_{i+1}$  is clearly not the right step. Instead we have to take a step back and construct a new variety  $X_i^+$  that sits in a *flip diagram*

$$\begin{array}{ccc}
 X_i & \overset{\phi_i}{\dashrightarrow} & X_i^+ \\
 \searrow p_i & & \swarrow p_i^+ \\
 & X_{i+1} &
 \end{array}$$

Geometrically, we start with  $X_i$ , find an especially badly behaving  $\mathbb{C}P^1 \cong C_i \subset X_i$  and remove it. Then we compactify the resulting  $X_i \setminus C_i$  by attaching another curve  $C_i^+ \cong \mathbb{C}P^1$  but differently. The key difference is a sign change:

$$\int_{C_i} c_1(X_i) > 0 \quad \text{but} \quad \int_{C_i^+} c_1(X_i^+) < 0.$$

This operation is called a *flip*. For more about flips, see [27, 37].

Flips are reminiscent of *Dehn surgery* in 3-manifold topology where we remove a circle and put it back differently.

In dimension 3 the existence of flips is proved in a very difficult paper by Mori [63], which completes the program in this case. A detailed description of 3-dimensional flips is given in [51]. The list is rather lengthy; this makes it unlikely that a similarly complete answer will ever be worked out in higher dimensions.

**8.4 (Log variants: Kawamata, Shokurov, 1984–1992).** The Iitaka program established that for many results one can work with cohomology classes in  $H^2(X, \mathbb{R})$  that are close enough to the first Chern class. This turned out to be a very powerful tool. By choosing the perturbations appropriately, we can focus our attention on one or another part of a variety. These are somewhat technical questions but by now we understand how to work with them and most applications of the Minimal Model Program use a perturbed case.

**8.5 (Abundance: Kawamata, Miyaoka, 1987–1992).** Even for surfaces, the Abundance Conjecture 6.7.2 is a rather subtle result. It is even harder for 3-folds. The proofs use many special properties of surfaces; this is why the higher dimensional cases are still not well understood. A rather complete account of the 3-dimensional methods is given in [38].

**8.6 (Inductive approach in low dimensions: Shokurov, 1992–2003).** In retrospect, the key development of the decade was an inductive approach to flips. A detailed treatment of the 3-dimensional case is given in [38]. For the rest of the nineties progress was slow, culminating in a treatment of 4-dimensional flips. There were many technical difficulties to overcome and the importance of these methods was not fully appreciated at first since the dimension reduction leads to a much more complicated problem that seems to fail in higher dimensions.

**8.7 (The Corti seminar, 2003–2005).** Over the course of several years a group led by Corti developed the previous ideas further and integrated them with the rest of the program [13]. This provided the bridge to the general case.

**8.8 (The general type case: Hacon and McKernan, 2005–2010).** The real breakthrough was achieved in [26] where the existence of flips in dimension  $n$  was reduced to an instance of the MMP in dimension  $n - 1$ . This left a series of global questions to resolve. The paper [9] settled everything for varieties of general type. A good introduction is in [14].

At about the same time Siu started to develop an analytic approach which aims to get  $I(X^m)$ , without going through the individual steps; see [70] for an overview. An algebraic variant of this is in [11].

**8.9 (Abundance: Hacon and Xu, 2012–).** Although the Abundance conjecture is known in very few cases, there has been significant progress when  $\dim I(X)$  is expected to be close to  $\dim X$ . The log version of the special case when  $\dim I(X) = \dim X$  is especially important for applications in moduli theory. These have been settled in [28, 68].

**8.10 (Positive characteristic, Hacon and Xu, Birkar, Patakfalvi, 2012–).** Mori's original works are very geometric and these ideas quickly lead to a simple proof of the 2-dimensional case of the Main Conjecture in positive characteristic. However, subsequent developments rely very heavily on Kodaira-type vanishing theorems that are known to fail in positive characteristic, although no actual failure is known in the cases used by the program. The 3-dimensional case was recently settled in [8, 29]. Substantial parts of the Iitaka program are proved in positive characteristic in [69].

**8.11 (Open problems).** From our point of view, the main open problem is to complete the missing parts of the Main Conjecture.

It is known that the MMP always runs, that is, the sequence of contractions and flips  $X = X_1 \dashrightarrow X_2 \dashrightarrow \dots$  exists. The problem is that it is not clear how to prove that the process eventually stops. In the 3–dimensional case, Mori’s approach provides a rather complete description of the steps of the MMP. This gives many ways to show that each step improves various invariants and that eventually the process stops. By contrast, the method of Hacon–McKernan produces the steps of the MMP in a rather indirect way. We have very little information about the steps beyond their existence.

### 9. Singularities of the minimal model program

So far we have been sweeping the singularities of the minimal models under a rug, but it is time for a look at them. Understanding the correct class of singularities is crucial in the development of the structure theory of algebraic varieties. This is a somewhat technical subject with many difficult questions and methods but by now we understand these singularities well enough that in many questions they do not cause any problems. A rather complete treatment is given in [47]. Here I focus on the main ideas behind the definitions.

Given a variety  $Y$ , one frequently looks at a resolution of singularities  $f : X \rightarrow Y$  as in Theorem 4.2 and translates problems on  $Y$  to questions on  $X$ . Then the hard part is to interpret the answer obtained on  $X$  in terms of  $Y$ . Here the key seems to be the inverse function theorem.

**9.1 (The inverse function theorem).** The classical inverse function theorem says that if  $\mathbf{f} := (f_1, \dots, f_n) : \mathbb{R}^n_x \rightarrow \mathbb{R}^n_y$  is a differentiable map then  $\mathbf{f}$  has a local inverse at a point  $p \in \mathbb{R}^n_x$  iff the Jacobian determinant

$$\text{Jac}(\mathbf{f}) := \det \left( \frac{\partial f_i}{\partial x_j} \right)$$

does not vanish at  $p$ . We can also think about it in terms of the “standard” volume forms  $\omega_x := dx_1 \wedge \dots \wedge dx_n$  and  $\omega_y := dy_1 \wedge \dots \wedge dy_n$ . Then

$$\mathbf{f}^* \omega_y = \text{Jac}(\mathbf{f}) \cdot \omega_x,$$

thus the vanishing/non-vanishing of the Jacobian tells us how the pull-back of the “standard” volume form of the target compares to the “standard” volume form of the source.

Note that the Jacobian itself depends on the choice of the coordinates, but its vanishing or non-vanishing depends only on  $\mathbf{f}$ .

In the complex analytic setting one can use the “standard” complex volume forms  $\omega_z := dz_1 \wedge \dots \wedge dz_n$  and  $\omega_w := dw_1 \wedge \dots \wedge dw_n$  on the unit balls  $\mathbb{B}^n_z \subset \mathbb{C}^n_z$  and  $\mathbb{B}^n_w \subset \mathbb{C}^n_w$ . Given a holomorphic map  $\mathbf{f} := (f_1, \dots, f_n) : \mathbb{B}^n_z \rightarrow \mathbb{B}^n_w$  we get that

$$\mathbf{f}^* \omega_w = \det \left( \frac{\partial f_i}{\partial z_j} \right) \cdot \omega_z =: \text{Jac}(\mathbf{f}) \cdot \omega_z,$$

and  $\mathbf{f}$  has a local inverse iff  $\text{Jac}(\mathbf{f})$  does not vanish at  $p$ .

**9.2 (The Jacobian in the singular case).** Let  $X$  be a normal algebraic variety and  $p \in X$  a singular point. It is quite easy to see that if  $\omega_1, \omega_2$  are two holomorphic volume forms on  $X \setminus \text{Sing } X$  in a neighborhood of a singular point  $p \in \text{Sing } X$  then there is a unique holomorphic function  $\phi$  such that  $\omega_1 = \phi \cdot \omega_2$  and  $\phi(p) \neq 0$ . Thus all holomorphic volume forms on  $X \setminus \text{Sing } X$  have the same asymptotic behavior near  $\text{Sing } X$ . The local existence of such forms is a slightly technical question, so let us just focus on an example. If  $Y = (f(w_1, \dots, w_{n+1}) = 0) \subset \mathbb{C}^{n+1}$  is a hypersurface then the “standard” volume form is given by

$$\omega_Y = (-1)^i \frac{dw_1 \wedge \dots \wedge dw_{i-1} \wedge dw_{i+1} \wedge \dots \wedge dw_{n+1}}{\partial f / \partial w_i}. \tag{9.2.1}$$

(It is easy to check that this is independent of  $i$ . Note also that  $\omega_Y$  is not defined when all of the  $\partial f / \partial w_i$  vanish; which happens exactly on  $\text{Sing } Y$ .) Thus if  $f : \mathbb{B}_{\mathbb{Z}}^n \rightarrow Y$  is holomorphic then we can define the *Jacobian* of  $f$  by the formula

$$\text{Jac}(f) := \frac{f^* \omega_Y}{\omega_{\mathbb{Z}}}.$$

Note that due to the denominators in (9.2.1), in general  $\text{Jac}(f)$  can have poles.

For example, consider the singularity  $Y_{n,d} := (w_1^d + \dots + w_n^d = w_{n+1}^d) \subset \mathbb{C}^{n+1}$  and a holomorphic map  $f : \mathbb{B}_{\mathbb{Z}}^n \rightarrow Y$  given by

$$f : (z_1, \dots, z_n) \rightarrow (z_1, z_1 z_2, \dots, z_1 z_n, z_1 \sqrt[d]{1 + z_2^d + \dots + z_n^d}). \tag{9.2.2}$$

Then  $\omega_{Y_{d,n}} = -d^{-1} w_1^{1-d} dw_2 \wedge \dots \wedge dw_{n+1}$  and we easily compute that the Jacobian of  $f$  has a zero/pole of order  $n - d$  along the hyperplane  $(z_1 = 0)$ .

As in the classical case, the Jacobian of  $f$  depends on the choice of the “standard” volume forms but the vanishing/non-vanishing or the order of vanishing of the Jacobian depends only on  $f$ .

We can now define terminal singularities; these form the smallest possible class needed for the Main Conjecture.

**Definition 9.3.** A normal variety  $Y$  has *terminal* singularities iff the inverse function theorem holds for  $Y$ . That is, if  $f : \mathbb{B}_{\mathbb{Z}}^n \rightarrow Y$  does not have a local inverse at  $p \in \mathbb{B}_{\mathbb{Z}}^n$  then  $\text{Jac}(f)$  vanishes at  $p$ . (There is a small problem when the exceptional set of  $f$  is too small, we can ignore it for now.)

For canonical models and for moduli questions, two more types of singularities are needed.

**Definition 9.4.** A normal variety  $Y$  has *canonical* singularities iff  $\text{Jac}(f)$  is holomorphic for every  $f : \mathbb{B}_{\mathbb{Z}}^n \rightarrow Y$  and *log-canonical* singularities iff  $\text{Jac}(f)$  has at most simple poles for every  $f$ .

The above computations suggest (and it is indeed true) that  $Y_{n,d}$  (as in 9.2.2) is terminal iff  $d < n$ , canonical iff  $d \leq n$  and log canonical iff  $d \leq n + 1$ .

**9.5 (Local volume of  $Y$  near  $\text{Sing } Y$ ).** A good way to think about these singularities is as follows. Pick a point  $p \in \text{Sing } Y$  and let  $\omega_Y$  be a “standard” local complex volume form. Then  $(\sqrt{-1}/2)^n \omega_Y \wedge \bar{\omega}_Y$  is a real volume form and we can ask about the *local volume of  $X$* , that is,  $\int_U (\sqrt{-1}/2)^n \omega_Y \wedge \bar{\omega}_Y$  for a suitably small neighborhood  $p \in U \subset X$ .

If  $Y$  has a canonical singularity near  $p$  then the local volume is finite. In the log-canonical case the local volume is infinite but barely. If  $g$  is any holomorphic function vanishing on  $\text{Sing } Y$  then  $\int_U |g|^\epsilon (\sqrt{-1}/2)^n \omega_Y \wedge \bar{\omega}_Y$  is finite for every  $\epsilon > 0$ .

**9.6 (Intermediate differential forms).** On an  $n$ -dimensional variety we have so far considered holomorphic  $n$ -forms only but for several questions one also needs to understand the pull-back  $f^*\eta$  of lower degree differential forms as well. This proved to be surprisingly difficult but almost all local questions were settled by Greb–Kebekus–Kovács–Peternell [23].

## 10. Moduli of varieties of general type

Let  $\mathbf{X}$  be a class of projective varieties, for instance curves or surfaces of a certain type. The theory of moduli aims to find “optimal” ways to write down all varieties in the class  $\mathbf{X}$ .

This is a large theory with many aspects. The 3 volumes of [21] contain surveys of most of the active areas. Here my aim is to focus on just one of them: the moduli of varieties of general type. Introductions are given in [25, 46, 57] while a detailed treatment should be in [48].

We start with the historically first example.

**Example 10.1** (Elliptic curves). They can all be given by an affine equation

$$E(a, b, c) := (y^2 = x^3 + ax^2 + bx + c) \subset \mathbb{C}^2;$$

the corresponding projective curve has a unique point  $[p]$  at infinity. Here  $c_1(E) = 0$ , so it is best to think of this as elliptic curves with a marked point  $[p]$ . The curve  $E(a, b, c)$  is smooth iff the discriminant of the cubic

$$\Delta(a, b, c) := 18abc - 4a^3c + a^2b^2 - 4b^3 - 27c^2 \quad \text{is not zero.}$$

Two such curves are isomorphic iff there is an affine-linear transformation  $(x, y) \mapsto (\alpha^2x + \beta, \alpha^3y)$  that transforms one equation into the other. All these transformations form a (2-dimensional) group  $G$ . Thus we get the following.

*Version 1.* The isomorphism classes of all elliptic curves are in one-to-one correspondence with the orbits of  $G$  on  $\mathbb{C}^3 \setminus (\Delta(a, b, c) = 0)$ .

Next we need to identify the  $G$ -orbits. The key is the  $j$ -invariant  $j(E(a, b, c)) := 2^8(a^2 - 3b)^3/\Delta(a, b, c)$ . (The factor  $2^8$  is not important for us, it is there for number-theoretic reasons.) It is not very hard to work out the following.

*Version 2.* Two elliptic curves are isomorphic iff they have the same  $j$ -invariant.

We can restate this as follows:

*Version 3.* The moduli space of elliptic curves is the complex line  $\mathcal{M}_1 \cong \mathbb{C}$  and the value  $j(E)$  of the  $j$ -invariant gives the point in  $\mathcal{M}_1$  that corresponds to  $E$ .

The only sensible compactification of  $\mathbb{C}$  is  $\mathbb{CP}^1$ , so what corresponds to the point at infinity? This should be a curve where the discriminant of the cubic  $x^3 + ax^2 + bx + c$  vanishes. That is, when  $x^3 + ax^2 + bx + c$  has a multiple root. There are 2 types of such cubics. If there is a triple root we get  $y^2 = x^3$ , a cuspidal curve. If there is a double root we get  $y^2 = x^3 + x^2$ , a nodal curve. In this case the correct choice is to go with the nodal curve.

**10.2** (The main steps of a moduli theory). We hope to do something similar with more general algebraic varieties. We proceed in several steps.

*Step 1.* Identify a class of projective varieties  $\mathbf{X}$  that should have a “good” moduli theory. We aim to prove that such a theory exists for negatively curved varieties as in (6.1). We allow canonical singularities, thus this includes canonical models of varieties of general type. (It seems that in most other cases there is no “good” compactified moduli theory, unless some additional structure is added on, for instance an ample divisor as in [2].)

*Step 2.* Add some extra data (also called rigidification) first. A typical extra datum is an embedding  $j : X \hookrightarrow \mathbb{P}^N$  for some  $N$ . Use the additional data to get a moduli space with a universal family

$$\mathbf{U}_{\mathbf{X},j} \subset \mathbb{P}^N \times \mathbf{M}_{\mathbf{X},j} \quad \text{with projection} \quad \pi_{\mathbf{X},j} : \mathbf{U}_{\mathbf{X},j} \rightarrow \mathbf{M}_{\mathbf{X},j}$$

such that every pair  $(X, j)$  occurs exactly once among the fibers of  $\pi_{\mathbf{X},j}$ . (It is not easy to show that one can choose a fixed  $N$  that works for all varieties in a given class. For smooth varieties this was proved by Matsusaka in 1972; the general case was settled recently by Hacon and Xu.)

*Step 3.* Next we get rid of the extra data. Usually we have to take a quotient by a Lie group like  $\mathrm{GL}(N+1, \mathbb{C})$ . This can be hard but, if everything works out, at the end we have

$$\mathbf{U}_{\mathbf{X}} := \mathbf{U}_{\mathbf{X},j}/\mathrm{GL}(N+1, \mathbb{C}), \quad \mathbf{M}_{\mathbf{X}} := \mathbf{M}_{\mathbf{X},j}/\mathrm{GL}(N+1, \mathbb{C})$$

and a morphism  $\pi_{\mathbf{X}} : \mathbf{U}_{\mathbf{X}} \rightarrow \mathbf{M}_{\mathbf{X}}$ . (See Step 6 for the possible dependence on  $N$ .)

*Step 4.* In almost all cases, the resulting spaces are not compact and compactifying them in a “good” way is difficult. The key step is to identify the limits of families of varieties in  $\mathbf{X}$  that should give a “good” compact moduli theory. There is no a priori reason to believe that such a choice exists or that it is unique. Finding the right choice in higher dimension was the last conceptual step in the program. For canonical models of varieties of general type we have the “right” answer, see (10.5) and (10.8).

*Step 5.* We have to go back and redo Steps 1–3 for this more general class of objects to get a compactified moduli theory

$$\bar{\pi}_{\mathbf{X}} : \bar{\mathbf{U}}_{\mathbf{X}} \rightarrow \bar{\mathbf{M}}_{\mathbf{X}}.$$

*Step 6.* An extra issue that arises is that the compactifications could also depend on the dimension of the  $\mathbb{P}^N$  chosen in Step 2. This does not seem to happen for  $\mathbf{M}_{\mathbf{X}}$  itself (at least for  $N$  large enough) but it does happen for  $\bar{\mathbf{M}}_{\mathbf{X}}$  for some of the proposed variants.

*Step 7.* Finally, if everything works out, we would like to study the properties of  $\mathbf{M}_{\mathbf{X}}$ ,  $\bar{\mathbf{M}}_{\mathbf{X}}$  and to use these to prove further theorems.

Next we review the historical development of the higher dimensional theory.

**10.3 (Geometric invariant theory: Mumford, 1965).** Riemann probably knew that all smooth, compact Riemann surfaces of a given genus  $g$  form a nice family, but the moduli spaces  $\mathcal{M}_g$  were first rigorously constructed by Teichmüller in 1940 as an analytic space

and by Mumford in 1965 as an algebraic variety. Mumford’s book [65] presents a program to construct moduli spaces under rather general conditions and uses it to obtain  $\mathcal{M}_g$ . Using these methods, moduli spaces were constructed for surfaces (Gieseker, 1977) and for higher dimensions (Viehweg, 1990).

The correct compactification of these moduli spaces was much less clear. In principle, GIT provides an answer, but the resulting compactification might depend on the embedding dimension chosen in (10.2.Step 2). Recently Wang–Xu [79] prove that, for surfaces and in higher dimensions, the GIT compactification does depend on the embedding dimension. (The current examples, however, do not exclude the possibility that some variant of the GIT approach does provide an answer that is independent of the embedding dimension.)

**10.4 (Compact moduli of curves: Deligne and Mumford, 1969).** The optimal compactification of  $\mathcal{M}_g$  is constructed in [17]. In the boundary  $\overline{\mathcal{M}}_g \setminus \mathcal{M}_g$  we should allow reducible curves  $C = \cup_i C_i$  that satisfy two restrictions.

(Local property)  $C$  has only nodes as singularities. In suitable local analytic coordinates these are given by an equation  $(xy = 0) \subset \mathbb{C}^2$ . As in (9.2) the “standard” volume form on a node is given by  $\frac{dx}{x}$  (on the line  $(y = 0)$ ) and by  $-\frac{dy}{y}$  (on the line  $(x = 0)$ ). These forms have a simple pole at the singularity, corresponding to the restriction on log canonical singularities in (9.4).

(Global property) Instead of each  $c_1(C_i)$  being negative, we assume that each  $c_1(C_i) - D_i$  is negative where  $D_i$  is the sum of the nodes that lie on  $C_i$ . (Thus we allow  $C_i \cong \mathbb{CP}^1$ , as long as at least 3 nodes also lie on  $C_i$ .)

**10.5 (Compact moduli of surfaces: Kollár and Shepherd–Barron, 1988).** It was clear from the Mumford–Gieseker approach that one should work with the *canonical models* of surfaces of general type (as in 4.6) in order to get a good moduli theory, but the correct class of singular limits was not known.

An approach using minimal models was proposed in [53]: given a family of canonical models over a punctured disc  $S^* \rightarrow \Delta^*$ , first construct any compactification whose central fiber is a reduced simple normal crossing divisor and then take the (relative) canonical model. It is not hard to see that this gives a unique limit. This says that at the boundary of the moduli space we should allow *stable surfaces*: reducible surfaces  $S = \cup_i S_i$  that satisfy two restrictions.

(Local property)  $S$  has so-called *semi-log canonical* singularities. What are these? First of all, aside from finitely many points  $S$  is either smooth or has two local branches meeting transversally, like  $(xy = 0) \subset \mathbb{C}^3$ . These are the natural generalizations of nodes. Then we can have log canonical singularities (9.4). Finally, it can happen that several  $S_i$  come together at a point and each of them has a log-canonical singularity there. An explicit list is given in [53].

(Global property) Instead of each  $c_1(S_i)$  being negative, we assume that each  $c_1(S_i) - D_i$  is negative where  $D_i$  is the sum of the double curves that lie on  $S_i$ .

Another interesting issue that arises is that not every deformation of such singular surfaces is allowed. It turns out that even basic numerical invariants of a surface can jump if we allow arbitrary deformations. To avoid this, [53] identifies a restricted deformation theory (called  $\mathbb{Q}G$ -condition) that produces the correct boundary.

This answers our second Main Question: First, the “simplest” families of surfaces of general type are families  $f : S_M \rightarrow M$  whose fibers are stable surfaces (and satisfy the  $\mathbb{Q}G$ -condition). Second, every family of surfaces of general type is birational to such a “simplest”

family, at least after a generically finite-to-one change of the base  $M$ .

The projectivity of the resulting moduli spaces was proved in [36].

**10.6 (Moduli of pairs: Alexeev, Kontsevich, 1994).** Frequently we are interested in understanding all subvarieties  $X$  of a given variety  $Y$ . All is well if  $X$  is smooth, but it is less clear how to handle singular subvarieties. Various methods have been proposed, going back to Cayley in 1860.

Alexeev proposed in [1] that instead of working with very singular subvarieties, one should look at morphisms  $X \rightarrow Y$  that mimic (10.5); see also [3]. Independently, Kontsevich developed this approach for curves [55]. The latter since became a standard tool in quantum cohomology theory.

**10.7 (Quotient theorems: Keel, Kollár, Mori, 1997).** Step 3 of (10.2) leads to the general problem of taking the quotient of a variety by a group. In our cases we have the extra information that every point has a finite stabilizer. In the sixties Artin and Seshadri proved several quotient theorems, especially when all stabilizers are trivial. The general results needed for the moduli theory were established in [35, 40]. This is a quite subtle subject since the resulting quotients are so called *algebraic spaces*, a concept somewhat more general than varieties (or even schemes). Using the ideas of [36] one can then show that, in the cases of interest to us, the quotients are in fact projective (Fujino, Kovács, McKernan).

**10.8 (Moduli in higher dimensions).** The general theory follows the outlines of (10.5) with some key differences.

First, when [53] was written, minimal models were known to exist only in dimension 3. The higher dimensional theory needs several results that were established only recently [28].

Second, it turned out to be quite difficult to understand how the irreducible components of a reducible variety  $X = \cup_i X_i$  glue together. For curves, as in (10.4) the well-defined residue of the 1-form  $\frac{dx}{x}$  is a key ingredient. The current approach in higher dimension relies on a new Poincaré-type residue theory for log canonical singularities; see [47, Chap.4]. The full theory should be written up in [48].

**10.9 (Explicit examples: Alexeev, 2002–).** While the above methods provide a complete answer in principle, it has been very difficult to work out a full description in concrete cases. The first such examples were Abelian varieties [2] and plane curves (Hacking, [24]). Recent surface examples are in [4].

**10.10 (Hyperbolicity: Kovács, Viehweg, Zuo, 2000–2010).** So far, very little is known about the moduli spaces of surfaces and higher dimensional varieties in general. The local structure of these spaces can be arbitrarily complicated [75]. Hyperbolicity properties of the moduli of smooth curves were conjectured by Shafarevich in 1962 and later extended to higher dimensions in [34, 56, 58, 76].

**10.11 (Degeneration of Fano varieties: Xu, 2007–).** We know much less about the moduli of Fano (=positively curved) varieties. Most of the geometric works deal with extending families  $g^* : X^* \rightarrow \Delta^*$  over a punctured disc across the puncture. Two questions turned out to be especially interesting: understanding the combinatorial structure of the central fiber  $X_0$  for arbitrary limits and finding limits where  $X_0$  is especially simple.

A series of papers [15, 31, 44] shows that the combinatorial structure of  $X_0$  is contractible; this answers an old conjecture of J. Ax. Recently, [60] shows that there are limits where  $X_0$  itself is a (singular) Fano variety, as conjectured by Tian.

**Acknowledgments.** I thank J. Fickenscher, A. Fulger, J.M. Johnson, S. Kovács, T. Murayama, Zs. Patakfalvi and N. Sheridan for helpful suggestions. Partial financial support was provided by the NSF under grant number DMS-0968337.

## References

- [1] Valery Alexeev, *Moduli spaces  $M_{g,n}(W)$  for surfaces*, Higher-dimensional complex varieties (Trento, 1994), de Gruyter, Berlin, 1996, pp. 1–22.
- [2] ———, *Complete moduli in the presence of semiabelian group action*, Ann. of Math. (2) **155** (2002), no. 3, 611–708.
- [3] ———, *Higher-dimensional analogues of stable curves*, International Congress of Mathematicians. Vol. II, Eur. Math. Soc., Zürich, 2006, pp. 515–536.
- [4] Valery Alexeev and Rita Pardini, *Non-normal abelian covers*, Compos. Math. **148** (2012), no. 4, 1051–1084.
- [5] Carolina Araujo and János Kollár, *Rational curves on varieties*, Higher dimensional varieties and rational points (Budapest, 2001), Bolyai Soc. Math. Stud., vol. 12, Springer, Berlin, 2003, pp. 13–68.
- [6] W. Barth, C. Peters, and A. Van de Ven, *Compact complex surfaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), vol. 4, Springer-Verlag, Berlin, 1984.
- [7] R. J. Berman and H. Guenancia, *Kähler–Einstein metrics on stable varieties and log canonical pairs*, ArXiv e-prints (2013).
- [8] C. Birkar, *Existence of flips and minimal models for 3-folds in char  $p$* , ArXiv e-prints (2013).
- [9] Caucher Birkar, Paolo Cascini, Christopher D. Hacon, and James McKernan, *Existence of minimal models for varieties of log general type*, J. Amer. Math. Soc. **23** (2010), no. 2, 405–468.
- [10] F. Campana, *Connexité rationnelle des variétés de Fano*, Ann. Sci. École Norm. Sup. (4) **25** (1992), no. 5, 539–545.
- [11] Paolo Cascini and Vladimir Lazić, *New outlook on the minimal model program, I*, Duke Math. J. **161** (2012), no. 12, 2415–2467.
- [12] Herbert Clemens, János Kollár, and Shigefumi Mori, *Higher-dimensional complex geometry*, Astérisque (1988), no. 166, 144 pp. (1989).
- [13] Alessio Corti, *Flips for 3-folds and 4-folds*, Oxford Lecture Ser. Math. Appl., vol. 35, Oxford Univ. Press, Oxford, 2007.
- [14] Alessio Corti, Paul Hacking, János Kollár, Robert Lazarsfeld, and Mircea Mustață, *Lectures on flips and minimal models*, Analytic and algebraic geometry, IAS/Park City Math. Ser., vol. 17, Amer. Math. Soc., Providence, RI, 2010, pp. 557–583.
- [15] Tommaso de Fernex, János Kollár, and Chenyang Xu, *The dual complex of singularities*, ArXiv e-prints (2012).
- [16] A. J. de Jong and J. Starr, *Every rationally connected variety over the function field of a curve has a rational point*, Amer. J. Math. **125** (2003), no. 3, 567–580.
- [17] P. Deligne and D. Mumford, *The irreducibility of the space of curves of given genus*,

- Inst. Hautes Études Sci. Publ. Math. (1969), no. 36, 75–109.
- [18] Alan H. Durfee, *Fifteen characterizations of rational double points and simple critical points*, Enseign. Math. (2) **25** (1979), no. 1-2, 131–163.
- [19] Hélène Esnault, *Varieties over a finite field with trivial Chow group of 0-cycles have a rational point*, Invent. Math. **151** (2003), no. 1, 187–191.
- [20] Philippe Eyssidieux, Vincent Guedj, and Ahmed Zeriahi, *Singular Kähler-Einstein metrics*, J. Amer. Math. Soc. **22** (2009), no. 3, 607–639.
- [21] Gavril Farkas and Ian Morrison (eds.), *Handbook of moduli I–III*, Advanced Lectures in Mathematics, vol. 24–26, International Press (Somerville, MA), 2013.
- [22] Tom Graber, Joe Harris, and Jason Starr, *Families of rationally connected varieties*, J. Amer. Math. Soc. **16** (2003), no. 1, 57–67 (electronic).
- [23] Daniel Greb, Stefan Kebekus, Sándor J. Kovács, and Thomas Peternell, *Differential forms on log canonical spaces*, Publ. Math. Inst. Hautes Études Sci. (2011), no. 114, 87–169.
- [24] Paul Hacking, *Compact moduli of plane curves*, Duke Math. J. **124** (2004), no. 2, 213–257.
- [25] Christopher D. Hacon and Sándor J. Kovács, *Classification of higher dimensional algebraic varieties*, Oberwolfach Seminars, vol. 41, Birkhäuser Verlag, Basel, 2010.
- [26] Christopher D. Hacon and James McKernan, *Extension theorems and the existence of flips*, Flips for 3-folds and 4-folds, Oxford Lecture Ser. Math. Appl., vol. 35, Oxford Univ. Press, Oxford, 2007, pp. 76–110.
- [27] ———, *Flips and flops*, Proceedings of the International Congress of Mathematicians. Volume II (New Delhi), Hindustan Book Agency, 2010, pp. 513–539.
- [28] Christopher D. Hacon and Chenyang Xu, *Existence of log canonical closures*, Invent. Math. **192** (2013), no. 1, 161–195.
- [29] ———, *On the three dimensional minimal model program in positive characteristic*, ArXiv e-prints (2013).
- [30] Joe Harris, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 133, Springer-Verlag, New York, 1995.
- [31] Amit Hogadi and Chenyang Xu, *Degenerations of rationally connected varieties*, Trans. Amer. Math. Soc. **361** (2009), no. 7, 3931–3949.
- [32] Kentaro Hori, Sheldon Katz, Albrecht Klemm, Rahul Pandharipande, Richard Thomas, Cumrun Vafa, Ravi Vakil, and Eric Zaslow, *Mirror symmetry*, Clay Mathematics Monographs, vol. 1, American Mathematical Society, Providence, RI, 2003.
- [33] Yujiro Kawamata, Katsumi Matsuda, and Kenji Matsuki, *Introduction to the minimal model problem*, Algebraic geometry, Sendai, 1985, Adv. Stud. Pure Math., vol. 10, North-Holland, Amsterdam, 1987, pp. 283–360.
- [34] Stefan Kebekus and Sándor J. Kovács, *Families of canonically polarized varieties over surfaces*, Invent. Math. **172** (2008), no. 3, 657–682.
- [35] Seán Keel and Shigefumi Mori, *Quotients by groupoids*, Ann. of Math. (2) **145** (1997), no. 1, 193–213.
- [36] János Kollár, *Projectivity of complete moduli*, J. Differential Geom. **32** (1990), no. 1,

- 235–268.
- [37] ———, *Flip and flop*, Proceedings of the International Congress of Mathematicians, Vol. I, II, (Kyoto, 1990), Tokyo, Math. Soc. Japan, 1991, pp. 709–714.
- [38] János Kollár (ed.), *Flips and abundance for algebraic threefolds*, Société Mathématique de France, 1992, Papers from the Second Summer Seminar on Algebraic Geometry held at the University of Utah, Salt Lake City, Utah, August 1991, Astérisque No. 211 (1992).
- [39] János Kollár, *Rational curves on algebraic varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge., vol. 32, Springer-Verlag, Berlin, 1996.
- [40] ———, *Quotient spaces modulo algebraic groups*, Ann. of Math. (2) **145** (1997), no. 1, 33–79.
- [41] ———, *Low degree polynomial equations: arithmetic, geometry and topology*, European Congress of Mathematics, Vol. I (Budapest, 1996), Progr. Math., vol. 168, Birkhäuser, Basel, 1998, pp. 255–288.
- [42] ———, *Rationally connected varieties over local fields*, Ann. of Math. (2) **150** (1999), no. 1, 357–367.
- [43] ———, *Which are the simplest algebraic varieties?*, Bull. Amer. Math. Soc. (N.S.) **38** (2001), no. 4, 409–433 (electronic).
- [44] ———, *A conjecture of Ax and degenerations of Fano varieties*, Israel J. Math. **162** (2007), 235–251.
- [45] ———, *Lectures on resolution of singularities*, Annals of Mathematics Studies, vol. 166, Princeton University Press, Princeton, NJ, 2007.
- [46] ———, *Moduli of varieties of general type*, Handbook of Moduli (Gavril Farkas and Ian Morrison, eds.), Advanced Lectures in Mathematics, International Press (Somerville, MA), 2013, pp. 131–158.
- [47] ———, *Singularities of the minimal model program*, Cambridge Tracts in Mathematics, vol. 200, Cambridge University Press, Cambridge, 2013, With the collaboration of Sándor Kovács.
- [48] ———, *Moduli of varieties of general type*, (book in preparation), 2015.
- [49] János Kollár, Yoichi Miyaoka, and Shigefumi Mori, *Rational connectedness and boundedness of Fano manifolds*, J. Differential Geom. **36** (1992), no. 3, 765–779.
- [50] ———, *Rationally connected varieties*, J. Algebraic Geom. **1** (1992), no. 3, 429–448.
- [51] János Kollár and Shigefumi Mori, *Classification of three-dimensional flips*, J. Amer. Math. Soc. **5** (1992), no. 3, 533–703.
- [52] ———, *Birational geometry of algebraic varieties*, Cambridge Tracts in Mathematics, vol. 134, Cambridge University Press, Cambridge, 1998, With the collaboration of C. H. Clemens and A. Corti, Translated from the 1998 Japanese original.
- [53] János Kollár and N. I. Shepherd-Barron, *Threefolds and deformations of surface singularities*, Invent. Math. **91** (1988), no. 2, 299–338.
- [54] János Kollár and Endre Szabó, *Rationally connected varieties over finite fields*, Duke Math. J. **120** (2003), no. 2, 251–267.
- [55] M. Kontsevich and Yu. Manin, *Gromov-Witten classes, quantum cohomology, and enu-*

- merative geometry*, *Comm. Math. Phys.* **164** (1994), no. 3, 525–562.
- [56] Sándor J. Kovács, *Algebraic hyperbolicity of fine moduli spaces*, *J. Algebraic Geom.* **9** (2000), no. 1, 165–174.
- [57] ———, *Young person's guide to moduli of higher dimensional varieties*, Algebraic geometry—Seattle 2005. Part 2, *Proc. Sympos. Pure Math.*, vol. 80, Amer. Math. Soc., Providence, RI, 2009, pp. 711–743.
- [58] Sándor J. Kovács and Max Lieblich, *Boundedness of families of canonically polarized manifolds: a higher dimensional analogue of Shafarevich's conjecture*, *Ann. of Math.* (2) **172** (2010), no. 3, 1719–1748.
- [59] László Lempert and Endre Szabó, *Rationally connected varieties and loop spaces*, *Asian J. Math.* **11** (2007), no. 3, 485–496.
- [60] Chi Li and Chenyang Xu, *Special test configurations and  $K$ -stability of Fano varieties*, ArXiv e-prints (2011) (*Annals of Math.*, to appear).
- [61] Shigefumi Mori, *Threefolds whose canonical bundles are not numerically effective*, *Ann. of Math.* (2) **116** (1982), no. 1, 133–176.
- [62] ———, *Classification of higher-dimensional varieties*, Algebraic geometry, Bowdoin, 1985 (Brunswick, Maine, 1985), *Proc. Sympos. Pure Math.*, vol. 46, Amer. Math. Soc., Providence, RI, 1987, pp. 269–331.
- [63] ———, *Flip theorem and the existence of minimal models for 3-folds*, *J. Amer. Math. Soc.* **1** (1988), no. 1, 117–253.
- [64] Andrei Moroianu, *Lectures on Kähler geometry*, London Mathematical Society Student Texts, vol. 69, Cambridge University Press, Cambridge, 2007.
- [65] David Mumford, *Geometric invariant theory*, *Ergebnisse der Mathematik und ihrer Grenzgebiete, Neue Folge, Band 34*, Springer-Verlag, Berlin, 1965.
- [66] ———, *Algebraic geometry. I*, Springer-Verlag, Berlin, 1976, *Complex projective varieties*, *Grundlehren der Mathematischen Wissenschaften*, No. 221.
- [67] Alan Michael Nadel, *The boundedness of degree of Fano varieties with Picard number one*, *J. Amer. Math. Soc.* **4** (1991), no. 4, 681–692.
- [68] Yuji Odaka and Chenyang Xu, *Log-canonical models of singular pairs and its applications*, *Math. Res. Lett.* **19** (2012), no. 2, 325–334.
- [69] Zsolt Patakfalvi, *On subadditivity of Kodaira dimension in positive characteristic*, ArXiv e-prints (2013).
- [70] Mihai Păun, *Quantitative extensions of twisted pluricanonical forms and non-vanishing*, *Proceedings of the International Congress of Mathematicians. Volume II* (New Delhi), Hindustan Book Agency, 2010, pp. 540–557.
- [71] Miles Reid, *Minimal models of canonical 3-folds*, Algebraic varieties and analytic varieties (Tokyo, 1981), *Adv. Stud. Pure Math.*, vol. 1, North-Holland, Amsterdam, 1983, pp. 131–180.
- [72] ———, *Young person's guide to canonical singularities*, Algebraic geometry, Bowdoin, 1985, (Brunswick, Maine, 1985), *Proc. Sympos. Pure Math.*, vol. 46, Amer. Math. Soc., Providence, RI, 1987, pp. 345–414.
- [73] Igor R. Shafarevich, *Basic algebraic geometry*, Springer-Verlag, New York, 1974, *Die Grundlehren der mathematischen Wissenschaften*, Band 213.

- [74] Kenji Ueno, *Classification theory of algebraic varieties and compact complex spaces*, Lecture Notes in Mathematics, Vol. 439, Springer-Verlag, Berlin, 1975, Notes written in collaboration with P. Cherenack.
- [75] Ravi Vakil, *Murphy's law in algebraic geometry: badly-behaved deformation spaces*, Invent. Math. **164** (2006), no. 3, 569–590.
- [76] Eckart Viehweg and Kang Zuo, *On the Brody hyperbolicity of moduli spaces for canonically polarized manifolds*, Duke Math. J. **118** (2003), no. 1, 103–150.
- [77] Claire Voisin, *Mirror symmetry*, SMF/AMS Texts and Monographs, vol. 1, American Mathematical Society, Providence, RI, 1999, Translated from the 1996 French original by Roger Cooke.
- [78] ———, *Rationally connected 3-folds and symplectic geometry*, Astérisque (2008), no. 322, 1–21, Géométrie différentielle, physique mathématique, mathématiques et société. II.
- [79] Xiaowei Wang and Chenyang Xu, *Nonexistence of asymptotic GIT compactification*, ArXiv e-prints (2012) (Duke Math. J., to appear).
- [80] Qi Zhang, *Rational connectedness of log  $\mathbf{Q}$ -Fano varieties*, J. Reine Angew. Math. **590** (2006), 131–142.

Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, USA

E-mail: kollár@math.princeton.edu



# Random Geometry on the Sphere

Jean-François Le Gall

**Abstract.** We introduce and study a universal model of random geometry in two dimensions. To this end, we start from a discrete graph drawn on the sphere, which is chosen uniformly at random in a certain class of graphs with a given size  $n$ , for instance the class of all triangulations of the sphere with  $n$  faces. We equip the vertex set of the graph with the usual graph distance rescaled by the factor  $n^{-1/4}$ . We then prove that the resulting random metric space converges in distribution as  $n \rightarrow \infty$ , in the Gromov-Hausdorff sense, toward a limiting random compact metric space called the Brownian map, which is universal in the sense that it does not depend on the class of graphs chosen initially. The Brownian map is homeomorphic to the sphere, but its Hausdorff dimension is equal to 4. We obtain detailed information about the structure of geodesics in the Brownian map. We also present the infinite-volume variant of the Brownian map called the Brownian plane, which arises as the scaling limit of the uniform infinite planar quadrangulation. Finally, we discuss certain open problems. This study is motivated in part by the use of random geometry in the physical theory of two-dimensional quantum gravity.

**Mathematics Subject Classification (2010).** Primary 05C80, 60D05; Secondary 05C12, 60F17

**Keywords.** Planar map, triangulation, graph distance, Gromov-Hausdorff convergence, Brownian map, geodesic, continuum random tree, Brownian plane.

## 1. Introduction

In the last ten years, there has been much interest in discrete and continuous models of random geometry in two dimensions. As a first naive attempt to construct a continuous model, one could imagine choosing at random a Riemannian metric on the sphere. However there seems to be no canonical way of making such a choice, and a better approach leading to a continuous object that is universal in some sense is to start from discrete models. Informally, we will consider a large graph drawn on the sphere and chosen at random in a suitable class: We then expect that the suitably rescaled graph distance on the vertex set will converge in an appropriate sense, when the size of the graph tends to infinity, to a random metric on the sphere.

More precisely, we consider planar maps, which are finite connected graphs embedded in the sphere  $\mathbb{S}^2$  – more precise definitions will be given in Section 2 below. One is interested in the “shape” of the graph, so that two planar maps are identified if the second one is the image of the first one under an orientation-preserving homeomorphism of the sphere. The faces of the map are the connected components of the complement of edges, and a planar map is called a triangulation if all faces are triangles, possibly with two edges glued together. For technical reasons, it is convenient to deal with *rooted* triangulations, meaning that there

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

is a distinguished oriented edge called the root edge, whose tail is the root vertex. Thanks to the preceding identification, for every fixed even integer  $n$ , there are only a finite number of rooted triangulations with  $n$  faces, and therefore it makes sense to choose one of them at random, which we denote by  $T_n$ . Then, supposing that there is a canonical way of embedding  $T_n$  in the sphere (recall that the embedding of a planar map is only defined up to orientation-preserving homeomorphisms), it seems reasonable to conjecture that the vertex set  $V(T_n)$  of  $T_n$  will become dense in  $\mathbb{S}^2$  when  $n \rightarrow \infty$ , and that the graph distance  $d_{\text{gr}}$  on this vertex set will converge, modulo a suitable rescaling, to a random metric on  $\mathbb{S}^2$ .

The preceding program is still the subject of active research – there are indeed (almost) canonical ways of embedding triangulations, using circle packings or the uniformization theorem for Riemann surfaces, see the discussion in Section 6 below. Here we will adopt a slightly different point of view. We consider the finite metric space  $(V(T_n), d_{\text{gr}})$  as a (random) element of the set  $\mathbb{K}$  of all isometry classes of compact metric spaces. The space  $\mathbb{K}$  is equipped with the Gromov-Hausdorff distance  $d_{GH}$  (cf. subsection 3.1), and  $(\mathbb{K}, d_{GH})$  is both separable and complete. A key result [31], which solves a conjecture of Oded Schramm [44], states that

$$(V(T_n), 6^{1/4}n^{-1/4}d_{\text{gr}}) \xrightarrow[n \rightarrow \infty]{(d)} (\mathbf{m}_\infty, D), \quad (1.1)$$

where the convergence holds in distribution in the space  $(\mathbb{K}, d_{GH})$ , and the limit  $(\mathbf{m}_\infty, D)$  is a random metric space, which is called the Brownian map after Marckert and Mokkadem [37]. Despite the somewhat unusual fact that we are dealing with random compact metric spaces, (1.1) is nothing but a particular case of the standard notion of convergence in distribution for random variables with values in a Polish space. Note that the constant  $6^{1/4}$  in (1.1) is just a normalization factor.

A very important feature of the convergence (1.1) is the fact that it holds for much more general random planar maps, with the *same* limiting space  $(\mathbf{m}_\infty, D)$  up to unimportant scaling constants. Recall that a planar map is a  $p$ -angulation if all faces have degree  $p$ . Then it was proved in [31] that an analog of (1.1) holds for  $p$ -angulations with a fixed number of faces, for every even integer  $p \geq 4$ . Note that the special case of quadrangulations, corresponding to  $p = 4$ , has been treated independently by Miermont [41] via a different method. Similarly, analogs of (1.1) hold for general planar maps with a fixed number of edges [11] or for bipartite planar maps with a fixed number of edges [1]. It is also possible to impose local constraints on the planar maps: A result similar to (1.1) holds for simple triangulations or quadrangulations [2], where no loops or multiple edges are allowed, or for quadrangulations with no pendant vertices [8]. One indeed expects that the Brownian map will appear as the scaling limit of very general random planar maps provided some bound holds on the distribution of the degree of a typical face (on the other hand, the paper [33] shows that different scaling limits may occur if one considers distributions that favour the appearance of “very large” faces).

The Brownian map thus appears as a *universal* object, in the sense that it is the scaling limit of many different models of random planar maps: This is of course similar to the case of Brownian motion, which is the universal scaling limit of many different random walks on the lattice. Just as Brownian motion can be viewed as a purely random continuous curve, the Brownian map seems to be the right model for a purely random surface. Note that the Brownian map  $(\mathbf{m}_\infty, D)$  is almost surely homeomorphic to the sphere  $\mathbb{S}^2$  [35] (see also [39]), even though its Hausdorff dimension is equal to 4. However, there is no canonical homeomorphism and we cannot a priori use the Brownian map to obtain a “canonical” random metric

on the sphere as suggested by preceding remarks.

The main goal of the present work is to present a general theorem of convergence toward the Brownian map (Theorem 3.1 below), which was derived in the series of papers [29–31]. We also give a detailed construction of the Brownian map, by showing that this random compact metric space can be obtained by gluing certain pairs of points in another famous probabilistic model, the Brownian continuum random tree or CRT, which was introduced and studied by Aldous [3, 4]. Note that the CRT itself is a universal scaling limit of discrete trees, in a sense analogous to (1.1). The best way to understand the construction of the Brownian map from the CRT is to start from the discrete bijections that exist between various classes of planar maps and corresponding classes of labeled trees. For this reason, we start in Section 2 by explaining these bijections in the particular case of quadrangulations (there exist similar bijections for triangulations, or more generally for  $p$ -angulations, but their description is more complicated). In Section 3, after stating the main theorem of convergence that extends (1.1), we give a precise definition of the CRT as the tree coded by a normalized Brownian excursion, and we construct the Brownian map via the above-mentioned gluing procedure. Note that, although the discrete bijections between planar maps and trees are very far from being sufficient to get results such as (1.1), they provide useful insight into the construction of the Brownian map. In Section 4, we describe the detailed results that are known about geodesics in the Brownian map. Section 5 is devoted to the Brownian plane, which can be viewed as an infinite-volume version of the Brownian map. The Brownian plane shares many properties of the Brownian map, but it enjoys an additional scale invariance property, which makes it more suitable for explicit calculations (we briefly present some recent results from [22]). One can also view the Brownian plane as a continuous analog of the infinite random lattices known as the uniform infinite planar triangulation and quadrangulation, which have been studied extensively in the recent years. Finally, Section 6 discusses what is probably the most important open problem in the area, namely finding a canonical construction of the Brownian map in terms of a random metric on the sphere. A related discussion can be found in Benjamini [9].

To complete this introduction, let us mention that, although our main motivation for the following results came from probability theory, there are important connections with several other areas of mathematics and physics. In particular, from the point of view of combinatorics, one can derive information about properties of “typical” large planar maps from the known results about the Brownian map. For instance, the fact that the Brownian map is homeomorphic to the sphere implies that for a large planar map chosen uniformly at random in a given class (e.g. a large triangulation) there cannot exist a cycle whose length is small in comparison with the diameter of the graph, such that both regions separated by this cycle have a macroscopic size (see [35]). Similarly, the statements about geodesics in the Brownian map show that, in a typical large planar map, the geodesic between two vertices chosen at random is “macroscopically unique”, meaning that the distance between two geodesics will be small in comparison with the diameter of the graph. Another strong motivation for this work came from physics, and particularly from the use of large random planar maps as models of random geometry in two-dimensional quantum gravity (see the book [5]). In this connection, we mention the important work of Bouttier and Guitter (see in particular [13–15]) which motivated part of our results. It is worth mentioning that a different mathematical approach to two-dimensional quantum gravity relying on the Gaussian free field has been given by Duplantier and Sheffield [24]. The construction of [24] seems quite different from our perspective, but the paper [45] gives a number of conjectures that would relate large ran-

dom planar maps and the Brownian map to the Gaussian free field approach. The very recent work of Miller and Sheffield [42] seems promising in this respect.

*Acknowledgements.* I thank Nicolas Curien and Igor Kortchemski for their help with figures.

## 2. Planar maps and bijections with trees

In this section, we recall the basic definitions concerning planar maps and we explain how planar maps can be coded by trees, in the particular case of quadrangulations. This coding is an important ingredient of the proofs, and it also helps to understand the definition of the Brownian map that will be given below.

**2.1. Planar maps.** Let us start with the definition of a planar map.

**Definition 2.1.** A planar map is a proper (without edge-crossing) embedding of a finite connected graph in the two-dimensional sphere  $\mathbb{S}^2$ . A rooted planar map is a planar map given with a distinguished oriented edge, which is called the root edge and whose tail vertex is called the root vertex. Two rooted planar maps are identified if they correspond via an orientation-preserving homeomorphism of  $\mathbb{S}^2$ .

We in fact allow loops and multiple edges in our graphs. Following the classical terminology found in combinatorics, the word “graph” should be replaced by “multigraph” in the preceding definition. The faces of a planar map are the connected components of the complement of the union of edges, and the degree of a face counts the number of edge sides that are incident to this face (in particular it may happen that both sides of an edge are incident to the same face, and then this edge is counted twice in the degree of the face).

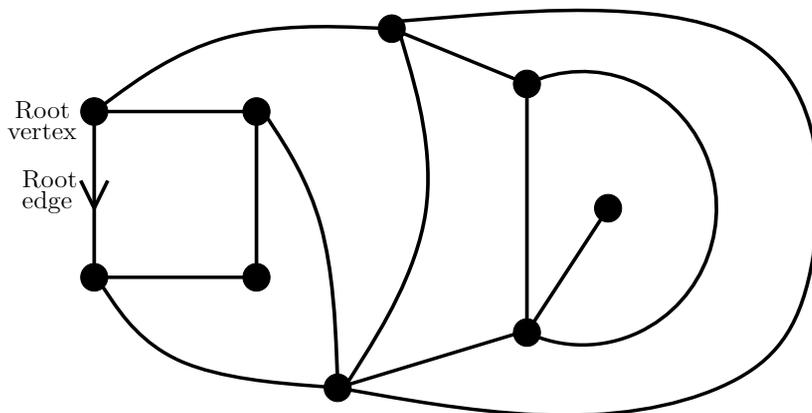


Figure 2.1. A rooted quadrangulation with 7 faces

For any integer  $p \geq 3$ , a planar map is called a  $p$ -angulation (a triangulation if  $p = 3$ , a quadrangulation if  $p = 4$ ) if all its faces have degree  $p$ . Fig. 1 shows an example of a quadrangulation with 7 faces. For every integer  $n \geq 1$ , we write  $\mathcal{M}_n^p$  for the set of all rooted  $p$ -angulations with  $n$  faces. When  $p$  is odd, the set  $\mathcal{M}_n^p$  is empty if (and only if)  $n$  is odd. So when we deal with odd values of  $p$  (in particular with triangulations), we will always assume that  $n$  is even.

Thanks to the identification in Definition 2.1, the sets  $\mathcal{M}_n^p$  are finite. Enumeration results for these sets were obtained by Tutte (see in particular [46]) in a series of important papers motivated by the four-color theorem.

We will use the notation  $d_{\text{gr}}^M$  for the graph distance on the vertex set  $V(M)$  of a planar map  $M$ .

**2.2. Bijections with trees.** In this section we explain the Cori-Vauquelin-Schaeffer bijection between rooted quadrangulations and well-labeled trees [18, 19]. We restrict ourselves to the special case of quadrangulations for the sake of simplicity, but similar bijections exist for  $p$ -angulations for any  $p$  (see in particular [12]).

First recall that a plane tree  $\tau$  is a (finite) rooted ordered tree: A way of specifying a plane tree is to represent each of its vertices by a finite word made of positive integers, in such a way that the empty word  $\emptyset$  corresponds to the root or ancestor of the tree, and that, for instance, the word 13 corresponds to the third child of the first child of the root (see the left side of Fig. 2).

Then a well-labeled tree is a plane tree  $\tau$ , with vertex set  $V(\tau)$ , whose vertices  $v$  are assigned labels  $(\ell_v)_{v \in V(\tau)}$ , in such a way that the following properties hold. First, the label of the root is 1, then the label  $\ell_v$  of any vertex  $v$  is a positive integer, and finally  $|\ell_v - \ell_{v'}| \leq 1$  if  $v$  and  $v'$  are adjacent vertices.

With any well-labeled tree  $(\tau, (\ell_v)_{v \in V(\tau)})$  with  $n$  edges, we can associate a rooted quadrangulation  $Q$  with  $n$  faces via the following construction. Suppose that the tree is embedded in the plane in the (obvious) way as suggested in the left part of Fig. 2 (in particular the successive children of a vertex appear from left to right). Then the vertex set of  $Q$  will be the union of the vertex set of  $\tau$  and of an extra vertex, and we now explain how to construct the edges of  $Q$ . First recall that a corner of the tree  $\tau$  is an angular sector between two successive edges of  $\tau$  around a given vertex. The set of all corners of  $\tau$  is given a cyclic ordering by moving clockwise around the tree. To construct the edges of  $Q$ , we first add an extra vertex  $\partial$  outside the tree, and we connect each corner of the tree  $\tau$  with label 1 to the vertex  $\partial$  by an edge starting from this corner. Then every corner of  $\tau$  with label  $k \geq 2$  is connected by an edge to the next corner (in cyclic ordering) with label  $k - 1$ . The construction can be made in a unique way so that edges do not cross and do not cross the edges of the tree. The resulting collection of edges forms a quadrangulation  $Q$  whose vertex set is  $V(Q) = V(\tau) \cup \{\partial\}$ . This quadrangulation is rooted at the edge connecting the first corner of the root of  $\tau$  to  $\partial$ , which is oriented so that  $\partial$  is the root vertex. See Fig. 2 for an example.

The previous construction yields a bijection from the set of all well-labeled trees with a fixed number  $n$  of edges onto the set  $\mathcal{M}_n^4$  of all rooted quadrangulations with  $n$  faces. This bijection is called the Cori-Vauquelin-Schaeffer bijection (the CVS bijection in short). Furthermore, the following important additional property holds. If  $(\tau, (\ell_v)_{v \in V(\tau)})$  is a well-labeled tree and  $Q$  is the associated quadrangulation defined as above, then, for every  $v \in V(Q) \setminus \{\partial\}$ ,

$$d_{\text{gr}}^Q(\partial, v) = \ell_v. \tag{2.1}$$

In other words, distances from the root vertex in the quadrangulation  $Q$  are given by labels on the tree associated with  $Q$  via the CVS bijection. There is no similar expression for  $d_{\text{gr}}^Q(v, v')$  when  $v$  and  $v'$  are two vertices other than  $\partial$ , but the following upper bound turns out to be important for our purposes. For every  $v, v' \in V(Q) \setminus \{\partial\}$ ,

$$d_{\text{gr}}^Q(v, v') \leq \ell_v + \ell_{v'} - 2 \max \left( \min_{w \in [v, v']} \ell_w, \min_{w \in [v', v]} \ell_w \right) + 2, \tag{2.2}$$



where the infimum is over all isometric embeddings  $\varphi_1 : E_1 \rightarrow E$  and  $\varphi_2 : E_2 \rightarrow E$  of  $E_1$  and  $E_2$  into the same metric space  $(E, d)$ , and  $d_{\text{Haus}}$  stands for the usual Hausdorff distance between compact subsets of  $E$ . If  $\mathbb{K}$  denotes the space of all isometry classes of compact metric spaces, then  $d_{GH}$  is a distance on  $\mathbb{K}$ , and moreover the metric space  $(\mathbb{K}, d_{GH})$  is Polish, that is, separable and complete. We refer to Chapter 7 of Burago, Burago and Ivanov [16] for a thorough discussion of the Gromov-Hausdorff distance.

**3.2. The main theorem.** Recall our notation  $\mathcal{M}_n^p$  for the space of all rooted  $p$ -angulations with  $n$  faces. Note that when  $p$  is odd we consider only even values of  $n$ . With every  $M \in \mathcal{M}_n^p$ , we can associate the metric space  $(V(M), d_{\text{gr}}^M)$ .

**Theorem 3.1** ([31]). *Suppose that either  $p = 3$  or  $p \geq 4$  is an even integer, and set*

$$c_p := \left( \frac{9}{p(p-2)} \right)^{1/4}$$

*if  $p$  is even, and*

$$c_3 := 6^{1/4}.$$

*For every integer  $n \geq 1$  (for every even integer  $n \geq 2$  if  $p = 3$ ), let  $M_n$  be uniformly distributed over  $\mathcal{M}_n^p$ . There exists a random compact metric space  $(\mathbf{m}_\infty, D)$  called the Brownian map, which does not depend on  $p$ , such that*

$$(V(M_n), c_p n^{-1/4} d_{\text{gr}}^{M_n}) \xrightarrow[n \rightarrow \infty]{(d)} (\mathbf{m}_\infty, D)$$

*where the convergence holds in distribution in the space  $(\mathbb{K}, d_{GH})$ .*

The role of the scaling constants  $c_p$  is only to ensure that the limit does not depend on  $p$ . The convergence in distribution of the theorem may be rephrased by saying that one can construct the full sequence  $(M_n)$  in such a way that the associated metric spaces  $(V(M_n), c_p n^{-1/4} d_{\text{gr}}^{M_n})$  converge to  $(\mathbf{m}_\infty, D)$  for the Gromov-Hausdorff distance, outside a set of zero probability. The name Brownian map is due to Marckert and Mokkadem [37], who proved a weak form of the theorem when  $p = 4$ . As was already pointed in the introduction, the case  $p = 4$  of the theorem has been derived independently by Miermont [41] using a different approach.

**Remark.** It is very likely that the convergence of the theorem also holds for odd values of  $p \geq 5$ , though additional technical difficulties appear in that case. Similarly, it is expected that a version of the convergence holds for Boltzmann distributed random planar maps, such that the probability of a given planar map  $M$  (with a fixed number  $n$  of vertices) will be proportional to

$$\prod_{f \text{ face of } M} w_{\text{degree}(f)}$$

where  $(w_k)_{k \geq 1}$  is a suitable sequence of weights. In the bipartite case where  $w_k = 0$  when  $k$  is odd, a version of the convergence of the theorem holds for such random planar maps [31] under appropriate assumptions. Other recent extensions of the theorem have been mentioned in the introduction above.

In the next two subsections, we will present a precise construction of the Brownian map as a quotient space of another (well-known) random compact metric space, which is the Brownian continuum random tree or CRT (see Aldous [3, 4]).

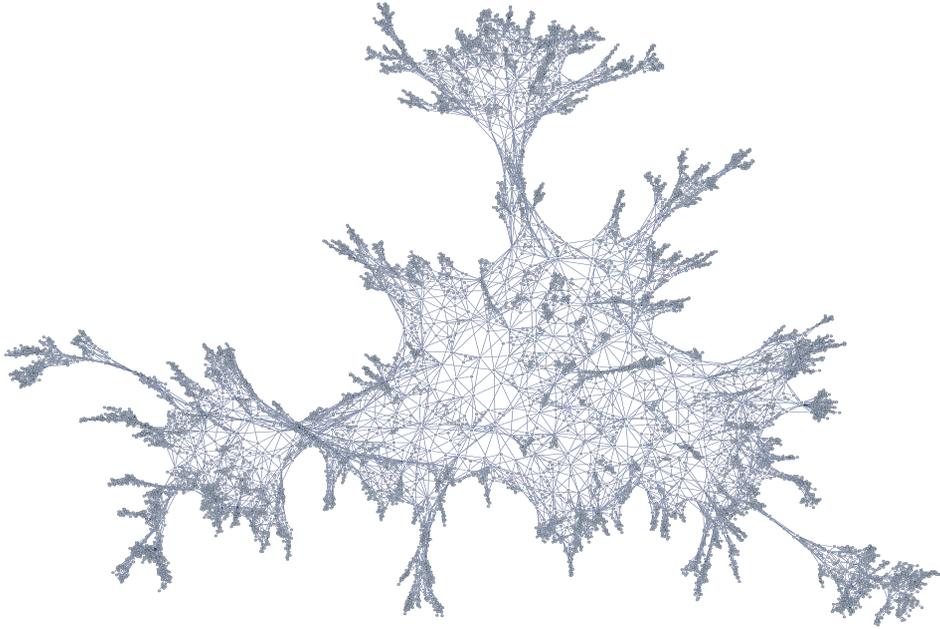


Figure 3.1. Simulation of a large triangulation of the sphere. Here only the graph structure is represented in three dimensions.

**3.3. The Brownian continuum random tree.** We first recall the notion of an  $\mathbb{R}$ -tree.

**Definition 3.2.** A metric space  $(\mathcal{T}, d)$  is an  $\mathbb{R}$ -tree if the following two properties hold for every  $a, b \in \mathcal{T}$ .

- (a) There is a unique isometric map  $f_{a,b}$  from  $[0, d(a, b)]$  into  $\mathcal{T}$  such that  $f_{a,b}(0) = a$  and  $f_{a,b}(d(a, b)) = b$ .
- (b) If  $q$  is a continuous injective map from  $[0, 1]$  into  $\mathcal{T}$ , such that  $q(0) = a$  and  $q(1) = b$ , we have

$$q([0, 1]) = f_{a,b}([0, d(a, b)]).$$

A rooted  $\mathbb{R}$ -tree is an  $\mathbb{R}$ -tree  $(\mathcal{T}, d)$  with a distinguished vertex  $\rho$  called the root.

We will be interested mainly in compact  $\mathbb{R}$ -trees. Informally, one should think of a compact  $\mathbb{R}$ -tree as a connected union of line segments in the plane with no loops, which is equipped with the appropriate (intrinsic) metric. For any two points  $a$  and  $b$  in the tree, there is a unique arc going from  $a$  to  $b$  in the tree, which is isometric to a line segment.

Rooted  $\mathbb{R}$ -trees can be coded by contour functions, in a way very similar to the well-known coding of plane trees by Dyck paths. Let  $g : [0, 1] \rightarrow \mathbb{R}_+$  be a nonnegative continuous function such that  $g(0) = g(1) = 0$  and, for every  $s, t \in [0, 1]$ , set

$$m_g(s, t) := \inf_{r \in [s \wedge t, s \vee t]} g(r),$$

and

$$d_g(s, t) := g(s) + g(t) - 2m_g(s, t).$$

It is easy to verify that  $d_g$  is a pseudo-metric on  $[0, 1]$ . As usual, we introduce the equivalence relation  $s \sim_g t$  if and only if  $d_g(s, t) = 0$  (or equivalently if and only if  $g(s) = g(t) = m_g(s, t)$ ). The function  $d_g$  induces a distance on the quotient space  $\mathcal{T}_g := [0, 1] / \sim_g$ , and we keep the notation  $d_g$  for this distance. We also write  $p_g$  for the canonical projection from  $[0, 1]$  onto  $\mathcal{T}_g$ . Then it is not hard to verify that the metric space  $(\mathcal{T}_g, d_g)$  is a compact  $\mathbb{R}$ -tree (see e.g. [25]), which by definition is rooted at  $\rho_g = p_g(0) = p_g(1)$ . Furthermore the mapping  $g \rightarrow \mathcal{T}_g$  is continuous with respect to the Gromov-Hausdorff distance, if the set of continuous functions  $g$  is equipped with the supremum distance.

The preceding coding induces a cyclic ordering on the tree  $\mathcal{T}_g$ , and it will be important for us to consider the corresponding intervals. By convention, if  $s, t \in [0, 1]$  are such that  $s > t$ , we set  $[s, t] := [s, 1] \cup [0, t]$ . Then we note that, for every  $a, b \in \mathcal{T}_g$  with  $a \neq b$ , there exists a smallest interval  $[s, t]$  such that  $p_g(s) = a$  and  $p_g(t) = b$ , and we define  $[a, b] := p_g([s, t])$ . Roughly speaking,  $[a, b]$  corresponds to the set of vertices that are visited when going from  $a$  to  $b$  in “clockwise order around the tree”.

Let  $\mathbf{e} = (\mathbf{e}_t)_{0 \leq t \leq 1}$  be a normalized Brownian excursion. Informally,  $\mathbf{e}$  behaves like a linear Brownian started from 0 at time 0, which is conditioned to stay positive over the time interval  $(0, 1)$  and to return to 0 at time 1 (of course these conditionings require special care, see e.g. [43, Chapter XII] for a rigorous definition and many properties of the Brownian excursion).

**Definition 3.3.** The CRT is the random  $\mathbb{R}$ -tree  $(\mathcal{T}_e, d_e)$  coded by the Brownian excursion  $\mathbf{e}$ .

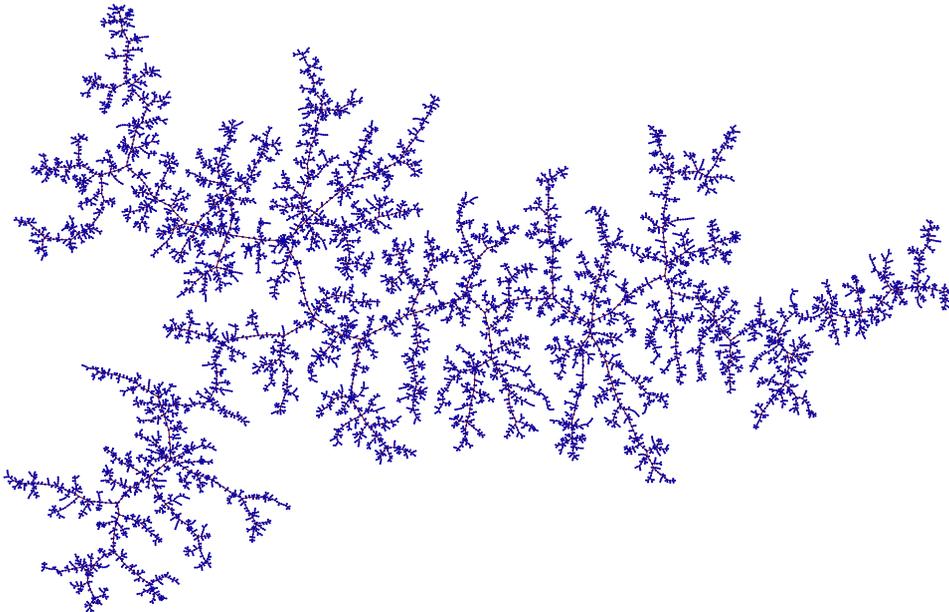


Figure 3.2. A large discrete random tree, which is an approximation of the CRT  $\mathcal{T}_e$ .

**3.4. Constructing the Brownian map.** Analogously to the construction of quadrangulations via the CVS bijection, we will need to introduce labels on  $\mathbb{R}$ -trees. Consider first a

deterministic  $\mathbb{R}$ -tree  $(\mathcal{T}, d)$ , which is rooted at  $\rho$ . We define Brownian motion indexed by  $\mathcal{T}$  as the centered Gaussian process  $(Z_a)_{a \in \mathcal{T}}$  whose distribution is characterised by the properties  $Z_\rho = 0$  and

$$E[(Z_a - Z_b)^2] = d(a, b)$$

for every  $a, b \in \mathcal{T}$ . For a general  $\mathbb{R}$ -tree,  $(Z_a)_{a \in \mathcal{T}}$  needs not have continuous paths. However, if  $\mathcal{T} = \mathcal{T}_g$  and the function  $g$  is Hölder continuous, then it is easy to verify that we can construct  $(Z_a)_{a \in \mathcal{T}_g}$  so that it has continuous sample paths. This applies in particular to (almost every realization of) the CRT  $\mathcal{T}_e$ .

The building blocks of our construction are first the CRT  $(\mathcal{T}_e, d_e)$  and then the process  $(Z_a)_{a \in \mathcal{T}_e}$  which, conditionally on the CRT, is distributed as Brownian motion indexed by  $\mathcal{T}_e$ . Note that there seems to be a technical difficulty here since we are considering a random process  $Z$  indexed by a random set  $\mathcal{T}_e$ : A simple way out is first to construct a process  $(Z_t)_{t \in [0,1]}$  as the “tip” of the Brownian snake driven by  $e$  (see [28]), and then to observe that  $Z_s = Z_t$  if  $s \sim_e t$ , which allows one to view  $Z$  as indexed by  $\mathcal{T}_e$ .

We set, for every  $a, b \in \mathcal{T}_e$ ,

$$D^\circ(a, b) := Z_a + Z_b - 2 \max \left( \min_{c \in [a,b]} Z_c, \min_{c \in [b,a]} Z_c \right). \tag{3.1}$$

Note that we are using the notion of a tree interval which was introduced in the previous subsection, and that  $D^\circ(a, b)$  is a continuous analog of the right side of (2.2).

The function  $D^\circ$  is symmetric but does not satisfy the triangle inequality. We then consider the largest symmetric function that is bounded above by  $D^\circ$  and satisfies the triangle inequality: for every  $a, b \in \mathcal{T}_e$ ,

$$D(a, b) := \inf \left\{ \sum_{i=1}^k D^\circ(a_{i-1}, a_i) \right\}, \tag{3.2}$$

where the infimum is over all choices of the integer  $k \geq 1$  and of the elements  $a = a_0, a_1, \dots, a_{k-1}, a_k = b$  of  $\mathcal{T}_e$ .

Then  $D$  is a pseudo-metric on  $\mathcal{T}_e$ , and we set

$$a \approx b \quad \text{if and only if} \quad D(a, b) = 0.$$

**Definition 3.4.** The Brownian map is the quotient space  $\mathbf{m}_\infty := \mathcal{T}_e / \approx$  equipped with the distance induced by  $D$ .

We will write  $\Pi$  for the canonical projection from  $\mathcal{T}_e$  onto  $\mathbf{m}_\infty$ , and we keep the notation  $D$  for the induced distance on  $\mathbf{m}_\infty$ , so that  $D(\Pi(a), \Pi(b)) = D(a, b)$  for every  $a, b \in \mathcal{T}_e$ .

One can prove [29] that the property  $D(a, b) = 0$  holds if and only if  $D^\circ(a, b) = 0$ , or equivalently

$$Z_a = Z_b = \max \left( \min_{c \in [a,b]} Z_c, \min_{c \in [b,a]} Z_c \right). \tag{3.3}$$

In other words, two vertices of the CRT are identified if they have the same label  $Z_a = Z_b$  and if one can go from  $a$  to  $b$  “around the tree” (in clockwise or counterclockwise order) encountering only vertices with label greater than or equal to  $Z_a$ . In a sense, not many pairs of vertices are identified. Only leaves of the CRT may be identified (a leaf is a vertex  $a$  of  $\mathcal{T}_e$  such that  $\mathcal{T}_e \setminus \{a\}$  remains connected), and the set of all vertices  $a$  that are identified to

another one has Hausdorff dimension 1 (whereas the CRT had dimension 2). Furthermore, there is only a countable collection of equivalence classes of  $\approx$  that contain 3 points, and no equivalence class contains more than 3 points. Still these identifications drastically change the topology.

**Remark.** As already mentioned above, the preceding construction of the Brownian map is analogous to the construction of quadrangulations via well-labeled trees in the CVS bijection (and to the construction of more general planar maps using other similar bijections). The pair  $(\mathcal{T}_e, (Z_a)_{a \in \mathcal{T}_e})$  is a kind of continuous analog of a well-labeled tree  $(\tau, (\ell_v)_{v \in V(\tau)})$  (there is a minor difference because we do not impose a positivity condition on the labels in the continuous setting, but we could have re-rooted  $\mathcal{T}_e$  at the vertex with minimal label and shifted all labels  $Z_a$  so that they become nonnegative, which would not have affected the preceding construction). Still one may notice that no identification of vertices is needed in the discrete setting. The reason why such identifications become necessary in the continuous setting can be explained intuitively as follows. In a large well-labeled tree, there will exist vertices  $u$  and  $v$ , which are at a “macroscopic distance” in the tree, and such that  $\ell_v = \ell_u - 1$  and  $\ell_w \geq \ell_u$  for every vertex  $w \in [u, v]$ , where the interval  $[u, v]$  is as in (2.2) (note that (3.3) is just a continuous version of these properties): According to the rules of the CVS bijection, two such vertices are linked by an edge of the quadrangulation, and asymptotically, recalling that the graph distance is rescaled by a factor tending to 0, this leads to an identification of two vertices in the tree.

**3.5. Properties of the Brownian map.** In this subsection, we describe a few properties of the Brownian map, which show that this random metric space behaves very differently from a smooth surface.

**Proposition 3.5.** (i) *The Brownian map  $(\mathbf{m}_\infty, D)$  is a.s. homeomorphic to the sphere  $\mathbb{S}^2$ .*  
 (ii) *The Hausdorff dimension of  $(\mathbf{m}_\infty, D)$  is a.s. equal to 4.*

Part (i) is proved in [35], and part (ii) was derived in [29]. In view of Proposition 3.5, one may look at the Brownian map as a sphere with a very singular metric. Property (ii) is of course reminiscent of the fact that the Hausdorff dimension of a Brownian path in  $\mathbb{R}^d$ ,  $d \geq 2$  is twice the dimension of a smooth curve.

We next turn to uniform estimates on the volume of balls. We first need to introduce the volume measure on  $\mathbf{m}_\infty$ , which in a sense is uniformly distributed over this random space. We set  $\mathbf{p} := \Pi \circ p_e$ , which maps  $[0, 1]$  onto  $\mathbf{m}_\infty$ . The volume measure  $\Lambda$  on the Brownian map is the image of Lebesgue measure on  $[0, 1]$  under  $\mathbf{p}$ . We could also have introduced this volume measure via the convergence in Theorem 3.1, as the limit of the normalized counting measure on the vertex set of  $M_n$  (to make this rigorous, one should state the latter convergence in the sense of the Gromov-Hausdorff topology for measured metric spaces).

If  $x \in \mathbf{m}_\infty$  and  $r > 0$ , we write  $B_D(x, r)$  for the closed ball of radius  $r$  centered at  $x$  in  $(\mathbf{m}_\infty, D)$ .

**Proposition 3.6.** *Let  $\delta > 0$ . There exist two (random) positive constants  $c_\delta$  and  $C_\delta$  such that, for every  $r \in (0, 1]$  and every  $x \in \mathbf{m}_\infty$ ,*

$$c_\delta r^{4+\delta} \leq \Lambda(B_D(x, r)) \leq C_\delta r^{4-\delta}.$$

This shows that in a way the Brownian map is “very regular in its irregularity”: The volume of any small ball of radius  $r$  is approximately  $r^4$ . The upper bound in Proposition

3.6 is proved in [30], and the lower bound is easy from the bound  $D \leq D^\circ$  and the Hölder continuity properties of the process  $Z$ .

Another way of quantifying the irregularity of the Brownian map is to look at connected components of the complement of a ball. If this complement is not empty, it will have typically infinitely many connected components. The following proposition from [32] gives a more precise estimate.

**Proposition 3.7.** *Suppose that  $U$  is a point of  $\mathbf{m}_\infty$  chosen at random according to the volume measure  $\Lambda$ . Let  $r > 0$  and, for every  $\varepsilon > 0$ , let  $N_\varepsilon^r$  be the number of connected components of  $B_D(U, r)^c$  that intersect  $B_D(U, r + \varepsilon)^c$ . Then,*

$$\varepsilon^3 N_\varepsilon^r \xrightarrow{\varepsilon \rightarrow 0} c_0 \lambda_r,$$

where the convergence holds in probability,  $c_0 > 0$  is an explicit constant and  $\lambda_r$  is the value at  $r$  of the continuous density of the profile of distances from  $U$ , which is the random measure  $\mathcal{I}$  on  $\mathbb{R}_+$  defined by

$$\mathcal{I}(A) = \int \Lambda(dx) \mathbf{1}_A(D(U, x)),$$

for any Borel subset  $A$  of  $\mathbb{R}_+$ .

### 4. Geodesics in the Brownian map

Although the Brownian map remains a mysterious object in many respects, there is detailed information about the structure of geodesics toward a typical point. In this section we describe these results following [30]. We rely on the construction given in subsection 3.4.

One can prove (see in particular [36]) that there exists a unique vertex  $\rho_* \in \mathcal{T}_e$  such that

$$Z_{\rho_*} = \min_{c \in \mathcal{T}_e} Z_c.$$

We set  $x_* := \Pi(\rho_*)$ . In what follows we discuss geodesics toward  $x_*$  in the Brownian map. It should be emphasized that  $x_*$  is not a special point of the Brownian map: If  $U$  is a point of the Brownian map chosen at random according to the volume measure, one can check [30] that the random pointed compact metric spaces  $(\mathbf{m}_\infty, D, U)$  and  $(\mathbf{m}_\infty, D, x_*)$  have the same distribution, so that our results also give information about geodesics toward the “typical” point  $U$ .

To simplify notation, we set

$$Z_* := \min_{c \in \mathcal{T}_e} Z_c.$$

We first observe that, for every  $a \in \mathcal{T}_e$ ,

$$D(x_*, \Pi(a)) = Z_a - Z_*. \tag{4.1}$$

Indeed, it immediately follows from (3.1) that  $D(x_*, \Pi(a)) \leq D^\circ(\rho_*, a) = Z_a - Z_*$ . On the other hand, by using the easy bound

$$D^\circ(a, b) \geq |Z_a - Z_b|, \tag{4.2}$$

which also follows from (3.1), we immediately get from (3.2) that the reverse inequality  $D(x_*, \Pi(a)) = D(\rho_*, a) \geq Z_a - Z_*$  holds.

Let  $x = \Pi(a)$ ,  $a \in \mathcal{T}_e$  be a point in the Brownian map. We can construct certain geodesics from  $x$  to  $x_*$  in the following manner. Choose  $t \in [0, 1)$  such that  $p_e(t) = a$ , and for every  $r \in [0, Z_a - Z_*]$ , set

$$\gamma_t(r) := \begin{cases} \min\{s \in [t, 1] : Z_{p_e(s)} = Z_a - r\} & \text{if } \{s \in [t, 1] : Z_{p_e(s)} = Z_a - r\} \neq \emptyset, \\ \min\{s \in [0, t] : Z_{p_e(s)} = Z_a - r\} & \text{otherwise.} \end{cases}$$

Informally,  $p_e(\gamma_t(r))$  is the first vertex with label  $Z_a - r$  that one encounters when exploring the tree clockwise starting from  $a$ . Note that, when  $a$  is not a leaf, there are more than one way of starting from  $a$ : This corresponds to the different possible choices of  $t$ . Since  $\rho_*$  is the unique vertex with minimal label, it is clear that  $p_e(\gamma_t(Z_a - Z_*)) = \rho_*$ .

**Lemma 4.1.** *Let  $a \in \mathcal{T}_e$  and  $t \in [0, 1)$  such that  $p_e(t) = a$ . Set  $\Gamma_t(r) = \mathbf{p}(\gamma_t(r))$  for every  $r \in [0, Z_a - Z_*]$ . Then  $(\Gamma_t(r))_{0 \leq r \leq Z_a - Z_*}$  is a geodesic from  $\Pi(a)$  to  $x_*$  in  $(\mathbf{m}_\infty, D)$ . Such geodesics are called simple geodesics.*

The proof is easy. If  $0 \leq r \leq r' \leq Z_a - Z_*$ , the bound  $D \leq D^\circ$  and the definition (3.1) immediately shows that  $D(\Gamma_t(r), \Gamma_t(r')) \leq D^\circ(p_e(\gamma_t(r)), p_e(\gamma_t(r'))) = r' - r$ . On the other hand, by (4.1),  $D(\Gamma_t(0), \Gamma_t(Z_a - Z_*)) = D(\Pi(a), x_*) = Z_a - Z_*$ , and the triangle inequality now gives  $D(\Gamma_t(r), \Gamma_t(r')) = r' - r$  for every  $0 \leq r \leq r' \leq Z_a - Z_*$ .

Simple geodesics are indeed analogs in our continuous setting of the discrete geodesics for quadrangulations that were briefly discussed at the end of Section 2. The following proposition [30] is the key result in our study of geodesics.

**Proposition 4.2.** *All geodesics to  $x_*$  in  $(\mathbf{m}_\infty, D)$  are simple geodesics.*

This proposition makes it possible to classify all the geodesics to  $x_*$ . Indeed, it is easy to count simple geodesics. If we fix a point  $x \in \mathbf{m}_\infty$ , a simple geodesic  $\gamma_t$  from  $x$  to  $x_*$  is obtained by choosing first  $a \in \mathcal{T}_e$  such that  $\Pi(a) = x$ , and then  $t \in [0, 1)$  such that  $p_e(t) = a$ . The choice of  $a$  is in fact irrelevant: If  $\Pi^{-1}(x)$  is not a singleton, then the vertices  $a$  in  $\Pi^{-1}(x)$  must be leaves, then, for each such  $a$ , there is a unique  $t \in [0, 1)$  with  $p_e(t) = a$ , and one immediately verifies that the associated simple geodesics coincide. On the other hand, if  $x = \Pi(a)$  and  $a$  is not a leaf of  $\mathcal{T}_e$ , then there are (2 or 3) values of  $t$  such that  $p_e(t) = a$ , and the corresponding simple geodesics are distinct. The preceding discussion leads to the following theorem.

**Theorem 4.3.** *Let  $\text{Sk}(\mathcal{T}_e)$  stand for the set of all vertices of the CRT that are not leaves, and  $\text{Skel} := \Pi(\text{Sk}(\mathcal{T}_e))$ . Then the restriction of  $\Pi$  to  $\text{Sk}(\mathcal{T}_e)$  is a homeomorphism, and the Hausdorff dimension of  $\text{Skel}$  is equal to 2. Furthermore, a.s. for every  $x \in \mathbf{m}_\infty$ ,*

- if  $x \in \mathbf{m}_\infty \setminus \text{Skel}$ , there is a unique geodesic from  $x$  to  $x_*$ ;
- if  $x \in \text{Skel}$ , the number of distinct geodesics from  $x$  to  $x_*$  is the multiplicity of  $x$  in  $\text{Skel}$ , that is, the number of connected components of  $\text{Skel} \setminus \{x\}$ . This multiplicity is either 2 or 3.

The set  $\text{Skel}$ , which is a dense subset of the Brownian map homeomorphic to a non-compact real tree, thus appears as the cut-locus of the Brownian map with respect to the point  $x_*$ : A point  $x$  belongs to  $\text{Skel}$  if and only if there are at least two distinct geodesics from

$x$  to  $x_*$ . Perhaps suprisingly, there is a strong analogy with classical results of differential geometry that go back to Poincaré. For a smooth surface homeomorphic to the sphere, the cut-locus is also a tree, and the number of distinct geodesics from a point of the cut-locus is equal to its multiplicity.

It is easy to verify that the set  $\text{Skel}$  has zero volume measure (in terms of Hausdorff dimension, it is already clear that  $\text{Skel}$  is a “small” subset of  $\mathbf{m}_\infty$ ). A consequence of Theorem 4.3 is thus the fact that, if one picks independently two points  $x$  and  $x'$  according to the volume measure on  $\mathbf{m}_\infty$ , then a.s. there is a unique geodesic between  $x$  and  $x'$  (see also Miermont [40] for a related result with a different approach). From this and Theorem 3.1, one can deduce a property of “macroscopic uniqueness” of discrete geodesics in large planar maps, which was already mentioned in the introduction. See [30] for more details.

Theorem 4.3 shows that our construction of the Brownian map as a quotient space of the CRT has a strong geometric meaning (although it is certainly not the only possible construction). Indeed the set  $\text{Skel}$ , which is a homeomorphic image of the skeleton of the CRT in our construction, is a geometric object defined intrinsically in terms of the Brownian map, and thus does not depend on the particular construction we have developed.

**Remark.** The version of Theorem 4.3 given in [30] applies to any random compact metric space which appears as a Gromov-Hausdorff limit in distribution of rescaled  $2p$ -angulations (the uniqueness of such a limit was not yet known). This description of geodesics was then a key ingredient of the proof of Theorem 3.1 in [31]

The following corollary gives a confluence property of geodesics, which easily follows from the fact that two simple geodesics will always merge before their endpoint.

**Corollary 4.4.** *Let  $\delta > 0$ . Then a.s. there exists  $\varepsilon > 0$  such that, whenever  $x$  and  $x'$  are two points of  $\mathbf{m}_\infty$  with  $D(x_*, x) \geq \delta$  and  $D(x_*, x') \geq \delta$ , then, if  $f$  is a geodesic from  $x_*$  to  $x$  and  $g$  is a geodesic from  $x_*$  to  $y$ , we have  $f(r) = g(r)$  for every  $r \in [0, \varepsilon]$ .*

Informally, there is only one way of leaving  $x_*$  along a geodesic. By the remarks of the beginning of this section, the same property holds for a typical point chosen according to the volume measure on  $\mathbf{m}_\infty$ .

## 5. Infinite volume limits and the Brownian plane

**5.1. UIPT and UIPQ.** Theorem 3.1 deals with the convergence of uniformly distributed rooted  $p$ -angulations with  $n$  faces in the Gromov-Hausdorff sense, provided that the graph distance is rescaled by the factor  $n^{-1/4}$  when  $n$  tends to infinity. Note that this rescaling is necessary if we want to obtain a compact limit. On the other hand, one may also consider the convergence of the same random objects without rescaling, but then in a different sense than the Gromov-Hausdorff convergence. This leads to infinite random lattices.

We consider possibly infinite (multi)graphs that are always connected, pointed (meaning that there is a distinguished vertex  $\rho$  called the root vertex) and locally finite in the sense that the degree of every vertex is finite. As previously, these graphs are equipped with the graph distance. A ball of radius  $k$  centered at a vertex  $v$  of the graph is then viewed as the subgraph consisting of all vertices at distance less than or equal to  $k$  from  $v$  and the edges connecting these vertices. A sequence  $(G_n, \rho_n)$  of pointed graphs is said to converge locally to a limiting pointed graph  $(G, \rho)$  if for every integer  $k \geq 0$ , for every  $n$  sufficiently large, the ball of radius

$k$  centered at  $\rho_n$  in  $G_n$  is equal to the ball of radius  $k$  centered at  $\rho$  in  $G$  (equality here is in the sense of isomorphism between finite graphs with a distinguished vertex).

The following result is due to Angel and Schramm [7] in the case of triangulations and to Krikun [27] (see also [23]) in the case of quadrangulations.

**Theorem 5.1.** *Let  $p = 3$  or  $p = 4$ , and let  $M_n$  be uniformly distributed over  $\mathcal{M}_n^p$ . There exists a random infinite graph  $M_\infty^{(p)}$  such that*

$$M_n \xrightarrow[n \rightarrow \infty]{(d)} M_\infty^{(p)},$$

where the convergence holds in distribution in the sense of the local convergence of graphs.

In this theorem, the convergence just means that, for every integer  $k \geq 0$ , the probability that the ball of radius  $k$  centered at the root vertex in  $M_n$  is equal to a given graph converges to the same probability for the limit  $M_\infty^{(p)}$ . A completely different construction of  $M_\infty^{(4)}$  based on a version of the CVS bijection for infinite trees was given in [17] (see [38] for a proof of the fact that the two constructions give the same object).

**Remark.** One may also define the limit in Theorem 5.1 as an infinite planar map, that is, with a given embedding in the sphere, and this makes it possible to give a stronger form of the local convergence (see [7] and [27]). Here for simplicity we avoid dealing with infinite planar maps, since only the graph structure will play a role in the subsequent statements.

The infinite random graph  $M_\infty^{(p)}$  is called the uniform infinite planar triangulation (UIPT) when  $p = 3$  and the uniform infinite planar quadrangulation (UIPQ) when  $p = 4$ . Properties of these infinite random graphs have been investigated in detail in the recent years. In particular, Angel [6] has studied percolation on the UIPT. The recurrence of simple random walk on the UIPT or the UIPQ has been obtained recently by Gurel-Gurevich and Nachmias [26]. See also Benjamini and Curien [10] for a proof of subdiffusivity of simple random walk on the UIPQ.

**5.2. Convergence to the Brownian plane.** The following results are taken from [21]. To simplify notation, we write  $Q_\infty = M_\infty^{(4)}$  for the UIPQ, which was introduced in the previous subsection. The next theorem shows that, if we rescale the graph distance on the UIPQ by a factor tending to 0, the resulting metric spaces converge (in a suitable sense) toward a limiting random non-compact metric space, which is called the Brownian plane. We recall that a pointed metric space is just a metric space equipped with a distinguished point.

**Theorem 5.2.** *Let  $V(Q_\infty)$  denote the vertex set of  $Q_\infty$ , which is equipped with the graph distance  $d_{gr}$ , and let  $\rho_{Q_\infty}$  stand for the root vertex of  $Q_\infty$ . There exists a random non-compact pointed metric space  $(\mathcal{P}, D_\infty, \rho_\infty)$  called the Brownian plane such that*

$$(V(Q_\infty), \lambda d_{gr}, \rho_{Q_\infty}) \xrightarrow[\lambda \rightarrow 0]{(d)} (\mathcal{P}, D_\infty, \rho_\infty),$$

where the convergence holds in distribution in the local Gromov-Hausdorff sense.

We refer to [16] for a precise definition of the local Gromov-Hausdorff convergence for (non-compact) pointed metric spaces. In the present setting, this means that, for every  $r > 0$ , the ball of radius  $r$  centered at  $\rho_{Q_\infty}$  in  $(V(Q_\infty), \lambda d_{gr})$  will converge in distribution in the

Gromov-Hausdorff sense to the ball of radius  $r$  centered at  $\rho_\infty$  in the limiting space  $(\mathcal{P}, D_\infty)$  – in fact we should use here the Gromov-Hausdorff distance between pointed compact metric spaces, which is defined by a minor modification of the definition in subsection 3.1.

The Brownian plane  $(\mathcal{P}, D_\infty, \rho_\infty)$  appears in several other limit theorems. In particular, if in the convergence of Theorem 3.1 for  $p = 4$  one replaces the scaling factor  $n^{-1/4}$  by a function  $\beta(n)$  tending to 0 but such that  $n^{1/4}\beta(n)$  tends to infinity, the convergence still holds in the local Gromov-Hausdorff sense and the limit is now the Brownian plane.

Alternatively, the Brownian plane can be viewed as the tangent cone in distribution of the Brownian map at a distinguished vertex  $U$  chosen at random according to the volume measure  $\Lambda$ . In fact, a stronger property holds: One can construct, on the same probability space, both the Brownian map  $\mathbf{m}_\infty$  and the Brownian plane  $\mathcal{P}$ , in such a way that, a.s., there exists a (random)  $\varepsilon > 0$  such that the balls of radius  $\varepsilon$  centered at the distinguished point in  $\mathbf{m}_\infty$  and in  $\mathcal{P}$  are isometric. The latter fact, together with the scale invariance property of  $\mathcal{P}$  (see subsection 5.3 below), allows one to derive many properties of the Brownian plane from those known for the Brownian map.

We will now give a precise construction of the Brownian plane. Not surprisingly, this construction is very similar to the one developed above for the Brownian map. We consider two independent three-dimensional Bessel processes  $R$  and  $R'$  started from 0 (see e.g. [43] for basic facts about Bessel processes). We then define a process  $Y = (Y_t)_{t \in \mathbb{R}}$  indexed by the real line, by setting

$$Y_t := \begin{cases} R_t & \text{if } t \geq 0, \\ R'_{-t} & \text{if } t \leq 0. \end{cases}$$

Then, for every  $s, t \in \mathbb{R}$ , we set

$$m_Y(s, t) := \inf_{r \in \overline{st}} Y_r,$$

with the notation  $\overline{st} = [s \wedge t, s \vee t]$  if  $st \geq 0$ ,  $\overline{st} = (-\infty, s \wedge t] \cup [s \vee t, \infty)$  if  $st < 0$ . We define a random pseudo-distance on  $\mathbb{R}$  by

$$d_Y(s, t) := Y_s + Y_t - 2m_Y(s, t)$$

and set  $s \sim_Y t$  if  $d_Y(s, t) = 0$ . The quotient space  $\mathcal{T}_\infty = \mathbb{R} / \sim_Y$  equipped with  $d_Y$  is a (non-compact) random real tree, which is sometimes called the infinite Brownian tree. We write  $p_\infty : \mathbb{R} \rightarrow \mathcal{T}_\infty$  for the canonical projection and set  $\rho_\infty := p_\infty(0)$ , which plays the role of the root of  $\mathcal{T}_\infty$ . The volume measure on  $\mathcal{T}_\infty$  is the image of Lebesgue measure on  $\mathbb{R}$  under  $p_\infty$ .

We next consider Brownian motion indexed by  $\mathcal{T}_\infty$ . Formally, we consider a real-valued process  $(Z_t^\infty)_{t \in \mathbb{R}}$  such that, conditionally given the process  $Y$ ,  $Z^\infty$  is a centered Gaussian process with covariance

$$E[Z_s^\infty Z_t^\infty \mid Y] = m_Y(s, t),$$

so that we have  $Z_0^\infty = 0$  and  $E[(Z_s^\infty - Z_t^\infty)^2 \mid Y] = d_Y(s, t)$ . It is not hard to verify that the process  $Z^\infty$  has a modification with continuous paths. Then a.s. we have  $Z_s^\infty = Z_t^\infty$  for every  $s, t \in \mathbb{R}$  such that  $d_Y(s, t) = 0$  and therefore we may view  $Z^\infty$  as indexed by  $\mathcal{T}_\infty$ .

For every  $s, t \in \mathbb{R}$ , we set

$$D_\infty^\circ(s, t) := Z_s^\infty + Z_t^\infty - 2 \min_{r \in [s \wedge t, s \vee t]} Z_r^\infty.$$

We extend the definition of  $D_\infty^\circ$  to  $\mathcal{T}_\infty \times \mathcal{T}_\infty$  by setting for  $a, b \in \mathcal{T}_\infty$ ,

$$D_\infty^\circ(a, b) := \min\{D_\infty^\circ(s, t) : s, t \in \mathbb{R}, p_\infty(s) = a, p_\infty(t) = b\}.$$

Finally, we set, for every  $a, b \in \mathcal{T}_\infty$ ,

$$D_\infty(a, b) := \inf_{a_0=a, a_1, \dots, a_k=b} \sum_{i=1}^k D_\infty^\circ(a_{i-1}, a_i)$$

where the infimum is over all choices of the integer  $k \geq 1$  and of the finite sequence  $a_0, a_1, \dots, a_k$  in  $\mathcal{T}_\infty$  such that  $a_0 = a$  and  $a_k = b$ . Then  $D_\infty$  is a pseudo-distance on  $\mathcal{T}_\infty$ , and we set  $a \approx b$  if  $D_\infty(a, b) = 0$  (this can be proved to be equivalent to the property  $D_\infty^\circ(a, b) = 0$ ). The Brownian plane is the quotient space  $\mathcal{P} = \mathcal{T}_\infty / \approx$ , which is equipped with the metric induced by  $D_\infty$  and with the distinguished point which is the equivalence class of  $\rho_\infty$  (without risk of confusion, we still write  $\rho_\infty$  for this equivalence class). The volume measure on  $\mathcal{P}$  is the image of the volume measure on  $\mathcal{T}_\infty$  under the canonical projection.

**5.3. Properties of the Brownian plane.** The Brownian plane is scale invariant, meaning that, for every  $\lambda > 0$ , the space  $(\mathcal{P}, \lambda D_\infty, \rho_\infty)$  has the same distribution as  $(\mathcal{P}, D_\infty, \rho_\infty)$ . This property can be derived from the explicit construction given above, or from the fact that the Brownian plane is a tangent cone in distribution to the Brownian map.

Other properties of the Brownian plane are very similar to those of the Brownian map. We state an analog of Proposition 3.5.

**Proposition 5.3.** *Almost surely, the Brownian plane  $(\mathcal{P}, D_\infty)$  is homeomorphic to the Euclidean plane  $\mathbb{R}^2$  and has Hausdorff dimension 4.*

From the construction of the previous subsection, the points of the Brownian plane inherit labels  $Z_x^\infty$  from the corresponding labels on the infinite Brownian tree  $\mathcal{T}_\infty$ . These labels can be interpreted as “distances from infinity” in the following sense. For every  $x, y \in \mathcal{P}$ ,

$$Z_x^\infty - Z_y^\infty = \lim_{z \rightarrow \infty} (D_\infty(x, z) - D_\infty(y, z)).$$

The existence of the preceding limit is related to a property of confluence of geodesic rays in the Brownian plane. Recall that a geodesic ray is a continuous path  $\gamma : [0, \infty) \rightarrow \mathcal{P}$  such that  $D_\infty(\gamma(t), \gamma(t')) = |t - t'|$  for every  $t, t' \geq 0$ .

**Proposition 5.4.** *Let  $\gamma$  and  $\gamma'$  be two geodesic rays in  $\mathcal{P}$ . Then there exists two reals  $\alpha$  and  $\beta \geq |\alpha|$  such that  $\gamma(t) = \gamma'(\alpha + t)$  for every  $t \geq \beta$ .*

This is a kind of version at infinity of Corollary 4.4. Note that an exact analog of Corollary 4.4 also holds for the Brownian plane, and is easy to prove by the coupling argument mentioned above.

The Brownian plane seems to be more tractable than the Brownian map for explicit calculations, partly because of its scale invariance. Let us discuss some recent results from [22] that shed light on the probabilistic structure of the Brownian plane. For every  $r > 0$ , we let  $B_r$  be the closed ball of radius  $r$  centered at  $\rho_\infty$  in  $\mathcal{P}$ , and we define the “extended ball”  $B_r^\bullet$  as the complement of the unbounded connected component of  $(B_r)^c$ . Informally,  $B_r^\bullet$  is obtained by “filling the holes” in  $B_r$ . We write  $|B_r^\bullet|$  for the volume of  $B_r^\bullet$ .

**Proposition 5.5.** *For every  $\lambda > 0$ ,*

$$E[\exp(-\lambda|B_r^\bullet|)] = \frac{3^{3/2} \cosh((2\lambda)^{1/4}r)}{\left(\cosh^2((2\lambda)^{1/4}r) + 2\right)^{3/2}}.$$

In fact, one can give a simple description of the whole process  $(|B_r^\bullet|)_{r \geq 0}$ . Set  $\psi(\lambda) := (\frac{8}{3})^{1/2} \lambda^{3/2}$  and recall that the continuous-state branching process with branching mechanism  $\psi$  is the Markov process in  $\mathbb{R}_+$  whose transition kernels are characterized by the Laplace transform

$$E[\exp(-\lambda X_t) \mid X_0 = x] = \exp(-x u_t(\lambda))$$

where the function  $u_t(\lambda)$  solves the differential equation  $\frac{du_t(\lambda)}{dt} = -\psi(u_t(\lambda))$  with initial condition  $u_0(\lambda) = \lambda$  (see e.g. [28]). Note that the sample paths of  $X$  are right-continuous with left limits and that  $X$  is absorbed at 0. We can define a process  $X^\infty = (X_t^\infty)_{t \leq 0}$ , which is indexed by negative times and corresponds to  $X$  “started from  $+\infty$  at time  $-\infty$  and conditioned to hit 0 at time 0” (formally, we start  $X$  at time 0 from  $x > 0$ , we then shift time so that the hitting time of 0 becomes 0, and we finally let  $x$  tend to  $+\infty$ ).

**Proposition 5.6.** *The process  $(|B_r^\bullet|)_{r \geq 0}$  has the same distribution as the process  $(W_r)_{r \geq 0}$  defined by*

$$W_r = \sum_{-r \leq u \leq 0} \xi_u (\Delta X_u^\infty)^2,$$

where  $\Delta X_u^\infty$  stands for the jump of  $X^\infty$  at time  $u$ , and, conditionally given  $X^\infty$ , the non-negative random variables  $\xi_u$  are independent and identically distributed with density

$$\frac{1}{\sqrt{2\pi}} x^{-5/2} e^{-1/2x}.$$

Informally, each jump of the process  $(|B_r^\bullet|)_{r \geq 0}$  corresponds to the creation of a new connected component of  $(B_r)^\complement$  (there are many such components, as shown by Proposition 3.7 in the Brownian map case), noting that this newly created connected component will be “swallowed” by the extended ball. For  $r \geq 0$ , the random variable  $X_{-r}^\infty$  should be interpreted as the length, in a generalized sense, of the boundary of  $B_r^\bullet$ . Note that, when a connected component is swallowed, the length of the boundary of  $B_r^\bullet$  has a negative jump. The distribution of the variables  $\xi_u$  then corresponds to the law of the volume of a connected component given that its boundary has length 1.

The preceding interpretation is closely related to some results of Krikun [27] in the discrete setting of the UIPQ. The description of the process  $(|B_r^\bullet|)_{r \geq 0}$  can also be interpreted in terms of the “peeling process” of the UIPT studied in [6].

### 6. Canonical embeddings and open questions

In this section, we come back to the questions that were discussed at the beginning of the introduction above. We would like to have a “canonical” construction of a random metric  $\Delta$  on the sphere  $\mathbb{S}^2$ , in such a way that

$$(\mathbf{m}_\infty, D) \stackrel{(d)}{=} (\mathbb{S}^2, \Delta).$$

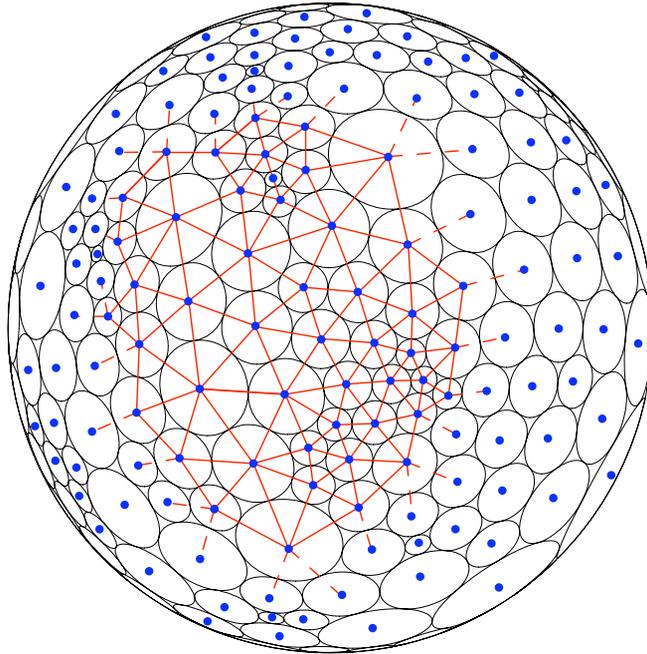


Figure 6.1. A circle packing of the sphere associated with a simple triangulation.

Furthermore we expect  $\Delta$  to behave well under the conformal transformations of the sphere.

Recall that the embedding of a planar map is defined up to orientation-preserving homeomorphisms of the sphere. However, there are (almost) canonical ways of choosing these embeddings. Consider the case of simple triangulations, that is, triangulations without loops or multiple edges. According the circle packing theorem, any such triangulation can be represented via a circle packing of the sphere, in such a way that the vertex set of the triangulation is the set of centers of all circles, and two vertices are linked by an edge if and only if the associated circles are tangent (see Fig. 5 for an example). This representation is in fact unique up to the conformal transformations of the sphere (the Möbius transformations).

Next suppose that, for every even integer  $n \geq 2$ , we have constructed a circle-packing embedding  $\mathcal{C}_n$  of a uniformly distributed simple triangulation with  $n$  faces. Write  $V(\mathcal{C}_n)$  for the vertex set of  $\mathcal{C}_n$  and  $d_{\text{gr}}^n$  for the graph distance on  $V(\mathcal{C}_n)$ . By Theorem 3.1, or more precisely by the extension of Theorem 3.1 to simple triangulations found in [2], we have

$$\left( V(\mathcal{C}_n), \left(\frac{3}{2}\right)^{1/4} n^{-1/4} d_{\text{gr}}^n \right) \xrightarrow[n \rightarrow \infty]{(d)} (\mathbf{m}_\infty, D)$$

in the Gromov-Hausdorff sense. Note that the constant  $(3/2)^{1/4}$  is different from the constant  $6^{1/4}$  in (1.1) or in Theorem 3.1, because we are dealing with simple triangulations.

**Conjecture.** *One can construct the circle packing embeddings  $\mathcal{C}_n$  in such a way that*

$$\sup_{x \in \mathbb{S}^2} \left( \min_{y \in V(\mathcal{C}_n)} |x - y| \right) \xrightarrow[n \rightarrow \infty]{} 0$$

*in probability, and there exists a continuous random process  $(\Delta(x, y))_{x, y \in \mathbb{S}^2}$ , which is nonzero*

outside the diagonal and such that

$$\sup_{x,y \in V(\mathcal{C}_n)} \left| \Delta(x,y) - \left(\frac{3}{2}\right)^{1/4} n^{-1/4} d_{\text{gr}}^n(x,y) \right| \xrightarrow{n \rightarrow \infty} 0$$

in probability.

If the conjecture holds, then it is not hard to verify that  $\Delta$  defines a random metric on the sphere and that  $(\mathbb{S}^2, \Delta)$  gives a representation of the Brownian map. We note that there are other ways of defining canonical embeddings of planar maps, which would lead to other versions of the conjecture. In particular, we can associate with a triangulation of the sphere a Riemann surface, which is obtained by viewing each face as an equilateral triangle with sides of length 1 and then gluing two adjacent triangles along their common edge. The resulting Riemann surface is homeomorphic to the sphere and thus the uniformization theorem yields a canonical embedding, again modulo Möbius transformations. See [20] for an application of these ideas to the conformal structure of random planar maps.

As a final remark, we note that a very recent work of Miller and Sheffield [42] has introduced a new growth process called Quantum Loewner Evolution (QLE), which might provide a direct construction of the Brownian map, or perhaps of the Brownian plane discussed in the previous section, with conformal invariance properties. Assuming that this construction goes through, one may expect new fascinating connections between the Brownian map on one hand, the Schramm-Loewner evolutions and the two-dimensional Gaussian free field on the other hand.

## References

- [1] Abraham, C., *Rescaled bipartite planar maps converge to the Brownian map*, Preprint (2013), available at: arXiv:1312.5959.
- [2] Addario-Berry, L. and Albenque, M., *The scaling limit of random simple triangulations and random simple quadrangulations*, Preprint (2013), available at: arXiv:1306.5227.
- [3] Aldous, D., *The continuum random tree I*, Ann. Probab., **19** (1991), 1–28.
- [4] ———, *The continuum random tree III*, Ann. Probab., **21** (1993), 248–289.
- [5] Ambjørn, J., Durhuus, B., and Jonsson, T., *Quantum Geometry*, A Statistical Field Theory Approach, Cambridge Monographs on Mathematical Physics, Cambridge Univ. Press, Cambridge, 1997.
- [6] Angel, O., *Growth and percolation on the uniform infinite planar triangulation*, Geomet. Funct. Anal. **3** (2003), 935–974.
- [7] Angel, O. and Schramm, O., *Uniform infinite planar triangulations*, Comm. Math. Phys., **241** (2003), 191–213.
- [8] Beltran, J. and Le Gall, J.-F., *Quadrangulations with no pendant vertices*, Bernoulli **19** (2013), 1150–1175.
- [9] Benjamini, I., *Random planar metrics*, Proceedings of the International Congress of Mathematicians, Vol. IV, 2177–2187, Hindustan Book Agency, New Delhi, 2010.
- [10] Benjamini, I. and Curien, N., *Simple random walk on the uniform infinite planar quadrangulation: subdiffusivity via pioneer points*, Geom. Funct. Anal., **23** (2013), 501–531.

- [11] Bettinelli, J., Jacob, E., and Miermont, G., *The Scaling Limit of Uniform Random Plane Maps, via the Ambjørn-Budd Bijection*, Preprint (2013), available at: arXiv:1312.5842.
- [12] Bouttier, J., Di Francesco, P., and Guitter, E., *Planar maps as labeled mobiles*, Electronic J. Combinatorics **11** (2004), #R69.
- [13] Bouttier, J. and Guitter, E., *Statistics in geodesics in large quadrangulations*, J. Phys. A **41** (2008), 145001, 30 pp.
- [14] ———, *The three-point function of planar quadrangulations*, J. Stat. Mech. Theory Exp. (2008), P07020.
- [15] ———, *Confluence of geodesic paths and separating loops in large planar quadrangulations*, J. Stat. Mech. Theory Exp. (2009), P03001.
- [16] Burago, D., Burago, Y., and Ivanov, S., *A Course in Metric Geometry*, Graduate Studies in Mathematics, vol. 33. Amer. Math. Soc., Boston, 2001.
- [17] Chassaing, P. and Durhuus, B., *Local limit of labeled trees and expected volume growth in a random quadrangulation*, Ann. Probab. **34** (2006), 879–917.
- [18] Chassaing, P. and Schaeffer, G., *Random planar lattices and integrated superBrownian excursion*, Probab. Th. Rel. Fields **128** (2004), 161–212.
- [19] Cori, R. and Vauquelin, B., *Planar maps are well labeled trees*, Canad. J. Math., **33** (1981), 1023–1042.
- [20] Curien, N., *A glimpse of the conformal structure of random planar maps*, Preprint (2013), available at: arXiv:1308.1807.
- [21] Curien, N. and Le Gall, J.-F., *The Brownian plane*, J. Theoret. Probab., to appear.
- [22] ———, in preparation.
- [23] Curien, N., Ménard, L., and Miermont, G., *A view from infinity of the uniform infinite quadrangulation*, ALEA Lat. Am. J. Probab. Math. Stat. **10** (2013), 45–88.
- [24] Duplantier, B. and Sheffield, S., *Liouville quantum gravity and KPZ*, Invent. Math., **185** (2011), 333–393.
- [25] Duquesne, T. and Le Gall, J.-F., *Probabilistic and fractal aspects of Lévy trees*, Probab. Th. Rel. Fields **131** (2005), 553–603.
- [26] Gurel-Gurevich, O. and Nachmias, A., *Recurrence of planar graph limits*, Ann. Math., **177** (2013), 761–781.
- [27] Krikun, M., *Local structure of random quadrangulations*, Preprint, math:PR/0512304.
- [28] Le Gall, J.-F., *Spatial branching processes, random snakes and partial differential equations*, Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 1999.
- [29] ———, *The topological structure of scaling limits of large planar maps*, Invent. Math. **169** (2007), 621–670.
- [30] ———, *Geodesics in large planar maps and in the Brownian map*, Acta Math. **205** (2010), 287–360.
- [31] ———, *Uniqueness and universality of the Brownian map*, Ann. Probab. **41** (2013), 2880–2960.
- [32] ———, *The Brownian cactus II. Upcrossings and local times of super-Brownian motion*, Preprint, arXiv:1308.6762.
- [33] Le Gall, J.-F., Miermont, G., *Scaling limits of random planar maps with large faces*,

- Ann. Probab. **39** (2011), 1–69.
- [34] ———, *Scaling limits of random trees and planar maps*, Probability and statistical physics in two and more dimensions, Clay Math. Proc., 15. Amer. Math. Soc., Providence, 2012, pp. 155–211.
- [35] Le Gall, J.-F., and Paulin, F., *Scaling limits of bipartite planar maps are homeomorphic to the 2-sphere*, Geomet. Funct. Anal. **18** (2008), 893–918.
- [36] Le Gall, J.-F., and Weill, M., *Conditioned Brownian trees*, Annales Inst. H. Poincaré Probab. Stat. **42** (2006), 455–489.
- [37] Marckert, J.-F. and Mokkadem, A., *Limit of normalized quadrangulations: the Brownian map*, Ann. Probab. **34** (2006), 2144–2202.
- [38] Ménard, L., *The two uniform infinite quadrangulations of the plane have the same law*, Ann. Inst. H. Poincaré Probab. Stat. **46** (2010), 190–208.
- [39] Miermont, G., *On the sphericity of scaling limits of random planar quadrangulations*, Electron. Commun. Probab., **13** (2008), 248–257.
- [40] ———, *Tessellations of random maps of arbitrary genus*, Ann. Sci. École Norm. Sup. **42** (2009), 725–781.
- [41] ———, *The Brownian map is the scaling limit of uniform random plane quadrangulations*, Acta Math. **210** (2013), 319–401.
- [42] Miller, J. and Sheffield, S., *Quantum Loewner Evolution*, Preprint, arXiv:1312.5745.
- [43] Revuz, D. and Yor, M., *Continuous Martingales and Brownian Motion*, Springer, Berlin, 1991.
- [44] Schramm, O., *Conformally invariant scaling limits: an overview and a collection of problems*, Proceedings of the International Congress of Mathematicians, (Madrid 2006), Vol. I, 513–543. European Math. Soc., Zürich, 2007.
- [45] Sheffield, S., *Conformal weldings of random surfaces: SLE and the quantum gravity zipper*, Preprint, arXiv:1012.4797.
- [46] Tutte, W.T., *A census of planar maps*, Canad. J. Math., **15** (1963), 249–271.

Université Paris-Sud, Mathématiques, Bât.425, 91405 Orsay Cédex, France

E-mail: jean-francois.legall@math.u-psud.fr

# Analytic low-dimensional dynamics: From dimension one to two

Mikhail Lyubich

**Abstract.** Let  $f : M \rightarrow M$  be an analytic (real or complex) self-map of a manifold, and let  $f^n$  stand for its  $n$ -fold iterate. The theory of Analytic Dynamical Systems with discrete time is concerned with understanding the asymptotic behavior of orbits ( $f^n x$ ). The main goal, as it was articulated in the second half of 20th century, is to describe, in probabilistic terms, the asymptotic distribution of typical orbits for typical systems. This goal is now achieved for unimodal one-dimensional maps, a great progress has been made in complex one-dimensional case, and a transition to the dissipative two-dimensional situation, real and complex, is underway. Renormalization ideas played a crucial role in this story. We will describe all these developments in their interplay.

**Mathematics Subject Classification (2010).** 37Exx (particularly, 05, 20, 30) and 37Fxx (particularly, 10, 15, 25, 30, 45).

**Keywords.** Hyperbolicity, structural stability, attractor, renormalization, homoclinic tangency, Julia set, Henon map, a priori bounds

## 1. Historical and conceptual background

**1.1. Homoclinic intersections: first glimpse into chaos.** In his fundamental memoirs on Celestial Mechanics, Poincaré came across a dynamical situation later called a *homoclinic intersection*, i.e., the intersection between the stable and unstable manifolds of a saddle periodic point. Attempts to comprehend it made him desperate: “One is struck by complexity of this figure that I am not even attempting to draw. Nothing can give us a better idea of the complexity of the three body problem and of all the problems of dynamics in general”.

**1.2. Hyperbolicity, structural stability, and combinatorial models.** In the 1960s Smale came up with a simple model, the *horseshoe*, that captures some complexity caused by *transverse* homoclinic intersections. This led to an idea of *hyperbolicity*, one of the central concepts of contemporary dynamics. Roughly speaking, it means that over a recurrent part of the phase space there exist two transverse invariant foliations, stable and unstable, which are (respectively) uniformly exponentially contracted and expanded by the dynamics. This implies that all recurrent trajectories are exponentially unstable, either in forward or in backward time. Remarkably, as Anosov demonstrated in 1967, hyperbolic systems are *structurally stable*,<sup>1</sup> i.e., all their complexity is qualitatively preserved under perturbations.

Also, an efficient combinatorial way of describing chaos<sup>2</sup> was developed. If we have a tiling of the phase space by  $k$  sets  $Y_i$ , then we can encode the orbits by sequences in  $k$

---

<sup>1</sup>Proceedings of the International Congress of Mathematicians, Seoul, 2014

symbols, turning the dynamics into the shift on some space of sequences. This produces a *combinatorial model* for the dynamics.

How good this model is depends on the quality of the sets, the character of their intersections, and the character of intersections  $f^{-1}(Y_i) \cap Y_j$ . A particularly nice situation is when the coding is *Markov*, i.e., the space of sequences is fully determined by admissible transitions  $i \rightsquigarrow j$ . It turns out that hyperbolic systems admit a nice *Markov model* that captures many essential features of the topological dynamics (Adler & Weiss, Sinai, around 1970). For instance, one important consequence is that *periodic orbits are equidistributed* with respect to a canonical measure called the *measure of maximal entropy* (Bowen, 1971).

For an introduction to the Hyperbolicity Theory, see e.g., Bowen's classical Lecture Notes [27].

**1.3. Newhouse phenomenon and Homoclinic tangencies.** On the other hand, it was soon discovered that hyperbolic systems are quite scarce, and in particular, not dense in the space of dynamical systems on a given manifold (except for the one-dimensional case, as we will see below). A most impressive manifestation of it was the discovery of the *Newhouse phenomenon* (1979) exhibiting infinitely many co-existing attracting cycles for a set of parameters that densely fill some domain in the parameter space. It revived the feeling that the world of dynamical systems is too complex to be understood in any comprehensive way.

The Newhouse phenomenon is intimately related to appearance of homoclinic intersections (again!) but this time *non-transverse*, see the book by Palis and Takens [98]. A quarter of a century later, Palis suggested that in dimension two lack of hyperbolicity is always related to this phenomenon:<sup>3</sup>

**Palis Conjecture** ([97]). *For a real two-dimensional surface  $M$ , hyperbolic maps together with maps with homoclinic tangencies form a dense subset of  $\text{Diff}^\omega(M)$ .*

In §5 we will describe a recent advance in the *complex analytic* version of the Palis Conjecture.

**1.4. Probabilistic viewpoint.** The above developments suggest that when dealing with chaotic dynamical systems (depending on some parameters), there is little chance to describe, even qualitatively, all trajectories of every single system in the class. Instead, one can try to look for typical phenomena within the class.

This immediately raises a question: What should be considered “typical”? Since the beginning of the 20th century there has existed two competing approaches to this issue, from the measure-theoretic (or rather, probabilistic) viewpoint and from the topological (Baire category) viewpoint. In his address to the ICM in Amsterdam in 1954, Kolmogorov compared these two viewpoints:

“Approach from the categorical side is interesting more like a tool of proving existence results..., while an approach from measure-theoretic side seems to be physically reasonable and natural... but faces the absence of a natural measure in functional spaces”. This viewpoint is generally accepted in our days. Though Dynamical Systems Theory is a highly non-homogeneous field with many different flavors, there seems to be a general agreement on what is its main goal: To study asymptotic behavior of *almost all* orbits for *almost any*

<sup>1</sup>This notion, under the name of *robustness* goes back to Andronov and Pontryagin (1937).

<sup>2</sup>with the idea going back to Hadamard (1897) and Morse (1921)

<sup>3</sup>This is the real analytic version of the conjecture that was originally concerned with  $C^r$ -diffeomorphisms.

parameter value in a *representative* finite parameter family of dynamical systems.

Of course, this formulation raises several questions: 1) in what terms the asymptotic behavior of orbits can be described? 2) “almost all” with respect to which measure? 3) what dynamical systems are “representative”?

**1.5. Physical and SRB measures.** Kolmogorov’s ideas were intensely developed in the 1970s, by means of *Ergodic Theory*, particularly by the Moscow School led by Arnold and Sinai. The *Birkhoff Ergodic Theorem* asserts that in the presence of an invariant ergodic measure<sup>4</sup>  $\mu$ , almost all orbits are equidistributed with respect to  $\mu$ , i.e., for  $\mu$ -a.e.  $x$ ,

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{f^k x} \rightarrow \mu \quad \text{as } n \rightarrow \infty, \quad (1.1)$$

where  $\delta_y$  is the delta-measure supported at the point  $y$ . (When (1.1) happens, one also says that  $\mu$  governs the behavior of the orbit of  $x$ .)

The drawback of this fundamental result is that the measure  $\mu$  can be *singular* with respect to the Lebesgue measure on our manifold  $M$  (so  $\mu$ -typical points form quite a thin set), while there may not be invariant measures in the Lebesgue measure class. This issue can be addressed by introducing the following notion: an invariant measure  $\mu$  is called *physical* if the set of points whose orbits are governed by  $\mu$  has *positive Lebesgue measure*.

At the same time, a good thought was given to the problem of relaxing the notion of hyperbolicity so that it could be more representative. In the late 1970s Oseledets and Pesin developed a very general Hyperbolicity Theory for an invariant measure  $\mu$ . Roughly speaking, this means the existence of transverse stable and unstable manifolds for  $\mu$ -typical points, which are exponentially contracted and expanded by the dynamics but perhaps in a *non-uniform* way. In particular, it yields positivity of the leading *Lyapunov exponent*,<sup>5</sup>

$$\chi(\mu) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \|Df^n(x)\| > 0, \quad (1.2)$$

and hence exponential instability, for  $\mu$ -almost all orbits.

We can now introduce a very important class of physical measures: a non-uniformly measure which is not supported on a cycle and whose conditional measures on the unstable foliation are absolutely continuous with respect to the leafwise Lebesgue measure is called an *SRB measure* (after Sinai, Ruelle and Bowen). Though the absolute continuity property sounds quite technical, it gives a deep insight into the geometric nature of an SRB measure.

We refer the reader to [25] for the background in Non-Uniformly Hyperbolic Dynamics.

**1.6. Attractors.** The notion of a “*strange*” attractor played an inspiring role in the 1970-80s. Any invariant set with somewhat complicated topology that attracts “many” points was regarded to be a strange attractor. Examples included the Smale solenoid, Lorenz, Hénon, and Feigenbaum attractors (see below). The notion itself was coined by Ruelle and Takens who proposed it as a mathematical foundation for the turbulence phenomenon.

However, mathematically the situation deteriorated fast because of the lack of agreement what exactly this notion means. Eventually, this issue was rigorously addressed from the

<sup>4</sup>All measures are assumed to be probabilistic. *Ergodicity* means that the phase space cannot be decomposed into two disjoint invariant subsets of positive measure.

<sup>5</sup>In the ergodic case, this limit exists a.e., and is independent of  $x$ .

measure-theoretic point of view by Milnor (1985) and Palis (1995). Let us give a definition that stems from that discussion.

A compact invariant set  $A$  is called a *physical attractor* if it supports a physical measure  $\mu$  (so that  $\text{supp } \mu = A$ ). If additionally,  $\mu$  is an SRB measure then  $A$  is called an *SRB attractor*. (We will also say that a physical attractor is *global* if it governs the behavior of almost all orbits of  $M$ .)

For hyperbolic systems, there exist finitely many attracting cycles and SRB attractors that govern the behavior of Lebesgue almost all orbits (Sinai-Ruelle-Bowen, see [27]). In general, both Milnor and Palis emphasized the problem of finiteness of the number of attractors. Moreover, Palis conjectured that systems with this property are *dense* in the space of all dynamical systems. Stemming from it, as well as from developments in the one-dimensional case, the following conjecture has been gradually shaped, particularly in the low-dimensional setting:

**Fundamental Conjecture.** *For a typical analytic dynamical system, there exist finitely many attracting cycles and SRB attractors that govern the behavior of Lebesgue almost all orbits.*

Here a “typical system” is understood in *Kolmogorov’s sense* as Lebesgue almost any system in a generic finite parameter family of systems. Of course, the quality of this notion depends on the exact meaning of “generic”. We will return to this issue later on.

The author believes that the Fundamental Conjecture is indeed the central problem in the contemporary Dynamical Systems Theory. The main theme of this article is to describe recent advances in this direction for low-dimensional systems, one-dimensional endomorphisms and two-dimensional automorphisms.

**1.7. Low-Dimensional phenomenon.** One of the most stimulating events in the 20th century Dynamics was the discovery of the *Lorenz attractor* (1963). It appeared in an innocently looking system of three ordinary differential equations approximating equations of gas dynamics. Computer experiments showed that for some parameter values the trajectories of this system converge to an attractor with an intricate structure. It demonstrated that you do not need to go to high-dimensional phase spaces in order to encounter chaos.

The next step was made by Hénon (1976) who considered a very simple discrete two-dimensional model

$$(x, y) \mapsto (x^2 + c - by, x), \quad (x, y) \in \mathbb{R}^2, \quad (1.3)$$

for which computer experiments again showed that for some parameter values the orbits of this system converge to a strange attractor.

Notice that  $b$  is equal to the Jacobian  $\det Df$  of  $f$ . If we let  $b \rightarrow 0$ , the Hénon family will degenerate to the one-dimensional quadratic family  $f_c : x \mapsto x^2 + c$ . These reductions suggested that the quadratic family can also exhibit some interesting chaotic features. At the same time R. May (1976) considered the quadratic family as a model for population dynamics, and this work ignited a great interest in this family. Probably nobody at that time could foresee how deep, beautiful and important the one-dimensional theory would grow. This story will be described in in §3.

**1.8. Complex One-dimensional dynamics.** Complex dynamics is concerned with iterates of holomorphic maps on complex manifolds. In dimension one, the most interesting case is that of rational endomorphisms of the Riemann sphere. This beautiful and deep theory was founded in the beginning of the 20th century by Fatou and Julia. However, it was almost

forgotten until the early 1980s when a fresh interest was sparked by fascinating computer images and by the realization of an intimate relation of the field to Teichmüller theory and Quasiconformal Analysis, Hyperbolic Geometry, Ergodic Theory, and Real One-Dimensional Dynamics. Computer images of the *Mandelbrot set* and discovery of *Sullivan's Dictionary* between the dynamics of rational maps and Kleinian groups were particular inspiring.

By the mid 1990s, complex one-dimensional dynamics proved to be a most powerful tool for crucial problems of real dynamics, fully confirming the classical Painlevé-Hadamard Principle that “Between two truths of the real domain, the easiest and shortest path quite often passes through the complex domain”. An interplay between real and complex worlds is one of the main themes of this article.

**1.9. Renormalization.** Renormalization is a higher structure in some “Universality class”  $\mathcal{S}$  of dynamical systems that controls small scale geometries of individual systems  $f \in \mathcal{S}$  and bifurcations in families  $(f_t)$ . In the best scenario, it implies the universality of these small scale geometries.

Let us describe an idea in a very general way. Let  $f : M \rightarrow M$  be a system of our class  $\mathcal{S}$ . Take some subspace  $N \subset M$  and consider the first return map  $g$  to  $N$ . It can happen that after performing some change of variable (“rescaling”), we obtain a system of the same class. Then we say that  $f$  is *renormalizable* and that  $g$  (up to that change of variable) is its renormalization,  $g = Rf$ . Thus, we obtain a (partially defined) *renormalization operator*  $R$  in  $\mathcal{S}$ .

If in turn,  $Rf$  is renormalizable, then  $f$  is *twice renormalizable*, with a well defined  $R^2f$ , etc. In this way, we come up with a notion of  $n$  times renormalizable maps, including  $n = \infty$ . This provides us with a dynamical system  $R$  acting on  $\mathcal{S}$  whose behavior can be translated to the small scale properties of systems  $f \in \mathcal{S}$ .

A crucial desirable quality of the renormalization is usually referred to as *a priori bounds*. This is a certain geometric control of the renormalized maps  $R^n f$  which is equivalent to *pre-compactness* of the orbits  $\{R^n f\}_{n=0}^{\infty}$ . This means that the small scale geometry of  $f$  *does not degenerate*.

Even better, it may happen that the renormalization orbits  $\{R^n f\}$  *converge to a unique fixed point*  $f_*$ . Then the small scale geometry of all infinitely renormalizable maps is *universal*, controlled by the geometry of  $f_*$ .

And the best possible scenario occurs when  $f_*$  is a *hyperbolic* fixed point of  $R$  with finite dimensional unstable manifold  $W^u(f_*)$ . Then this manifold represents a *universal bifurcation scenario* in finite-parameter families  $(f_t) \subset \mathcal{S}$ .

At first glance, it is unconceivable that such a wonderful picture can ever occur in real life. However, we will see that in dimension one, both real and complex, it actually does, exceeding all expectations in its beautiful completeness. In dimension two, the situation becomes more complicated but still, some key features of the picture survive.

The Renormalization phenomenon is in the very heart of almost all other matters that will be discussed in this article.

**1.10. Note on related themes.** Much of the theory discussed below can be extended to the  $C^r$ -category with  $r$  sufficiently big (usually,  $r \geq 2$  or  $3$  is sufficient). As a straightforward example, density of hyperbolicity in the smooth one-dimensional category follows immediately from the analytic result. On the other hand, it is more difficult to generalize the Renormalization Theory and the Regular or Stochastic Dichotomy (see §3.7), though a

number of important steps have been made in this direction.

Let us also mention that in the  $C^1$ -case, the story is completely different. It is becoming quite apparent that the  $C^r$ -category with  $r$  sufficiently big is a natural extension of the  $C^\omega$ -rather than  $C^1$ -category.

In this article, as the title suggests, we will set smooth category aside, and *will always assume, unless otherwise is explicitly stated, that the maps under consideration are analytic.*

Nor will we touch upon the very important world of *Conservative Dynamics*. It exhibits completely new phenomena (*KAM Theory*) and is based upon different techniques which are farther away from the one-dimensional world (as one should transit from Jacobian 0 all the way to Jacobian 1).

In fact, a natural transition from our discussion to the conservative dynamics would go through *Circle Dynamics*, but this important theme will be set aside as well.

Yet another classical theme which is quite popular these days but is completely omitted in our discussion is *Transcendental Dynamics*, e.g., the dynamics of  $z \mapsto e^z$  in the complex plane.

## 2. Complex One-dimensional dynamics

This is one of the fields where the conjecture on density of hyperbolic maps stands the chance, and in fact, in this situation, it goes back to Fatou. A good advance in this direction was made in the early 1980s. It was also discovered that in the quadratic case, *Fatou's Conjecture* would follow from the *MLC Conjecture* on local connectivity of the Mandelbrot set. In turn, the latter turned out to be intimately related to the *Complex Renormalization Theory*. These are the main themes of this chapter.

**2.1. Julia, Fatou and Mandelbrot sets.** In this section,  $f : \mathbb{C} \rightarrow \mathbb{C}$  will stand for a polynomial of degree  $d \geq 2$ . The *basin of infinity*  $U(f)$ , the *filled Julia set*  $K(f)$  and the *Julia set*  $J(f)$  are defined as follows:

$$U(f) = \{z \in \mathbb{C} : f^n z \rightarrow \infty \text{ as } n \rightarrow +\infty\},$$

$$K(f) = \mathbb{C} \setminus U(f), \quad J(f) = \partial U(f) = \partial K(f).$$

The *Fatou set*<sup>6</sup> is defined as the complement of the Julia set:

$$F(f) = \mathbb{C} \setminus J(f) = U(f) \cup \text{int } K(f).$$

The *postcritical set*  $\mathcal{O}_f$  is the closure of union of the orbits of the critical points. Its structure largely determines the global dynamics of  $f$ . For instance, a classical theorem asserts that *the Julia set  $J(f)$  is connected if and only if none of the critical points escapes to  $\infty$ .* In the quadratic case  $f_c : z \mapsto z^2 + c$ , the set of parameters  $c$  for which  $f_c$  is connected is called the *Mandelbrot set*.<sup>7</sup>

**2.2. Periodic points and the measure of maximal entropy.** Let us consider a periodic point  $\alpha$  of period  $p$ , i.e.,  $f^p \alpha = \alpha$  (and assume  $p$  is the smallest positive moment with this

<sup>6</sup>The Fatou and Julia sets can also be defined for general rational endomorphisms  $f : \hat{C} \rightarrow \hat{C}$  of the Riemann sphere, but  $\infty$  would not play a special role in these definitions.

<sup>7</sup>Probably, the first image of this set, albeit rough, was produced by Brooks and Matelski in 1978.

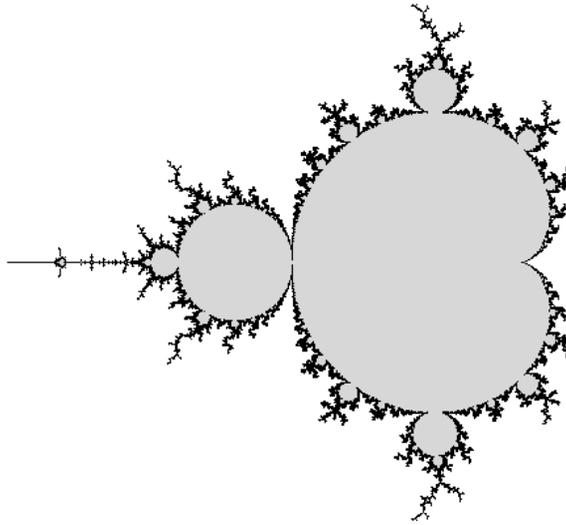


Figure 2.1. Mandelbrot set. It encodes in one picture all beauty and subtlety of the complex quadratic family.

property). Its multiplier  $\lambda$  is defined as  $(f^p)'(\alpha)$ , and  $\alpha$  is called *attracting (sink)*, *repelling* or *neutral* according to  $|\lambda| < 1$ ,  $|\lambda| > 1$ , or  $\lambda = e^{2\pi i\theta}$ , where  $\theta \in \mathbb{R}/\mathbb{Z}$  is called the *rotation number* of  $\alpha$ . In the latter case, if  $\theta$  is rational then  $\alpha$  is called *parabolic*. In the irrational case,  $\alpha$  is called *Siegel* if  $f$  is linearizable near  $\alpha$ , and is called *Cremers* otherwise.

Attracting points lie in the Fatou set, while repelling and parabolic lie in the Julia set. Moreover, by a classical Fatou-Julia Theorem, *repelling points are dense in  $J(f)$* . In fact, they are equidistributed with respect to a canonical invariant measure:

**Theorem 2.1** ([31, 72]). *Periodic points (and repelling periodic, too) are equidistributed with respect to the harmonic measure on  $J(f)$  which coincides with the unique measure of maximal entropy (whose entropy is equal to  $\log d$ ).*

This result gives a good sense of the chaotic nature of the dynamics on the Julia set.

**2.3. Dynamics on the Fatou set.** For each attracting cycle  $\alpha = \{f^n \alpha\}_{n=0}^{p-1}$ , there is an open set of points containing  $\alpha$ , whose orbits converge to  $\alpha$ . It is called the *attracting basin*,  $D(\alpha)$ . By a Fatou-Julia Theorem, each attracting basin contains at least one critical point, and hence, *the total number of attracting cycles in  $\mathbb{C}$  is bounded by  $d - 1$* .

Similarly, for each parabolic cycle  $\alpha$ , there is an open set of points containing  $\alpha$  on its boundary, whose orbits converge to  $\alpha$ . It is called the *parabolic basin*,  $D(\alpha)$ .

There is one more type of dynamics on the Fatou set: there may exist periodic components of  $\text{int } K(f)$  on which the dynamics is conformally conjugate to an irrational rotation of the disk  $\mathbb{D}$ . Such components are called *Siegel disks*.

By the classical work of Fatou, Julia and Siegel, any periodic component of the Fatou set belongs to one of the above three types (and there are only finitely many periodic components). In the early 1980s, Sullivan [104] proved a celebrated result that *any component of the Fatou set is eventually periodic*, thus completing a description of the dynamics on the

Fatou set:

**Theorem 2.2.** *The Fatou set of a polynomial  $f : \mathbb{C} \rightarrow \mathbb{C}$  (of degree  $d \geq 1$ ) is the union of finitely many attracting basins, parabolic basins, and Siegel basins.*

**Remark 2.1.** For rational endomorphisms of the Riemann sphere, one more type of periodic components can appear, *Herman rings*, on which the dynamics is conformally conjugate to an irrational rotation of an annulus. Together, Siegel disk and Herman rings are called *rotation domains*.

**2.4. Hyperbolic maps.** A polynomial map  $f : \mathbb{C} \rightarrow \mathbb{C}$  is called *hyperbolic* if it is uniformly expanding on the Julia set. It happens if and only if all critical points converge to attracting cycles. Moreover, in this case all points in the Fatou set converge to attracting cycles, while the Julia set has zero area. At the same time, on the Julia set there is a good Markov model for the dynamics. So, hyperbolic maps form a very well understood class of maps.

Note that a *quadratic polynomial is hyperbolic if and only if it has an attracting cycle*.

**2.5. Combinatorial model.** In fact, there is a general way of encoding points of any *connected* Julia set  $J(f)$ . Namely, in this case, the basin of infinity  $U(f)$  can be uniformized by the complement of the unit disk,  $\mathbb{C} \setminus \bar{\mathbb{D}}$ , conjugating  $f$  to  $z \mapsto z^2$ . The latter map has two invariant foliations, by round circles centered at 0 and by orthogonal straight rays. The corresponding foliations on  $U(f)$  are  $f$ -invariant and their leaves are called *equipotentials* and *external rays* for  $f$ . Moreover, the equipotentials are naturally labeled by their *external angles*.

It is not difficult to show that any rational ray lands at some periodic or pre-periodic points on  $J(f)$ . This generates an equivalence relation on  $\mathbb{Q}/\mathbb{Z}$ , namely  $\theta \sim \theta'$  if the corresponding rays land at the same point. Taking the quotient of  $\mathbb{T} \approx \mathbb{R}/\mathbb{Z}$  by the closure of this relation, we obtain a *combinatorial model* for  $J(f)$ . There is a natural dynamics on the model, induced by the doubling map  $\theta \mapsto 2\theta \pmod{1}$ .

Two polynomials with connected Julia set are called *combinatorially equivalent* if they have the same model.

It turns out that the actual Julia set can be equivariantly projected onto its model. Moreover, the classical *Carathéodory Theorem* implies that this projection is a homeomorphism if and only if the Julia set is *locally connected*.

Remarkably, the above equivalence relation can be explicitly described (see Douady [41] and Thurston [111]), providing us with an *explicit topological model for locally connected Julia sets!*

**2.6. Holomorphic motions.** Let  $\Lambda$  be a complex manifold, and let  $X_\lambda$ ,  $\lambda \in \Lambda$ , be a family of subsets of  $\mathbb{C}$ . We say that  $X_\lambda$  moves under a *holomorphic motion* if there is a family  $\mathcal{G}$  of disjoint graphs of holomorphic functions  $\phi : \Lambda \rightarrow \mathbb{C}$  such that  $X_\lambda = \{\phi(\lambda) : \phi \in \mathcal{G}\}$ . Then, given a base point  $\lambda_0 \in \Lambda$ , we obtain a family of injections  $h_\lambda : X_{\lambda_0} \rightarrow X_\lambda$  holomorphically depending on  $\lambda$ .

Holomorphic motions of subsets of  $\mathbb{C}$  possess remarkable properties that are usually summarized under the name of  $\lambda$ -lemma:

**$\lambda$ -lemma.** *A holomorphic motion of any subset  $X_\lambda \subset \mathbb{C}$  extends to a holomorphic motion of the whole plane  $\mathbb{C}$ . Moreover, the maps  $h_\lambda : \mathbb{C} \rightarrow \mathbb{C}$  of this motion are automatically quasiconformal homeomorphisms.*

**Remark 2.2.** Continuous extension to the closure  $\bar{X}_\lambda$  was observed in [73, 86], extension to the whole plane was constructed in [22, 103, 107], quasiconformality appeared in [86].

If we have a holomorphic family of polynomial maps  $f_\lambda : \mathbb{C} \rightarrow \mathbb{C}$  then a holomorphic motion  $\mathcal{G}$  is called *equivariant* if it is invariant under the fibered map  $\Lambda \times \mathbb{C} \rightarrow \Lambda \times \mathbb{C}$ ,  $(\lambda, z) \mapsto (\lambda, f_\lambda(z))$ . Equivalently, the maps  $h_\lambda$  conjugate  $f_{\lambda_0}|_{X_{\lambda_0}}$  to  $f_\lambda|_{X_\lambda}$ .

**2.7.  $J$ -stability.** A map  $f_{\lambda_0}$  in a holomorphic family  $(f_\lambda)_{\lambda \in \Lambda}$  is called  *$J$ -stable* if all nearby maps  $f_\lambda$  restricted to their Julia sets  $J_\lambda$  are topologically conjugate to  $f_{\lambda_0}|_{J_{\lambda_0}}$ .

**Theorem 2.3** ([73, 86]). *For any holomorphic family  $(f_\lambda)_{\lambda \in \Lambda}$  of polynomials  $f_\lambda : \mathbb{C} \rightarrow \mathbb{C}$ , the set of  $J$ -stable parameters is open and dense in  $\Lambda$ . Moreover, the Julia set moves holomorphically over the stability region.*

The stability set  $\mathcal{S}$  can be identified as the set of parameters  $\lambda$  such that there is a neighborhood  $\Lambda' \subset \Lambda$  of  $\lambda$  over which all repelling periodic points can be followed holomorphically, without bifurcations. By the  $\lambda$ -lemma, this holomorphic motion can be extended to the Julia set, implying  $J$ -stability.

The complement of  $\mathcal{S}$ , called the *bifurcation locus*  $\mathcal{B}$ , coincides with the closure of parabolic parameters.<sup>8</sup> Since parabolic points can be perturbed to attracting ones, the nowhere density of  $\mathcal{B}$  easily follows from the classical Fatou-Julia bounds on the number of attracting cycles.

In case of the quadratic family  $f_c : z \mapsto z^2 + c$ ,  $c \in \mathbb{C}$ , the bifurcation locus coincides with the boundary of the Mandelbrot set  $M$ . The stability region comprises  $\mathbb{C} \setminus M$  and components of  $\text{int } M$ . Conjecturally, all of these components are *hyperbolic*, i.e., the maps inside have an attracting cycle. This is a central open problem in the field which is often referred to as the *Fatou Conjecture*.

**2.8. MLC Conjecture .** In the early 1980s, Douady and Hubbard undertook a deep analysis of the structure of the Mandelbrot set that appeared in celebrated Orsay Notes [42]. It led to an explicit *topological model* for the Mandelbrot set *as long as the latter is locally connected* (similarly to what was described above for Julia sets). This prompted the most famous *Conjecture* in Holomorphic Dynamics abbreviated as the MLC (“Mandelbrot is Locally Connected”). It was also shown in [42] that the MLC would imply the Fatou Conjecture mentioned above.

The MLC Conjecture can be also formulated as the *Combinatorial Rigidity Conjecture* asserting that *two non-hyperbolic combinatorially equivalent quadratic polynomials  $f_c$  and  $f_{\bar{c}}$  are the same*. In turn, this can be reduced to the assertion *that combinatorially equivalent quadratic polynomials are quasiconformally conjugate*.

These rigidity properties sound in spirit very similar to the *Mostow Rigidity* in hyperbolic geometry. And indeed, there is an intimate connection between the two phenomena.

**2.9. Complex renormalization.** One of the prominent features of the Mandelbrot set readily observable on its computer images is that it contains all over the place little copies of itself that look exactly like the whole set (except that the little copy may miss the cusp at its root point – such a copy is called *satellite*; otherwise it is called *primitive*). To explain this phenomenon, Douady and Hubbard introduced a notion of *complex renormalization* acting in

<sup>8</sup>We assume for simplicity that in our family there are no persistently parabolic periodic points.

the space of quadratic-like maps.

A *quadratic-like map* is a degree two holomorphic branched covering  $f : U \rightarrow V$  between two conformal disks  $U \Subset V \subset \mathbb{C}$ . The set of non-escaping points,

$$K(f) = \{z : f^n(z) \in U, n = 0, 1, \dots\}$$

is called the *filled Julia set* of  $f$ . The *Julia set* of  $f$  is the boundary of the filled Julia set,  $J(f) = \partial K(f)$ .

Roughly speaking, a quadratic-like map is *renormalizable* if there is a topological disk  $U' \ni c_0$  and a period  $p > 1$  such that the map  $f^p : U' \rightarrow f^p(U')$  is quadratic-like with connected “little” Julia set  $J'$ . It comes together with its *combinatorics*: the way the little Julia sets  $J_k := f^k(J')$ ,  $k = 0, \dots, p - 1$ , are located in the big one. The main theorem of [43] asserts that *the parameters  $c$  for which the quadratic map  $f_c$  is renormalizable with a given combinatorics form a topological copy of the Mandelbrot set  $M$* . Moreover, this Mandelbrot copy is primitive if and only if the little Julia sets  $J_k$  are pairwise disjoint. In this case, the renormalization type is also called *primitive*.

We can now naturally define maps which are several times renormalizable, including *infinitely renormalizable maps*. For such a map, there is a sequence of periods  $p_n \rightarrow \infty$ , such that  $p_{n+1}$  is divisible by  $p_n$ , and a nested sequence of little Julia sets  $J^n$  corresponding to quadratic-like renormalizations  $f^{p_n} : U^n \rightarrow V^n$ .

Let us call the invariant set

$$A_f = \bigcap_{n=0}^{\infty} \bigcup_{k=0}^{p_n-1} f^k(J^n) \tag{2.1}$$

the *postcritical core* of  $f$ . It contains the postcritical set  $\mathcal{O}_f$ , and in the case when  $\text{diam } f^k(J^n) \rightarrow 0$  as  $n \rightarrow \infty$  (uniformly in  $k$ ),  $A_f = \mathcal{O}_f$  is a Cantor set. Moreover, in this case the dynamics on  $A_f$  is topologically conjugate to the *adding machine* (i.e., to a transitive<sup>9</sup> translation on a compact group), and in particular, it has a unique invariant measure  $\mu$ , the *canonical measure* on  $A_f$ .

The ratios  $p_{n+1}/p_n$  are called *relative renormalization periods*. If they are bounded, then  $f$  is called *infinitely renormalizable of bounded type*. If all the renormalizations have the same combinatorics then  $f$  is called of *stationary type*.

**2.10. Yoccoz puzzle.** The first breakthroughs in the MLC Conjecture came around 1990 in the work by Yoccoz (see [56]) who proved<sup>10</sup> that *for any quadratic polynomial  $f_c$  which is not infinitely renormalizable and does not have neutral cycles*

- *The Julia set  $f_c$  is locally connected* (Dynamical part);
- *the Mandelbrot set is locally connected at  $c$*  (Parameter part).

This result revealed deep connection between the MLC Conjecture and the Renormalization Theory. Not less important than the result itself is the technique introduced in this work. It provides us with a nest of dynamically defined *Markov tilings* of Julia sets and the corresponding tilings of the Mandelbrot set that capture the combinatorics and geometry of the sets in question. These tilings are called *Yoccoz puzzles*.

<sup>9</sup>We use “transitive” for “topologically transitive”, meaning the existence of a dense orbit.

<sup>10</sup>In the context of cubic polynomials with one escaping critical point, analogous results had been obtained by Branner and Hubbard [28].

Note that the renormalization idea leads, once again, to an efficient way of exploring the small scale geometry of the puzzles. A relevant renormalization scheme was designed in [76, 77, 81] under the name of *Generalized Renormalization*.

Yoccoz's geometric argument used in some serious way the *degree two assumption*. It took quite long before new analytic machinery was developed in [61] to cover the general unicritical case  $z \mapsto z^d + c$  [2, 62]. Then it was extended further to the multicritical case in [68].

**2.11. Beau bounds.** Let us say that an infinitely renormalizable quadratic-like map  $f : U \rightarrow V$  has *a priori bounds* if there exists an  $\epsilon > 0$  such that for any  $n \in \mathbb{N}$ ,

$$\text{mod}(V^n \setminus U^n) \geq \epsilon, \quad (2.2)$$

where  $U^n$  and  $V^n$  are the domains and ranges for the renormalizations  $R^n f : U_n \rightarrow V_n$ .

Given a family  $\mathcal{F}$  of polynomial-like maps as above, we say that they have *a priori beau bounds* if there exist an  $\epsilon > 0$  such that for any map  $f : U \rightarrow V$  in  $\mathcal{F}$  there is an  $N$  (depending only on a lower bound for  $\text{mod}(V \setminus U)$ ) with the property that (2.2) hold for all  $n \geq N$ .

Due to developments of the 1990s, all key problems of the Renormalization Theory were tightly linked to the existence of beau bounds.

Motivated by the work of Douady and Hubbard, this notion was introduced in the early 1990s by Dennis Sullivan who established beau bounds for any class  $\mathcal{I}^p$  of real infinitely renormalizable maps whose type is bounded by  $p$  (see [92]). For all infinitely renormalizable real maps, beau bounds were then proved in [70, 85] (compare §3.3 below).

In [77] beau bounds were also established for a certain class of complex maps, but then there had been no further progress in the problem until a breakthrough by Jeremy Kahn [60] who proved *beau bounds for complex infinitely renormalizable unicritical maps of bounded primitive type*. These two cases were then unified in a joint work by Kahn and the author [63].

We moved on to prove beau bounds in one new and quite different case when the little Mandelbrot copies encoding the relative renormalization types go to the cusp [64].

All these results heavily rely upon new analytic tools, the *Quasi-Additivity Law* and the *Covering Lemma*, developed in [61] (and already mentioned in §2.10) which allow one to compare various geometric moduli on nearly degenerate bordered Riemann surfaces.

In all of the above complex (!) cases, the *a priori* bounds imply the MLC at the corresponding parameters.

Note in conclusion that the satellite renormalization is very different from the primitive one: in particular, *a priori* bounds generally fail for infinitely satellite renormalizable maps. A different approach to the MLC in this case based upon the has been recently developed by Levin [69] and Cheraghi & Shishikura (conference announcements, 2012–13). However, so far it does not cover any maps with bounded combinatorics.

This is where the MLC Conjecture currently stands.

Looking beyond MLC, one can ask *whether the boundary of the Mandelbrot set,  $\partial M$ , has zero area*, and hence (given MLC), *whether hyperbolic parameters  $c$  form a full measure set<sup>11</sup> in  $\mathbb{C}$* ? If to believe in the Fundamental Conjecture, this should be the case. Indeed, by [7, 109], the set of at most finitely renormalizable parameters in  $\partial M$  has zero area (albeit

<sup>11</sup>An analogous statement is known to be false in the space of *rational* maps, as the corresponding set of parameters has positive Lebesgue measure [102]

Hausdorff dimension two [108]). On the other hand, it is safe to conjecture that an infinitely renormalizable map cannot have an SRB attractor. (It can have a physical attractor, though, as we will see below.)

## 2.12. Area and Hausdorff dimension.

**2.12.1. Area (Lebesgue measure).** The problem of the area of Julia sets goes back to Fatou who observed that certain hyperbolic Cantor Julia sets have zero area. In fact, it is easy to show that any hyperbolic Julia set has zero area. Going beyond hyperbolicity, it was proved by the author that  $\text{area } J(f) = 0$  for parabolic and Misiurewicz maps. These facts follow from the following general observation, which was one of the first application of the *Koebe Distortion Principle* to Dynamics:

**Lemma 2.4** ([71]). *Assume<sup>12</sup>  $\text{area } J(f) > 0$ . Then for almost all  $z \in J(f)$ , the limit set  $\omega(z)$  is contained in the postcritical set  $\mathcal{O}_f$ .*

In particular, any physical invariant measure  $\mu$  in  $J(f)$  is supported in the postcritical set  $\mathcal{O}_f$ .

Further examples of zero area Julia sets were of Yoccoz type considered in §2.10 [74, 109], of Collet-Eckmann type [51, 101], and of Siegel type with “sufficiently Diophantine” rotation numbers [99, 100].

However, by the mid 1990s a feeling started to grow that there could exist Julia sets of positive area. One kind of candidates for this role were *Fibonacci maps* (see [81]) of sufficiently high degree. To test it, a *random walk* computer experiment was run in 1993 by Scott Sutherland and the author. It indicated that this could be the case indeed, but this problem remains open until now.

At about the same time, a totally different program of constructing positive area Julia sets was designed by Douady. The idea was to construct a sequence of Siegel quadratic maps  $f_n : z \mapsto e^{2\pi i\theta_n} z + z^2$  with a definite area  $K(f_n) \geq \delta > 0$  converging to a Cremer map  $f$ . Then  $\text{area } K(f) > 0$  by semicontinuity, while for Cremer maps, we have  $K(f) = J(f)$ . In a remarkable development ten years later, this program was carried through by Buff and Cheritat [17].

Besides Cremer examples, the Douady-Buff-Cheritat strategy produces examples of two more types: *some Siegel examples and some infinitely renormalizable of highly unbounded satellite type*. Since these Julia sets are topologically quite wild (e.g., Cremer Julia sets are never locally connected), there was a perception that positive area of a Julia set is related to its topological complexity. In the next section, we will see that this is not the case: there exist topologically tame Julia sets of positive area!

In conclusion, let us mention a parallel great story, but with a different outcome. According to Sullivan’s Dictionary between rational maps and Kleinian groups, Julia sets correspond to limit sets  $\Lambda(\Gamma)$  of Kleinian groups  $\Gamma$ , i.e., finitely generated discrete groups of Möbius transformations of  $\hat{\mathbb{C}}$ . It was conjectured by Ahlfors in the 1960s that any limit set which is not the whole  $\hat{\mathbb{C}}$  has zero area. It took about 40 years to confirm this conjecture, through the work of Ahlfors himself, Thurston, Bonahon, Canary, Gabai & Calegry, and Agol. At this point it might look like the Dictionary started to break down.

<sup>12</sup>For general rational maps, one should also assume that  $J(f) \neq \hat{\mathbb{C}}$ .

**2.12.2. Hausdorff dimension (HD).** In the mid 1990s it was demonstrated by Bishop & Jones that for a Kleinian group  $\Gamma$  with nowhere dense limit set  $\Lambda$ , we have  $\text{HD}(\Lambda) < 2$  iff  $\Gamma$  is *geometrically finite*. In Sullivan's Dictionary, geometric finiteness naturally corresponds to hyperbolic and parabolic rational maps, so one could naively anticipate a similar criterion for rational maps. However, here a less coherent picture started to emerge. Many non-hyperbolic and non-parabolic rational maps mentioned above were shown to have the Julia set  $J$  with  $\text{HD}(J) < 2$  [51, 91, 101], while according to Shishikura's celebrated result,  $\text{HD}(J) = 2$  for a generic Yoccoz quadratic map [108].

Also, a geometrically very interesting class of maps, Feigenbaum ones, was singled out, that was well placed, due to McMullen [90], in Sullivan's Dictionary. Based on this analogy, it was strongly anticipated that Feigenbaum Julia set should have zero area but Hausdorff dimension two. The situation turned out to be more complicated, as we will see momentarily.

**2.13. Geometry of Feigenbaum maps.** A *Feigenbaum polynomial* is an infinitely renormalizable quadratic polynomial with bounded combinatorics and *a priori* bounds.

**2.13.1. Geometric Trichotomy.** Let us roughly classify all the plane sets into three types:

*Lean Case:*  $\text{HD}(J) < 2$ ;

*Balanced Case:*  $\text{HD}(J) = 2$  but  $\text{area } J = 0$ ;

*Black Hole Case:*  $\text{area } J > 0$ .

In [3] we demonstrated that *there exist lean Julia Feigenbaum sets with stationary combinatorics*. We also proved by an Interpolation argument that *if the Black Hole case is realizable then the Balanced Case is also realizable* (albeit, not necessarily with stationary combinatorics). Finally, recently we have constructed examples of *Feigenbaum black holes* [5].

The idea is to consider a (non-linear) *random walk* on  $\mathbb{Z}$  that tracks down transitions from one renormalization level to another. If the probability of escaping to a shallower level dominates (with a sufficiently big constant) the probability of going down to a deeper level, then typical orbits escape, and we are in the Lean case. If the probabilities are comparable then typical orbits oscillate back and forth, and we are in the Balanced case. Finally, if the probability of going down dominates the escaping probability, typical orbits go down to deep renormalization levels being swallowed by a "Black Hole".

We see that the lean and black hole examples are related to the *asymmetry* of the corresponding random walks. By contrast, Kleinian groups are balanced because they produce *reversible* dynamics (consider the associated geodesic flow or Brownian motion). So, Sullivan's Dictionary does not break but rather Kleinian groups have a symmetry that makes them special.

**2.13.2. New features.** While the Douady-Buff-Cheritat strategy is based upon a *Liouvillean mechanism* our Black Holes are of *Diophantine type*, and as such have very different virtues. Here are a few new features of our examples:

*Tameness:* They are *locally connected*, and hence admit a precise topological model.

*Parameter Visibility:* The corresponding set of parameters  $c$  has *positive Hausdorff dimension* (in fact, it is at least  $1/2$ ). [By contrast, it is known that Cremer parameters form a set of zero Hausdorff dimension, and it is unknown what is the dimension of all Buff-Cheritat parameters.]

*Primitive vs satellite:* Our maps are *primitively* renormalizable: this is kind of renormaliza-

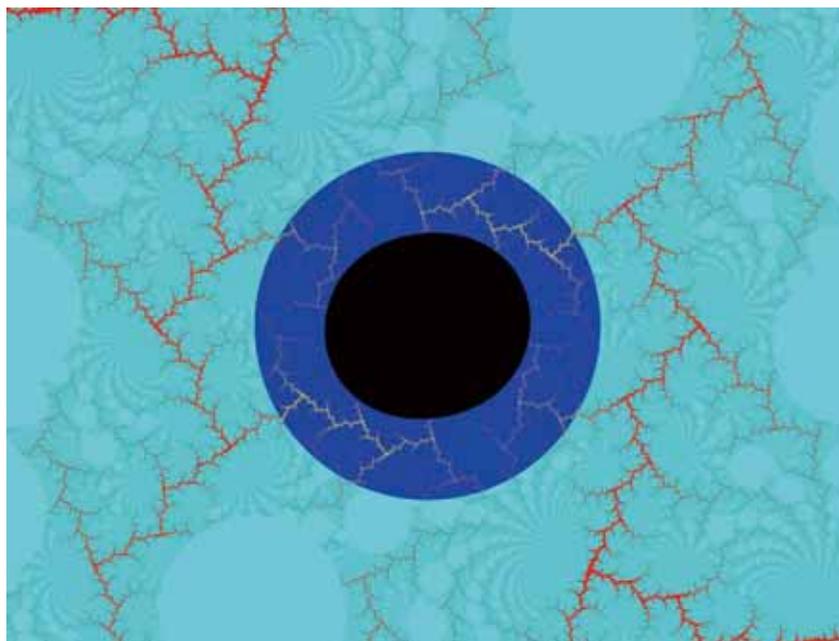


Figure 2.2. Black hole. It shows the renormalization domain for a quadratic polynomial that clearly swallows most of the nearby matter: only the light rays can escape this fate.

tion for which “wild phenomena” were less anticipated.

*Measurable Dynamics* on the Julia set has a clear nature: it is *ergodic*, and almost all orbits converge to the *physical Cantor attractor*  $A_f$  (2.1). On the other hand, the dynamics on the Julia set  $J(f)$  is transitive, so in the sense of Baire category, typical orbits are dense in  $J(f)$ . Such an attractor is called *wild* (compare with Remark 3.5 below). [Note that in the Buff-Cheritat examples of Cremer type, the “physical” measure is supported on the Cremer fixed point, which is not a physical attractor [1].]

*Spectral gap*: The *hyperbolic dimension* of  $J(f)$  (i.e., the supremum of Hausdorff dimensions over all hyperbolic subsets) is *strictly less than the full Hausdorff dimension*:  $\text{HD}_{\text{hyp}}(J) < 2 = \text{HD}(J)$ . This is the first known example exhibiting this property.

*Computer images*: It is not so hard to localize Black Hole parameters  $c$  of the type we construct, and to produce pictures of the corresponding Julia sets, see Figure 2.2. [On the other hand, so far nobody has seen a picture of a Cremer Julia set.]

Let us emphasize that both Buff-Cheritat’s and our constructions are based upon *several* renormalization theories, especially, the *Siegel Renormalization Theory* (see [91, 114]) and the *Parabolic Renormalization Theory*, with a remarkable recent advance by Inou and Shishikura [57].

Let us finish with a couple of open problems. Can maps of the following types have positive area Julia sets? 1) *real* (infinitely renormalizable) quadratic polynomials?<sup>13</sup>; 2) *non-renormalizable* unicritical polynomials of higher degree (e.g., Fibonacci maps)? Can *Cremer* quadratic polynomials have zero area Julia set?

<sup>13</sup>Computer experiments indicate that the doubling Feigenbaum maps has the Julia set of zero area

### 3. Real one-dimensional dynamics

The main themes of this chapter will be *Density of Hyperbolicity*, the *Universality* phenomenon, and the *Regular or Stochastic Dichotomy* for unimodal maps (which is the first occasion when the Fundamental Conjecture is confirmed).

**3.1. Setting.** Let  $I$  be a closed interval. An analytic map  $f : I \rightarrow I$  is called *unimodal* if it has one critical point, this point belongs to  $\text{int } I$ , and is an extremum. The main example is given by the *real quadratic family* (also called the *logistic family*)

$$f_c : [-\beta_c, \beta_c] \rightarrow [-\beta_c, \beta_c], \quad f_c : x \mapsto x^2 + c, \quad c \in [-2, 1/4], \quad (3.1)$$

where  $\beta_c > 0$  is the positive fixed point of  $f_c$ .

Much of the one-dimensional story discussed below develops as follows. First, the quadratic family is studied using its global holomorphic extension and the full power of Complex Dynamics. Then, using universal properties of this family and renormalization ideas, the theory is transferred to the real analytic case with quadratic critical point. Most of these developments happened in the last decade of the 20th century.

The quadratic case has some special geometric features (alluded already in §2.10) that made an extension to the higher degree case problematic. However, in the first decade of this century the unicritical case  $x \mapsto x^d + c$  of an arbitrary even degree was fully studied as well (with the real analytic case followed up).

Some important results were carried all the way to the general multimodal case, but this theory is not complete yet.

**3.2. Kneading theory.** The Milnor-Thurston Kneading Theory<sup>14</sup> provides us with a combinatorial classification of the one-dimensional dynamical systems  $f : I \rightarrow I$ . The combinatorial model for  $f$  is obtained by coding the orbits by means of the tiling of  $I$  into the monotonicity intervals. It turns out that the model is fully determined by the symbolic sequences of the critical values that form the *kneading invariant*  $\kappa(f)$  of  $f$ . Milnor and Thurston described all admissible kneading invariants providing us with a full combinatorial classification of the one-dimensional dynamical systems in question.

An important special case of the theory covers the dynamics of unimodal maps. A remarkable conclusion of the Kneading Theory is that the real quadratic family  $f_c : x \mapsto x^2 + c$ ,  $c \in [-2, 1/4]$ , is *full* in the space of unimodal maps, in the sense that any unimodal map is combinatorially equivalent to some  $f_c$ . This put polynomials into a special position in the dynamical world.

A beautiful problem was raised in a preliminary version of [94]: Does the kneading invariant  $\kappa(f_c)$  depend monotonically on  $c$  (with respect to a natural “twisted lexicographic order” on the space of kneading sequences)? The problem was resolved affirmatively in the final version of the paper. The proof is based on ideas of Holomorphic Dynamics, more precisely, on *Thurston’s Rigidity Theorem* (see [44]) asserting that a superattracting parameter  $c \in [-2, 1/4]$  (i.e., such that the critical point is periodic) is determined by its kneading invariant (which is finite in this case). It was the first deep application of ideas of Holomorphic Dynamics to Real Dynamics.

<sup>14</sup> A preliminary version of the Kneading Theory had been developed by Metropolis, Stein & Stein (1973).

**3.3. Density of hyperbolicity.** A more general *Rigidity Conjecture* that naturally emerged from the above theory is that every infinite kneading invariant is also realized by a *single* quadratic map  $f_c$ ,  $c \in [-2, 1/4]$ . (Note that finite kneading sequences correspond to hyperbolic maps that fill intervals of parameters.) This conjecture was quickly recognized as equivalent to the *Real Fatou Conjecture* on density of hyperbolic maps in the real quadratic family. It was proved in the mid 1990s [50, 77]. Methods of Holomorphic Dynamics play a crucial role in the proof, and until now no purely real argument has been found.

Alike the complex MLC Conjecture, the real Fatou Conjecture follows from the Rigidity assertion that *any two real quadratic maps with the same kneading invariant are quasimetrically conjugate*. This result is based upon complex *a priori* bounds for real infinitely renormalizable maps [70, 85] (which is the main point that needs reality of the maps) and certain geometric bounds for the Yoccoz puzzle. The latter bounds used in the original argument were special for the quadratic case. It took 10 years before these methods were sufficiently improved, by Kozlovski, Shen and van Strien (2007), to cover the higher degree case:

**Theorem 3.1** ([66]). *Hyperbolic maps are dense in the space of real polynomials of any degree with real critical points.*

Carrying the quadratic Fatou Conjecture further, Kozlovski proved that *Hyperbolicity is dense in the space of real analytic unimodal maps* ([65], 1998). The idea was to renormalize an analytic map with quadratic critical point to a quadratic-like map (in an ordinary or generalized sense), which allows one to apply the previous machinery. The general real analytic multimodal case was handled in 2007 [67], thus wrapping up the story on density of hyperbolicity in real dimension one.

However, as we will see momentarily, the complementary set of non-hyperbolic maps cannot be disregarded...

**3.4. Abundance of stochastic maps.** Let us say that a smooth map  $f : M \rightarrow M$  is *stochastic* if it has a unique global SRB attractor. In the case of a unimodal map  $f : I \rightarrow I$ , such an attractor has a simple description: it is the cycle of a periodic interval,

$$A = \bigcup_{n=0}^{p-1} f^n(J),$$

and  $\mu$  is an absolutely continuous invariant measure<sup>15</sup> with  $\text{supp } \mu = A$ . Moreover, this measure is *non-uniformly hyperbolic* (i.e., it has the positive Lyapunov exponent (1.2)). It provides us with an excellent picture of the dynamics from the probabilistic point of view. (See [24, 87] for a discussion of the measure-theoretic structure of unimodal maps.)

Around 1980 Jakobson proved that *the set of stochastic parameters  $c \in [-2, 1/4]$  in the real quadratic family has positive Lebesgue measure* [58].

An important sufficient condition for stochasticity was then proposed by *Collet and Eckmann*: there exist  $\lambda > 1$  and  $q > 0$  such that

$$|Df^n(f(0))| \geq q\lambda^n, \quad n \in \mathbb{N},$$

i.e.,  $f$  is exponentially expanding along the critical orbit. Benedicks and Carleson [17] went

<sup>15</sup>Note, however, that the density of this measure is usually highly singular.

on to prove that *the set of Collet-Eckmann parameters*  $c \in [-2, 1/4]$  *has positive Lebesgue measure.*

Collet-Eckmann maps exhibit non-uniform hyperbolicity that comes as close as one can get to the uniform one.

**3.5. Universality phenomenon.** A remarkable discovery was made in the mid 1970s by Feigenbaum and independently by Couillet and Tresser: some dynamical and parameter objects (within a certain class) have a universal geometry, independent of specific maps and families under consideration. As it had been already known (Myrberg, 1962) the quadratic family begins (as  $c$  decreases from  $1/4$ ) with the cascade of doubling bifurcations  $c_n$ : at this moment the attracting periodic cycle of period  $2^n$  bifurcates into the attracting cycle of period  $2^{n+1}$ . Moreover,  $c_n \rightarrow c_\infty$ , where  $f_{c_\infty}$  is an infinitely renormalizable map with all relative periods 2. Feigenbaum observed that the parameters  $c_n$  converge to  $c_\infty$  at exponential rate,

$$c_n - c_\infty \sim q\lambda^{-n}, \quad \lambda > 1,$$

with *the scaling factor*  $\lambda$  *independent of the particular family under consideration*, as long as it “looks like the quadratic family” (for instance, one can consider  $a \sin \pi x$ ). Couillet and Tresser observed that *the small-scale geometry of the postcritical core*  $A_{c_\infty}$  (2.1) *is also universal.* (Notice that for an infinitely renormalizable unimodal map, the postcritical core is a Cantor set. Moreover, it is a physical (but not SRB) attractor called the *Feigenbaum attractor*.)

To explain these surprising phenomena, the authors laid down a real one-dimensional dynamical Renormalization Theory, which was later complexified to the theory discussed in §2.9. The real definition is even simpler: A unimodal map  $f$  is called *renormalizable* if it has a periodic interval  $J \ni 0$  of period  $p > 1$ . Then the *renormalization*  $Rf$  of  $f$  is defined as the first return map  $f^p : J \rightarrow J$  considered up to rescaling. The *combinatorics* of the renormalization is the order of the intervals  $f^n(J)$ ,  $n = 0, 1, \dots, p - 1$ , on the real line.

Then we can proceed with the definition of the real renormalization operator  $R$  as in the complex setting of §2.9. The Feigenbaum-Couillet-Tresser *Renormalization Conjecture* asserted that *for any given combinatorics, the renormalization operator has a unique fixed point*  $f_*$  (i.e., it satisfies the Cvitanović-Feigenbaum functional equation  $Rf_* = f_*$ ), and  $R$  *is hyperbolic at this fixed point with unstable direction of dimension one.*

This conjecture would imply the above universalities: the universal scaling in the parameter plane would be controlled by the unstable eigenvalue of  $DR(f_*)$ , while the universal geometry of the Feigenbaum attractor  $A_f$  for all infinitely renormalizable maps would be determined by the geometry of  $A_{f_*}$ .

**Remark 3.1.** The Renormalization Conjecture admits an immediate generalization to any *real stationary* combinatorics, e.g., to the real triplings. In the complex plane, the Universal scaling law, in the case of triplings, was first observed by Golberg, Sinai and Khanin in the early 1980s [49]. The complex renormalization from §2.9 lays down a conjectural foundation for this phenomenon.

**3.6. Proof of the real Renormalization Conjecture.** In his address to the ICM in Berkeley [105], Dennis Sullivan proposed a program to approach the Feigenbaum-Couillet-Tresser Renormalization Conjecture based upon Teichmüller theory. The idea was to supply the

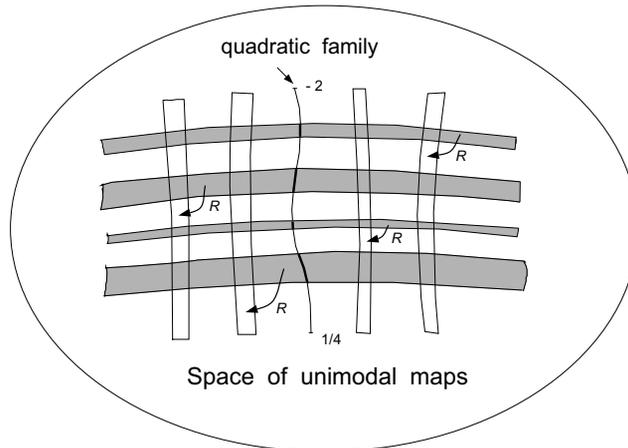


Figure 3.1. Full Renormalization Horseshoe. Various universality features of unimodal families are encoded in this image. Each horizontal strip represents a subspace of unimodal maps that are renormalizable with a certain combinatorics. Under the renormalization operator  $R$ , it is mapped to a vertical strip, being contracted horizontally and stretched vertically.

corresponding *hybrid class*<sup>16</sup> with a natural *Teichmüller metric* that would make the renormalization operator strongly contracting.

This idea was partially realized in [106]: it was proved that the renormalization operator is weakly contracting, which implied the existence of a unique renormalization fixed point  $f_*$  and convergence to it of all the orbits  $\{R^n f\}$  within the hybrid class. It did not imply exponential convergence, though. Then McMullen developed another method to give a new proof of the above results, accompanied with the exponential convergence [90].

More recently, in the work of Avila and the author [4], a new simple and natural approach to the problem was implemented. Instead of the Teichmüller metric, it makes use of a *Carathéodory metric* on the hybrid classes. It also uses some ideas from the theory of *almost periodic representations of semigroups*.

The proof of the Renormalization Conjecture was completed in [79] in the mid 1990s by demonstrating that the fixed point  $f_*$  is actually hyperbolic with one-dimensional unstable manifold. Along the lines, it was shown that the hybrid classes form a *lamination* in the space of quadratic-like maps, with complex codimension-one leaves. Moreover, the quadratic family is the global transversal to this lamination, which illuminated further the special role of this family.

Finally, in [80], the Renormalization Conjecture was extended to *all* real combinatorics simultaneously. It asserts that in the space of quadratic-like maps there is an invariant subset  $\mathcal{A}$  (the *Renormalization Horseshoe*) on which the renormalization operator  $R$  acts as the two-sided shift with infinitely many symbols (corresponding to various finite kneading sequences), and this action is uniformly hyperbolic with one-dimensional unstable direction. In [4], this result was generalized to the higher degree unicritical case, with a substantially improved argument.

**Remark 3.2.** The status of the *Complex Renormalization Conjecture* depends on the avail-

<sup>16</sup>Two quadratic-like maps are called *hybrid equivalent* if they are conjugate by a quasiconformal map  $h$  such that  $\bar{\partial}h = 0$  a.e. on the filled Julia set.

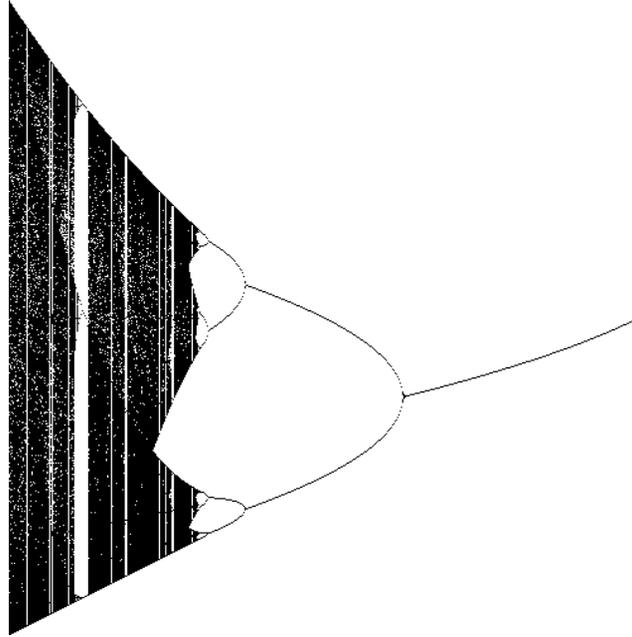


Figure 3.2. Real quadratic family. This iconic picture presents how the limit set of the orbit  $\{f_c^n(0)\}_{n=0}^{\infty}$  bifurcates as the parameter  $c$  changes from  $1/4$  on the right to  $-2$  on the left. Two types of regimes are intertwined in an intricate way. The gaps correspond to the regular regimes. The black regions correspond to the stochastic regimes (though of course there are many narrow invisible gaps therein). In the beginning (on the right) you can see the cascade of doubling bifurcations.

ability of the *a priori* beau bounds discussed in §2.11. For instance, in the complex stationary case, the Conjecture is established exactly for the *primitive* renormalization type. In particular, the Golberg-Sinai-Khanin scaling law for the *satellite* complex triplings still remains conjectural!

**3.7. Regular or stochastic dichotomy.** The following result was the first instance when the Fundamental Conjecture was confirmed:

**Theorem 3.2** ([80]). *Almost any real quadratic map  $f_c : x \mapsto x^2 + c$ ,  $c \in [-2, 1/4]$ , is either regular or stochastic.*

In the former case, it has an attracting cycle that attracts almost all orbits; in the latter, it has an absolutely continuous invariant measure that governs the behavior of almost all orbit. One can also say that *almost any quadratic map is hyperbolic, either uniformly or non-uniformly.*

There are two main parts of this picture:

- It was proved in [78] that almost any real quadratic map which is at most finitely renormalizable is a stochastic map satisfying the Martens-Nowicki criterion [89]. The proof is based upon geometric analysis of the *parapuzzle* (the Yoccoz puzzle in the parameter plane) by means of the *generalized renormalization*.

**Remark 3.3.** There exist non-renormalizable quadratic maps which are *not stochastic* (Johnson [59]). In fact, there are such maps without any physical measure or with the “physical”

measure supported on a repelling fixed point (Hofbauer and Keller [55]). As the result of [78] shows, all these examples are neglectable from the probabilistic viewpoint.

- The Full Renormalization Horseshoe from [80] easily implies that the set of infinitely renormalizable parameters has zero length.

Then Avila and Moreira went on to prove that *almost all stochastic quadratic maps are Collet-Eckmann* [8].

At this point the reader can wonder: OK, using full power of the complex machinery, one can deal with one special family, the quadratic family, but how about other families of unimodal maps? This is the moment when the quadratic family fully demonstrated its universal role: the above results can be *transferred* from the quadratic family to *any generic real analytic family of quasiquadratic*<sup>17</sup> maps. It was done by Avila, de Melo and the author in [6].

The main step is to prove that non-hyperbolic topological classes form a lamination<sup>18</sup> with codimension one analytic leaves for which the quadratic family as a global transversal. Then one can use the holonomy along this lamination to transfer the results.

**Remark 3.4.** The  $\lambda$ -lemma ensures “for free” that this holonomy is *quasisymmetric*. However, one should be careful in using it for transferring “almost everywhere statements” as it is *singular* on the set of non-renormalizable parameters (see Avila and Moreira [9]).

Moreover, “generic” assumes a very precise meaning in this context. All what is needed is that *the family in question is not contained in a single topological class*.

The Regular or Stochastic Theorem was recently generalized to the unicritical families  $x \mapsto x^d + c$  of an arbitrary even degree  $d$  (with many non-trivial issues to address) [4, 7], with the corresponding real analytic case following up [37].

**Remark 3.5.** Let us also mention the work [32] preceding [4, 7] where it had been shown that for almost all real  $c$ , the unimodal map  $x \mapsto x^d + c$  has a unique physical attractor. Along with regular and stochastic attractors, this attractor could be a Feigenbaum adding machine as well as a “wild Cantor attractor”. The existence of the latter for sufficiently high  $d$  was demonstrated in [30]. The work [4, 7] shows that the latter two situations are probabilistically neglectable.

In the *multimodal* case, it has been known since the mid 1980s [23]<sup>19</sup> that there are finitely many attractors  $A_i$  in the sense of Milnor such that for Lebesgue a.e.  $x$ , we have  $\omega(x) = A_i$  for some  $i$ . In this case, the Fundamental Conjecture asserts that for a typical  $f$ , each  $A_i$  is either an attracting cycle or a cycle of intervals supporting an ergodic absolutely continuous invariant measure. This conjecture is still open.

## 4. Real Two-Dimensional dynamics

In this section we will describe some geometric properties of the real Hénon family (1.3). This simple global analytic family remains the most popular model in 2D Dynamics.

---

<sup>17</sup>Meaning that the maps in question have quadratic critical point and are topologically conjugate to quadratic maps.

<sup>18</sup>Some parabolic topological classes are also excluded.

<sup>19</sup>This work treats maps with negative Schwarzian derivative. For the general case, see [88] and [75].

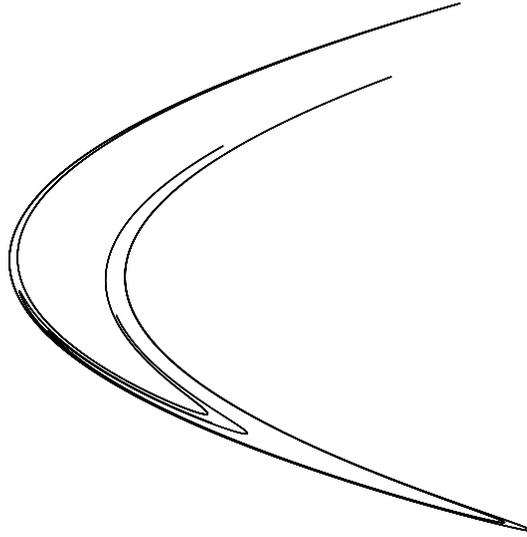


Figure 4.1. Hénon attractor: this swallow has become a symbol of chaos

**4.1. Abundance of stochastic Hénon maps.** Recall that a real Hénon map is *stochastic* if its restriction to a bounded domain in  $\mathbb{R}^2$  has a global SRB attractor. The following statement represents an accumulative breakthrough work by Benedicks, Carleson, Lai-Sang Young, and Viana from the 1990s:

**Theorem 4.1** ([18–20]). *In the real Hénon family (1.3), there is a positive measure set of stochastic parameters  $(c, b)$ .*

These parameters are obtained by perturbing a *Misiurewicz*<sup>20</sup> quadratic map  $x \mapsto x^2 + c$ . Note that the value of the the Jacobian  $b$  in this perturbation is tiny, very far from the values suggested by experiments.

The general outline of the argument follows the one-dimensional case [17]. A major challenge is related to the *thickness* of the horseshoes involved that makes a notion of “critical point” poorly defined. At the same time, the number of these “points” becomes exponentially big in deep scales. Handling of these (and many other) technical problems is an impressive *tour de force*. (See Tsujii [110] and Berger [21] for further refinements of these results.)

Currently, the author is working with Marco Martens on designing a complex Hénon puzzle techniques that could make these problems more manageable, as well as lead to a number of further consequences. We analyze the small scale geometry of the puzzle by means of a generalized renormalization combining one-dimensional machinery with the Hénon renormalization discussed below.

Let us note in conclusion that nearly homoclinic tangencies can be renormalized to produce Hénon-like families to which the Benedicks-Carleson techniques applies yielding a strange SRB attractor (Mora and Viana [95]). In fact, infinitely many co-existing attractors can be produced in this way [39]. The Fundamental Conjecture suggests, however, that the latter should be neglectable from the probabilistic viewpoint.

---

<sup>20</sup>meaning that the critical point is preperiodic

**4.2. Hénon renormalization and probabilistic rigidity.** Computer experiments of the 1980s (see [40]) indicated that the Universality phenomenon is not special to the one-dimensional situation only. In particular, if to take a small Jacobian  $b$  in the real Hénon family and to vary  $c$  from  $1/4$  down to an approximate Feigenbaum parameter  $c_*$ , the doubling bifurcations were observed with the *same scaling law* as in the one-dimensional case.

To justify this observation, Collet, Eckmann and Koch [38] set up a renormalization scheme in the space of real Hénon-like maps and proved that the one-dimensional renormalization fixed point  $f_*$  (viewed as a degenerate Hénon-like map) remains a hyperbolic fixed point for the Hénon renormalization (with the same one-dimensional unstable manifold). Gambaudo, van Strien and Tresser [48] followed up to show that the Cantor set  $A_f$  (2.1) survives Hénon perturbations  $f \in W^s(f_*)$ . It was later proved that  $A_f$  remains to be the global physical attractor for  $f$  [83].

Initially, this led to a belief that this Cantor set  $A_f$  for a Hénon map  $f \in W^s(f_*)$  has the same universal geometry as  $A_{f_*}$ . However, it was disproved by de Carvalho, Martens and the author [36]: the geometry changes as the Jacobian  $b$  varies, and the natural conjugacy between  $f|_{A_f}$  and  $f_*|_{A_{f_*}}$  can be  $1/2$ -Hölder at best. In fact, *for typical Jacobians, the Cantor set has unbounded geometry* [53]. At that point it started to look like the Dynamical Universality collapses in dimension two. However, the probabilistic point of view saved the day once again:

**Theorem 4.2** ([82]). *For real Hénon  $f \in W^s(f_*)$  sufficiently close to  $f_*$ , the Cantor attractor  $A_f$  is probabilistically rigid in the sense that the conjugacy  $A_f \rightarrow A_{f_*}$  is  $C^{1+\alpha}$ -differentiable almost everywhere with respect to the canonical measure  $\nu_f$  on  $A_f$ .*

Another striking difference with the one-dimensional situation is that the maps within the stable manifold  $W^s(f_*)$  are not globally topologically equivalent. In fact, a new rigidity phenomenon takes place:

**Theorem 4.3** ([83]). *If two real Hénon maps  $f, \tilde{f} \in W^s(f_*)$  are topologically conjugate then they have the same Jacobian.*

Note in conclusion that the renormalization scheme used in [36] is different from that in [38], and in particular, the renormalization changes of variable in [36] are *non-linear*. Hénon-like maps in [36] are normalized as follows:

$$f : (x, y) \mapsto (f(x) - \epsilon(x, y), x),$$

where  $f(x)$  is a one-dimensional unimodal map. The Jacobian  $\text{Jac } f = \partial\epsilon/\partial y$  does not have a dynamical significance and should be replaced with the *average Jacobian* over the canonical measure on  $A_f$ :

$$b = \exp \int \log \text{Jac } f \, d\nu.$$

It is interesting to explore the above rigidity phenomena within the codimension-one submanifold of Hénon-like maps  $f \in W^s(f_*)$  that have the *same* average Jacobian. For instance, under this assumption: 1) Can the Probabilistic Universality be replaced with a stronger property? 2) Can one characterize topological conjugacy classes in terms of homoclinic/heteroclinic webs?

Let us finally mention one interesting potential consequence of the Hénon renormalization theory in the spirit of the Fundamental Conjecture. It is natural to conjecture that

Theorem 4.1 is still valid for small perturbations of *Collet-Eckmann* maps. By [8], the latter form a set of full measure among non-hyperbolic maps. Then the above Renormalization Theory would imply that *Hénon parameters in  $W^s(f_*)$  are density points for the union of regular and stochastic Hénon maps.*

**4.3. Concluding remark.** The real Hénon family, albeit capturing many very interesting dynamical features, cannot play the same universal role in 2D dynamics as the quadratic family in 1D. The latter is combinatorially full, while *in 2D there is no full finite-parameter Hénon-like family*, see [54]. A “Kneading Theory” for Hénon-like dynamics that attempts to address this kind of problems, was developed by Cvitanović et al [34] and de Carvalho & Hall [35].

### 5. Complex Two-Dimensional dynamics

Global complex two-dimensional dynamics is a fairly recent field that emerged in the late 1980s. A pioneering work by Friedland & Milnor, Hubbard & Oberst-Vorth, Bedford & Smillie, Fornaess & Sibony, followed by many others, turned it into an active flourishing area, see the survey [10, 52]. Deep connections to the pluripotential theory and algebraic geometry, as well as applications to mathematical physics magnified this effect. We will not attempt to survey all these developments focusing on several recent advances that are close to the main themes of this article. Still, we have to start with describing the basic dynamical structure.

Below we will assume that  $f : \mathbb{C}^2 \rightarrow \mathbb{C}^2$  is a generalized complex Hénon map

$$f : (x, y) \mapsto (p(x) - by, x),$$

where  $p$  is a one-dimensional polynomial of degree  $d \geq 2$ , though all of our results apply to any “non-elementary” polynomial automorphism of  $\mathbb{C}^2$ . Recall that  $b = \text{Jac } f = \det Df$ . The map  $f$  is called *dissipative* if  $|\text{Jac } f| < 1$ .

**5.1. Basic dynamical structure.** Since the map  $f$  is invertible, basic dynamical objects exist in two incarnations, forward and backward. The *forward escaping basin*, the *filled Julia set*, and the *Julia set* are respectively defined as

$$U^+ = \{z \in \mathbb{C}^2 : f^n z \rightarrow \infty \text{ as } n \rightarrow +\infty\},$$

$$K^+ = \mathbb{C}^2 \setminus U^+, \quad J^+ = \partial K^+.$$

Similarly, one defines the backward objects,  $U^-$ ,  $K^-$  and  $J^-$ . Notice that in the dissipative case,  $K^-$  has always empty interior, so  $J^- = K^-$ .

Next, let us introduce *small* and *big Julia sets*:

$$J = J^+ \cap J^-, \quad \widehat{J} = J^+ \cup J^-.$$

In dimension two, any periodic point  $\alpha$  has two multipliers (i.e., eigenvalues of  $Df^p(\alpha)$ ),  $\lambda_1$  and  $\lambda_2$ . Let  $|\lambda_1| \leq |\lambda_2|$ . In the dissipative case  $|\lambda_1| < 1$ . Then the point  $\alpha$  is called *attracting*, *saddle*, or *semi-parabolic* according as  $|\lambda_2| < 1$ ,  $|\lambda_2| > 1$ , or  $\lambda_2 = e^{2\pi i p/q}$ .

Finally, let us introduce an important set  $J^* \subset J$  defined as the closure of the set of saddles. It also coincides with the closure of homoclinic intersections for any saddle, so

according to the general dynamical terminology, it can be called the *homoclinic class* of  $f$ . By analogy with the one-dimensional situation, Hubbard conjectured about 25 years ago that  $J^* = J$ , but this problem is still open.

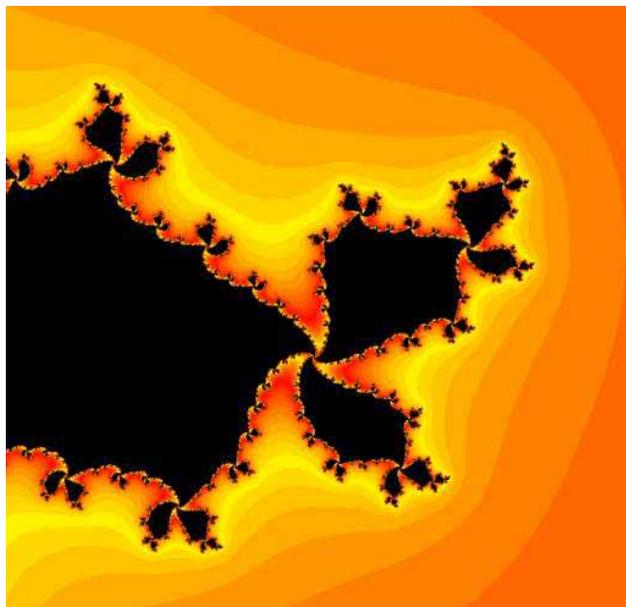


Figure 5.1. The unstable manifold of the hyperbolic fixed point for the Hénon map  $H(x, y) = (x^2 + c - ay, x)$  which is a small perturbation of the “Douady rabbit” (an unreal quadratic polynomial with periodic critical point of period 3). Generated with FractalStream.

**5.2. Periodic points and measure of maximal entropy.** As in dimension one, the set  $J^*$  supports a canonical measure  $\mu$  that gives an asymptotic distribution for periodic points. This measure appeared as the “pluri-harmonic measure” on  $J$  around 1990 (see Bedford & Smillie [11] and Fornaess & Sibony [47]), and it was then re-interpreted in terms of the *Entropy Theory*:

**Theorem 5.1** ([14, 15]). *The measure  $\mu$  is the unique measure of maximal entropy for  $f$ . Periodic points (and saddles as well) are equidistributed with respect to  $\mu$ . Moreover,  $\text{supp } \mu = J^*$ .*

**5.3. Classification of periodic Fatou components.** Let us say that a Hénon map is *moderately dissipative* if  $|\text{Jac } f| < d^{-2}$ .

**Theorem 5.2** ([84]). *For a moderately dissipative Hénon map  $f$ , any periodic component  $D$  of  $\text{int } K^+$  is either an attracting basin, or a semi-parabolic basin, or a rotation basin.*

Notice that an attracting or semi-parabolic basin  $D$  is a *Fatou-Bieberbach domain*, i.e., it is biholomorphically equivalent to  $\mathbb{C}^2$ . Moreover, if the multipliers of the corresponding attracting point are different in absolute value<sup>21</sup>, i.e.,  $|\lambda_1| < |\lambda_2|$ , then  $D$  is endowed with

<sup>21</sup>In the parabolic case, it is automatically so.

the *strong stable foliation*  $\mathcal{F}^{ss}$  with  $\mathbb{C}$ -leaves, which get contracted by the forward iterates at exponential rate  $\lambda_1^n$  (see [112, 113]).

Being diffeomorphisms, Hénon maps do not possess critical points per se. Somewhat surprising, this makes it more difficult to treat them, as we cannot spot the precise place that determines the global dynamics. So, a substantial effort has been made to define appropriate analogues of the critical points. Inside of a basin, it can be done quite naturally.

Let us say that  $c$  is a *critical point* in an attracting or parabolic basin  $D$  associated to a saddle  $\alpha$  if it is a point of tangency between the unstable manifold  $W^u(\alpha)$  and the strong stable foliation  $\mathcal{F}^{ss}$  in  $D$ . Now we can formulate an important complement to Theorem 5.2:

**Theorem 5.3** ([46]). *Let  $f$  be a moderately dissipative Hénon map. Then any attracting<sup>22</sup> basin with different multipliers and any parabolic basin contains a critical point associated to any saddle.*

The condition of being moderately dissipative in the above results is related to an application of the classical *Wiman* and *Denjoy-Carleman-Ahlfors Theorems* on the behavior of entire or subharmonic functions of slow growth.

**5.4. Branched holomorphic motions.** Let  $\Lambda$  be a complex manifold, and let  $X_\lambda, \lambda \in \Lambda$ , be a family of subsets of  $\mathbb{C}^d$ . We say that  $X_\lambda$  moves under a *branched holomorphic motion* if there is a family  $\mathcal{G}$  of holomorphic graphs  $\phi : \Lambda \rightarrow \mathbb{C}^d$  such that  $X_\lambda = \{\phi(\lambda) : \phi \in \mathcal{G}\}$ .

If we also have a holomorphic family of polynomial automorphisms  $f_\lambda : \mathbb{C}^d \rightarrow \mathbb{C}^d$  then the branched holomorphic motion  $\mathcal{G}$  is called *equivariant* if  $\mathcal{G}$  is invariant under the fibered map  $\Lambda \times \mathbb{C}^d \rightarrow \Lambda \times \mathbb{C}^d, (\lambda, z) \mapsto (\lambda, f_\lambda(z))$ .

**5.5. Weak stability.** A holomorphic family  $(f_\lambda)_{\lambda \in \Lambda}$  is called *weakly stable* if any saddle  $\alpha_\lambda$  can be followed holomorphically over  $\Lambda$ . Equivalently, no saddle can become semi-parabolic under deformation.

A map  $f_0 \equiv f_{\lambda_0}$  (and the corresponding parameter  $\lambda_0 \in \Lambda$ ) is called *weakly stable* if the family  $(f_\lambda)$  is weakly stable over some neighborhood of  $\lambda_0$ .

The set of weakly unstable parameters  $\mathcal{B} \subset \Lambda$  is called the *bifurcation locus*. It follows from the definitions that  $\mathcal{B}$  coincides with the *closure of semi-parabolic parameters*<sup>23</sup>.

**Proposition 5.4** ([46]). *Weak stability of a family  $(f_\lambda)_{\lambda \in \Lambda}$  is equivalent to the following properties:*

- *The set  $J_\lambda^*$  moves under an equivariant branched holomorphic motion over  $\Lambda$ ;*
- *The set  $J_\lambda^*$  depends continuously on  $\lambda$  in the Hausdorff topology;*
- *The big Julia set  $\widehat{J}_\lambda$  moves under an equivariant branched holomorphic motion over  $\Lambda$ .*

**5.6. Newhouse phenomenon.** By a simple analysis of the one-dimensional argument, one can arrive to the following observation:

**Proposition 5.5** ([46]). *The union of stable and Newhouse parameters is dense in  $\Lambda$ .*

<sup>22</sup>In this case, we need to require that  $|\text{Jac } f| < d^{-4}$ .

<sup>23</sup>As in dimension one, we assume for simplicity that in our family there are no persistently semi-parabolic periodic points.

Thus, int  $\mathcal{B}$  is densely filled with Newhouse parameters. On the other hand, it is unknown at this stage whether there might exist stable Newhouse parameters. (Of course, this should not be the case if to believe in the Fundamental Conjecture).

Let us also mention that it is known that there exist complex Newhouse polynomials of sufficiently high degree [33], but it is an open problem for degree two.

**5.7. Homoclinic tangencies.** The following result partly confirms a complex version of the Palis Conjecture:

**Theorem 5.6** ([46]). *In any moderately dissipative holomorphic family  $(f_\lambda)_{\lambda \in \Lambda}$ , the set of parameters with homoclinic tangencies is dense in the bifurcation locus  $\mathcal{B}$ .*

The strategy is to perturb a semi-parabolic parameter to a homoclinic tangency. It is based on a two-dimensional version of the Douady-Sentenac Tour de Valse (Parabolic Bifurcation Theory) [16, 46] and Theorem 5.3 on the existence of critical points in parabolic basins.

**5.8. Concluding remarks.** The reader can notice that unlike the discussion in the previous sections, this section does not rely on the renormalization ideas. In fact, the Renormalization Theory in complex dimension two has not been yet properly developed. A natural starting point would be to complexify the Period Doubling Theory from §4.2 and to study the associated parameter and dynamical self-similarities. There is a natural notion of complex *Hénon-like map* [45] that would replace Douady-Hubbard's quadratic-like maps. There are several major difficulties, though, e.g.:

- As we have mentioned above, the Hénon family is not full, so precise *little copies of the parameter space* would probably not appear in this family;
- *Connectivity* is not so frequently observed in the Hénon family, so perhaps, the precise definition of renormalization should not be based on this property. (See [12] on the connectivity in dimension two.)

Along with the *Hénon Puzzle* (compare [13] and §4.1), developing a *Hénon Renormalization Theory* looks to the author as key challenges of contemporary 2D Complex Dynamics. With these tools, the field may start playing a similar important role in Real Dynamics as it has happened in dimension one.

Note also that no examples of complex Hénon maps with SRB attractors are known. However, perturbing Feigenbaum Julia sets of positive area (see §2.13), one can produce a complex Hénon map with a *physical attractor* (work in progress, joint with Artur Avila). Examples of *SRB attractors* for 2D complex *endomorphisms* have been discovered by Bonifant and Milnor [26].

**Acknowledgments.** I would like to thank Artur Avila for a substantial discussion of the content of this article. I also thank Romain Dujardin, Yuri Lyubich, Marco Martens, Wellington de Melo, John Milnor, and Scott Sutherland for reading the manuscript and making useful comments. Thanks also go to Remus Radu and Raluka Tanase for helping to produce Figure 6.1.

## References

- [1] A. Avila and D. Cheraghi, *Statistical properties of quadratic polynomials with a neutral fixed point*, arXiv:1211.4505v1 [math.DS] (2012).
- [2] A. Avila, J. Kahn, M. Lyubich, and W. Shen, *Combinatorial rigidity for unicritical polynomials*, *Annals Math.* **170** (2009), 783–797.
- [3] A. Avila and M. Lyubich, *Hausdorff dimension and conformal measures of Feigenbaum Julia sets*, *J. of the AMS*, **21** (2008), 305–383.
- [4] ———, *The full renormalization horseshoe for unimodal maps of higher degree: exponential contraction along hybrid classes*, *Publ. Math. IHES* **114** (2011), 171–223.
- [5] ———, *Feigenbaum Julia sets of positive area*, Manuscript in preparation (2011).
- [6] A. Avila, M. Lyubich, and W. de Melo, *Regular or stochastic dynamics in real analytic families of unimodal maps*. *Invent. Math.*, **154** (2003), 451–550.
- [7] A. Avila, M. Lyubich, and W. Shen, *Parapuzzle of the Multibrot set and typical dynamics of unimodal maps*, *Journal of European Math Soc.*, **13** (2011), 27–56.
- [8] A. Avila and C. G. Moreira, *Statistical properties of unimodal maps: the quadratic family*, *Annals Math.*, **161** (2005), 831–881.
- [9] ———, *Statistical properties of unimodal maps: periodic orbits, physical measures and pathological laminations*, *Publications Math IHES*, **101** (2005), 1–67.
- [10] E. Bedford, *Dynamics of rational surface automorphisms*, In: “Holomorphic dynamical systems”, 57–104. *Lecture Notes in Math.*, 1998, Springer, Berlin, 2010.
- [11] E. Bedford and J. Smillie, *Polynomial diffeomorphisms of  $\mathbb{C}^2$ : currents, equilibrium measure and hyperbolicity*, *Invent. Math.*, **103** (1991), 69–99.
- [12] ———, *Polynomial diffeomorphisms of  $\mathbb{C}^d$ . VI. Connectivity of  $J$* , *Ann. of Math. (2)* **148** (1998), 695–735.
- [13] ———, *The Hénon family: the complex horseshoe locus and real parameter space*, In: *Contemporary Math.*, AMS, **396** (2006), 21–36.
- [14] E. Bedford, M. Lyubich, and J. Smillie, *Polynomial diffeomorphisms of  $\mathbb{C}^2$ . IV. The measure of maximal entropy and laminar currents*, *Invent. Math.*, **112** (1993), 77–125.
- [15] ———, *Distribution of periodic points of polynomial diffeomorphisms of  $\mathbb{C}^2$* , *Invent. Math.*, **114** (1993), 277–288.
- [16] E. Bedford, J. Smillie and T. Ueda, *Parabolic bifurcations in complex dimension two*, Preprint 2011.
- [17] M. Benedicks and L. Carleson., *On iterations of  $1 - ax^2$  on  $(-1, 1)$* , *Ann. Math.*, **122** (1985), 1–25.
- [18] ———, *On dynamics of the Hénon map*, *Ann. Math.*, **133** (1991), 73–169.
- [19] M. Benedicks and M. Viana, *Solution of the basin problem for Hénon-like attractors*, *Invent. Math.*, **143** (2001), 375–434.
- [20] M. Benedicks and L.-S. Young, *Sinai-Åñ-Bowen-Ruelle measures for certain Hénon maps*, *Invent. Math.*, **112** (1993), 541–576.
- [21] P. Berger, *Abundance of one dimensional non uniformly hyperbolic attractors for sur-*

- face endomorphisms*, arXiv:0903.1473 [math.DS] (2011).
- [22] L. Bers and H.L. Royden, *Holomorphic families of injections*, Acta Math., **157** (1986), 259–286.
- [23] A. Blokh and M. Lyubich, *Attractors of transformations of the interval*, Funct. Anal. and Appl., **21** (1987), 70–71.
- [24] ———, *Measurable dynamics of  $S$ -unimodal maps of the interval*, Ann. Sci. Éc. Norm. Sup., **24** (1991), 545–573.
- [25] C. Bonatti, L. Diaz and M. Viana, *Dynamics beyond uniform hyperbolicity*, Encyclop. Math. Sci., **102**, Springer, 2005.
- [26] A. Bonifant and J. Milnor, *Elliptic curves as attractors in  $\mathbb{P}^2$ . I Dynamics*, Experim. Math., **16** (2007), 385–420.
- [27] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math., **470** (1975), Springer-Verlag.
- [28] B. Branner and J. Hubbard, *The iteration of cubic polynomials*, Part II, Acta Math., **169** (1992), 229–325.
- [29] X. Buff and A., *Cheritat. Examples of Julia sets with positive area*, Annals Math., **176** (2012), 673–746.
- [30] H. Bruin, G. Keller, T. Nowicki, and S. van Strien, *Wild Cantor attractors exist*, Annals Math., **143** (1996), 97–130.
- [31] H. Brolin, *Invariant sets under iteration of rational functions*, Arkiv für Math., **6** (1965), 103–144.
- [32] H. Bruin, W. Shen and S. van Strien, *Existence of unique SRB-measures is typical for real unicritical polynomial families*, Ann. Sci. École Norm. Sup., **39** (2006), 381–414.
- [33] G. Buzzard, *Infinitely many periodic attractors for holomorphic maps of 2 variables*, Annals Math., **145** (1997), 389–417.
- [34] P. Cvitanović, G. Gunaratne, and I. Procaccia, *Topological and metric properties of Hénon-type strange attractors*, Phys. Rev. A, **38**(3) (1988), 1503–1520.
- [35] A. de Carvalho and T. Hall, *How to prune the horseshoe*, Nonlinearity, **15** (2002), R19–R68.
- [36] A. de Carvalho, M. Lyubich, and M. Martens, *Renormalization in the Hénon family, I. Universality but non-rigidity*, J. Stat. Phys., **5** (2005), 611–669.
- [37] T. Clark, *Real and complex dynamics of unicritical maps*, Thesis, Stony Brook 2010.
- [38] P. Collet, J.P. Eckmann, and H. Koch, *Period doubling bifurcations for families of maps on  $\mathbb{R}^n$* , J. Stat. Physics, **25** (1980), 1–15.
- [39] E. Colli, *Infinitely many coexisting strange attractors*, Ann. Inst. H. Poincaré. Anal. Non Linéaire, **15** (1988), 539–579.
- [40] P. Cvitanović (editor), *Universality in Chaos*, Institute of Physics Publishing, 1989.
- [41] A. Douady, *Description of compact sets in  $\mathbb{C}$* , In: “Topological Methods in Modern Mathematics, A Symposium in Honor of John Milnor’s 60th Birthday”, Publish or Perish, 1993.
- [42] A. Douady and J. H. Hubbard, *Étude dynamique des polynômes complexes. Parties I et II*, Publications Mathématiques d’Orsay, 84-2 & 85-4.

- [43] ———, *On the dynamics of polynomial-like maps*, Ann. Sc. Éc. Norm. Sup., **18** (1985), 287–343.
- [44] ———, *A proof of Thurston’s topological characterization of rational functions*, Acta Math., **171** (1993), 263–297.
- [45] R. Dujardin, *Hénon-like mappings in  $\mathbb{C}^2$* , Amer. J. Math., **126** (2004), 439–472.
- [46] R. Dujardin and M. Lyubich, *Stability and bifurcations of dissipative polynomial automorphisms of  $\mathbb{C}^2$* , Preprint IMS at Stony Brook, # 1 (2013).
- [47] J.-E. Fornæss and N. Sibony, *Complex Hénon mappings in  $\mathbb{C}^d$  and Fatou-Bieberbach domains*, Duke Math. J., **65** (1992), 345–380.
- [48] J.-M. Gambaudo, S. van Strien, and C. Tresser, *Hénon-like maps with strange attractors: there exist  $C^1$  Kupka-Smale diffeomorphisms on  $S^2$  with neither sinks nor sources*, Nonlinearity, **2** (1989), 287–304.
- [49] A.I. Golberg, Ya.G. Sinai, and K.M. Khanin, *Universal properties of a sequence of period-tripling bifurcations*, Russian Math. Surveys, **38** (1983), 187–188.
- [50] J. J. Graczyk and G. Świątek, *Generic hyperbolicity in the logistic family*, Ann. of Math. **146** (1997), 1–52.
- [51] J. Graczyk and S. Smirnov, *Non-uniform hyperbolicity in complex dynamics*, Invent. Math., **175** (2009), 335–415.
- [52] V. Guedj, *Propriétés ergodiques des applications rationnelles*, arXiv:math/0611302 (2008).
- [53] P. Hazard, M. Lyubich, and M. Martens, *Renormalizable Hénon-like maps and unbounded geometry*, Nonlinearity, **25** (2012), 397–420.
- [54] P. Hazard, M. Martens, and C. Tresser, *Zero entropy Hénon-like maps depend on infinitely many parameters*, Manuscript in preparation (2013).
- [55] F. Hofbauer and G. Keller, *Quadratic maps without asymptotic measures*, Comm. Math. Phys., **127** (1990), 319–337.
- [56] J.H. Hubbard, *Local connectivity of Julia sets and bifurcation loci: three theorems of J.-C. Yoccoz*, In: “Topological Methods in Modern Mathematics, A Symposium in Honor of John Milnor’s 60th Birthday”, Publish or Perish, 1993.
- [57] H. Inou and M. Shishikura, *The renormalization for parabolic fixed points and their perturbations*, Preprint mitsu@kum.kyoto-u.ac.jp.
- [58] M. Jakobson, *Absolutely continuous invariant measures for one-parameter families of one-dimensional maps*, Comm. Math. Phys., **81** (1981), 39–88.
- [59] S. Johnson, *Singular measures without restrictive intervals*, Comm. Math. Phys., **110** (1987), 185–190.
- [60] J. Kahn, *A priori bounds for some infinitely renormalizable quadratics: I. Bounded primitive combinatorics*, Preprint IMS at Stony Brook, # 5 (2006).
- [61] J. Kahn and M. Lyubich, *Quasi-Additivity Law in conformal geometry*, Annals of Math., **169** (2009), 561–593.
- [62] ———, *Local connectivity of Julia sets for unicritical polynomials*, Annals of Math., **170** (2009).
- [63] ———, *A priori bounds for some infinitely renormalizable quadratics: II. Decora-*

- tions, *Ann. Sci. École Norm. Sup.*, **41** (2008), 57–84.
- [64] ———, *A priori bounds for some infinitely renormalizable quadratics: IV. Elephants' eyes*, Manuscript in preparation (2011).
- [65] O. Kozlovski, *Axiom A maps are dense in the space of unimodal maps in the  $C^k$  topology*, *Annals Math.*, **157** (2003), 1–43.
- [66] O. Kozlovski, W. Shen, and S. van Strien, *Rigidity for real polynomials*, *Annals Math.*, **165** (2007), 749–841.
- [67] ——— Strien., *Density of hyperbolicity in dimension one*, *Annals Math.*, **166** (2007), 145–182.
- [68] O. Kozlovski and S. van Strien, *Local connectivity and quasi-conformal rigidity of non-renormalizable polynomials*, *Proc. London Math. Soc.*, **99** (2009), 275–296.
- [69] G. Levin, *Rigidity and non-local connectivity of Julia sets of some quadratic polynomials*, *Comm. Math. Phys.*, **304** (2011), 295–328.
- [70] G. Levin and S. van Strien, *Local connectivity of Julia sets of real polynomials*, *Annals Math.*, **147** (1998), 471–541.
- [71] M. Lyubich., *Typical behaviour of trajectories of a rational mapping of the sphere*, *Dokl. Akad. Nauk SSSR*, **268** (1982), 29–32.
- [72] ———, *The measure of maximal entropy of a rational endomorphism of the Riemann sphere*, *Funct. Anal. and Appl.*, **16** (1982), 78–79.
- [73] ———, *Some typical properties of the dynamics of rational mappings*, *Russian Math. Surveys* **38** (1983), no. 5, 154–155.
- [74] ———, *On the Lebesgue measure of the Julia set of a quadratic polynomial*, Preprint IMS at Stony Brook, # 1991/10.
- [75] ———, *Ergodic theory for smooth one-dimensional dynamical systems*, Preprint IMS at Stony Brook, no 11 (1991).
- [76] ———, *Combinatorics, geometry and attractors of quasi-quadratic maps*, *Annals of Math.*, **140** (1994), 347–404.
- [77] ———, *Dynamics of quadratic polynomials, I-II*, *Acta Math.*, **178** (1997), 185–297.
- [78] ———, *Dynamics of quadratic polynomials, III, Parapuzzle and SBR measure*. *Asterisque*, **261** (2000), 173–200.
- [79] ———, *Feigenbaum-Couillet-Tresser Universality and Milnor's Hairiness Conjecture*, *Ann. Math.*, **149** (1999), 319–420.
- [80] ———, *Almost every real quadratic map is either regular or stochastic*, *Annals Math.*, **156** (2002), 1–78.
- [81] M. Lyubich and J. Milnor, *The Fibonacci unimodal map*, *Journal of AMS*, **6** (1993), 425–457.
- [82] M. Lyubich and M. Martens, *Renormalization in the Hénon family, II: the heteroclinic web*, *Invent. Math.*, **186** (2011), 115–189.
- [83] ———, *Probabilistic universality in two-dimensional dynamics*, Preprint IMS at Stony Brook, no 2 (2011).
- [84] M. Lyubich and H. Peters, *Classification of invariant Fatou components for dissipative Hénon maps*, Preprint IMS at Stony Brook, # 7 (2012).

- [85] M. Lyubich and M. Yampolsky, *Complex bounds for real maps*, Ann. Inst. Fourier., **47** (1997), 1219–1255.
- [86] R. Mañé, P. Sad, and D. Sullivan, *On the dynamics of rational maps*, Ann. Sci. École Norm. Sup., **16** (1983), 193–217.
- [87] M. Martens, *Distortion results and invariant Cantor sets for unimodal maps*, Erg. Th. & Dyn. Syst., **14** (1994), 331–349.
- [88] M. Martens, W. de Melo, and S. van Strien, *Julia-Fatou-Sullivan theory for real one-dimensional dynamics*, Acta Math., **168** (1992), 273–318.
- [89] M. Martens and T. Nowicki, *Invariant measures for Lebesgue typical quadratic maps*, Asterisque, **261** (2000), 239–252.
- [90] C. McMullen, *Renormalization and three manifolds which fiber over the circle*, Princeton University Press, 1996.
- [91] McMullen, *Self-similarity of Siegel disks and Hausdorff dimension of Julia sets*, Acta Math., **180** (1998), 247–292.
- [92] W. de Melo and S. van Strien, *One-dimensional dynamics*, Springer-Verlag, Berlin, 1993.
- [93] J. Milnor, *On the concept of attractor*, Comm. Math. Physics., **99** (1985), 177–195.
- [94] J. Milnor and W. Thurston, *On iterated maps of the interval*, “Dynamical Systems”, Proc. U. Md., 1986-87, ed. J. Alexander, Lect. Notes Math., **1342** (1988), 465–563.
- [95] L. Mora and M. Viana, *Abundance of strange attractors*, Acta Math, **171** (1993), 1–71.
- [96] S. Newhouse, *Diffeomorphisms with infinitely many sinks*, Topology, **13** (1974), 9–18.
- [97] J. Palis, *A global view of dynamics and a Conjecture of the denseness of finitude of attractors*, Asterisque, **261** (2000), 335–348.
- [98] J. Palis and F. Takens, *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations*, Cambridge University Press, 1993.
- [99] C. Petersen, *Local connectivity of some Julia sets containing a circle with an irrational rotation*, Acta Math., **177** (1996), 163–224.
- [100] C. Petersen and S. Zakeri, *On the Julia set of a typical quadratic polynomial with a Siegel disk*, Ann. of Math., **159** (2004), 1–52.
- [101] F. Przytycki and S. Rohde, *Porosity of Collet-Eckmann Julia sets*, Fund. Math., **155** (1998), no. 2, 189–199.
- [102] M. Rees, *Positive measure sets of ergodic rational maps*, Ann. Sci. École Norm. Sup., **19** (1986), 383–407.
- [103] Z. Ślodkowski, *Holomorphic motions and polynomial hulls*, Proc. AMS, **111** (1991), 347–355.
- [104] D. Sullivan, *Quasiconformal homeomorphisms and dynamics I, solution of the Fatou-Julia problem on wandering domains*, Annals Math., **122** (1985), 401–418.
- [105] D. Sullivan, *Quasiconformal homeomorphisms and dynamics, topology and geometry*, Proc. ICM-86, Berkeley, v. II, 1216–1228.
- [106] ———, *Bounds, quadratic differentials and renormalization conjectures*, AMS Centennial Publ. 2: Mathematics into Twenty-first century, 1992.

- [107] D. Sullivan and W. Thurston, *Extending holomorphic motions*, Acta Math. **157** (1986), 243–257.
- [108] M. Shishikura, *The Hausdorff dimension of the boundary of the Mandelbrot set and Julia sets*, Ann. of Math. (2) **147** (1998), no. 2, 225–267.
- [109] ———, *Topological, geometric and complex analytic properties of Julia sets*, In Proceedings of the International Congress of Mathematicians (Zürich, 1994), pp. 886–895. Birkhäuser, Basel (1995).
- [110] M. Tsujii, *Physical measures for partially hyperbolic surface endomorphisms*, Acta Math. **194** (2005), 37–132.
- [111] W. Thurston, *On the geometry and dynamics of iterated rational maps*, In: Complex Dynamics: Friends and Families, pp. 3–110. Editor: D. Schleicher. A.K. Peters.
- [112] T. Ueda, *Local structure of analytic transformations of two complex variables. I.*, J. Math. Kyoto Univ., **26** (1986), 2, 233–261.
- [113] ———, *Local structure of analytic transformations of two complex variables. II.*, J. Math. Kyoto Univ., **31** (1991), 695–711.
- [114] M. Yampolsky, *Siegel Disks and Renormalization Fixed Points*, Fields Inst. Comm., **53** (2008), 377–393.

Department of Mathematics Stony Brook University Stony Brook, NY 11794-3651, USA

E-mail: mlyubich@math.sunysb.edu

# Asymptotics for critical nonlinear dispersive equations

Frank Merle

**Abstract.** We consider various examples of critical nonlinear partial differential equations which have the following common features: they are Hamiltonian, reversible in time, of dispersive nature, have a conservation law invariant by scaling, and have solutions of nonlinear type (their asymptotic behavior in time differs from the behavior of solutions of linear equations). The main questions concern the possible behaviors one can expect asymptotically in time. Are there many possibilities, or on the contrary very few universal behaviors depending on the type of initial data?

We shall see that the asymptotic behavior of solutions starting with general or constrained initial data is related to very few special solutions of the equation. This will be illustrated through different examples related to classical problems.

For a given equation, the first challenge is to construct solutions with a given behavior, including solutions with interactions between different types of waves (localized/localized or localized/non-localized) leading to nonlinear behavior or blow-up. In many of these problems, a formal guess is made based on a better understanding of the hidden laws of interaction between these waves. Then, from this guess, the questions are how to construct such examples, and why other behaviors in different regimes cannot appear. In particular, these questions are related to finding irreversibility in Hamiltonian systems, and to why oscillations of the solution can be controlled in time. We will see that universality is deeply related to stability or instability of the blow-up regime and the asymptotic behavior.

**Mathematics Subject Classification (2010).** Primary 35B40, 35B44; Secondary 35B33, 35Q53, 35Q55, 35L70.

**Keywords.** Dispersive nonlinear P.D.E., criticality, asymptotics, blow-up, global solution, soliton.

## 1. General setting and universality questions

Nonlinear partial differential equations with Hamiltonian structure appear in models of wave propagation in physics or geometry. In the 1980s, basic properties of these equations were established, notably the existence and stability of special solutions called solitons. In the 1990s, tools from harmonic analysis led to a refined understanding of properties of the corresponding linear equations and how to extend these properties to nonlinear equations. In particular, the notion of criticality appeared. There remained the problem of understanding the dynamics related to nonlinear objects (or special solutions). These questions have attracted considerable interest in the last fifteen years, and yet we are just beginning to have a rough picture of the subject. More precisely, the questions are what to expect in this context, what can be proved, and with which patterns or tools can one approach these problems.

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

In this talk, I will consider a few classical models which I hope will give a picture of the state-of-the-art:

- the  $L^2$  (mass) critical Nonlinear Schrödinger Equation (cNLS)

$$\begin{cases} i\partial_t u + \Delta u + |u|^{\frac{4}{N}} u = 0, & (t, x) \in [0, T) \times \mathbb{R}^N, \\ u|_{t=0} = u_0, \end{cases} \quad (1.1)$$

- the  $L^2$  (mass) critical Korteweg–de Vries Equation (cKdV)

$$\begin{cases} \partial_t u + \partial_x(\partial_x^2 u + u^5) = 0, & (t, x) \in [0, T) \times \mathbb{R}, \\ u|_{t=0} = u_0, \end{cases} \quad (1.2)$$

- the energy critical Nonlinear Wave Equation (ecNLW)

$$\begin{cases} \partial_t^2 u = \Delta u + |u|^{\frac{4}{N-2}} u, & (t, x) \in [0, T) \times \mathbb{R}^N, \\ (u|_{t=0}, \partial_t u|_{t=0}) = (u_0, u_1). \end{cases} \quad (1.3)$$

These are special cases of the Nonlinear Schrödinger Equation (NLS), the generalized Korteweg–de Vries Equation (gKdV), and the Nonlinear Wave Equation (NLW) with power nonlinearities:

$$i\partial_t u + \Delta u \pm |u|^{p-1} u = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^N, \quad (1.4)$$

$$\partial_t u + \partial_x(\partial_x^2 u \pm u^p) = 0, \quad (t, x) \in [0, T) \times \mathbb{R}, \quad (1.5)$$

$$\partial_t^2 u = \Delta u \pm |u|^{p-1} u, \quad (t, x) \in [0, T) \times \mathbb{R}^N, \quad (1.6)$$

where the equation with a  $-$  sign in front of the nonlinear term is called defocusing and is expected to have only linear behavior, while the equation with a  $+$  sign is called focusing and is expected to have nonlinear behavior (the nonlinear effect balances the linear effect).

**1.1. Local Cauchy theory.** Given initial data at time  $t = 0$ , the general problem is to understand the behavior of the solution  $u(t)$  for  $t > 0$  (and  $t < 0$ ; note that all equations considered are time reversible). First, in the 1980s and 90s, the existence and uniqueness of local solutions in time were clarified using Strichartz estimates on the linear equation, and fixed point arguments to treat the nonlinear term in a perturbative way. Many authors contributed to these developments; pioneering works are [6, 21, 26, 28, 39], and many others. For the above equations, we have the following results.

The mass critical Nonlinear Schrödinger Equation (cNLS) and the mass critical Korteweg–de Vries Equation (cKdV) are both locally well-posed (exhibiting existence and uniqueness of a maximal solution) on  $[0, T)$  (similarly on  $(T_-, 0]$ ) in  $L^2$  (for  $u_0 \in L^2$ ) and  $H^1$  (for  $u_0 \in H^1$ ) where  $H^1 = \{f : f \text{ and } \nabla f \in L^2\}$ . Either  $T = +\infty$  (and the solution is said to be global), or  $T < +\infty$ , and then if  $u_0 \in H^1$ ,  $\lim_{t \rightarrow T} \|\nabla u(t)\|_{L^2} = +\infty$  (the solution is said to blow up in finite time). Note that the value of  $T$  is in fact independent of the space, and in  $L^2$  the blow-up criterion is given by a Strichartz norm (See Kenig/Ponce/Vega [28] Cazenave/Weissler [6]). Moreover, one has the following conservation laws for solutions  $u(t)$  of Eq. (cNLS) and Eq. (cKdV) (mass and energy), for all  $t \in [0, T)$ :

$$\text{for } u_0 \in L^2, \quad M(u(t)) = \int |u|^2(t, x) dx = M(u_0), \quad (1.7)$$

$$\begin{aligned} \text{for } u_0 \in H^1, \quad E(u(t)) &= \frac{1}{2} \int |\nabla u|^2(t, x) dx - \frac{1}{2 + \frac{4}{N}} \int |u|^{2 + \frac{4}{N}}(t, x) dx = E(u_0), \\ \text{or } E(u(t)) &= \frac{1}{2} \int (\partial_x u)^2(t, x) dx - \frac{1}{6} \int u^6(t, x) dx = E(u_0). \end{aligned}$$

We have, in addition to the standard invariance by translation in space and time of the equation (with phase and Galilean invariance for NLS), the scaling symmetry of the solution: if  $u(t, x)$  is a solution of Eq. (cNLS), then for  $\lambda > 0$ ,

$$u_\lambda(t, x) = \lambda^{\frac{N}{2}} u(\lambda^2 t, \lambda x)$$

is also a solution, and for solutions of Eq. (cKdV),

$$u_\lambda(t, x) = \lambda^{\frac{1}{2}} u(\lambda^3 t, \lambda x)$$

is also a solution. These transformations leave invariant the  $L^2$  norm of the solution, so that both problems are called *mass critical*. Note that in the defocusing case, the  $-$  sign in the energy becomes a  $+$  sign and the energy is coercive.

The energy critical wave equation has properties similar to those of the mass critical NLS equation in  $L^2$ , in the energy space:  $(u(t), \partial_t u(t)) \in \dot{H}^1 \times L^2$  where  $\dot{H}^1 = \{f : \nabla f \in L^2\}$  (where Galilean invariance is replaced by Lorentz invariance). There is a unique solution on  $[0, T)$  in  $\dot{H}^1 \times L^2$ . Either the solution is global, or blows up in finite time ( $T < +\infty$ ) and the blow-up criterion is given by a Strichartz norm  $\|u\|_{L^{\frac{2(N+1)}{N-2}}((0, T) \times \mathbb{R}^N)} = +\infty$ . Note that,

since the space  $\dot{H}^1 \times L^2$  is critical, blow-up in finite time does not imply blow-up of the norm  $\dot{H}^1 \times L^2$ , and one expects blow-up to occur in some cases by concentration or bubbling with bounded norm. Moreover, we have the following energy conservation for solutions  $u(t)$  of Eq. (ecNLW):

$$E(u(t), \partial_t u(t)) = \frac{1}{2} \int |\nabla u|^2(t, x) - \frac{N-2}{2N} \int |u|^{\frac{2N}{N-2}}(t, x) = E(u_0, u_1), \quad (1.8)$$

and if  $u(t, x)$  is a solution, then for  $\lambda > 0$ ,  $u_\lambda(t, x) = \lambda^{\frac{N-2}{2}} u(\lambda t, \lambda x)$  is also a solution which leaves invariant the  $\dot{H}^1 \times L^2$ -norm of the solution so that the problem is *energy critical*.

Here one can see a dichotomy between low and high regularity spaces in which the equation is solved. We shall see that both notions can be useful. To some extent, low regularity says that near a solution which has a linear behavior as time goes to infinity (scattering), oscillations in the initial data do not change the asymptotics. In high regularity spaces, on one hand, we'll see that near a solution with nonlinear behavior, small oscillations can have a dramatic effect on the long term behavior (instability). On the other hand, the solution has more conservation laws, hence it is more constrained, and universality (properties somehow independent of the initial data) may be found.

**1.2. Challenges related to the problem of asymptotic behavior.** We first have the following classical examples:

- (i) *Small data result.* If the solution is small in the critical space (with a constant depending on a Strichartz inequality) then the solution is global and scatters (has linear

behavior) as time approaches infinity. Let  $S(t)v_0$  and  $S(t)(v_0, v_1)$  be the solutions of the corresponding linear equations.

There is a  $\delta > 0$ , such that in the case of Eqs. (cNLS) and (cKdV), if  $|u_0|_{L^2} < \delta$ , then the solution is global and there are  $v_{\pm} \in L^2$  such that

$$|u(t) - S(t)v_{\pm}|_{L^2} \rightarrow 0, \text{ as } t \rightarrow \pm\infty. \tag{1.9}$$

For Eq. (ecNLW), if  $|(u_0, u_1)|_{\dot{H}^1 \times L^2} < \delta$ , then the solution is global and there are  $(v_{0\pm}, v_{1\pm}) \in \dot{H}^1 \times L^2$  such that

$$|(u(t), \partial_t u(t)) - S(t)(v_{0\pm}, v_{1\pm})|_{\dot{H}^1 \times L^2} \rightarrow 0, \text{ as } t \rightarrow \pm\infty. \tag{1.10}$$

- (ii) *Nonlinear objects such as periodic-in-time or stationary solutions.* In the focusing situation, we have simple nonlinear objects which are stationary solutions up to the invariance of the equation. More precisely, we have for Eq. (cKdV) a traveling wave solution of the form  $u(t, x) = Q(x - t)$  where  $Q$  is the one-dimensional solution in  $H^1$  of

$$Q_{xx} - Q + Q^5 = 0. \tag{1.11}$$

For Eq. (cNLS), the periodic solution is of the form  $u(t, x) = e^{it}Q(x)$ , where  $Q$  is the ground state solution in  $H^1$  of

$$\Delta Q - Q + |Q|^{\frac{4}{N}}Q = 0 \tag{1.12}$$

(cf. [2]; note that excited states may also be considered).

For Eq. (ecNLW), we have stationary solutions  $u(t, x) = V(x)$  where  $V$  is the solution in  $\dot{H}^1$  of

$$\Delta V + |V|^{\frac{4}{N-2}}V = 0, \tag{1.13}$$

e.g. its ground state solution (Talenti [81]):

$$W(x) = \frac{1}{\left(1 + \frac{|x|^2}{N(N-2)}\right)^{\frac{N-2}{2}}}. \tag{1.14}$$

- (iii) *The so-called self-similar solution.* This is an expected typical example of blow-up solutions. We consider solutions of the form (up to some time-dependent translation and phase) for the mass critical KdV (Eq. (cKdV)),

$$u(t, x) = \frac{1}{(T-t)^{\frac{1}{6}}} F\left(\frac{x}{(T-t)^{\frac{1}{3}}}\right),$$

for the mass critical NLS (Eq. (cNLS)),

$$u(t, x) = \frac{1}{(T-t)^{\frac{N}{4}}} F\left(\frac{x}{(T-t)^{\frac{1}{2}}}\right),$$

and for the energy critical wave equation (Eq. (ecNLW)),

$$u(t, x) = \frac{1}{(T-t)^{\frac{N-2}{2}}} F\left(\frac{x}{T-t}\right).$$

From the criticality and the conservation laws, for Eqs. (cNLS), (cKdV), and (ecNLW), we will exclude such self-similar blow-up (in a more general form) for solutions in the critical space. *Nonexistence of self-similar blow-up* is one of the features of critical equations compared to subcritical equations. Indeed, one can see that  $F$  satisfies an equation and that the behavior of the solution as the space variable goes to infinity shows that the solution is not in the critical space. To exclude self-similar-like behavior is in general a challenge deeply related to the nature of each equation. It corresponds to replacing an ordinary differential equation analysis by a partial differential equation analysis. This is an essential step toward classification results for the asymptotic behavior of solutions of these equations. Surprisingly, all objects appearing in such classifications are simply the ones presented in (ii), above.

Other examples of solutions are generally challenging to construct and are in some respects a combination of the previous examples. They involve understanding the nature of dispersion at infinity in space and its coupling with the nonlinear dynamics. In particular, to construct a blow-up solution with a precise behavior is extremely complicated, even at a formal level. We can classify the interactions as being one of the following three types:

- (i) *Interaction of nonlinear/linear dynamics*: This is the main situation we consider. Typically, dynamics of the solution is (up to scaling) asymptotic to a simple nonlinear object as defined before. In the global case, it is an example of asymptotic stability. In the blow-up case, we obtain a bubbling solution with a universal profile. In the examples considered, dynamics near a soliton are quite degenerate (having more degenerate directions than those given by the symmetries of the equation) and small perturbations in a regular space can dramatically change the global nonlinear behavior. In particular, behavior of initial data at infinity (tails) is essential. To get to a formal understanding of these dynamics and to rigorously establish the formal picture was a challenge and required a new set of ideas. In general, understanding these interactions will lead to a classification of the possible dynamics. The two main problems of this type that we considered were *to understand blow-up behavior for the mass critical NLS, and to prove blow-up for the the mass critical KdV*. These questions were open for several decades, and their resolution has a number of consequences in different contexts. Our strategy is to see that in each situation deep knowledge of the dispersion is related to a monotonicity formula (or sets of monotonicity formulas) which encodes notions of irreversibility. Note that when we speak about monotonicity formulas in these problems of time oscillatory integrals, we mean to have a decreasing quantity up to terms of lower order which are controlled. As we shall see, this monotonicity gives stability properties of the resulting dynamics.
- (ii) *Interaction of nonlinear/nonlinear dynamics*: This is the second situation where the interaction takes place between (spatially) decoupled nonlinear objects. We will mostly comment on the gKdV Equation.
- (iii) *General decomposition and interaction*: The question is now to understand the nonlinear dynamics for general, not prepared, large data. In the theory of PDE, only two situations are known where one can answer this question: in the integrable case (Eq. (gKdV) with  $p = 2$  where the nonlinear equation can be reduced to a linear equation) and in the parabolic situation (where irreversibility is natural). Since the 1970s, there has been a widespread belief in the mathematical physics community that, for large global solutions of dispersive equations, the evolution asymptotically decouples

for large time into a sum of modulated solitons and a free radiation term (the *soliton resolution conjecture*). In our context, the challenge is to find a situation that affirms this conjecture.

We will now restrict attention to the physically relevant space dimension associated with each model. This is dimension one for the mass critical KdV (Eq. (cKdV)), dimension two for the mass critical NLS (Eq. (cNLS)), and dimension three for the energy critical NLW equation (Eq. (ecNLW)).

## 2. The nonlinear Schrödinger equation

**2.1. History of the problem.** We focus in this section on Eq. (cNLS) in the physically relevant dimension  $N = 2$ . We will work in the energy space, assuming that  $u_0 \in H^1$ . From an obstructive identity related to a pseudo-conformal invariance, it is known since the 1970s (cf. [23] [83]) that if

$$E(u_0) < 0 \text{ and } u_0 \in \Sigma = H^1 \cap \{xu_0 \in L^2\}, \tag{2.1}$$

then the solution blows up in finite time (without information on the structure of the blow-up). In addition, the conformal invariance (if  $u(t, x)$  is a solution then  $\frac{1}{|t|}u\left(\frac{1}{t}, \frac{x}{|t|}\right)e^{i\frac{|x|^2}{4t}}$  is a solution) applied to the explicit solution  $Q(x)e^{it}$  (where  $Q$  is the ground state solution of Eq. (1.12)) generates a blow-up solution:

$$S(t, x) = \frac{1}{|t|^{\frac{N}{2}}}Q\left(\frac{x}{|t|}\right)e^{i\frac{|x|^2}{4t} + \frac{i}{|t|}} \text{ with } |S|_{L^2} = |Q|_{L^2}, |\nabla S(t)|_{L^2} \sim \frac{1}{|t|}. \tag{2.2}$$

It was realized that this explicit type of blow-up was not generic, whereas the generic (stable) blow-up scenario was left open for several decades. In the 1980s, a series of formal and numerical works led to different predictions by Landman, Papanicolaou, Sulem, and Sulem [38] of the so called “loglog” law (a “loglog” correction of the self-similar rate:  $|\nabla u(t)|_{L^2} \sim \sqrt{\frac{\log|\log(T-t)|}{T-t}}$ ) governing the stable generic singularity formation and a so called “log” law (a “log” correction of the self-similar rate:  $|\nabla u(t)|_{L^2} \sim \sqrt{\frac{|\log(T-t)|}{T-t}}$ ) predicted by another school.

Variational arguments yield that blow-up is related to bubbling (and no blow-up occurs for  $|u_0|_{L^2} < |Q|_{L^2}$ , Weinstein [84]). Then in the early 1990s, the following precursor result containing a rigidity notion for Hamiltonian dynamics was proved:

**Theorem 2.1** (Dynamical characterization of  $S$  (2.2) and  $Q$  (1.12), Merle [54, 56, 57]).  
 Let  $u_0 \in H^1$ ,  $|u_0|_{L^2} = |Q|_{L^2}$  and  $u(t, x)$  be the solution of Eq. (cNLS) with initial data  $u_0$ , then:

- either  $u$  is equal to  $S$  or to  $Q(x)e^{it}$ , up to the symmetries of the equation,
- or  $u$  is global, and scatters as  $t \rightarrow \pm\infty$  if  $u_0 \in \Sigma$ .

The first step of the proof is to show, using the minimality of the mass, that the solution is either scattering or nondispersive. Then variational arguments, estimates on tails, and

conformal invariance lead to the result. Now the solutions  $S, Q$  can be seen as the only solutions which have a nonlinear dynamics at the critical mass level  $|Q|_{L^2}$ .

The next challenge was then to understand the dynamics in the context of a nonlinear/linear interaction, and a natural setting for this was small nonlinear data theory: for a small  $0 < \alpha^* \ll 1$  and  $u_0 \in H^1$  with small supercritical mass:

$$|Q|_{L^2}^2 < |u_0|_{L^2}^2 < |Q|_{L^2}^2 + \alpha^*. \tag{2.3}$$

**2.2. Loglog blow-up and classification (the Merle/Raphaël theory).** The starting point of this program was the Martel/Merle theory introducing rigidity notions for general data and dynamical application of these (see next section). We are now considering dynamics close to  $Q$  up to renormalization. Here the linearized problem around  $Q$  is very degenerate (having a higher degree of degeneracy than invariances of the equation) and the picture even at the formal level is not given by the linear theory.

The idea is the following: we consider, near  $Q$ , a family of nonlinear objects related to self-similar blow-up with a small time-dependent parameter  $b(t)$ . (On bounded sets, these self-similar solutions look like  $Q$ , but have a tail at infinity and thus fail (just barely) to belong to  $L^2$ .) Next, we consider  $Q_b$ , a regularization at infinity of this family which is minimal in some sense. At this point, the idea is to find irreversibility through the time evolution of the parameter  $b(t)$  from a monotonicity formula in  $b(t)$  (recall that this problem originally involves oscillatory integrals in time).

The algebra related to  $Q_b$  gives a formal proof of the loglog rate (related to cancellations at any polynomial order in the equation of the parameter  $b(t)$ ). These notions based on monotonicity formulas do yield a rigorous proof of stable blow-up. The remarkable fact is that this theory works in  $H^1$  and leads finally to the following theorem including a classification result:

**Theorem 2.2** ( $L^2$  critical blow-up, Merle, Raphaël [20, 59–63, 74]).

Let  $u_0 \in H^1$  with small supercritical mass (2.3) and  $u \in \mathcal{C}([0, T], H^1)$  be the corresponding solution to Eq. (cNLS). Then:

- (i) Sufficient condition for loglog law: If  $E(u_0) < 0$ , or  $E(u_0) = 0$  and  $u \neq Q$ , then  $u$  blows up in finite time with the loglog law

$$|\nabla u(t)|_{L^2} \sim \sqrt{\frac{\log |\log(T - t)|}{2\pi(T - t)}} \text{ as } t \rightarrow T. \tag{2.4}$$

- (ii) Stability of loglog blow-up: The set of  $H^1$  initial data  $u_0$  such that  $u(t)$  blows up in finite time with the loglog law (2.4) is open in  $H^1$ .

- (iii) Universality of the bubble profile and classification of the blow-up rate: If  $T < +\infty$ , then there exist parameters  $(\lambda(t), x(t), \gamma(t))$  and  $u^* \in L^2$  such that:

$$u(t, x) - \frac{1}{\lambda(t)} Q \left( \frac{x - x(t)}{\lambda(t)} \right) e^{i\gamma(t)} \rightarrow u^* \text{ in } L^2, \tag{2.5}$$

where  $Q$  is defined in Eq. (1.12),  $x(t) \rightarrow x(T)$ , and  $\lambda(t) \sim \frac{1}{|\nabla u(t)|_{L^2}}$  when  $t \rightarrow T$ , and the speed of blow-up satisfies

- either the loglog law (2.4),
- or is bounded from below by the pseudo-conformal speed:

$$|\nabla u(t)|_{L^2} \geq \frac{1}{T-t} \text{ as } t \rightarrow T. \tag{2.6}$$

We remark that self-similar blow-up ( $|\nabla u(t)|_{L^2} \sim \frac{1}{(T-t)^{\frac{1}{2}}}$ ), and the log correction of the self-similar rate are excluded. In the blow-up situation after renormalization,  $Q$  is indeed the universal object which appears (related to the *dynamical characterization* of  $Q$  in the set of initial data  $E(u_0) = 0$ ). Moreover the loglog regime exists and is a *stable regime* in the energy space. In dimension one, see also Galina Perelman [73], in which a special family of solutions with the loglog law is constructed.

**2.3. Threshold solutions.** Theorem 2.2 yields the existence of an  $H^1$ -open set of loglog blow-up solutions. In the neighborhood of  $Q$ , there are at least two other regimes: scattering solutions displaying an  $H^1$ -stable dynamics, and the solutions constructed by Bourgain/Wang [4] Krieger/Schlag [34] which scatter to  $S$ :

$$u(t, x) - S(t, x) \rightarrow u^* \text{ in } H^1 \text{ when } t \rightarrow 0 \tag{2.7}$$

and which saturate the lower bound (2.6):  $|\nabla u(t)|_{L^2} \sim \frac{1}{|t|}$  when  $t \rightarrow 0$ .

Such solutions are constructed by canceling interactions between  $S(t)$  and  $u^*$ , taking  $u^*$  to be very flat near the zero. Therefore, instability of such a solution is expected. In [64], adapting monotonicity properties to a mass constraint, one sees that the solutions (2.7) have an unstable *threshold* dynamics:

**Theorem 2.3** (Instability of  $S$ -type solutions (2.7), Merle, Raphaël, Szeftel [64]).

*The Bourgain/Wang solutions are the threshold dynamics for Eq. (cNLS) and lie on the boundary of both  $H^1$ -open sets of solutions which scatter linearly as time goes to infinity, and solutions which blow up in finite time in the loglog regime.*

**2.4. Other applications of this approach.** There are spectacular applications of this approach to the construction of blow-up solutions with a given behavior. This point of view involving monotonicity properties in problems of oscillatory integrals has been successfully used to solve some classical critical problems.

The first step is to perform a formal analysis, where one considers specific localization of a self-similar profile (or its development with respect to a small parameter) and obtains, by computing the nonlinear equation of this reduction, a nonlinear finite-dimensional reduction of the problem. In all cases, we obtain the derivation of a monotonicity formula in a specific regime which ultimately leads to a rigorous proof of the dynamics. It also shows that the infinite-dimensional part of the solution is controlled by the finite-dimensional parameters. Here, in most cases, we use high regularity theory to obtain such monotonicity formulas via specific properties of the equation considered. A byproduct of the proof is a stability property with respect to the initial data of the dynamics in this higher regularity space (where one has the monotonicity formula).

At this level, general classification is still out of reach and is a real challenge in most cases. Let me cite a few of these problems (see [75] for more details or examples). We have the focusing energy critical wave and Schrödinger equations in dimension three (or their

geometric counterparts in the critical dimension two) which can be reduced in the case of symmetry to the following equation:

- (i) *The Wave maps into the sphere  $S^2$* : (Raphaël/Rodnianski [76], see also Rodnianski/Sterbentz [77], Krieger/Schlag/Tataru [36]):

$$\partial_t^2 u = \Delta u + (|\nabla u|^2 - |\partial_t u|^2)u, \text{ in } \mathbb{R}^2, \tag{2.8}$$

- (ii) *The Schrödinger maps into the sphere  $S^2$* : (Merle/Raphaël/Rodnianski [65]):

$$\partial_t u \wedge u = \Delta u + |\nabla u|^2 u, \text{ in } \mathbb{R}^2. \tag{2.9}$$

### 3. Generalized Korteweg–de Vries equation

Equation (cKdV) admits the same conservation laws and scaling invariance as the cNLS equation and is *mass critical*. The problem of blow-up for Eq. (cKdV) was considered as a classical and natural question, since it has the same features as the mass critical NLS but no conformal invariance (or associated virial identity which leads to a simple obstruction argument to global existence). For small  $0 < \alpha^* \ll 1$ , we consider data such that

$$|u_0|_{L^2}^2 < |Q|_{L^2}^2 + \alpha^*. \tag{3.1}$$

**3.1. Subcritical and critical Martel/Merle theory for the generalized Korteweg–de Vries equation.** This problem was thoroughly studied by Martel and Merle in the early 2000s. Following the dynamical characterization of  $S$  for the mass critical NLS, where the nondispersive character of solutions follows from the mass constraint of the initial data, and the work of Gnanogbo, Merle [22], where for *general data* a minimality of an asymptotic dynamical property shows the nondispersive character of solutions, the set of results presented in this subsection is the next major breakthrough in the application of the notion of nondispersive solutions.

The idea is to find a contradiction from energy constraints ( $E(u_0) < 0$ ) and the exact asymptotic behavior of the solution in the critical situation. For this purpose, a method was introduced to produce irreversibility and rigidity in the problem. This method has as a byproduct the spectacular application in the subcritical case where solitons are stable (up to symmetry). Let us start with the simpler configuration.

- (i) *The Subcritical case ( $1 < p < 5$ )*:

In this subsection, we consider Eq. (1.5) for  $1 < p < 5$ . Solutions are global in  $H^1$  and the solution  $u(t, x) = Q(x - t)$  where  $Q$  is a solution of Eq. (1.12) and is stable up to translation in time. Note that by scaling, for  $c > 0$ ,

$$u(t, x) = Q_c(x - ct) \tag{3.2}$$

is also a solution, where  $Q_c(x) = c^{\frac{1}{p-1}} Q(c^{\frac{1}{2}} x)$ . The main question was the asymptotic stability of the soliton  $Q$ : for initial data  $u_0$  initially close to  $Q$  in the energy space, does the solution centered at a suitably chosen  $x(t)$  converge to  $Q_c$  locally in space, as time goes to infinity? The main approach is to introduce rigidity, breaking

the reversibility of the equation. For this purpose, we consider a new entire solution  $v(t)$  with initial data asymptotic to  $u(t_n, x_n + \cdot)$  locally in space for some  $x_n$ , where  $t_n$  goes to infinity:

$$u(t_n + t, x_n + x) \rightarrow v(t, x) \text{ locally in } L^2. \tag{3.3}$$

Then from a family of monotonicity formulas of the mass on half-lines, we are able to break the reversible character of the solution  $v(t)$  and to prove elliptic exponential estimates in  $x$ , uniform in time, on  $v(t, x + y(t))$  for some  $y(t)$ . Thus  $v(t)$  is a nondispersive solution of the equation and we are able to conclude using dispersive properties that  $v(t, x)$  is exactly  $Q(x - t)$  up to symmetry of the equation.

**Theorem 3.1** (Asymptotic stability of  $Q$ , Martel, Merle [42]).

Assume that  $1 < p < 5$ . If  $\|u_0 - Q\|_{H^1} < \delta \ll 1$ , and  $u(t, x)$  is the solution of Eq. (1.5) with initial data  $u_0$ , then there exists  $c^+$  close to one, such that

$$u(t) - Q_{c^+}(\cdot - x(t)) \rightarrow 0 \text{ in } H^1(x > \frac{10}{t}) \text{ as } t \rightarrow +\infty,$$

where  $Q_{c^+}$  is defined in Eq. (3.2).

(ii) *The Critical case ( $p = 5$ ):*

The situation in the critical case is much more delicate than in the subcritical case because of the possible oscillation in time of the scaling of the soliton. Nevertheless, through a use of irreversibility we are able to prove in the energy space the following:

**Theorem 3.2** ( $L^2$  critical blow-up for (KdV), Martel, Merle [42–44, 58]).

Let  $u_0 \in H^1$  satisfying (3.1) and  $u \in C([0, T], H^1)$  be the corresponding solution to the cKdV equation. Then:

(i) *Negative energy gives blow-up: If the initial data is such that  $E(u_0) < 0$ , then the solution blows up with  $T$  finite or infinite ( $\|\nabla u(t)\|_{L^2} \rightarrow \infty$  as  $t \rightarrow T$ ).*

(ii) *No self-similar blow-up: There are no solutions such that  $T < +\infty$  and*

$$\|\nabla u(t)\|_{L^2} \sim \frac{1}{(T - t)^{\frac{1}{3}}} \text{ when } t \rightarrow T. \tag{3.4}$$

(iii) *Universality of the bubble of concentration: There exist  $(\lambda(t), x(t))$  such that: for  $A > 0$ ,*

$$u(t, x) - \frac{1}{\lambda(t)^{\frac{1}{2}}} Q\left(\frac{x - x(t)}{\lambda(t)}\right) \rightarrow 0 \text{ in } L^2 \text{ for } \{|x - x(t)| < A\lambda(t)\}, \tag{3.5}$$

where  $\lambda(t) \sim \frac{1}{\|\nabla u(t)\|_{L^2}}$  when  $t \rightarrow T$ .

We remark that blow-up is in fact a consequence of asymptotic stability and energy constraints. Let  $E(u_0) < 0$ . The proof of blow-up goes along the following lines (arguing by contradiction): If the solution does not blow up, we are able to prove that  $u(t_n)$  satisfies (3.5) with a sequence  $\lambda(t_n) > c > 0$ . Using  $E(Q_c) = 0$  and the coercivity of the energy for small mass, we obtain that the energy computed on this time sequence  $E(u(t_n))$  is positive, which contradicts the conservation of the energy.

**3.2. Critical Martel/Merle/Raphaël theory.** Another piece of Martel/Merle theory is this: Space decay of the initial data with negative energy leads to blow-up in finite time. Moreover, an estimate on the blow-up rate was obtained (see [44]). But clearly, compared to the mass critical NLS, one piece is missing in the full description of the blow-up (optimal lower bounds on the blow-up rate).

Recently, we came back to this problem and achieved a much more ambitious goal: we were able to completely understand all solutions and their asymptotics for initial data near the ground state with decay (including blow-up rate/stability/instability/universality questions). This was set forth in the series of papers [48–50] by Martel/Merle/Raphaël. Finally, we end up with a *complete nonlinear finite dimensional description of the dynamical picture* (despite the high degeneracy of the equation near the ground state). This is the only such situation known in the literature. The expectation is that the picture obtained is canonical and should be extended to different contexts.

More precisely, consider the set of initial data for  $\alpha_0$  small,

$$\mathcal{A} = \left\{ u_0 = Q + \epsilon_0 \text{ with } |\epsilon_0|_{H^1} < \alpha_0 \text{ and } \int_{x>0} x^{10} \epsilon_0^2 < 1 \right\}, \tag{3.6}$$

and consider the  $L^2$  neighborhood around the family of solitary waves

$$\mathcal{T}_{\alpha^*} = \left\{ u \in H^1 \text{ with } \inf_{c_0>0, x_0 \in \mathbb{R}} |u - Q_{c_0}(\cdot - x_0)|_{L^2} < \alpha^* \right\} \tag{3.7}$$

such that for  $\alpha_0 < \alpha^*$ ,  $\mathcal{A} \subset \mathcal{T}_{\alpha^*}$ . One first has the rigidity of the dynamics for data in  $\mathcal{A}$ :

**Theorem 3.3** (Rigidity of the flow in  $\mathcal{A}$  (3.6), Martel, Merle, Raphaël [48]).

Let  $0 < \alpha_0 \ll \alpha^* \ll 1$  and  $u_0 \in \mathcal{A}$ . Let  $u \in \mathcal{C}([0, T], H^1)$  be the corresponding solution to Eq. (cKdV). Then one of the following three scenarios occurs:

(Blow-up): The solution blows up in finite time  $T > 0$  with the universal regime

$$|u(t)|_{H^1} \sim \frac{\ell(u_0)}{T - t} \text{ as } t \rightarrow T, \text{ with } \ell(u_0) > 0, \tag{3.8}$$

(Soliton): The solution is global ( $T = +\infty$ ) and converges asymptotically to a solitary wave  $Q_{c(u_0)}$ .

(Exit): The solution leaves the tube  $\mathcal{T}_{\alpha^*}$  (3.7) at some time  $0 < t^*(u_0) < +\infty$ .

Moreover, the scenarios (Blow-up) and (Exit) are stable under small perturbations of the initial data in  $\mathcal{A}$ .

This is a complete classification of solutions with data in  $\mathcal{A}$  which remain close in the  $L^2$  sense to the manifold of solitary waves. Again, a monotonicity formula (not in the energy space but in a norm related to  $\mathcal{A}$ ) is a crucial step in the proof of this result. As for the cNLS, we have the following dynamical characterization of  $Q$  (1.12): if  $E(u_0) \leq 0$ ,  $u_0 \in \mathcal{A}$  and  $u \neq Q$ , then  $u$  blows up in finite time on both sides in time with the blow-up law (3.8).

It remains to understand the long-time dynamics in the (Exit) regime. The first step is the *existence and uniqueness of a minimal blow-up element* which is the generalization of the  $S(t)$  dynamics for the cNLS. This result is a surprise since it was thought to be specific to the mass critical NLS and linked to the conformal invariance. A key to this existence result is the above classification result on localized initial data (even if this special solution has a spatially slow decay at infinity), and for the uniqueness a set of monotonicity properties.

**Theorem 3.4** (Existence and uniqueness of the minimal mass blow-up element, Martel, Merle, Raphaël [49]).

*There exists a unique solution (up to symmetries of the equation)  $\tilde{S}(t)$  in  $H^1$  of Eq. (cKdV) with minimal mass  $|\tilde{S}(t)|_{L^2} = |Q|_{L^2}$  which blows up at  $T = 0$ .*

*Moreover,  $\tilde{S}(t)$  is globally defined for positive time.*

We next proved the relevance of this unstable solution  $\tilde{S}(t)$  and the classification at minimal mass through a result which links  $\tilde{S}(t)$  to a stable scenario (see also special examples of this fact in [56] and [64] for the mass critical NLS, and for the critical wave equation in Krieger/Nakanishi/Schlag [33, 72], where they obtained a related classification of the flow near the solitary wave involving a description of the scattering zone and its boundary through a non-return lemma). The solution  $\tilde{S}$  is the *universal attractor* of all solutions in the (Exit) regime:

**Proposition 3.5** (Description of the (Exit) scenario, Martel, Merle, Raphaël [49]).

*Let  $u(t)$  be a solution in the (Exit) scenario of Theorem 3.3 and let  $t^*$  be the corresponding exit time.*

(i) *Then there exist  $\tau^* = \tau^*(\alpha^*)$  and  $(\lambda^*, x^*)$  such that*

$$|(\lambda^*)^{\frac{1}{2}}u(t^*, \lambda^*x + x^*) - \tilde{S}(\tau^*, x)|_{L^2} < \delta(\alpha_0) \rightarrow 0 \text{ as } \alpha_0 \rightarrow 0.$$

(ii) *Assume that the solution  $\tilde{S}(t)$  scatters as  $t \rightarrow +\infty$ , then any solution in the (Exit) scenario is global for positive time and scatters as  $t \rightarrow +\infty$ .*

Note that it is natural to expect  $\tilde{S}(t)$  to scatter as  $t \rightarrow +\infty$  from the situation for cNLS and ecNLW (see below) where it is proved.

It is important to notice that the above results rely on the *explicit* computation of some parametrization of the solution for initial data in  $\mathcal{A}$ , and not on algebraic virial type identities. One may justify the following procedure: introduce the nonlinear decomposition of the flow

$$u(t, x) = \frac{1}{\lambda(t)^{\frac{1}{2}}}(Q + \epsilon) \left( t, \frac{x - x(t)}{\lambda(t)} \right)$$

where  $\epsilon$  is small and show that to leading order,  $\lambda(t)$  obeys the dynamical system

$$\lambda_{tt}(t) = 0, \quad \lambda(0) = 1. \tag{3.9}$$

The three regimes (Exit), (Blow-up), and (Soliton) now correspond at the formal level respectively to  $\lambda_t(0) > 0$ ,  $\lambda_t(0) < 0$ , and  $\lambda_t(0) = 0$ . The main and deep part is a monotonicity formula in the original variable.

We now consider initial data with slowly decaying tails interacting with the solitary wave which lead to new exotic singular regimes:

**Proposition 3.6** (Exotic blow-up regimes for Eq. (cKdV), Martel, Merle, Raphaël [50]).

*There are solutions  $u \in H^1$  of Eq. (cKdV), with initial data arbitrarily close in  $H^1$  to  $Q$ ,*

(i) *which blow up at  $t = 0$  with speed  $|\partial_x u(t)|_{L^2} \sim t^{-\nu}$  as  $t \rightarrow 0^+$ , for  $\nu > \frac{11}{13}$ .*

(ii) *which blow up at  $+\infty$  with  $|\partial_x u(t)|_{L^2} \sim t^\nu$  as  $t \rightarrow +\infty$ , for  $\nu > 0$ .*

This shows that universality is lost without decay of the initial data ( $u_0 \notin \mathcal{A}$ ). In particular, the  $H^1$  Martel/Merle theory is still relevant and optimal for solutions only in the energy space without strong decay.

**3.3. Subcritical large data examples, D-solitons.** In this subsection, we are interested in the problem for large data and in a subcritical situation. We consider the quartic KdV equation (Eq. (1.5) with  $p = 4$ ), where we have no blow-up in the energy space and a nonintegrable situation.

The question is how to construct solutions with interaction between two or  $D$  nonlinear objects, which is a natural step toward the soliton resolution conjecture. Such solutions exist in the integrable case by algebraic formulas. Is this a general fact? We are able to obtain the following (see also Merle [55])

**Theorem 3.7** (Multi-solitons at  $t = +\infty$ , Martel, Merle and Tsai [51], Martel [41]).

Consider Eq. (1.5) with exponent  $p = 4$ ,  $0 < c_1 < \dots < c_D$  and  $x_1, \dots, x_D \in \mathbb{R}$ .

- (i) Existence, uniqueness of pure multi-solitons at  $t = +\infty$ : *There exists a unique  $H^1$  solution  $U_{x_1, c_1, \dots, x_D, c_D}(t)$  (called a  $D$ -soliton) such that*

$$\lim_{t \rightarrow +\infty} |U_{x_1, c_1, \dots, x_D, c_D}(t) - [Q_{c_1}(\cdot - c_1 t - x_1) + \dots + Q_{c_D}(\cdot - c_D t - x_D)]|_{H^1} = 0,$$

where  $Q_c$  is defined in Eq. (3.2).

- (ii) Stability, asymptotic stability of multi-solitons in  $H^1$ : *For  $\alpha_0 > 0$ , there is a  $\delta_0 > 0$  such that : If  $u(t, x)$  is a solution with  $|u(0) - U_{x_1, c_1, \dots, x_D, c_D}(0)|_{H^1} \leq \delta_0$ , then, for some  $x_1(t), \dots, x_D(t)$ , we have*

$$\sup_{t \geq 0} |u(t) - U_{x_1(t), c_1, \dots, x_D(t), c_D}(t)|_{H^1} \leq \alpha_0, \tag{3.10}$$

and there exists  $c_i^+$  close to  $c_i$  such that

$$u(t) - U_{x_1(t), c_1^+, \dots, x_D(t), c_D^+}(t) \rightarrow 0 \text{ in } H^1(x > \frac{c_1}{10}t) \text{ as } t \rightarrow +\infty.$$

The remarkable result above is (ii). Indeed, in the subcritical case, at the formal level near a soliton, we have one direction to control in addition to the distance to the soliton. Thus, we need two conservation laws. In the case of a  $D$ -soliton, which is roughly a superposition of  $D$  solitons, we have  $D$  directions of potential instability. So at the formal level, we would need  $D + 1$  conservation laws! To overcome this obstruction, we use the two conservation laws and  $D - 1$  monotonicity formulas. One can see that if the formal trajectories of the solitons do not collide, the analysis can make the formal picture rigorous. Once we have constructed the  $D$ -soliton for  $t > 0$ , and since the solution is defined for all time, one can ask about the behavior as  $t \rightarrow -\infty$ . In the integrable case, we again recover a  $D$ -soliton. We prove that this fact is specific to the integrable case: soliton collision produces dispersion of energy in the nonintegrable case.

**Theorem 3.8** (Martel, Merle [45–47]).

Consider the 2-soliton for  $t > 0$   $U_{x_1, c_1, x_2, c_2}(t)$  constructed in Theorem 3.7 above. For sufficiently small  $\delta > 0$ , if

$$\frac{c_1}{c_2} < \delta \text{ or } \frac{3}{4} < \frac{c_1}{c_2} < 1, \tag{3.11}$$

then  $U_{x_1, c_1, x_2, c_2}(t)$  is not a 2-soliton as  $t \rightarrow -\infty$ .

In addition, we have for  $c_1 < \delta c_2$  or  $1 - \delta < \frac{c_1}{c_2} < 1$  a complete description of the solution for all  $(t, x)$  including the interaction region. Here, the method and the challenge is to find a way to compute the solution. In one case, a wavelet type computation is carried out. In the other case, a delicate tail interaction is performed. Then, a monotonicity formula again justifies the formal asymptotics. In this case, we can prove that the solution asymptotically decomposes as  $t \rightarrow -\infty$  into a 2-soliton and a dispersive tail (with linear behavior) of a precise size.

The proof in the case  $\frac{3}{4} < \frac{c_1}{c_2} < 1 - \delta$  is of a different nature. We argue by contradiction and understand the solution independently of its behavior in  $|x| + |t| < A$  for  $A$  large and we find a contradiction outside this region. It is a powerful method which gives explicit constraints on the ratio of the speeds, but does not describe the collision.

#### 4. Large data results: The case of the energy critical wave equation

To develop a theory for non-prepared data (not small or close to a given profile), we are going to consider in this section the focusing energy critical Nonlinear Wave Equation (ecNLW) in dimension  $N = 3$ . All the following theory is a *critical* theory invariant by scaling. An important related problem is the energy critical Nonlinear Schrödinger Equation (ecNLS):

$$i\partial_t u + \Delta u \pm |u|^4 u = 0, x \in \mathbb{R}^3. \tag{4.1}$$

Historically, defocusing nonlinear wave equations were first considered in the 1980s and 90s by a number of authors including Grillakis [24], Shatah, and Struwe [78], who prove that for any initial data, solutions exist globally and scatter. Similar results were proved in breakthrough papers for the defocusing energy critical Nonlinear Schrödinger Equation in the late 1990s and early 2000s by Bourgain in [3] and Colliander/Keel/Staffilani/Takaoka/Tao in [7]. The methodology is to prove, through various scenarios related to time oscillations, that if blow-up occurs, some of the energy must concentrate, which contradicts a monotonicity formula (Morawetz, 1961).

In the focusing case, the situation is different: nonlinear objects such as  $W$  given by Eq. (1.14) exist and by the classical obstruction identity, blow-up can occur (Levine, 1974). Let us recall for the subconformal nonlinear wave equation in dimension  $N$ , universality was proved for blow-up solutions (without size conditions) following the work of Giga/Kohn in the parabolic context. Indeed, a monotonicity formula in the self-similar variable leads to a proof that all blow-up is self-similar:

**Theorem 4.1** (Self-similar blow-up in the subconformal situation for solutions of Eq. (1.6), Merle, Zaag [68]).

Let  $1 < p < 1 + \frac{4}{N-1}$ . If  $u$  is a solution of Eq. (1.6) which blows up at time  $T$ , then, as  $t \rightarrow T$ :

$$\sup_x \left( |\partial_t u(t)|_{L^2(B(x, T-t))} + |\nabla u(t)|_{L^2(B(x, T-t))} \right) \sim (T - t)^{-\left(\frac{2}{p-1} + 1 - \frac{N}{2}\right)}. \tag{4.2}$$

In addition, universality of the profile is proved in dimension one, in a series of papers (Merle/Zaag [69–71]).

The power  $p = 5$  is not covered by this theorem (there is no monotonicity formula), and the situation is different: universality is lost. Indeed, there are different types of blow-up:

the self-similar one given by the ODE  $u_{tt} = u^5$ , and bubble-type blow-up (so-called type II blow-up, with solutions  $u$  satisfying  $\sup_{0 < t < T} |\nabla u(t)|_{L^2} + |\partial_t u(t)|_{L^2} < \infty$ ):

**Theorem 4.2** (Existence of type II blow-up, Krieger, Schlag, Tataru [35–37]).  
*For all  $\nu > 0$ , there exists a blow-up solution of Eq. (ecNLW), such that*

$$(u(t), \partial_t u(t)) = \left( \frac{1}{(1-t)^{\frac{1+\nu}{2}}} W \left( \frac{x}{(1-t)^{1+\nu}} \right), 0 \right) + \eta(t, x),$$

with  $\eta$  continuous in  $\dot{H}^1 \times L^2$  up to  $T = 1$  and  $W$  defined in Eq. (1.14).

**4.1. Critical theory and the Kenig/Merle approach.** For the focusing equation (ecNLW), before 2006 only the small data theory was available. The smallness was measured in terms of Strichartz estimates and was unrelated to the size of the nonlinear stationary solution of the equation given by the elliptic theory (in connection with the best constant for Sobolev embedding), where the explicit solution  $W$  appears as a ground state (see [81]). So nothing was known for general nonlinear dynamics. A first step taken by Kenig/Merle was to relate these two theories using the criticality of the problem and to obtain an optimal small data theory (see [27] for a detailed review of the Kenig/Merle approach related to critical equations). This theory has its inspiration in variational problems where critical points correspond to nondispersive solutions. Nonvanishing or compactness properties for nonlinear solutions of Eq. (ecNLW) are obtained from the criticality using refined Strichartz estimates (see Bahouri/Gérard/Merle/Vega [1, 67], and other work in the 1980s for the elliptic counterpart: Talenti, Trudinger, Aubin, Schoen, Sachs, Uhlenbeck, Brezis, Coron, P.-L. Lions, Gérard, etc. [5, 40] and references therein). The difficult part (rigidity) is then to relate nondispersive solutions to stationary solutions in the energy space.

The breakthrough for this problem was, as before, the elimination of the case of energy-bounded, self-similar blow-up. Let us outline the methodology used to prove this fact (which is shared by all results of this type in other contexts):

- (i) Extraction of a nondispersive object  $v$  which is a solution of Eq. (ecNLW) with self-similar blow-up.
- (ii) Gain of decay on  $v$  (in this case: compact support in the space of the solution).
- (iii) Use of a monotonicity formula employed by Merle/Zaag which can be applied because of the gain of decay.
- (iv) As a consequence of the parabolic effect of this formula, we obtain a nonzero solution of an elliptic equation with a high degree of flatness at the boundary.
- (v) Then a contradiction follows from a unique continuation result on the elliptic equation.

As a consequence, we obtain the optimal small data theory and the classification of dynamics at the critical energy level.

**Theorem 4.3** (Dynamical characterization of  $W$  (1.14), Kenig, Merle [29, 30]).

*If  $u(t, x)$  is a solution of Eq. (ecNLW) with  $E(u, \partial_t u) < E(W, 0)$  (where  $E$  is defined in Eq. (1.8)), then:*

- (i) *If  $|\nabla u_0|_{L^2} < |\nabla W|_{L^2}$ ,  $u(t)$  is global and scatters.*

(ii) If  $|\nabla u_0|_{L^2} > |\nabla W|_{L^2}$ , then  $T_+, T_- < \infty$ .

Note the case  $|\nabla u_0|_{L^2} = |\nabla W|_{L^2}$  is impossible from the energy constraint and the constant in the Sobolev inequality ([81]). The analog of the  $L^2$  minimal mass characterization of  $S$  for the mass critical (NLS) is the following

**Theorem 4.4** (Dynamics at critical energy level, Duyckaerts, Merle [18, 19]). *There exist solutions  $W_-, W_+$  of Eq. (ecNLW) with  $E(W_{\pm}, W_{\pm t}) = E(W, 0)$  (where  $E$  is defined in Eq. (1.8)) such that*

- (i)  $W_-$  scatters at  $-\infty$  to  $W$  defined in Eq. (1.14) and at  $+\infty$  to a linear solution,
- (ii)  $W_+$  scatters at  $-\infty$  to  $W$  and  $T_+(W_+) < \infty$ .

Furthermore, up to the symmetry of the equation, if  $E(u, \partial_t u) = E(W, 0)$ :

- (i) If  $|\nabla u_0|_{L^2} < |\nabla W|_{L^2}$ , then  $u$  is globally defined, and  $u$  scatters to a linear solution at  $\pm\infty$ , or  $u = W_-$ .
- (ii) If  $|\nabla u_0|_{L^2} = |\nabla W|_{L^2}$ , then  $u = W$ .
- (iii) If  $|\nabla u_0|_{L^2} > |\nabla W|_{L^2}$ , then, either  $T_+, T_- < \infty$ , or  $u = W_+$ .

Other spectacular applications of this theory have been carried out, especially for the defocusing/focusing problems of the mass critical and the energy critical NLS (See Dodson/Killip/Tao/Visan [11, 12, 31, 32, 82]).

**4.2. The large data case following Duyckaerts/Kenig/Merle.** The final design of the Kenig/Merle method was to attack the problem of asymptotic soliton decomposition for general data. The goal is to prove that in the energy space, for a global solution in time, there exist an integer  $J$ ,  $J$  stationary solutions  $V^i$  of Eq. (ecNLW), and  $J$  parameters  $|l_i| < 1$  such that

$$u(t, x) - \left( \sum_{j=1}^J \frac{1}{\lambda_j(t)^{\frac{1}{2}}} V_{l_j}^j \left( 0, \frac{x - x_j(t)}{\lambda_j(t)} \right) + v_{lin}(t, x) \right) \rightarrow 0 \text{ as } t \rightarrow +\infty \tag{4.3}$$

in the energy space, where  $V_l$  is a traveling wave in the direction  $l$  defined by the Lorentz transformation from a stationary solution  $V$  of Eq. (ecNLW) for  $|l| < 1$ :

$$V_l(t, x) = V \left( \left( -\frac{t}{\sqrt{1 - |l|^2}} + \frac{1}{|l|^2} \left( \frac{1}{\sqrt{1 - |l|^2}} - 1 \right) l \cdot x \right) l + x \right), \tag{4.4}$$

and where  $v_{lin}$  is a solution of the linear wave equation and each  $V_{l_j}^j$  does not interact with the others.

This is a challenging question left open for decades for any equation in the nonintegrable case. Note that for large data we do not know anything about the solution (not even that it is bounded in the energy space). This conjecture can be decomposed into three different steps of varying difficulty:

Step (i) Let  $u$  be a solution of Eq. (ecNLW) and assume that  $u$  is bounded and does not scatter as  $t \rightarrow +\infty$ . Then there are  $(x_n, t_n), \lambda_n, t_n \rightarrow \infty$  such that we have the following: for a stationary solution  $V$  and  $l \in \mathbb{R}^3$  with  $|l| < 1$ ,

$$\lambda_n^{\frac{3}{2}} \nabla_{t,x} u(t_n, x_n + \lambda_n x) - \nabla_{t,x} V_l(0, x) \rightarrow 0 \text{ in } L^2 \text{ locally as } n \rightarrow +\infty, \tag{4.5}$$

where  $\nabla_{t,x} u = (\partial_t u, \partial_{x_1} u, \dots, \partial_{x_N} u)$  (Duyckaerts/Kenig/Merle [16]). In other words, the profile of the soliton appears recurrently localized in space as  $t_n \rightarrow \infty$ . (See also Christodoulou, Tahvildar-Zadeh [10], Struwe [80] Sterbenz, Tataru [79] for the wave map case.)

Part of the proof is the following characterization of nondispersive solutions: in the radial case or under a nondegeneracy condition, *a solution is nondispersive if and only if there exists a stationary solution  $V$  and  $|l| < 1$  such that  $u = V_l$  where  $V_l$  is defined by Eq. (4.4) (cf. [13, 17]).*

Step (ii) Obtain this decomposition result for a sequence  $t_n \rightarrow \infty$  (see [13] and Cote, Kenig, Lawrie, Schlag [8, 9]).

Step (iii) Obtain this decomposition result for the full set  $t \rightarrow \infty$ .

Note that going from step (ii) to step (iii) is challenging, and requires understanding collisions of solitons.

We now have the *full decomposition result in the radial situation* (which gives the first result in the nonintegrable case, see [13–15]):

**Theorem 4.5** (Radial soliton resolution, Duyckaerts, Kenig, Merle [15]).

*Let  $u$  be a radial solution of Eq. (ecNLW). Then one of the following holds:*

- (i) Type I blow-up:  $T < \infty$  and  $\lim_{t \rightarrow T} \|(u(t), \partial_t u(t))\|_{\dot{H}^1 \times L^2} = +\infty$ .
- (ii) Type II blow-up:  $T < \infty$  and there exist  $(v_0, v_1) \in \dot{H}^1 \times L^2$ , an integer  $J > 0$ ,  $\iota_j \in \{\pm 1\}$ , and functions  $\lambda_j(t)$  such that

$$\lambda_1(t) \ll \lambda_2(t) \ll \dots \ll \lambda_J(t) \ll T - t \text{ as } t \rightarrow T \tag{4.6}$$

$$\left\| (u(t), \partial_t u(t)) - \left( v_0 + \sum_{j=1}^J \frac{\iota_j}{\lambda_j^{1/2}(t)} W\left(\frac{x}{\lambda_j(t)}\right), v_1 \right) \right\|_{\dot{H}^1 \times L^2} \xrightarrow{t \rightarrow T} 0, \tag{4.7}$$

where  $W$  is defined in Eq. (1.14).

- (iii) Global solution:  $T = +\infty$  and there exist a solution  $v_{lin}$  of the linear wave equation, an integer  $J \geq 0$ ,  $\iota_j \in \{\pm 1\}$ , and functions  $\lambda_j(t)$  such that

$$\lambda_1(t) \ll \lambda_2(t) \ll \dots \ll \lambda_J(t) \ll t \text{ as } t \rightarrow +\infty \tag{4.8}$$

$$\left\| (u(t), \partial_t u(t)) - \left( v_{lin}(t) + \sum_{j=1}^J \frac{\iota_j}{\lambda_j^{1/2}(t)} W\left(\frac{x}{\lambda_j(t)}\right), \partial_t v_{lin}(t) \right) \right\|_{\dot{H}^1 \times L^2} \xrightarrow{t \rightarrow +\infty} 0. \tag{4.9}$$

One important ingredient of the proof is the following simultaneous dispersive property of radial solutions of Eq. (ecNLW): Any solution that is different from 0 and  $\pm W$  up to scaling has the following property on the existence of a channel of energy approaching infinity

starting from  $t = 0$ :  $\int_{|x|>R+|t|} |\nabla_{x,t} u(x, t)|^2 dx \geq \eta > 0$  for all  $t \geq 0$  or all  $t \leq 0$ . This leads to surprising decoupled estimates on the interactions of solitons. The proof of this property has a flavor similar to the moving plane techniques used in elliptic PDE to prove symmetry properties.

### 5. A supercritical example

One can see that the analysis of the previous examples is reduced to a *critical* situation where there is a *conservation law at the level of the critical space*. We are now going to illustrate an example in a *supercritical situation*, considering the construction of a blow-up solution for the supercritical NLS. Let us first briefly recall previous results about construction of blow-up solutions in a supercritical situation.

The only known example was the parabolic setting where the following equation was considered for  $p > \frac{N+2}{N-2}$ :

$$\partial_t u = \Delta u + |u|^{p-1}u, \quad x \in \mathbb{R}^N.$$

Herrero/Velázquez [25] were able to construct a blow-up solution related to a stationary solution for  $N > N(p)$  where  $N(p)$  is explicit. For the previous equation, we have a complete classification in the radial case of the different asymptotic behaviors of the solution (see Matano/Merle [52, 53]).

The problem is to initiate this program in the Hamiltonian situation by constructing a blow-up solution related to a stationary solution for  $N > N(p)$ . We will consider the supercritical nonlinear Schrödinger equation, for  $p$  an odd integer with  $p > \frac{N+2}{N-2}$ :

$$i\partial_t u + \Delta u + |u|^{p-1}u = 0, \quad u|_{t=0} = u_0, \quad x \in \mathbb{R}^N. \tag{5.1}$$

Our goal is to see if the previous method described in 2.2 is flexible enough to adapt to a supercritical situation. This includes getting formal asymptotics and a rigorous analysis for a blow-up solution.

For Eq. (5.1), we have a notion of critical space where the size of the initial data does not change with the scaling, which in the supercritical case is above the energy space. The scaling symmetry  $u_\lambda(t, x) = \lambda^{\frac{2}{p-1}} u(\lambda^2 t, \lambda x)$  for  $\lambda > 0$  is such that  $|u_\lambda(t, \cdot)|_{\dot{H}^{s_c}} = |u(\lambda^2 t, \cdot)|_{\dot{H}^{s_c}}$ , where  $s_c = \frac{N}{2} - \frac{2}{p-1} > 1$ .

We consider the stationary radial solution  $V^*$  of

$$\Delta V^* + |V^*|^{p-1}V^* = 0, \tag{5.2}$$

which has a slow decay at infinity ( $V^*(r) \sim r^{\frac{-2}{p-1}}$ ).  $V^*$  is not in the energy space but in  $\dot{H}^s$  for  $s > s_c > 1$ . There is an explicit  $\alpha = \alpha(p, N) > 0$  such that the following is true:

**Theorem 5.1** (Type II blow-up for the supercritical NLS Eq., Merle, Raphaël, Rodnianski [66]). *Let  $p$  be an odd integer and  $N > N(p)$ . For any integer  $\ell > \frac{\alpha}{2}$ , under some generic conditions, there exist  $s(\ell) > s_c$  and  $u_0 \in H^{s(\ell)}$  such that the solution  $u(t)$  of Eq. (5.1) blows up in finite time  $0 < T < +\infty$  by concentrating to the soliton profile:*

$$|\lambda(t)^{\frac{2}{p-1}} u(t, \lambda(t)r) - V^*(r)|_{\dot{H}^s} \rightarrow 0 \text{ as } t \rightarrow T \text{ for } s_c < s \leq s(\ell), \tag{5.3}$$

where  $V^*$  is the solution of Eq. (5.2), with:

- (i) Blow-up speed:  $\lambda(t) \sim (T - t)^{\frac{\ell}{\alpha}}$  and  $|u(t)|_{\dot{H}^s} \sim \frac{1}{(T-t)^{\frac{\ell(s-s_c)}{\alpha}}}$  for  $s_c < s \leq s(\ell)$ ,
- (ii) Behavior of the critical norm:  $|u(t)|_{\dot{H}^{s_c}} \sim \sqrt{|\log(T - t)|}$ , and
- (iii) Boundedness below scaling:  $\limsup_{t \uparrow T} |u(t)|_{H^s} < +\infty$  for  $0 \leq s < s_c$ .

This shows that the previous approaches are not specific to critical situations and the problems related to supercriticality can be overcome by adapting the critical method to the supercritical context. In particular,  $V^*$  does not have to belong to the critical space to be used as a bubbling profile. Properties of slow decay of the ground state are linked to a finite-dimensional nonlinear dynamical system, where we look for an *Ansatz* with a specific behavior. Finally, a key idea is to use estimates in high regularity spaces (even if we have no conservation laws in these spaces). Interpolation and cancellations allow us to obtain estimates in low regularity spaces.

## 6. Conclusion

We have illustrated *universality* for some canonical problems arising in different physical contexts. In these various Hamiltonian equations, we have found similarities in the results, and similarities in the approaches (monotonicity formulas, few parameters related to special solutions with nonlinear interactions where cancellations play a basic role, etc.). In conclusion, one can see a unity in these problems without apparent links *a priori*.

Nevertheless, it is clear that the proofs have to use the specifics of the each equation. Finally, we observe that there are many wide open directions of research related to these approaches.

**Acknowledgements.** Supported by The E.R.C. Advanced Grant 291214 BLOWDISOL.

## References

- [1] H. Bahouri and P. Gérard, *High frequency approximation of solutions to critical nonlinear wave equations*, Amer. J. Math. **121** (1999), 131–175.
- [2] H. Berestycki and P.-L. Lions, *Nonlinear scalar field equations. I. Existence of a ground state*, Arch. Rational Mech. Anal. **82** (1983), 313–345.
- [3] J. Bourgain, *Global well-posedness of defocusing critical nonlinear Schrödinger equation in the radial case*, J. Am. Math. Soc. **12** (1999), 145–171.
- [4] J. Bourgain and W. Wang, *Construction of blowup solutions for the nonlinear Schrödinger equation with critical nonlinearity*, Ann. S. Nor. Pisa **25** (1998), 197–215.
- [5] J. M. Coron and H. Brezis, *Convergence of solutions of H-systems or how to blow bubbles*, Arch. Rational Mech. Anal. **89** (1985), 21–56.
- [6] T. Cazenave and F. Weissler, *Some remarks on the nonlinear Schrödinger equation in the critical case*, Nonlinear semigroups, partial differential equations and attractors, 18–29, Lecture Notes in Math., 1394, Springer, Berlin, 1989.
- [7] J. Colliander, M. Keel, G. Staffilani, H. Takaoke, and T. Tao, *Global well-posedness*

- and scattering for the energy-critical nonlinear Schrödinger equation in  $\mathbb{R}^3$ , *Ann. of Math.* **167** (2008), 767–865.
- [8] R. Côte, *Soliton resolution for equivariant wave maps to the sphere*, arXiv:1305.5325.
- [9] R. Côte, C. Kenig, A. Lawrie, and W. Schlag, *Profiles for the radial focusing 4d energy-critical wave equation*, arXiv:1402.2307.
- [10] D. Christodoulou and A. Shadi Tahvildar-Zadeh, *On the asymptotic behavior of spherically symmetric wave maps*, *Duke Mathematical Journal* **71** (1993), 31–69.
- [11] B. Dodson, *Global well-posedness and scattering for the defocusing,  $L^2$ -critical nonlinear Schrödinger equation when  $d \geq 3$* , *J. Amer. Math. Soc.* **25** (2012), 429–463.
- [12] ———, *Global well-posedness and scattering for the defocusing, mass - critical generalized KdV equation*, arXiv:1304.8025.
- [13] T. Duyckaerts, C. Kenig and F. Merle, *Profiles of bounded radial solutions of the focusing, energy-critical wave equation*, *Geom. Funct. Anal.* **22** (2012), 639–698.
- [14] ———, *Universality of the blow-up profile for small type II blow-up solutions of the energy-critical wave equation: the nonradial case*, *J. Eur. Math. Soc.* **14** (2012), 1389–1454.
- [15] ———, *Classification of the radial solutions of the focusing, energy-critical wave equation*, *Cambridge Journal of Math.* **1** (2013), 75–144.
- [16] ———, *Profiles for bounded solutions of dispersive equations with applications to energy-critical wave and Schrödinger equations*, arXiv:1311.0665.
- [17] ———, *Solutions of the focusing nonradial critical wave equation with the compactness property*, arXiv:1402.0365.
- [18] T. Duyckaerts and F. Merle, *Dynamics of threshold solutions for energy-critical wave equation*, *Int. Math. Res. Pap. IMRP* 2007, Art. ID rpn002, 67 pp. (2008).
- [19] ———, *Dynamic of threshold solutions for energy-critical NLS*, *Geom. Funct. Anal.* **18** (2009), 1787–1840.
- [20] G. Fibich, F. Merle, and P. Raphaël, *Proof of a spectral property related to the singularity formation for the critical NLS*, *Phys. D* **220** (2006), 1–13.
- [21] J. Ginibre and G. Velo, *Generalized Strichartz inequalities for the wave equation*, *J. Funct. Anal.* **133** (1995), 50–68.
- [22] L. Glangetas and F. Merle, *A geometrical approach of existence of blow-up solution in  $H^1$  for nonlinear Schrödinger equations*, *Publications du Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie*, (1995).
- [23] R. Glassey, *On the blowing up of solutions to the Cauchy problem for nonlinear Schrödinger equations*, *J. Math. Phys.* **18** (1977), 1794–1797.
- [24] M. Grillakis, *Regularity and asymptotic behaviour of the wave equation with a critical nonlinearity*, *Ann. of Math.* **132** (1990), 485–509.
- [25] M. Herrero and J. Velázquez, *Explosion de solutions d'équations paraboliques semi-linéaires supercritiques*, *C. R. A. S.* **319** (1994), 141–145.
- [26] T. Kato, *On nonlinear Schrödinger equations*. *Ann. Inst. H. Poincaré Phys. Théor.* **46** (1987), 113–129.

- [27] C. Kenig, *Recent developments on the global behavior to critical nonlinear dispersive equations*, Proceedings of the International Congress of Mathematicians Volume I, 326–338, Hindustan Book Agency, New Delhi, 2010.
- [28] C. Kenig, G. Ponce, and L. Vega, *Well-posedness and scattering results for the generalized Korteweg–de Vries equation via the contraction principle*, Comm. Pure Appl. Math. **46** (1993), 527–620.
- [29] C. Kenig and F. Merle, *Global well-posedness, scattering and blow-up for the energy-critical focusing non-linear wave equation*. Acta Math. **201** (2008), 147–212.
- [30] ———, *Global well-posedness, scattering and blow-up for the energy-critical, focusing, non-linear Schrödinger equation in the radial case*, Invent. Math. **166** (2006), 645–675.
- [31] R. Killip, T. Tao, and M. Visan, *The cubic nonlinear Schrödinger equation in two dimensions with radial data*, J. Eur. Math. Soc. **11** (2009), 1203–1258.
- [32] R. Killip and M. Visan, *The focusing energy-critical nonlinear Schrödinger equation in dimensions five and higher*, Amer. J. Math. **132** (2010), 361–424.
- [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), 935–965.
- [34] J. Krieger and W. Schlag, *Non-generic blow-up solutions for the critical focusing NLS in 1-D*, Jour. Eur. Math. Soc. **11** (2009), 1–125.
- [35] ———, *Full range of blow up exponents for the quintic wave equation in three dimensions*, arXiv:1212.3795.
- [36] J. Krieger, W. Schlag, and D. Tataru, *Renormalization and blow-up for charge one equivariant critical wave maps*. Invent. Math. **171** (2008), 543–615.
- [37] ———, *Slow blow-up solutions for the  $H^1(\mathbb{R}^3)$  critical focusing semilinear wave equation*, Duke Math. J. **147** (2009), 1–53.
- [38] M. J. Landman, G. C. Papanicolaou, C. Sulem, and P.-L. Sulem, *Rate of blowup for solutions of the nonlinear Schrödinger equation at critical dimension*. Phys. Rev. A **38** (1988), 3837–3843.
- [39] H. Lindblad and C. D. Sogge, *On existence and scattering with minimal regularity for semilinear wave equations*. J. Funct. Anal. **130** (1995), 357–426.
- [40] P.-L. Lions, *The concentration-compactness principle in the calculus of variations. The limit case. I and II*, Rev. Mat. Ibero. **1** (1985), 45–121 and 145–201.
- [41] Y. Martel, *Asymptotic  $N$ -soliton-like solutions of the subcritical and critical generalized Korteweg–de Vries equations*, Amer. J. Math. **127** (2005), 1103–1140.
- [42] Y. Martel and F. Merle, *A Liouville theorem for the critical generalized Korteweg–de Vries equation*, J. Math. Pures Appl. **79** (2000), 339–425.
- [43] ———, *Stability of blow-up profile and lower bounds for blow-up rate for the critical generalized KdV equation*, Ann. of Math. **155** (2002), 235–280.
- [44] ———, *Blow-up in finite time and dynamics of blow-up solutions for the  $L^2$ -critical generalized KdV equation*, J. Amer. Math. Soc. **15** (2002), 617–664.
- [45] ———, *Description of two soliton collision for the quartic gKdV equation*. Ann. of

- Math. **174** (2011), 757–857.
- [46] ———, *Inelastic interaction of nearly equal solitons for the quartic gKdV equation*, Invent. Math. **183** (2011), 563–648.
- [47] ———, *On the Nonexistence of Pure Multi-solitons for the Quartic gKdV Equation*, Int Math Res Notices (2013), To appear.
- [48] Y. Martel, F. Merle, and P. Raphaël, *Blow-up for critical gKdV equation I: Dynamics near the soliton*, arXiv:1204.4625 To appear in Acta Math..
- [49] ———, *Blow-up for critical gKdV equation II: minimal mass solution*, arXiv:1204.4624.
- [50] ———, *Blow-up for critical gKdV equation III: exotic regimes*, arXiv:1209.2510 To appear in Annali Scuola Norm. Sup. di Pisa.
- [51] Y. Martel, F. Merle, and T.-P. Tsai, *Stability and asymptotic stability in the energy space of the sum of  $N$  solitons for subcritical gKdV equations*. Comm. Math. Phys. **231** (2002), 347–373.
- [52] H. Matano and F. Merle, *On nonexistence of type II blowup for a supercritical nonlinear heat equation*, Comm. Pure Appl. Math. **57** (2004), 1494–1541.
- [53] ———, *Classification of type I and type II behaviors for a supercritical nonlinear heat equation*, J. Funct. Anal. **256** (2009), 992–1064.
- [54] F. Merle, *Determination of blow-up solutions with minimal mass for nonlinear Schrödinger equations with critical power*. Duke Math. J. **69** (1993), 427–454.
- [55] ———, *Construction of solutions with exactly  $k$  blow-up points for the Schrödinger equation with critical nonlinearity*. Comm. Math. Phys. **129** (1990), 223–240.
- [56] ———, *On uniqueness and continuation properties after blow-up time of self-similar solutions of nonlinear Schrödinger equation with critical exponent and critical mass*, Comm. Pure Appl. Math. **45** (1992), 203–254.
- [57] ———, *Blow-up phenomena for critical nonlinear Schrödinger and Zakharov equations*, Proceedings of the International Congress of Mathematicians Volume III, 57–66, Doc. Math., Berlin, 1998.
- [58] ———, *Existence of blow-up solutions in the energy space for the critical generalized KdV equation*. J. Amer. Math. Soc. **14** (2001), 555–578.
- [59] F. Merle and P. Raphaël, *Sharp upper bound on the blow-up rate for the critical nonlinear Schrödinger equation*, Geom. Func. Anal. **13** (2003), 591–642.
- [60] ———, *On universality of blow-up profile for  $L^2$  critical nonlinear Schrödinger equation*. Invent. Math. **156** (2004), 565–672.
- [61] ———, *The blow-up dynamics and upper bound on the blow-up rate for the critical nonlinear Schrödinger equation*, Ann. of Math. **161** (2005), 157–222.
- [62] ———, *Profiles and quantization of the blow-up mass for critical nonlinear Schrödinger equation*, Commun. Math. Phys. **253** (2005), 675–704.
- [63] ———, *On a sharp lower bound on the blow-up rate for the  $L^2$  critical nonlinear Schrödinger equation*. J. Amer. Math. Soc. **19** (2006), 37–90.
- [64] F. Merle, P. Raphaël, and J. Szeftel, *The instability of Bourgain-Wang solutions for the*

- $L^2$  critical NLS, Amer. Jour. Math. **135** (2013), 967–1017.
- [65] F. Merle, P. Raphaël, and I. Rodnianski, *Blow-up dynamics for smooth data equivariant solutions to the energy critical Schrödinger map problem*, Invent. Math. **193** (2013), 249–365.
- [66] ———, *Type II blow up for the energy supercritical NLS*, preprint.
- [67] F. Merle and L. Vega, *Compactness at blow-up time for  $L^2$  solutions of the critical nonlinear Schrödinger equation in 2D*, Internat. Math. Res. Notices **8** (1998), 399–425.
- [68] F. Merle and H. Zaag, *Determination of the blow-up rate for the semilinear wave equation*, Amer. J. Math. **125** (2003), 1147–1164.
- [69] F. Merle and H. Zaag, *Existence and classification of characteristic points at blow-up for a semilinear wave equation in one space dimension*, Amer. J. Math. **134** (2012), 581–648.
- [70] ———, *Isolatedness of characteristic points for a semilinear wave equation in one space dimension*, Duke Math. J. **161** (2012), 2837–2908.
- [71] ———, *On the stability of the notion of non-characteristic point and blow-up profile for semilinear wave equations*, (2013), arXiv:1309.7760.
- [72] K. Nakanishi and W. Schlag, *Global dynamics above the ground state energy for the cubic NLS equation in 3D*, Arch. Ration. Mech. Anal. **203** (2012), 809–851.
- [73] G. Perelman, *On the formation of singularities in solutions of the critical nonlinear Schrödinger equation*, Ann. Henri Poincaré **2** (2001), 605–673.
- [74] P. Raphaël, *Stability of the log-log bound for blow-up solutions to the critical nonlinear Schrödinger equation*. Math. Ann. **331** (2005), 577–609.
- [75] ———, *Blow up bubbles in Hamiltonian evolution equations: a quantitative approach*, *Proceedings of the International Congress of Mathematicians*, (2014), To appear.
- [76] 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 Etudes Sci. **115** (2012), 1–122.
- [77] I. Rodnianski and J. Sterbenz, *On the formation of singularities in the critical  $O(3)$   $\sigma$ -model*, Ann. of Math. **172** (2010), 187–242.
- [78] J. Shatah and M. Struwe, *Regularity results for nonlinear wave equations*, Ann. of Math. **138** (1993), 503–518.
- [79] J. Sterbenz and D. Tataru, *Regularity of wave-maps in dimension  $2 + 1$* , Comm. Math. Phys. **298** (2010), 139–230.
- [80] M. Struwe, *Radially symmetric wave maps from  $(1 + 2)$ -dimensional Minkowski space to the sphere*, Math. Z. **242** (2002), 407–414.
- [81] G. Talenti, *Best constant in Sobolev inequality*. Ann. Mat. Pura Appl. **110** (1976), 353–372.
- [82] T. Tao, M. Visan, and X. Zhang, *Minimal-mass blowup solutions of the mass-critical NLS*, Forum Math. **20** (2008), 881–919.
- [83] S. Vlasov, V. Petrishchev, and V. Talanov, *Averaged description of wave beams in linear*

- and nonlinear media*, Radiophysics and Quantum Electronics **14** (1971), 1062–1070.
- [84] M. I. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates*, Comm. Math. Phys. **87** (1983), 567–576.

Université de Cergy-Pontoise, Mathématiques, CNRS, F-95000 Cergy-Pontoise, FRANCE; Institut des Hautes Études Scientifiques, 35 Route de Chartres, 91440 Bures-sur-Yvette, FRANCE.

E-mail: merle@math.u-cergy.fr; merle@ihes.fr

# Wild harmonic bundles and twistor $\mathcal{D}$ -modules

Takuro Mochizuki

**Abstract.** The notion of twistor structure is a generalization of that of Hodge structure. Harmonic bundles and twistor  $\mathcal{D}$ -modules are the counterparts of polarized variations of Hodge structure and Hodge modules in the context of twistor structures. The study on harmonic bundles with wild singularity and twistor  $\mathcal{D}$ -modules lead us to an interesting interaction between global analysis and algebraic analysis. It has resulted in significant progress in the theory of holonomic  $\mathcal{D}$ -modules also in the context of irregular singularities. We will report on these developments.

**Mathematics Subject Classification (2010).** 14F10, 32C38, 32G20, 32S40, 53C07.

**Keywords.** Variation of Hodge structure, twistor structure, holonomic  $\mathcal{D}$ -module, stokes structure, singularity.

## 1. Introduction

Recall a theorem of Corlette [15]: a vector bundle with a flat connection on a complex projective manifold is semisimple if and only if it admits a pluri-harmonic metric. It is a deep theorem obtained with methods in global analysis. One of our main results is a generalization of the theorem of Corlette in the context of algebraic holonomic  $\mathcal{D}$ -modules.

The notion of a holonomic  $\mathcal{D}$ -module is a generalization of that of a vector bundle with a flat connection, and has been fundamental in contemporary mathematics. Holonomic  $\mathcal{D}$ -modules, unlike vector bundles, can have singularities. Hence, passing from the vector bundle case to that of holonomic  $\mathcal{D}$ -modules requires significant effort to deal with the singularities of pluri-harmonic metrics and holonomic  $\mathcal{D}$ -modules.

“Simpson’s Meta-Theorem” is an important guiding principle in our study of the singularities. C. Simpson [99] introduced the notion of twistor structure as a generalization of that of Hodge structure. He found that a flat bundle with a pluri-harmonic metric, that is, a harmonic bundle, can be regarded as a “polarized variation of pure twistor structure”. He proposed a principle that he calls Meta-Theorem, which roughly says that any theorem concerning Hodge structures should have a counterpart in the context of twistor structures.

The principle turns out to be useful in the study of the asymptotic behaviour of harmonic bundles near their singularities, although there are some peculiar phenomena in the twistor case. Indeed, our study [65, 72] is inspired by the study of the asymptotic behaviour of polarized variation of pure Hodge structure due to E. Cattani, A. Kaplan, W. Schmid [9–11, 92] and M. Kashiwara, T. Kawai [42, 46].

The principle also suggests that the meaning of a “pluri-harmonic metric” for a holonomic  $\mathcal{D}$ -module  $M$  could be clarified as a “twistor version of polarized pure Hodge mod-

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

ule" over  $M$ . This led to the notion of a polarized pure twistor  $\mathcal{D}$ -modules. The study of twistor  $\mathcal{D}$ -modules was initiated by C. Sabbah [81, 83] and developed by him and the author [65, 72, 73], on the basis of the theory of Hodge modules due to M. Saito [87, 88].

Besides pursuing a twistor version of the Hodge theory suggested by Simpson's Meta-Theorem, we also need some advances in the general theory concerning the structure of irregular singularities of meromorphic flat bundles. In the one dimensional case, there is the classical theory. It consists of (i) the existence theorem of a nice formal decomposition on a ramified covering, (ii) the asymptotic analysis in order to lift the formal decomposition to convergent flat decompositions on small sectors, which leads us the concept of the Stokes structure, (iii) the classification of meromorphic flat bundles by local systems with Stokes structure. We can say that it is appropriately generalized in the higher dimensional case, after the study by Y. André [1], K. Kedlaya [49, 50], H. Majima [56], B. Malgrange [58], Sabbah [80] and the author [72]. It is fundamental not only for our study but also in the theory of holonomic  $\mathcal{D}$ -modules.

Why is the characterization interesting? It is mainly because of the nice functorial behaviour of twistor  $\mathcal{D}$ -modules with respect to the standard operations. This implies some highly non-trivial results on semisimple holonomic  $\mathcal{D}$ -modules.

Actually, the most interesting consequence of our study is a solution of a conjecture of Kashiwara [44]. As in the Hodge case, the Hard Lefschetz theorem is shown to hold for polarizable pure twistor  $\mathcal{D}$ -modules. Said briefly, it means the functoriality of polarizable pure twistor  $\mathcal{D}$ -modules with respect to the push-forward by any projective morphisms, and the decomposition theorem for the underlying holonomic  $\mathcal{D}$ -modules. Together with the characterization of semisimplicity for algebraic holonomic  $\mathcal{D}$ -modules, we obtain the following [72]: Let  $f : X \rightarrow Y$  be a morphism of complex projective manifolds, and let  $M$  be a semisimple holonomic  $\mathcal{D}_X$ -module. Then, the  $j$ -th cohomology of the push-forward  $f_+^j M$  is semisimple, and we have an isomorphism  $f_+ M \simeq \bigoplus f_+^j M[-j]$  in the derived category of holonomic  $\mathcal{D}_Y$ -modules. This statement was conjectured by Kashiwara.

In the following, we shall explain the above results with more details. An outline is as follows. In §2, we recall the definition of a harmonic bundle. In §3, we review the fundamental results on harmonic bundles over complex projective manifolds, as a motivation for our study. Roughly, it is our purpose to develop a theory of singular harmonic bundles. As already mentioned, twistor structure is the key notion which is explained in §4. We explain in §5 some basic results on the local structure of meromorphic flat bundles, which are fundamental for our analysis of the singularities of harmonic bundles and meromorphic flat bundles. It is necessary to consider an additional structure on meromorphic flat bundles, called a parabolic structure or equivalently a filtered bundle, which is explained in §6. We give an overview of the study on wild harmonic bundles in §7, where the concepts and the results in §4–§6 appear naturally. We shall briefly describe the theory of twistor  $\mathcal{D}$ -modules in §8.

## 2. Harmonic bundles

We begin with some basic concepts in differential geometry of complex manifolds, and fix our notation. For any complex manifold  $X$ , let  $C^\infty(X)$  be the space of  $C^\infty$ -functions on  $X$ . Let  $\Omega_X^{p,q}$  denote the bundle of  $(p, q)$ -forms on  $X$ . For any  $C^\infty$ -vector bundle  $E$  on  $X$ , let  $A^{p,q}(E)$  denote the space of  $C^\infty$ -sections of  $E \otimes \Omega_X^{p,q}$ . We set  $A^r(E) := \bigoplus_{p+q=r} A^{p,q}(E)$ .

A connection of  $E$  is a differential operator  $\nabla : A^0(E) \rightarrow A^1(E)$  satisfying the Leibniz rule  $\nabla(fs) = f\nabla(s) + s \otimes df$  for any  $f \in C^\infty(X)$  and  $s \in A^0(E)$ . The induced operators  $A^p(E) \rightarrow A^{p+1}(E)$  are also denoted by  $\nabla$ . It is well known that the curvature  $R(\nabla) := \nabla \circ \nabla$  is an element of  $A^2(\text{End}(E))$ . The connection is called flat if  $R(\nabla) = 0$ . In that case,  $(E, \nabla)$  is called a flat bundle.

A holomorphic structure of  $E$  is a differential operator  $\bar{\partial}_E : A^0(E) \rightarrow A^{0,1}(E)$  such that (i)  $\bar{\partial}_E(fs) = f\bar{\partial}_E s + s \otimes \bar{\partial}f$  for any  $f \in C^\infty(X)$  and  $s \in A^0(E)$ , (ii)  $\bar{\partial}_E \circ \bar{\partial}_E = 0$ , where the induced operators  $A^{p,q}(E) \rightarrow A^{p,q+1}(E)$  are also denoted by  $\bar{\partial}_E$ . A vector bundle with a holomorphic structure is called holomorphic vector bundle. It is well known that this definition is equivalent to another definition as patched objects of trivial bundles by holomorphic transformations.

Let  $(E, \bar{\partial}_E)$  be a holomorphic vector bundle. If we are given a Hermitian metric  $h$  of  $E$ , we have a unique unitary connection  $\nabla_h$  such that the  $(0, 1)$ -part of  $\nabla_h$  is equal to  $\bar{\partial}_E$ . It is called the Chern connection of  $(E, \bar{\partial}_E, h)$ .

Let  $\Omega_X^p$  denote the sheaf of holomorphic  $p$ -forms on  $X$ . The corresponding holomorphic vector bundle is also denoted by  $\Omega_X^p$ . A Higgs field on a holomorphic vector bundle  $(E, \bar{\partial}_E)$  on  $X$  is a holomorphic section  $\theta$  of  $\text{End}(E) \otimes \Omega_X^1$  satisfying  $\theta \wedge \theta = 0$ . In other words, a Higgs field is  $\theta \in A^{1,0}(\text{End}(E))$  such that  $\theta \wedge \theta = \bar{\partial}\theta = 0$ .

**Harmonic bundles.** Let  $(E, \bar{\partial}_E, \theta)$  be a Higgs bundle with a metric  $h$ . We have the Chern connection  $\nabla_h$ . We also have the adjoint  $\theta_h^\dagger \in A^{0,1}(\text{End}(E))$  of  $\theta$  with respect to  $h$ . If the connection  $\mathbb{D}_h^1 := \nabla_h + \theta + \theta_h^\dagger$  is flat, the metric  $h$  is called a pluri-harmonic metric of the Higgs bundle, and  $(E, \bar{\partial}_E, \theta, h)$  is called a harmonic bundle [97]. More explicitly, for the decomposition  $\nabla_h = \bar{\partial}_E + \partial_{E,h}$  into the  $(0, 1)$ -part and the  $(1, 0)$ -part, the condition is described as  $\partial_{E,h}\theta = \bar{\partial}\theta_h^\dagger = R(\nabla_h) + [\theta, \theta_h^\dagger] = 0$ . If  $\dim X = 1$ , the equalities  $\partial_{E,h}\theta = \bar{\partial}\theta_h^\dagger = 0$  hold trivially, and the equation  $R(\nabla_h) + [\theta, \theta_h^\dagger] = 0$  is often called the Hitchin equation.

Let  $(V, \nabla)$  be a flat bundle. If we are given a Hermitian metric  $h$  of  $V$ , we have a unique decomposition  $\nabla = \nabla_h + \Phi_h$ , where  $\nabla_h$  is a unitary connection and  $\Phi_h \in A^1(\text{End } E)$  is self-adjoint. We also have the decompositions  $\nabla_h = \bar{\partial}_{V,h} + \partial_{V,h}$  and  $\Phi_h = \theta_h^\dagger + \theta_h$  into the  $(0, 1)$ -part and the  $(1, 0)$ -part. If  $(V, \bar{\partial}_{V,h}, \theta_h)$  is a Higgs bundle, then  $h$  is called a pluri-harmonic metric of  $(V, \nabla)$ , and  $(V, \nabla, h)$  is called a harmonic bundle [97]. The condition is described as  $\bar{\partial}_{V,h} \circ \partial_{V,h} = \bar{\partial}_{V,h}\theta_h = \theta_h \wedge \theta_h = 0$ . Indeed, the condition  $\bar{\partial}_{V,h}\theta = 0$  implies the others.

If  $h$  is a pluri-harmonic metric for a Higgs bundle  $(E, \bar{\partial}_E, \theta)$ , it is a pluri-harmonic metric for the induced flat bundle  $(E, \mathbb{D}_h^1)$ . If  $h$  is a pluri-harmonic metric for a flat bundle  $(V, \nabla)$ , it is a pluri-harmonic metric for the induced Higgs bundle  $(V, \bar{\partial}_{V,h}, \theta_h)$ . The constructions are mutually inverse.

### 3. Corlette-Simpson correspondence

The following theorem is one of the earliest but also the most interesting in the theory of harmonic bundles. It is due to S. K. Donaldson [25] and N. Hitchin [36] in the case  $\dim X = 1$  and  $\text{rank } E = 2$ , and due to K. Corlette [15] and C. Simpson [94] in the general case. It is a really deep result in the crossroad between algebraic geometry and differential geometry.

**Theorem 3.1.** *Let  $X$  be an  $n$ -dimensional connected complex projective manifold with an ample line bundle  $L$ . We have the correspondences of harmonic bundles, semisimple flat bundles, and  $\mu_L$ -polystable Higgs bundles  $(E, \bar{\partial}_E, \theta)$  with  $\text{ch}_1(E)c_1(L)^{n-1} = \text{ch}_2(E)c_1(L)^{n-2} = 0$ .*

We explain the details of the statement. A flat bundle is called irreducible if it does not have any non-trivial flat subbundles, and it is called semisimple if it is a direct sum of irreducible ones. For any coherent torsion-free  $\mathcal{O}_X$ -module  $\mathcal{F}$ , we set

$$\mu_L(\mathcal{F}) := (\text{rank } \mathcal{F})^{-1} \int_X c_1(\mathcal{F})c_1(L)^{n-1}$$

A Higgs bundle  $(E, \bar{\partial}_E, \theta)$  is  $\mu_L$ -stable (resp.  $\mu_L$ -semistable) if  $\mu_L(\mathcal{F}) < \mu_L(E)$  (resp.  $\mu_L(\mathcal{F}) \leq \mu_L(E)$ ) holds for any saturated  $\mathcal{O}_X$ -subsheaf  $\mathcal{F}$  of  $E$  with  $\theta(\mathcal{F}) \subset \mathcal{F} \otimes \Omega_X^1$  and  $0 < \text{rank } \mathcal{F} < \text{rank } E$ . A  $\mu_L$ -semistable Higgs bundle is called  $\mu_L$ -polystable if it is a direct sum of  $\mu_L$ -stable ones. (The condition is also called  $\mu$ -stability, slope-stability, etc.)

On one side of Theorem 3.1, for any harmonic bundle on  $X$ , the underlying flat bundle is semisimple, and the underlying Higgs bundle  $(E, \bar{\partial}_E, \theta)$  is  $\mu_L$ -polystable with

$$\text{ch}_1(E)c_1(L)^{n-1} = \text{ch}_2(E)c_1(L)^{n-2} = 0 \text{ (actually } \text{ch}_j(E) = 0 \text{ for } j > 0\text{)}.$$

We have orthogonal decompositions into irreducible ones or stable ones. The other side of Theorem 3.1 says: if a flat bundle  $(V, \nabla)$  on  $X$  is irreducible, there exists a pluri-harmonic metric  $h$  for  $(V, \nabla)$ ; if a Higgs bundle  $(E, \bar{\partial}_E, \theta)$  is  $\mu_L$ -stable with  $\text{ch}_1(E)c_1(L)^{n-1} = \text{ch}_2(E)c_1(L)^{n-2} = 0$  then there exists a pluri-harmonic metric  $h$  for  $(E, \bar{\partial}_E, \theta)$ . The metrics are unique up to the multiplication by a positive constant. These existence results are hard to prove, and are obtained through powerful methods developed in global analysis, mentioned later.

It is instructive to recall an easy consequence of the theorem. Let  $F : Y \rightarrow X$  be a morphism of projective manifolds. Then, for any semisimple flat bundle  $(V, \nabla)$  on  $X$ , the pull back  $F^*(V, \nabla)$  is also semisimple. Indeed, take a pluri-harmonic metric  $h$  on  $(V, \nabla)$ . It is easy to check that  $F^*h$  is a pluri-harmonic metric on  $F^*(V, \nabla)$ , and hence  $F^*(V, \nabla)$  is semisimple. The author does not know an elementary proof of this fact.

The theorem consists of two correspondences: Corlette [15] proved the correspondence between harmonic bundles and flat bundles, and Simpson [94] proved the correspondence between harmonic bundles and Higgs bundles.

Corlette studied harmonic metrics on flat bundles on Riemannian manifolds  $Y$ . (See also the previous works [22] and [25]). A Hermitian metric on a flat bundle on  $Y$  is defined to be a harmonic metric if the induced map from a universal covering of  $Y$  to the space of Hermitian matrices is a harmonic map. With the method of heat equations developed by S. K. Donaldson [23] and J. Eells-J. Sampson[29], he established that a flat bundle on a compact Riemannian manifold is semisimple if and only if it admits a harmonic metric. Moreover, he proved that a harmonic metric on  $(V, \nabla)$  on a compact Kähler manifold is always *pluri-harmonic*, by effectively applying the Bochner-Sampson-Siu technique in the theory of harmonic maps. Thus, he arrived at the above correspondence. Note that Corlette studied more general objects, and he obtained various impressive applications including remarkable results for rigidity of lattices in some Lie groups. See [16] for details.

Simpson studied the Kobayashi-Hitchin correspondence for Higgs bundles. Let  $X$  be a smooth projective variety with an ample line bundle  $L$  and a Kähler form  $\omega$  representing

$c_1(L)$ . A Hermitian metric  $h$  on a Higgs bundle  $(E, \bar{\partial}_E, \theta)$  on  $X$  is called Hermitian-Einstein if  $\Lambda_\omega(\mathbb{D}_h^1 \circ \mathbb{D}_h^1) \in A^0(\text{End}(E))$  is the multiplication by a constant, where  $\Lambda_\omega$  denotes the adjoint of the multiplication of  $\omega$ . He established that  $(E, \bar{\partial}_E, \theta)$  is  $\mu_L$ -polystable if and only if it admits a Hermitian-Einstein metric, by developing the method of Donaldson [23, 24], Hitchin [36], K. Uhlenbeck and S. T. Yau [104]. The origin of the subject can trace back to the work of M. S. Narasimhan and C. S. Seshadri [75] on unitary flat bundles over compact Riemann surfaces, and the introduction of the Einstein condition by S. Kobayashi [51]. See [55] for more details on the Kobayashi-Hitchin correspondence. Moreover, Simpson proved  $\int_X ((2 \text{rank } E)^{-1} c_1(E)^2 - \text{ch}_2(E)) \omega^{n-2} = C \int_X |(\mathbb{D}_h^1 \circ \mathbb{D}_h^1)^\perp|_{h, \omega}^2 \geq 0$  for any Hermitian-Einstein metric  $h$ , where  $(\cdot)^\perp$  denotes the trace free part. It is the Bogomolov-Gieseker inequality in the Higgs case, and the vanishing condition for the Chern character of  $E$  implies  $\mathbb{D}_h^1 \circ \mathbb{D}_h^1 = 0$ , i.e.,  $h$  is pluri-harmonic.

It would be valuable to mention that Simpson developed his arguments even for non-compact  $X$  under some conditions, which was applied in the curve case [94, 95]. He also found that his argument for Higgs bundles can be adapted for flat bundles [97]. They were crucial for our study [64, 66, 72]. His result was also efficiently used in a recent application to monopoles [14].

**Polarized variation of Hodge structure.** We can appreciate Simpson’s work as a very exciting fusion of interesting branches of mathematics. As explained above, his study takes place in an interesting stream of global analysis. It is also a significant bridge with another apparently completely different field of mathematics, that is the Hodge theory.

To a smooth projective morphism of complex manifolds  $f : X \rightarrow S$ , we associate the local system  $R^i f_* \mathbb{Q}_X$  on  $S$ . The associated holomorphic vector bundle  $R^i f_* \mathbb{Q}_X \otimes_{\mathbb{Q}_S} \mathcal{O}_S$  with the induced flat connection is called the Gauss-Manin connection. P. Griffiths introduced the concept of polarized variation of pure Hodge structure by abstracting the properties of the induced structure on the Gauss-Manin connection, and opened a new research area. See his survey [32], for example.

We only recall the concept of complex variation of Hodge structure [94] for which we consider two filtrations instead of a filtration with a  $\mathbb{Q}$ -structure. First, a complex Hodge structure is a  $\mathbb{C}$ -vector space  $H$  with two finite decreasing filtrations  $F = (F^p \mid p \in \mathbb{Z})$  and  $G = (G^p \mid p \in \mathbb{Z})$ . Then,  $(H, F, G)$  is called pure of weight  $n$  if we have  $F^p \cap G^q = 0$  for  $p + q > n$  and  $H = \bigoplus_{p+q=n} F^p \cap G^q$ . A polarization  $S$  of  $(H, F, G)$  is a Hermitian or anti-Hermitian form of  $H$  depending whether  $n$  is even or odd, such that (i) the decomposition  $H = \bigoplus_{p+q=n} F^p \cap G^q$  is orthogonal with respect to  $S$ , (ii)  $(\sqrt{-1})^{p-q} S$  on  $F^p \cap G^q$  is positive definite.

A variation of complex Hodge structure on a complex manifold  $X$  is a flat bundle  $(V, \nabla)$  on  $X$  equipped with two decreasing filtrations  $F$  and  $G$  by vector subbundles satisfying the Griffiths transversality conditions  $\nabla A^0(F^p) \subset A^{0,1}(F^p) \oplus A^{1,0}(F^{p-1})$  and  $\nabla A^0(G^q) \subset A^{1,0}(G^q) \oplus A^{0,1}(G^{q-1})$ . It is called pure of weight  $n$  if the restriction  $(V, F, G)|_P$  for any  $P \in X$  is pure Hodge structure of weight  $n$ . A polarization of  $(V, \nabla, F, G)$  is a  $\nabla$ -flat sesqui-linear form  $S$  on  $V$ , such that  $S|_P$  is a polarization of  $(V, F, G)|_P$  for any  $P \in X$ .

A harmonic bundle naturally appears from a polarized variation of Hodge structure. Indeed, the positive definite Hermitian pairing in the definition of polarization is a pluri-harmonic metric of the flat bundle. Equivalently, the associated graded holomorphic vector bundle  $\text{Gr}_F(V)$  is equipped with the Higgs field induced by the sum of  $\text{Gr}_F(\nabla) : \text{Gr}_F^p(V) \rightarrow \text{Gr}_F^{p-1}(V) \otimes \Omega^1$ , and the induced metric on  $\text{Gr}_F(V)$  is a pluri-harmonic

metric of the Higgs bundle.

Conversely, suppose we are given a Hodge bundle i.e., a Higgs bundles  $(E, \theta)$  with a grading  $E = \bigoplus E_p$  such that  $\theta(E_p) \subset E_{p-1} \otimes \Omega^1$ . If  $(E, \theta)$  has a pluri-harmonic metric  $h$  for which the decomposition  $\bigoplus E_p$  is orthogonal, then the harmonic bundle  $(E, \theta, h)$  comes from a polarized complex variation of Hodge structure. It is more convenient to replace the orthogonality condition by the equivariance with respect to the  $S^1$ -action on  $E$  given by  $g_t = \bigoplus t^p \text{id}_{E_p}$ .

By applying his correspondence, Simpson obtained a construction of a polarized variation of Hodge structure, not through the Gauss-Manin connections.

**Theorem 3.2** ([94]). *Suppose that we are given a Hodge bundle with trivial Chern character on a projective manifold with an ample line bundle  $L$ . It is  $\mu_L$ -polystable if and only if it comes from a polarized complex variation of Hodge structure.*

See [77, 94, 96, 97, 100, 107], and their references for more applications of harmonic bundles on projective manifolds.

**Afterward.** Through the work of Simpson, it turned out that harmonic bundles and polarized variations of pure Hodge structure share many important properties. For example, Simpson [97] developed the harmonic theory for harmonic bundles, and established the Hard Lefschetz theorem for the cohomology groups associated to harmonic bundles on projective manifolds, as in the Hodge case due to Deligne. Because any semisimple flat bundle comes from a harmonic bundle, we may say that Hodge-like property is rather common. This is a new point of view.

The subsequent study is roughly divided into two branches. One is directed to the interest in the moduli spaces of Higgs (flat) bundles on curves, which are interesting in relation with many subjects including the Langlands theory, the non-abelian Hodge theory, the mirror symmetry, the integrable systems, etc. Rich structures on the moduli spaces revealed by Hitchin and Simpson are intriguing for many researchers. Moreover, delicate issues to deal with the moduli spaces caused the development of the theory of  $n$ -stacks. The other is directed to the interest in harmonic bundles with singularities particularly in the higher dimensional case, that is the main topic in the rest of this text.

## 4. Twistor structure

A twistor structure, as introduced by Simpson, is a holomorphic vector bundle on  $\mathbb{P}^1$ . A mixed twistor structure is a twistor structure  $V$  with a finite increasing filtration  $W = (W_k \mid k \in \mathbb{Z})$  such that  $\text{Gr}_k^W(V) \simeq \mathcal{O}_{\mathbb{P}^1}(k)^{\oplus r_k}$  for some  $r_k \geq 0$ . If  $\text{Gr}_k^W(V) = 0$  unless  $k = w$ , it is called pure of weight  $w$ .

The relation with Hodge structure is given by the Rees construction. From a vector space  $H$  with a Hodge filtration  $F$ , we obtain a  $\mathbb{C}[\lambda]$ -module  $\xi(H; F) = \sum F^j(H)\lambda^{-j}$ , which gives a vector bundle on  $\text{Spec } \mathbb{C}[\lambda]$ . Another filtration  $G$  on  $H$  induces  $\xi(H; G)$  on  $\mathbb{C}[\lambda^{-1}]$ . By gluing them, we obtain a twistor structure. It is naturally equipped with an action of  $G_m = \text{Spec } \mathbb{C}[\lambda, \lambda^{-1}]$ . Thus, we obtain a functor  $\xi$  from the category of complex Hodge structures to the category of  $G_m$ -equivariant twistor structures.

**Theorem 4.1** (Simpson). *The functor  $\xi$  is an equivalence. It induces an equivalence between the categories of mixed Hodge structures and  $G_m$ -equivariant mixed twistor structures.  $\square$*

Simpson found that mixed twistor structures share some of important properties with mixed Hodge structures. The concepts of Tate twist, polarization, etc., for Hodge structures are naturally translated to those for twistor structures. A morphism of mixed twistor structures is always strict with respect to the weight filtrations, from which we obtain that the category of mixed twistor structures is abelian, as is the category of mixed Hodge structure.

Another important discovery due to Simpson is that harmonic bundles can be regarded as polarized variations of pure twistor structure. As he emphasized, this is the “raison d’être” of mixed twistor structures. It establishes a similarity between harmonic bundles and polarized variations of Hodge structure at the level of definitions, which explains the previously known resemblances. This gives us a drastic change of view to harmonic bundles. It indicates how theorems concerning variation of Hodge structure could be generalized to those in the context of harmonic bundles. Our study of the asymptotic behaviour of wild harmonic bundles owes much to this idea (see §7). It also predicted a twistor version of the theory of Hodge modules, realized as the theory of twistor  $\mathcal{D}$ -modules (see §8).

We explain the correspondence only in the case of weight 0, although we can freely shift the weight.

Let  $\sigma : \mathbb{P}^1 \rightarrow \mathbb{P}^1$  be the anti-holomorphic involution defined by  $\sigma(\lambda) = -\bar{\lambda}^{-1}$ . Let  $V$  be any twistor structure. Then,  $\sigma^*V$  is naturally a twistor structure with the action of  $\mathcal{O}_{\mathbb{P}^1}$  on  $\sigma^*V$  given by  $f \sigma^*v := \sigma^*(\sigma^*(f)v)$ . We have a natural identification  $\sigma^*\mathcal{O}_{\mathbb{P}^1} \simeq \mathcal{O}_{\mathbb{P}^1}$  given by  $\sigma^*(f) \mapsto \sigma^*(\bar{f})$ .

Let  $V$  be a pure twistor structure of weight 0. A morphism  $\mathcal{S} : V \otimes \sigma^*V \rightarrow \mathcal{O}_{\mathbb{P}^1}$  is called symmetric, if we have  $\sigma^*\mathcal{S}(\sigma^*v \otimes u) = \mathcal{S}(u \otimes \sigma^*v)$ . A symmetric pairing  $\mathcal{S}$  induces a Hermitian pairing  $H^0(\mathcal{S})$  on the vector space  $H^0(\mathbb{P}^1, V)$ . We say that  $\mathcal{S}$  is a polarization of  $V$  if  $H^0(\mathcal{S})$  is positive definite. It is easy to see that  $H^0$  induces an equivalence between pure twistor structures of weight 0 with a polarization and vector spaces with a Hermitian metric. A quasi-inverse is given as follows. A vector space  $U$  induces a pure twistor structure  $V = U \otimes \mathcal{O}_{\mathbb{P}^1}$  of weight 0. A Hermitian metric  $h$  of  $U$  induces a polarization  $\mathcal{S}_h : V \otimes \sigma^*V \rightarrow \mathcal{O}_{\mathbb{P}^1}$  given by  $\mathcal{S}_h(fu, \sigma^*(gv)) = f \sigma^*(\bar{g})h(u, v)$  for  $u, v \in U$  and local sections  $f$  and  $g$  of  $\mathcal{O}_{\mathbb{P}^1}$ .

Let  $Y$  be any complex manifold. Let  $\mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$  be the sheaf of  $C^\infty$ -functions which are holomorphic in the  $\mathbb{P}^1$ -direction. Let  $p_i$  ( $i = 1, 2$ ) denote the projections of  $\mathbb{P}^1 \times Y$  onto the  $i$ -th component. We set  $\xi\mathcal{A}_Y^1 := \left( p_2^{-1}\Omega_Y^{0,1} \otimes p_1^{-1}\mathcal{O}_{\mathbb{P}^1}(\{\infty\}) \right) \oplus \left( p_2^{-1}\Omega_Y^{1,0} \otimes p_1^{-1}\mathcal{O}_{\mathbb{P}^1}(\{0\}) \right)$ . We have a naturally defined differential operator  $d''_{\mathbb{P}^1} : \xi\mathcal{A}_Y^1 \rightarrow \xi\mathcal{A}_Y^1 \otimes p_1^{-1}\Omega_{\mathbb{P}^1}^{0,1}$ . We use the same symbol  $\xi\mathcal{A}_Y^1$  to denote the sheaf of  $C^\infty$ -sections  $f$  of  $\xi\mathcal{A}_Y^1$  such that  $d''_{\mathbb{P}^1}f = 0$ . The exterior derivative in the  $Y$ -direction naturally gives  $d : \mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}} \rightarrow \xi\mathcal{A}_Y^1$ .

A variation of twistor structure on  $Y$  is a locally free  $\mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$ -module  $V^\Delta$  with a differential operator  $\mathbb{D}^\Delta : V^\Delta \rightarrow V^\Delta \otimes \xi\mathcal{A}_Y^1$  satisfying the conditions (i)  $\mathbb{D}^\Delta(fs) = s \otimes df + f\mathbb{D}^\Delta(s)$ , where  $f$  and  $s$  are local sections of  $\mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$  and  $V^\Delta$  respectively, (ii)  $\mathbb{D}^\Delta \circ \mathbb{D}^\Delta = 0$ , where the induced operators  $V^\Delta \otimes \wedge^i \xi\mathcal{A}_Y^1 \rightarrow V^\Delta \otimes \wedge^{i+1} \xi\mathcal{A}_Y^1$  are denoted by the same symbol. A variation of twistor structure  $(V^\Delta, \mathbb{D}^\Delta)$  is called pure of weight  $n$  if the induced twistor structures  $V_{|\mathbb{P}^1 \times \{P\}}^\Delta$  are pure of weight  $n$  for any  $P \in Y$ . Note that the poles of  $\mathbb{D}^\Delta$  along  $\{0, \infty\} \times Y$  correspond to the Griffiths transversality condition.

The involution  $\sigma \times \text{id}$  on  $\mathbb{P}^1 \times Y$  is also denoted by  $\sigma$ . We have a natural identification of  $\mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$ -modules  $\sigma^*(\Omega^{0,1} \otimes p_2^{-1}\mathcal{O}_{\mathbb{P}^1}(\{\infty\})) \simeq \Omega^{1,0} \otimes p_2^{-1}\mathcal{O}_{\mathbb{P}^1}(\{0\})$  given by  $\sigma^*\omega \mapsto \overline{\sigma^*\omega}$

which induces  $\sigma^*\xi\mathcal{A}_Y^1 \simeq \xi\mathcal{A}_Y^1$ . For any variation of twistor structure  $(V^\Delta, \mathbb{D}^\Delta)$ , we have an induced variation of twistor structure  $\sigma^*(V^\Delta, \mathbb{D}^\Delta)$  where  $\sigma^*\mathbb{D}^\Delta$  is the induced operator  $\sigma^*V^\Delta \rightarrow \sigma^*V^\Delta \otimes \sigma^*\xi\mathcal{A}_Y^1 \simeq \sigma^*V^\Delta \otimes \xi\mathcal{A}_Y^1$ . When  $(V^\Delta, \mathbb{D}^\Delta)$  is pure of weight 0, a  $\mathbb{D}^\Delta$ -flat morphism  $\mathcal{S} : V^\Delta \otimes \sigma^*V^\Delta \rightarrow \mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$  is called a polarization if  $\mathcal{S}|_{\mathbb{P}^1 \times P}$  are polarizations of  $V|_{\mathbb{P}^1 \times P}$  for any  $P \in Y$ .

Let  $(E, \bar{\partial}_E, \theta, h)$  be a harmonic bundle on  $Y$ . We set  $\mathcal{E}^\Delta := p_2^{-1}E$  and  $\mathbb{D}^\Delta := \bar{\partial}_E + \lambda\theta^\dagger + \partial_{E,h} + \lambda^{-1}\theta$ . We have  $\mathcal{S}_h : \mathcal{E}^\Delta \otimes \sigma^*\mathcal{E}^\Delta \rightarrow \mathcal{C}_{\mathbb{P}^1 \times Y}^{\infty \text{ an}}$  induced by  $h$ . Then,  $(\mathcal{E}^\Delta, \mathbb{D}^\Delta)$  is naturally a variation of pure twistor structure with a polarization  $\mathcal{S}_h$ .

**Theorem 4.2** (Simpson). *The above correspondence gives an equivalence between harmonic bundles and polarized variations of pure twistor structure of weight 0.*

A variation of twistor structure  $(V^\Delta, \mathbb{D}^\Delta)$  is described as a patched object as follows. We regard  $\mathbb{P}^1$  as the gluing of  $\mathbb{C}_\lambda$  and  $\mathbb{C}_\mu$  by  $\lambda\mu = 1$ . We identify  $\xi\mathcal{A}_{Y|\mathbb{C}_\lambda \times Y}^1 \simeq p_2^{-1}(\Omega_Y^{1,0} \oplus \Omega_Y^{0,1})$  by multiplying  $\lambda$  on the  $(1, 0)$ -part. Then,  $\mathbb{D}|_{\mathbb{C}_\lambda \times Y}$  induces a family of flat  $\lambda$ -connections  $\mathbb{D}$  on  $V = V|_{\mathbb{C}_\lambda \times Y}$ , i.e., a differential operator satisfying the twisted Leibniz rule  $\mathbb{D}(fs) = s \otimes (\lambda\partial_Y + \bar{\partial}_Y)f + f\mathbb{D}(s)$ , and the integrability condition  $\mathbb{D} \circ \mathbb{D} = 0$ . Similarly,  $\mathbb{D}|_{\mathbb{C}_\mu \times Y}$  induces an operator  $\mathbb{D}^\dagger$  on  $V^\dagger = V|_{\mathbb{C}_\mu \times Y}$  satisfying  $\mathbb{D}^\dagger(fs) = s \otimes (\mu\bar{\partial}_Y + \partial_Y)f + f\mathbb{D}^\dagger(s)$ , and  $\mathbb{D}^\dagger \circ \mathbb{D}^\dagger = 0$ . We can regard  $\mathbb{D}^\Delta$  as the patched objects of  $(V, \mathbb{D})$  and  $(V^\dagger, \mathbb{D}^\dagger)$  by an isomorphism on  $\mathbb{C}^* \times Y$  compatible with the associated flat connections. Note that  $(V^\dagger, \mathbb{D}^\dagger)$  can be regarded as a holomorphic vector bundle with a family of flat  $\mu$ -connections on  $\mathbb{C}_\mu \times Y^\dagger$ , where  $Y^\dagger$  denote the complex manifold conjugate of  $Y$ .

## 5. Local structure of meromorphic flat bundles

Let  $X$  be a complex manifold with a hypersurface  $D$ . For simplicity,  $D$  will be assumed to be simply normal crossing. Let  $\mathcal{O}_X(*D)$  denote the sheaf of meromorphic functions on  $X$  whose poles are contained in  $D$ . A meromorphic bundle on  $(X, D)$  means a locally free  $\mathcal{O}_X(*D)$ -module. A meromorphic bundle on  $(X, D)$  with a flat connection is called a meromorphic flat bundle on  $(X, D)$ .

We review the basic theory on the structure of meromorphic flat bundles near their poles. It is a generalization of the well known theory in the curve case.

**Regular singular meromorphic flat bundle.** Let  $X$  be a complex manifold with a simply normal crossing hypersurface  $D$ . Let  $(\mathcal{V}, \nabla)$  be a meromorphic flat bundle on  $(X, D)$ . It is called regular singular if there exists a lattice  $V \subset \mathcal{V}$  for which  $\nabla$  is logarithmic, i.e.,  $\nabla(V) \subset V \otimes \Omega_X^1(\log D)$ . See [19] for basic facts on regular singular meromorphic flat bundles. We mention some well known properties. Suppose that  $X$  is equipped with a holomorphic coordinate system  $(z_1, \dots, z_n)$  such that  $D = \bigcup_{i=1}^\ell \{z_i = 0\}$ . We can take a local frame  $v$  of  $\mathcal{V}$  for which the connection form of  $\nabla$  is  $\sum_{j=1}^\ell A_j dz_j/z_j$ , where  $A_j$  are constant matrices. Let  $\mathcal{V}|_{\hat{O}}$  denote the formal completion  $\mathcal{V} \otimes_{\mathcal{O}_X} \mathbb{C}[[z_1, \dots, z_n]]$  at  $O = (0, \dots, 0)$ . Then, any  $\hat{v} \in \mathcal{V}|_{\hat{O}}$  with  $\nabla(\hat{v}) = 0$  is convergent, i.e., we have a section  $v$  of  $\mathcal{V}$  on a neighbourhood of  $O$  which induces  $\hat{v}$ . In a more general situation, the isomorphism classes of the regular singular meromorphic flat bundles on any  $(X, D)$  can be classified by the associated local systems on  $X \setminus D$ . In particular, the study on flat bundles on quasi-

projective manifolds is reduced to the study of regular singular meromorphic flat bundles on projective manifolds.

**Unramifiedly good elementary meromorphic flat bundle.** For any section of  $f \in \mathcal{O}_X(*D)$ , let  $L(f)$  denote the meromorphic flat bundle  $(\mathcal{O}_X(*D), d+df)$ . We regard that meromorphic flat bundles of the form  $\bigoplus_{f \in I} R_f \otimes L(f)$  are easy to understand, where  $I$  is a finite set of sections of  $\mathcal{O}_X(*D)$ , and  $R_f$  are regular singular meromorphic flat bundles. Such meromorphic flat bundles are called elementary. We can understand the local structure of a general meromorphic flat bundle by measuring how it is far from an elementary one. It is convenient to impose some technical condition on the index set  $I$ .

Let  $\mathcal{U}$  be a neighbourhood of  $O = (0, \dots, 0)$  in  $\mathbb{C}^n$ . Let  $(z_1, \dots, z_n)$  be the standard coordinate system of  $\mathbb{C}^n$ . Set  $D := \mathcal{U} \cap \bigcup_{i=1}^n \{z_i = 0\}$ . Let  $f$  be a section of  $\mathcal{O}_{\mathcal{U}}(*D)$ . Suppose that there exists  $\mathbf{m} = (m_i) \in \mathbb{Z}_{>0}^n$  such that (i)  $z^{\mathbf{m}} f = \prod z_i^{m_i} f$  is holomorphic, (ii) if  $\mathbf{m} \neq (0, \dots, 0)$ , we have  $(z^{\mathbf{m}} f)(O) \neq 0$ . Then, we set  $\text{ord}(f) := -\mathbf{m}$ . In general, such  $\mathbf{m}$  does not exist. For any holomorphic function  $f$ , we have  $\text{ord}(f) = (0, \dots, 0)$ . If  $\text{ord}(g)$  exists for  $g \in \mathcal{O}_{\mathcal{U}}(*D)$ , then  $\text{ord}(g + f) = \text{ord}(g)$  for any holomorphic function  $f$ . Hence, we can consider  $\text{ord}$  for sections of  $\mathcal{O}_{\mathcal{U}}(*D)/\mathcal{O}_{\mathcal{U}}$ .

A finite subset  $\mathcal{I} \subset \mathcal{O}_{\mathcal{U}}(*D)_O/\mathcal{O}_{X,O}$  is called a good set of irregular values, if the following conditions are satisfied: (i)  $\text{ord}(f)$  exists for any  $f \in \mathcal{I}$ , (ii)  $\text{ord}(f - g)$  exists for any  $f, g \in \mathcal{I}$ , (iii) the set  $\{\text{ord}(f - g) \mid f, g \in \mathcal{I}\} \subset \mathbb{Z}^n$  is totally ordered. Here, we use the partial order  $\leq$  on  $\mathbb{Z}^n$  given by  $\mathbf{m} \leq \mathbf{n} \stackrel{\text{def}}{\iff} m_i \leq n_i$  for any  $i$ .

An elementary meromorphic flat bundle  $\bigoplus_{f \in I} R_f \otimes L(f)$  on  $(\mathcal{U}, D)$  is called unramifiedly good, if the image of  $I \rightarrow \mathcal{O}_{\mathcal{U}}(*D)_O/\mathcal{O}_{\mathcal{U},O}$  is a good set of irregular values. We shall not so carefully distinguish  $I$  from its image in  $\mathcal{O}_{\mathcal{U}}(*D)_O/\mathcal{O}_{\mathcal{U},O}$ .

**Good meromorphic flat bundle.** Let  $X$  be a complex manifold with a simply normal crossing hypersurface  $D$ . Let  $(\mathcal{V}, \nabla)$  be a meromorphic flat bundle on  $(X, D)$ . The formal completion of  $(\mathcal{V}, \nabla)$  at  $P$  is denoted by  $(\mathcal{V}, \nabla)_{|\hat{P}}$ . We say that  $(\mathcal{V}, \nabla)$  is unramifiedly good at  $P$  if we have (i) a good set of irregular values  $\text{Irr}(\nabla, P) \subset \mathcal{O}_X(*D)_P/\mathcal{O}_{X,P}$ , (ii) an unramifiedly good elementary meromorphic flat bundle  $(\mathcal{V}^{(P)}, \nabla^{(P)}) = \bigoplus_{\alpha \in \text{Irr}(\nabla, P)} (\mathcal{V}_{\alpha}^{(P)}, \nabla_{\alpha}^{(P)})$  on a neighbourhood of  $P$ , and (iii) an isomorphism  $\hat{\psi}_P : (\mathcal{V}, \nabla)_{|\hat{P}} \simeq (\mathcal{V}^{(P)}, \nabla^{(P)})_{|\hat{P}}$ . We say that  $(\mathcal{V}, \nabla)$  is good at  $P$  if we have a ramified covering  $\varphi_P : (X'_P, D'_P) \rightarrow (X_P, D_P)$  of a neighbourhood  $X_P$  of  $P$  such that  $\varphi_P^*(\mathcal{V}, \nabla)$  is unramifiedly good at  $\varphi^{-1}(P)$ . We say that  $(\mathcal{V}, \nabla)$  is (unramifiedly) good on  $(X, D)$  if it is (unramifiedly) good at any  $P \in D$ .

The classical Hukuhara-Levelt-Turrittin theorem says that any meromorphic flat bundle on a curve is good. According to [58] (see also [72]), for any meromorphic flat bundle  $(\mathcal{V}, \nabla)$  on  $(X, D)$ , there exists a closed analytic subset  $Z \subset D$  which is nowhere dense, such that  $(\mathcal{V}, \nabla)_{|X \setminus Z}$  is good on  $(X \setminus Z, D \setminus Z)$ .

**Resolution of turning points.** In general, a meromorphic flat bundle  $(\mathcal{V}, \nabla)$  on  $(X, D)$  may have points called turning points [1], at which  $(\mathcal{V}, \nabla)$  is not good. An easy non-trivial example is given as follows. We have a meromorphic flat bundle  $(\mathcal{V}_0, \nabla_0)$  on  $(\mathbb{P}^1, 0)$  which is ramified at 0. Let  $\varphi$  be a rational map from  $\mathbb{C}^2$  to  $\mathbb{P}^1$  given by  $\varphi(z, w) = [z : w]$ . Then,  $(0, 0)$  is a turning point of  $\psi^*(\mathcal{V}_0, \nabla_0)$ .

Sabbah conjectured the existence of a projective birational morphism  $F : (X', D') \rightarrow (X, D)$  such that  $F^*(\mathcal{V}, \nabla)$  is good, called a resolution of turning points. In the algebraic

case, it was proved by Kedlaya [49, 50] and the author [67, 72].

**Theorem 5.1.** *If  $X$  and  $(\mathcal{V}, \nabla)$  are algebraic, there exists a resolution of turning points for  $(\mathcal{V}, \nabla)$ .*

As for the complex analytic case, Kedlaya (loc.cit) proved the existence of local resolutions.

**Theorem 5.2.** *Let  $P$  be any point of  $D$ . There exists a neighbourhood  $X_P$  such that a resolution of turning points of  $(\mathcal{V}, \nabla)|_{X_P}$  exists.*

These theorems are important in the study of meromorphic flat bundles, because resolutions of turning points play the role of resolutions of singularity in algebraic geometry.

**Asymptotic analysis.** The asymptotic analysis for meromorphic flat bundles on curves can be generalized to that for good meromorphic flat bundles. A systematic study was started by Majima [56], revisited by Sabbah [80] in a different framework, and later by the author [72].

Let  $\tilde{X}(D_i)$  be the oriented real blow up of  $X$  along  $D_i$ , and let  $\tilde{X}(D) \xrightarrow{\pi} X$  denote the fiber product of  $\tilde{X}(D_i)$  over  $X$ . A  $C^\infty$ -function  $f$  on an open subset  $U \subset \tilde{X}(D)$  is called holomorphic if its restriction to  $U \setminus \pi^{-1}(D)$  is holomorphic. In the one dimensional case, it is a holomorphic function with asymptotic expansion. Let  $\mathcal{O}_{\tilde{X}(D)}$  denote the sheaf of holomorphic functions on  $\tilde{X}(D)$ .

Let  $(\mathcal{V}, \nabla)$  be an unramifiedly good meromorphic flat bundle on  $(X, D)$ . We set  $(\tilde{\mathcal{V}}, \tilde{\nabla}) := \pi^{-1}(\mathcal{V}, \nabla) \otimes_{\pi^{-1}\mathcal{O}_X} \mathcal{O}_{\tilde{X}(D)}$ . For any  $P \in D$ , we have  $(\mathcal{V}^{(P)}, \nabla^{(P)})$ ,  $\text{Irr}(\nabla, P)$  and  $\hat{\psi}_P$  as above. Then, for any  $Q \in \pi^{-1}(P)$  there exist a neighbourhood  $U_Q \subset \tilde{X}(D)$  and an isomorphism  $\psi_{U_Q} : (\tilde{\mathcal{V}}, \tilde{\nabla})|_{U_Q} \simeq (\tilde{\mathcal{V}}^{(P)}, \tilde{\nabla}^{(P)})|_{U_Q}$  compatible with  $\pi^* \hat{\psi}_P$ .

The isomorphism  $\psi_{U_Q}$  is not unique. Instead, we have the canonically defined Stokes filtrations. For any  $Q \in \pi^{-1}(P)$ , we have the partial order  $\leq_Q$  on  $\text{Irr}(\nabla, P)$  given as follows:  $\mathfrak{a} \leq_Q \mathfrak{b} \stackrel{\text{def}}{\iff} -\text{Re}(\mathfrak{a} - \mathfrak{b}) \leq 0$  holds on any small neighbourhood of  $Q$ . (We permit  $-\infty$ .) We set  $\mathcal{F}_\mathfrak{a}^Q(\tilde{\mathcal{V}}_Q) = \bigoplus_{\mathfrak{b} \leq_Q \mathfrak{a}} \psi_{U_Q}^{-1}(\tilde{\mathcal{V}}_\mathfrak{b}^{(P)})_Q$ . Here  $(\cdot)_Q$  denotes the stalk of  $(\cdot)$  at  $Q$ . It is independent of the choice of  $U_Q$  and the decomposition as above. Thus, we obtain a filtration of  $\tilde{\mathcal{V}}_Q$  indexed by the partially ordered set  $(\text{Irr}(\nabla, P), \leq_Q)$ . If  $Q_1 \in \pi^{-1}(P)$  is sufficiently close to  $Q$ ,  $\mathcal{F}^Q$  induces a filtration on  $\tilde{\mathcal{V}}_{Q_1}$ , and we have the compatibility condition  $\mathcal{F}_\mathfrak{a}^{Q_1} = \mathcal{F}_\mathfrak{a}^Q + \mathcal{F}_{<\mathfrak{a}}^{Q_1}$ .

We set  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\mathcal{V}}_Q) := \mathcal{F}_\mathfrak{a}^Q(\tilde{\mathcal{V}}_Q) / \mathcal{F}_{<\mathfrak{a}}^Q(\tilde{\mathcal{V}}_Q)$ . By the above compatibility condition, we can glue  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\mathcal{V}}_Q)$  ( $Q \in \pi^{-1}(P)$ ), and obtain an  $\mathcal{O}_{\tilde{X}(D)}(*D)$ -module  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\mathcal{V}})$  on a neighbourhood of  $\pi^{-1}(P)$  with an induced flat connection  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\nabla})$ . By construction,  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\mathcal{V}}, \tilde{\nabla})$  is naturally isomorphic to  $(\tilde{\mathcal{V}}_\mathfrak{a}^{(P)}, \tilde{\nabla}_\mathfrak{a}^{(P)})$ . The push-forward  $\text{Gr}_\mathfrak{a}^{\mathcal{F}}(\mathcal{V}, \nabla) := \pi_* \text{Gr}_\mathfrak{a}^{\mathcal{F}}(\tilde{\mathcal{V}}, \tilde{\nabla})$  is naturally isomorphic to  $(\mathcal{V}_\mathfrak{a}^{(P)}, \nabla_\mathfrak{a}^{(P)})$ .

Although we can find  $(\mathcal{V}^{(P)}, \nabla^{(P)})$  directly from  $(\mathcal{V}, \nabla)|_{\hat{P}}$ , the above construction as the grading with respect to the Stokes filtrations is useful when we are interested in some induced structure on  $X \setminus D$ .

**Riemann-Hilbert-Birkhoff correspondence.** We can classify meromorphic flat bundles on curves by local systems with Stokes structure, according to P. Deligne, B. Malgrange [57] and Y. Sibuya [93]. This is naturally generalized for good meromorphic flat bundles as in [72] (see a survey [71]). See also [80, 84].

We continue to use the above notation. Let  $\mathcal{L}'$  be the local system on  $X \setminus D$  associated to  $(\mathcal{V}, \nabla)|_{X \setminus D}$ . It can be extended to a local system  $\mathcal{L}$  on  $\tilde{X}(D)$ . We also have the local system  $\mathcal{L}^{(P)} = \bigoplus \mathcal{L}_\alpha^{(P)}$  associated to  $(\mathcal{V}^{(P)}, \nabla^{(P)})$ . The isomorphism  $(\tilde{\mathcal{V}}, \tilde{\nabla})|_{U_Q} \simeq (\tilde{\mathcal{V}}^{(P)}, \tilde{\nabla}^{(P)})|_{U_Q}$  induces an isomorphism  $\mathcal{L}|_{U_Q} \simeq \mathcal{L}^{(P)}|_{U_Q}$ . We obtain a filtration  $\mathcal{F}^Q$  on the stalk  $\mathcal{L}_Q$  indexed by  $(\text{Irr}(\nabla, \pi(Q)), \leq_Q)$  as before. The family of filtrations  $\{\mathcal{F}^Q\}$  satisfies the following compatibility condition. If  $P' \in D$  is sufficiently close to  $P$ , we have the map  $\text{Irr}(\nabla, P) \rightarrow \mathcal{O}_X(*D)_{P'}/\mathcal{O}_{X,P'}$ , and the image equals  $\text{Irr}(\nabla, P')$ . If  $Q'$  is sufficiently close to  $Q$ , the map  $(\text{Irr}(\nabla, \pi(Q)), \leq_Q) \rightarrow (\text{Irr}(\nabla, \pi(Q')), \leq_{Q'})$  is order preserving. So,  $\mathcal{F}^Q$  of  $\mathcal{L}_Q$  indexed by  $(\text{Irr}(\nabla, \pi(Q)), \leq_Q)$  induces a filtration  $\mathcal{F}^{Q'}$  of  $\mathcal{L}_{Q'}$  indexed by  $(\text{Irr}(\nabla, \pi(Q')), \leq_{Q'})$ . (See [71] for the procedure, for example.) It turns out that it is equal to  $\mathcal{F}^{Q'}$ . A family of the filtrations satisfying the compatibility condition is called a Stokes structure of  $\mathcal{L}$ . Unramifiedly good meromorphic flat bundles are classified by local systems with a Stokes structure. This is the Riemann-Hilbert-Birkhoff correspondence in [72].

**Deformation.** A good system of irregular values  $\mathcal{I}$  on  $(X, D)$  is defined to be a family of good set of irregular values  $(\mathcal{I}_P \mid P \in D)$  such that for any  $P$  there exists a neighbourhood  $U$  and  $\mathcal{I}_{P'}$  ( $P' \in U \cap D$ ) are the image of the natural map  $\mathcal{I}_P \rightarrow \mathcal{O}_X(*D)_{P'}/\mathcal{O}_{X,P'}$ . We say that  $(\mathcal{V}, \nabla)$  is  $\mathcal{I}$ -unramifiedly good if  $\text{Irr}(\nabla, P) \subset \mathcal{I}_P$  for any  $P \in D$ . A Stokes structure  $\{\mathcal{F}^Q \mid Q \in \pi^{-1}(D)\}$  of a local system  $\mathcal{L}$  is called  $\mathcal{I}$ -Stokes structure if the index sets of  $\mathcal{F}^Q$  are contained in  $\mathcal{I}_{\pi(Q)}$ . Then,  $\mathcal{I}$ -unramifiedly good meromorphic flat bundles are classified by local systems with  $\mathcal{I}$ -Stokes structure.

We have iso-Stokes deformations associated to a variation of the index set in the higher dimensional case. For simplicity, let  $(X_1, D_1) := (X, D) \times Y$ , where  $Y$  is a simply connected complex manifold. Let  $\mathcal{I}$  be a good system of irregular values on  $(X_1, D_1)$ . For any  $y \in Y$ , we set  $(X_1^y, D_1^y) := (X, D) \times \{y\}$ . By the restriction, we obtain a good system of irregular values  $\mathcal{I}^y$  on  $(X_1^y, D_1^y)$ . We have the categorical equivalence between local systems on  $\tilde{X}_1(D_1)$  and  $\tilde{X}_1^y(D_1^y)$ . Let  $\mathcal{L}$  be a local system on  $\tilde{X}_1(D_1)$ , and  $\mathcal{L}^y$  its restriction to  $\tilde{X}_1^y(D_1^y)$ .

**Theorem 5.3.** *The restriction induces an equivalence between  $\mathcal{I}$ -Stokes structures on  $\mathcal{L}$  and  $\mathcal{I}^y$ -Stokes structures on  $\mathcal{L}^y$ . Hence, the restriction induces an equivalence between  $\mathcal{I}$ -unramifiedly good meromorphic flat bundles on  $(X_1, D_1)$  and  $\mathcal{I}^y$ -unramifiedly good meromorphic flat bundles on  $(X_1^y, D_1^y)$ .*

Let us look at a special case. Let  $T > 0$ . Let  $\mathcal{I}$  be a good system of irregular values on  $(X, D)$ . We set  $\mathcal{I}_P^{(T)} := \{T\alpha \mid \alpha \in \mathcal{I}_P\}$ . Then,  $\mathcal{I}^{(T)} = (\mathcal{I}_P^{(T)} \mid P \in D)$  is a good system of irregular values. By considering the case where  $Y \subset \mathbb{C}$  is a neighbourhood of the segment connecting 1 and  $T$ , we obtain a deformation of any  $\mathcal{I}$ -unramifiedly good meromorphic flat bundle  $(\mathcal{V}, \nabla)$  to a  $\mathcal{I}^{(T)}$ -unramifiedly good meromorphic flat bundle  $(\mathcal{V}^{(T)}, \nabla^{(T)})$ . We can also obtain the deformation directly as follows. For each  $Q \in \pi^{-1}(D)$  we set  $\mathcal{F}_{T\alpha}^{(T)Q} := \mathcal{F}_\alpha^Q$ . Then, the family of filtrations gives a Stokes structure on  $\mathcal{L}$ . Thus, we obtain the corresponding deformation of an unramifiedly good meromorphic flat bundle  $(\mathcal{V}, \nabla)$  to  $(\mathcal{V}^{(T)}, \nabla^{(T)})$ .

## 6. Filtered bundles

Let  $X$  be a Riemann surface, and let  $(E, \bar{\partial}_E)$  be a holomorphic vector bundle on  $X$ . Let  $P \in X$ . A parabolic structure of  $E$  at  $P$  is a filtration  $E|_P = F_1(E|_P) \supset \cdots \supset F_\ell(E|_P)$  with a sequence  $0 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_\ell < 1$ . This simple notion turns out to be quite significant in various aspects of the study of vector bundles. It first appeared in the study of V. Mehta and C. S. Seshadri [63]. They defined the stability condition by introducing the degree for parabolic bundles. They found that the stability condition fits with the geometric invariant theory so that they obtained the moduli spaces. (See [59, 105] for the generalization of parabolic structure and the construction of the moduli spaces in the higher dimensional case.) They also generalized the theorem of Narasimhan and Seshadri: the correspondence between unitary flat bundles and polystable parabolic bundles with slope 0.

For the Kobayashi-Hitchin correspondence, it is more convenient to consider the essentially equivalent notion of filtered bundle introduced in the curve case [94, 95], and considered in the higher dimensional case [65]. We consider the growth order along each irreducible component of the hypersurface, and we impose some compatibility condition at the intersection of the components of the hypersurfaces. We explain it by following [39]. See also [8] and [35].

Let  $X$  be a complex manifold, and  $D$  a simply normal crossing hypersurface with the irreducible decomposition  $D = \bigcup_{i \in \Lambda} D_i$ . A filtered bundle  $\mathcal{P}_* \mathcal{V}$  consists of a locally free  $\mathcal{O}_X(*D)$ -module  $\mathcal{V}$  together with an increasing sequence of locally free  $\mathcal{O}_X$ -submodules  $\mathcal{P}_a \mathcal{V} \subset \mathcal{V}$  ( $a \in \mathbb{R}^\Lambda$ ) satisfying (i)  $\mathcal{P}_a \mathcal{V}(*D) = \mathcal{V}$ , (ii)  $\mathcal{P}_a \mathcal{V} \subset \mathcal{P}_b \mathcal{V}$  if  $a_i \leq b_i$  for any  $i \in \Lambda$ , (iii) for any  $\mathbf{n} \in \mathbb{Z}^\Lambda$  we have  $\mathcal{P}_{\mathbf{a}+\mathbf{n}} \mathcal{V} = \mathcal{P}_a \mathcal{V} \otimes \mathcal{O}(\sum n_i D_i)$ , (iv)  $\mathcal{P}_* \mathcal{V}$  is locally abelian, that is, around any  $P \in D$ , setting  $\Lambda_P := \{i \in \Lambda \mid P \in D_i\}$ , there exists a local frame  $v = (v_j \mid j = 1, \dots, \text{rank } \mathcal{V})$  of  $\mathcal{V}$  around  $P$  with tuples of numbers  $\mathbf{a}(v_j) \in \mathbb{R}^{\Lambda_P}$  such that  $\mathcal{P}_b \mathcal{V} = \bigoplus_{j=1}^{\text{rank } \mathcal{V}} \mathcal{O}_X(\mathbf{n}(v_j, \mathbf{b}) \cdot D) v_j$  for any  $\mathbf{b} \in \mathbb{R}^\Lambda$ . Here, put  $n_i(v_j, \mathbf{b}) := \max\{n \in \mathbb{Z} \mid n + a_i(v_j) \leq b_i\}$  for  $i \in \Lambda_P$ , and set  $\mathbf{n}(v_j, \mathbf{b}) \cdot D := \sum_{i \in \Lambda_P} n_i(v_j, \mathbf{b}) D_i$ .

Among various operations for filtered bundles, we explain the descent in a local situation. Suppose that  $X$  is an open subset in  $\mathbb{C}^n$  and  $D = X \cap \bigcup_{j=1}^\ell \{z_j = 0\}$ . Let  $\varphi : \mathbb{C}^n \rightarrow \mathbb{C}^n$  be given by  $\varphi(\zeta_1, \dots, \zeta_n) = (\zeta_1^{k_1}, \dots, \zeta_\ell^{k_\ell}, \zeta_{\ell+1}, \dots, \zeta_n)$ . We set  $X' := \varphi^{-1}(X)$  and  $D' := \varphi^{-1}(D)$ . Let  $\mathcal{P}_* \mathcal{V}'$  be a filtered bundle on  $(X', D')$ . Then we have a natural filtered bundle  $\mathcal{P}_* \varphi_* \mathcal{V}'$  over  $\varphi_* \mathcal{V}'$  given by  $\mathcal{P}_a \varphi_* \mathcal{V}' := \varphi_* \mathcal{P}_{\mathbf{k}a} \mathcal{V}'$ , where  $\mathbf{k}a = (k_i a_i)$ . Let  $G$  denote the Galois group of  $\varphi$ . When  $\mathcal{P}_* \mathcal{V}'$  is  $G$ -equivariant, the descent of  $\mathcal{P}_* \mathcal{V}'$  is defined to be the invariant part of  $\mathcal{P}_* \varphi_* \mathcal{V}'$  with respect to the induced  $G$ -action.

**Good filtered  $\lambda$ -flat bundles.** Let  $\mathcal{V}$  be a meromorphic bundle on  $(X, D)$ . We introduce a compatibility condition for a filtered bundle  $\mathcal{P}_* \mathcal{V}$  and a flat connection over  $\mathcal{V}$ , or more generally a flat  $\lambda$ -connection ( $\lambda \in \mathbb{C}$ ). Here, a flat  $\lambda$ -connection is a  $\mathbb{C}$ -linear map  $\mathbb{D}^\lambda : \mathcal{V} \rightarrow \mathcal{V} \otimes \Omega_X^1$  satisfying (i)  $\mathbb{D}^\lambda(fs) = f\mathbb{D}^\lambda s + \lambda df \otimes s$  for any local sections  $f \in \mathcal{O}_X(*D)$  and  $s \in \mathcal{V}$ , (ii)  $\mathbb{D}^\lambda \circ \mathbb{D}^\lambda = 0$ . If  $\lambda \neq 0$ , it is equivalent to a flat connection. If  $\lambda = 0$ , it is a Higgs field. When we consider a variation of twistor structure, a family of flat  $\lambda$ -connections naturally appears (see §4).

Let  $\mathcal{P}_* \mathcal{V}$  be a filtered bundle over a meromorphic bundle  $\mathcal{V}$  on  $(X, D)$ . Let  $\mathbb{D}^\lambda$  be a flat  $\lambda$ -connection of  $\mathcal{V}$ . We say that  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^\lambda)$  is regular if  $\mathbb{D}^\lambda$  is logarithmic with respect to each  $\mathcal{P}_a \mathcal{V}$ . We say that  $\mathbb{D}^\lambda$  is unramifiedly good if the following holds: For any  $P \in D$ , we have a good set of irregular values  $\mathcal{I}_P$  at  $P$  and a decomposition  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^\lambda)|_{\hat{P}} =$

$\bigoplus_{\mathfrak{a} \in \mathcal{I}_P} (\mathcal{P}_* \widehat{\mathcal{V}}_{\mathfrak{a}}, \widehat{\mathbb{D}}_{\mathfrak{a}}^{\lambda})$  such that  $(\mathcal{P}_* \widehat{\mathcal{V}}_{\mathfrak{a}}, \widehat{\mathbb{D}}_{\mathfrak{a}}^{\lambda} - d\tilde{\mathfrak{a}})$  are regular. We say that  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^{\lambda})$  is good if it is the descent of an unramifiedly good filtered  $\lambda$ -flat bundle on a small neighbourhood of each  $P \in D$ .

**Stability.** The characteristic classes of filtered bundles  $\mathcal{P}_* \mathcal{V}$  on  $(X, D)$  were systematically studied in [39]. The first and the second Chern characters in [64] are particularly important for our purpose. For any  $\mathcal{V}' \subset \mathcal{V}$ , we obtain a sequence of  $\mathcal{O}_X$ -modules  $\mathcal{P}_{\mathfrak{a}} \mathcal{V}' := \mathcal{P}_{\mathfrak{a}} \mathcal{V} \cap \mathcal{V}'$ . We can define  $c_1(\mathcal{P}_* \mathcal{V}')$  similarly although  $\mathcal{P}_* \mathcal{V}'$  is not necessarily a filtered bundle in the above sense.

Suppose that  $X$  is irreducible projective with an ample line bundle  $L$ . Set  $n = \dim X$ . Let  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^{\lambda})$  be a good filtered  $\lambda$ -flat bundle on  $(X, D)$ . For any  $\mathcal{V}' \subset \mathcal{V}$  such that  $\mathbb{D}^{\lambda} \mathcal{V}' \subset \mathcal{V}' \otimes \Omega_X^1$ , put  $\mu_L(\mathcal{P}_* \mathcal{V}') := (\text{rank } \mathcal{V}')^{-1} \int_X c_1(\mathcal{P}_* \mathcal{V}') c_1(L)^{n-1}$ . Then, the  $\mu_L$ - (poly, semi)stability condition for  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^{\lambda})$  is defined in the standard manner. We have the Bogomolov-Gieseker inequality for  $\mu_L$ -stable regular filtered  $\lambda$ -flat bundles [64, 66]. (See [76] for an interesting application.) It can be generalized for  $\mu_L$ -stable good filtered  $\lambda$ -flat bundles. (See [72] for the case  $\lambda = 1$ .)

### 7. Wild harmonic bundles

We will use harmonic bundles on  $X \setminus D$  to study meromorphic flat bundles on  $(X, D)$ . We introduce some conditions on the behaviour of a harmonic bundle  $(E, \bar{\partial}_E, \theta, h)$  around  $D$  which are intended to correspond to “meromorphic” and “regular singular”. The underlying Higgs bundle induces an  $\mathcal{O}_{\Omega^1(X \setminus D)}$ -module. The support  $\Sigma_{\theta}$  of the  $\mathcal{O}_{\Omega^1(X \setminus D)}$ -module is called the spectral variety.

**Definition 7.1.**  $(E, \bar{\partial}_E, \theta, h)$  is called *wild* on  $(X, D)$  if there exists a large integer  $N_P$  for each  $P \in X$  such that the closure of  $\Sigma_{\theta}$  in  $\Omega_X^1(\log D) \otimes \mathcal{O}(N_P D)$  is proper over  $X$  around  $P$ . It is called *tame* on  $(X, D)$  if the closure of  $\Sigma_{\theta}$  in the logarithmic cotangent bundle  $\Omega_X^1(\log D)$  is proper over  $X$ . (Tameness is a stronger condition.)

**(Unramifiedly) Good wild harmonic bundles.** Because general wild harmonic bundles are not easy to study directly, we introduce some more conditions. Let  $X$  be any complex manifold with a normal crossing hypersurface  $D$ . A wild harmonic bundle  $(E, \bar{\partial}_E, \theta, h)$  on  $(X, D)$  is called *unramifiedly good* at  $P$  if the following holds on a holomorphic coordinate neighbourhood  $(U_P, z_1, \dots, z_n)$  of  $P$  such that  $D \cap U_P = \bigcup_{i=1}^{\ell} \{z_i = 0\}$ :

- There exist a good set of irregular values  $\mathcal{I}$  and a decomposition on  $U_P \setminus D$

$$(E, \bar{\partial}_E, \theta) = \bigoplus_{\mathfrak{a} \in \mathcal{I}} (E_{\mathfrak{a}}, \bar{\partial}_{E_{\mathfrak{a}}}, \theta_{\mathfrak{a}}), \tag{7.1}$$

such that for the expression  $\theta_{\mathfrak{a}} - d\tilde{\mathfrak{a}} \text{id}_{E_{\mathfrak{a}}} = \sum_{i=1}^{\ell} f_{\mathfrak{a},i} dz_i/z_i + \sum_{i=\ell+1}^n g_{\mathfrak{a},i} dz_i$ , the coefficients of the polynomials  $\det(t \text{id}_{E_{\mathfrak{a}}} - f_{\mathfrak{a},i})$  and  $\det(t \text{id}_{E_{\mathfrak{a}}} - g_{\mathfrak{a},i})$  in  $t$  are holomorphic on  $X$ . Moreover, the coefficients of  $\det(t \text{id}_{E_{\mathfrak{a}}} - f_{\mathfrak{a},i})|_{z_i=0}$  are constant. Here  $\tilde{\mathfrak{a}} \in \mathcal{O}_X(*D)$  is any lift of  $\mathfrak{a}$ .

A wild harmonic bundle is called good at  $P$ , if there exists a ramified covering

$$\varphi : (U'_P, \zeta_1, \dots, \zeta_n) \longrightarrow (U_P, z_1, \dots, z_n); \varphi(z_1, \dots, z_n) = (z_1^k, \dots, z_{\ell}^k, z_{\ell+1}, \dots, z_n)$$

such that  $\varphi^{-1}(E, \bar{\partial}_E, \theta, h)$  is unramifiedly good at  $\varphi^{-1}(P)$ . A wild harmonic bundle is called (unramifiedly) good on  $(X, D)$  if it is (unramifiedly) good at any  $P \in D$ . The following proposition ensures that it is enough to study (unramifiedly) good wild harmonic bundles. (It is adopted as the definition of wildness in [72].)

**Proposition 7.2** ([73]). *Let  $X$  be a complex manifold with a normal crossing hypersurface  $H$ . There exists a proper birational morphism  $\varphi : X' \rightarrow X$  such that (i)  $H' = \varphi^{-1}(H)$  is simply normal crossing, (ii) it induces an isomorphism  $X' \setminus H' \simeq X \setminus H$ , (ii)  $\varphi^*(E, \bar{\partial}_E, \theta, h)$  is good on  $(X', H')$ .*

Unramifiedly good wild harmonic bundles on  $(X, D)$  are a priori given only on  $X \setminus D$ , not on  $X$ . We need to extend the associated variation of twistor structure on  $X \setminus D$  to some meromorphic object on  $X$ , as explained in §7.1. We show in §7.2 that the meromorphic object has a nice property along  $D$  described in terms of twistor structure. In §7.3, we explain the Kobayashi-Hitchin correspondence for good wild harmonic bundles and the characterization of semisimplicity of algebraic meromorphic flat bundles.

**7.1. Local properties I: Prolongation.**

**Prolongation of the  $\lambda$ -flat bundle.** Let us consider the case where  $X$  is a neighbourhood of  $O = (0, \dots, 0)$  in  $\mathbb{C}^n$  and  $D = \bigcup_{i=1}^\ell \{z_i = 0\}$ . Let  $(E, \bar{\partial}_E, \theta, h)$  be an unramifiedly good wild harmonic bundle on  $(X, D)$  with the decomposition (7.1). For any complex number  $\lambda$ , we have the holomorphic bundle  $\mathcal{E}^\lambda = (E, \bar{\partial}_E + \lambda\theta^\dagger)$  with the flat  $\lambda$ -connection  $\mathbb{D}^\lambda = \bar{\partial}_E + \lambda\theta^\dagger + \lambda\partial_E + \theta$  on  $X \setminus D$ . For any  $\mathbf{a} \in \mathbb{R}^\ell$  and any open subset  $U \subset X$ , let  $\mathcal{P}_\mathbf{a}\mathcal{E}^\lambda(U)$  denote the space of holomorphic sections  $f$  of  $\mathcal{E}^\lambda$  on  $U \setminus D$  such that  $|f|_h = O\left(\prod_{i=1}^\ell |z_i|^{-a_i - \epsilon}\right)$  ( $\forall \epsilon > 0$ ) holds locally around any  $P \in D \cap U$ . Thus, we obtain an  $\mathcal{O}_X$ -module  $\mathcal{P}_\mathbf{a}\mathcal{E}^\lambda$  for each  $\mathbf{a} \in \mathbb{R}^\ell$ . We also obtain an  $\mathcal{O}_X(*D)$ -module  $\mathcal{P}\mathcal{E}^\lambda = \varinjlim \mathcal{P}_\mathbf{a}\mathcal{E}^\lambda$ . We have the following.

**Theorem 7.3** ([65, 72]).  *$\mathcal{P}_*\mathcal{E}^\lambda = (\mathcal{P}_\mathbf{a}\mathcal{E}^\lambda \mid \mathbf{a} \in \mathbb{R}^\ell)$  is a filtered bundle on  $(X, D)$  over a meromorphic bundle  $\mathcal{P}\mathcal{E}^\lambda$ .*

This is technically an important step. We began with an asymptotic orthogonality of the decomposition and an estimate of the Higgs field, called the wild version of Simpson’s main estimate. Because the curvature of the Chern connection of  $(\mathcal{E}^\lambda, h)$  equals  $-(1 + |\lambda|^2)[\theta, \theta^\dagger]$ , we obtain that the curvature is bounded with respect to  $h$  and the Poincaré like metric of  $X \setminus D$ . Such holomorphic bundles with a Hermitian metric are called acceptable bundles. After the study by M. Cornalba, P. Griffiths [17], Simpson [94, 95] and the author [65, 72], it turns out that any acceptable bundle can be prolonged to a filtered bundle by the above procedure.

The flat  $\lambda$ -connection  $\mathbb{D}^\lambda$  of  $\mathcal{E}^\lambda$  can also be prolonged on  $\mathcal{P}\mathcal{E}^\lambda$ . Moreover, it satisfies the compatibility conditions with the filtered bundle  $\mathcal{P}_*\mathcal{E}^\lambda$ .

**Theorem 7.4** ([72]).  *$(\mathcal{P}_*\mathcal{E}^\lambda, \mathbb{D}^\lambda)$  is an unramifiedly good filtered  $\lambda$ -flat bundle. We have  $\text{Irr}(\mathbb{D}^\lambda, O) = \{(1 + |\lambda|^2)\mathbf{a} \mid \mathbf{a} \in \mathcal{I}\}$ .*

**Prolongation of variation of twistor structure.** As in §4, we have the variation of pure twistor structure  $(\mathcal{E}^\Delta, \mathbb{D}^\Delta)$  with the polarization  $\mathcal{S}_h$  associated to  $(E, \bar{\partial}_E, \theta, h)$ . We have the holomorphic vector bundle  $\mathcal{E}$  with the family of flat  $\lambda$ -connections  $\mathbb{D}$  on  $\mathbb{C}_\lambda \times (X \setminus D)$

obtained as the restriction of  $(\mathcal{E}^\Delta, \mathbb{D}^\Delta)$ . We also have  $(\mathcal{E}^\dagger, \mathbb{D}^\dagger)$  on  $\mathbb{C}_\mu \times (X^\dagger \setminus D^\dagger)$ . It is important to obtain their meromorphic extensions on  $\mathbb{C}_\lambda \times X$  and  $\mathbb{C}_\mu \times X^\dagger$ .

The family  $\mathcal{P}\mathcal{E}^\lambda$  ( $\lambda \in \mathbb{C}$ ) is not appropriate, because  $(1 + |\lambda|^2)^{\mathfrak{a}}$  in Theorem 7.4 is not holomorphic with respect to  $\lambda$  unless  $\mathfrak{a} = 0$ . We have to deform the family. If  $\lambda \neq 0$ , a flat  $\lambda$ -connection is equivalent to a flat connection in an obvious way. We can apply the deformation procedure explained in §5 with  $T(\lambda) = (1 + |\lambda|^2)^{-1}$ , and obtain  $(\mathcal{Q}\mathcal{E}^\lambda, \mathbb{D}^\lambda) := (\mathcal{P}\mathcal{E}^\lambda, \mathbb{D}^\lambda)^{(T(\lambda))}$ . If  $\lambda = 0$ , we set  $(\mathcal{Q}\mathcal{E}^0, \mathbb{D}^0) := (\mathcal{P}\mathcal{E}^0, \mathbb{D}^0)$ .

**Theorem 7.5** ([65, 72]). *There exists a unique meromorphic bundle  $\mathcal{Q}\mathcal{E}$  on  $\mathbb{C}_\lambda \times (X, D)$  with a family of flat  $\lambda$ -connections  $\mathbb{D}$  such that  $(\mathcal{Q}\mathcal{E}, \mathbb{D})|_{\{\lambda\} \times X} = (\mathcal{Q}\mathcal{E}^\lambda, \mathbb{D}^\lambda)$  for any  $\lambda \in \mathbb{C}$ .*

We have a similar meromorphic extension  $\mathcal{Q}\mathcal{E}^\dagger$  of  $\mathcal{E}^\dagger$  on  $\mathbb{C}_\mu \times (X^\dagger, D^\dagger)$ . Although  $\mathcal{Q}\mathcal{E}$  and  $\mathcal{Q}\mathcal{E}^\dagger$  do not glue as a sheaf on  $\mathbb{P}^1 \times X$ , we can prove that for any  $\lambda = \mu^{-1} \neq 0$ ,  $(\mathcal{Q}\mathcal{E}, \mathbb{D})|_{\lambda \times X}$  and  $(\mathcal{Q}\mathcal{E}^\dagger, \mathbb{D}^\dagger)|_{\mu \times X^\dagger}$  correspond to the same local system with the Stokes structure. Indeed, the local systems underlying  $(\mathcal{E}, \mathbb{D})|_{\lambda \times (X \setminus D)}$  and  $(\mathcal{E}^\dagger, \mathbb{D}^\dagger)|_{\mu \times (X^\dagger \setminus D^\dagger)}$  are the same by construction. It turns out that the Stokes structures of  $\mathcal{Q}\mathcal{E}$  and  $\mathcal{Q}\mathcal{E}^\dagger$  are characterized by the growth order with respect to  $h$ , and hence the Stokes structures on the local system are the same. In this sense, they give a meromorphic extension of  $(\mathcal{E}^\Delta, \mathbb{D}^\Delta)$ . The polarization  $\mathcal{S}_h$  can also be extended in an appropriate sense.

### 7.2. Local properties II: Reductions.

**Reduction to tame harmonic bundles.** We continue to use the notation in §7.1. We explain a fundamental sequence of reductions of unramifiedly good wild harmonic bundles, given in terms of twistor structure. It shows the nice properties of our meromorphic object. We shall shrink  $X$  without mentioning it.

We begin with the reduction with respect to the Stokes structure. It turns out that we have a decomposition  $(\mathcal{Q}\mathcal{E}, \mathbb{D})|_{\widehat{\mathbb{C}_\lambda \times \mathcal{O}}} = \bigoplus_{\mathfrak{a} \in \text{Irr}(\theta)} (\mathcal{Q}\widehat{\mathcal{E}}_\mathfrak{a}, \widehat{\mathbb{D}}_\mathfrak{a})$  such that  $\mathcal{Q}\widehat{\mathcal{E}}_\mathfrak{a}$  locally has a lattice for which  $\widehat{\mathbb{D}}_\mathfrak{a} - d\mathfrak{a}$  is logarithmic. Here,  $|\widehat{\mathbb{C}_\lambda \times \mathcal{O}}$  means the formal completion along  $\mathbb{C}_\lambda \times \mathcal{O}$ . It turns out that  $(\mathcal{Q}\widehat{\mathcal{E}}_\mathfrak{a}, \widehat{\mathbb{D}}_\mathfrak{a})$  is convergent, i.e., there exists  $(\mathcal{Q}\mathcal{E}_\mathfrak{a}, \mathbb{D}_\mathfrak{a})$  on  $\mathbb{C}_\lambda \times X$  with  $(\mathcal{Q}\widehat{\mathcal{E}}_\mathfrak{a}, \widehat{\mathbb{D}}_\mathfrak{a}) \simeq (\mathcal{Q}\mathcal{E}_\mathfrak{a}, \mathbb{D}_\mathfrak{a})|_{\widehat{\mathbb{C}_\lambda \times \{\mathcal{O}\}}}$ . Indeed, we can canonically construct it as the grading with respect to the Stokes structure. (See §5 for a related construction.) We also have the formal decomposition  $(\mathcal{Q}\mathcal{E}^\dagger, \mathbb{D}^\dagger)|_{\widehat{\mathbb{C}_\mu \times \mathcal{O}}} = \bigoplus_{\bar{\mathfrak{a}} \in \text{Irr}(\theta^\dagger)} (\mathcal{Q}\widehat{\mathcal{E}}^\dagger_{\bar{\mathfrak{a}}}, \widehat{\mathbb{D}}^\dagger_{\bar{\mathfrak{a}}})$ , and we can canonically construct  $(\mathcal{Q}\mathcal{E}^\dagger_{\bar{\mathfrak{a}}}, \mathbb{D}^\dagger_{\bar{\mathfrak{a}}})$  on  $\mathbb{C}_\mu \times X^\dagger$  with an isomorphism  $(\mathcal{Q}\widehat{\mathcal{E}}^\dagger_{\bar{\mathfrak{a}}}, \widehat{\mathbb{D}}^\dagger_{\bar{\mathfrak{a}}})|_{\widehat{\mathbb{C}_\mu \times \mathcal{O}}} \simeq (\mathcal{Q}\mathcal{E}^\dagger_{\bar{\mathfrak{a}}}, \mathbb{D}^\dagger_{\bar{\mathfrak{a}}})$  as the grading of the Stokes structure. Because the Stokes structures of  $(\mathcal{Q}\mathcal{E}, \mathbb{D})$  and  $(\mathcal{Q}\mathcal{E}^\dagger, \mathbb{D}^\dagger)$  are essentially the same as mentioned, we have a natural identification of  $(\mathcal{Q}\mathcal{E}_\mathfrak{a}, \mathbb{D}_\mathfrak{a})|_{\mathbb{C}_\lambda^* \times (X \setminus D)}$  and  $(\mathcal{Q}\mathcal{E}^\dagger_{\bar{\mathfrak{a}}}, \mathbb{D}^\dagger_{\bar{\mathfrak{a}}})|_{\mathbb{C}_\mu^* \times (X^\dagger \setminus D^\dagger)}$ , and obtain a variation of twistor structure  $(\mathcal{E}_\mathfrak{a}^\Delta, \mathbb{D}_\mathfrak{a}^\Delta)$ , which is also equipped with a pairing  $\mathcal{S}_\mathfrak{a}$  induced by  $\mathcal{S}_h$ .

**Theorem 7.6** ([72]).  *$(\mathcal{E}_\mathfrak{a}^\Delta, \mathbb{D}_\mathfrak{a}^\Delta, \mathcal{S}_\mathfrak{a})$  is a polarized variation of twistor structure.*

Because the specialization of  $(\mathcal{E}_\mathfrak{a}^\Delta, \mathbb{D}_\mathfrak{a}^\Delta)$  to  $\{0\} \times (X \setminus D)$  is the Higgs bundle  $(E_\mathfrak{a}, \bar{\partial}_{E_\mathfrak{a}}, \theta_\mathfrak{a})$  in (7.1), the theorem implies that  $(E_\mathfrak{a}, \bar{\partial}_{E_\mathfrak{a}}, \theta_\mathfrak{a})$  has a naturally induced pluri-harmonic metric  $h_\mathfrak{a}$ . Note that  $(E_\mathfrak{a}, \bar{\partial}_{E_\mathfrak{a}}, \theta_\mathfrak{a} - d\mathfrak{a}, h_\mathfrak{a})$  is tame. Once we understand tame harmonic bundles, we can deduce much important information about the original harmonic bundle  $(E, \bar{\partial}_E, \theta, h)$ .

**Reduction to polarized mixed twistor structures.** We explain the reduction of tame harmonic bundles to polarized mixed twistor structures in [65], inspired by [95] and [99]. (See [68] and [69].) Suppose that  $(E, \bar{\partial}_E, \theta, h)$  is tame. Let  $\mathbf{a} \in \mathbb{R}^\ell$ . We define  $\text{Gr}_\mathbf{a}^\mathbb{Q}(\mathcal{Q}\mathcal{E}^\lambda, O)$  as the cokernel of the natural map  $\bigoplus_{b \leq \mathbf{a}} \mathcal{Q}_b \mathcal{E}^\lambda|_O \rightarrow \mathcal{Q}_\mathbf{a} \mathcal{E}^\lambda|_O$ . The endomorphisms  $\text{Res}_i(\mathbb{D}^\lambda)$  ( $i = 1, \dots, \ell$ ) of  $\text{Gr}_\mathbf{a}^\mathbb{Q}(\mathcal{Q}\mathcal{E}^\lambda, O)$  are obtained as the residues of  $\mathbb{D}^\lambda$ . Because they pairwise commute, we have a decomposition  $\text{Gr}_\mathbf{a}^\mathbb{Q}(\mathcal{Q}\mathcal{E}^\lambda, O) = \bigoplus_{\alpha \in \mathbb{C}^\ell} \text{Gr}_{\mathbf{a}, \alpha}^{\mathbb{Q}, \mathbb{E}}(\mathcal{Q}\mathcal{E}^\lambda, O)$ , where  $\text{Res}_i(\mathbb{D}^\lambda) - \alpha_i$  are nilpotent on  $\text{Gr}_{\mathbf{a}, \alpha}^{\mathbb{Q}, \mathbb{E}}(\mathcal{Q}\mathcal{E}^\lambda, O)$ . Let  $N_i$  denote the nilpotent endomorphisms obtained as the nilpotent part of  $\text{Res}_i(\mathbb{D}^\lambda)$ . Let  $\mathfrak{k}(\lambda) : \mathbb{R} \times \mathbb{C} \rightarrow \mathbb{R} \times \mathbb{C}$  be given by  $\mathfrak{k}(\lambda, a, \alpha) = (a + 2 \text{Re}(\lambda \bar{\alpha}), \alpha - a\lambda - \bar{\alpha}\lambda^2)$ . It induces a map  $\mathbb{R}^\ell \times \mathbb{C}^\ell \rightarrow \mathbb{R}^\ell \times \mathbb{C}^\ell$ , also denoted by  $\mathfrak{k}(\lambda)$ . For  $\mathbf{u} \in \mathbb{R}^\ell \times \mathbb{C}^\ell$ , we set  $\mathcal{G}_\mathbf{u}^\lambda(E) := \text{Gr}_{\mathfrak{k}(\lambda, \mathbf{u})}^{\mathbb{Q}, \mathbb{E}}(\mathcal{Q}\mathcal{E}^\lambda, O)$ . By using the prolongation in family with an additional consideration on the filtrations in [65, 95], it turns out that the family  $\mathcal{G}_\mathbf{u}(E) = (\mathcal{G}_\mathbf{u}^\lambda(E) \mid \lambda \in \mathbb{C})$  is naturally a holomorphic vector bundle on  $\mathbb{C}_\lambda$  with nilpotent endomorphisms  $N_i$  ( $i = 1, \dots, \ell$ ).

Applying the same procedure to  $(E, \partial_E, \theta^\dagger, h)$  on  $X^\dagger \setminus D^\dagger$ , we obtain holomorphic vector bundles  $\mathcal{G}_{\mathbf{u}^\dagger}^\dagger(E)$  on  $\mathbb{C}_\mu$  with nilpotent endomorphisms  $N_i^\dagger$  defined as the nilpotent part of  $\text{Res}_i(\mathbb{D}^\dagger)$ .

We have the bijection  $\mathbb{R} \times \mathbb{C} \rightarrow \mathbb{R} \times \mathbb{C}$  given by  $(a, \alpha) \mapsto (-a, \bar{\alpha})$ , which induces  $\mathbb{R}^\ell \times \mathbb{C}^\ell \simeq \mathbb{R}^\ell \times \mathbb{C}^\ell$  denoted by  $\mathbf{u} \mapsto \mathbf{u}^\dagger$ . By considering the spaces of the multivalued flat sections of  $(\mathcal{E}, \mathbb{D})|_{\lambda \times (X \setminus D)} = (\mathcal{E}^\dagger, \mathbb{D}^\dagger)|_{\mu \times (X^\dagger \setminus D^\dagger)}$  ( $\lambda = \mu^{-1}$ ) with the filtration induced by the growth order of the flat sections and the generalized eigen decompositions of the monodromy, we obtain an isomorphism  $\mathcal{G}_\mathbf{u}(E)|_{\mathbb{C}_\lambda^*} \simeq \mathcal{G}_{\mathbf{u}^\dagger}^\dagger(E)|_{\mathbb{C}_\mu^*}$  under which we have  $\lambda^{-1} N_i = -\mu^{-1} N_i^\dagger$ . By gluing, we obtain a holomorphic vector bundle  $S_\mathbf{u}^{\text{can}}(E)$  on  $\mathbb{P}^1$  with a commuting tuple  $\mathcal{N} = (\mathcal{N}_i \mid i = 1, \dots, \ell)$  of morphisms  $\mathcal{N}_i : S_\mathbf{u}^{\text{can}}(E) \rightarrow S_\mathbf{u}^{\text{can}}(E) \otimes \mathcal{O}_{\mathbb{P}^1}(2)$ . It turns out that the monodromy weight filtration  $W$  of  $\sum_{i=1}^\ell \mathcal{N}_i$  is a filtration by subbundles. We also have the induced symmetric pairing  $\mathcal{S}_\mathbf{u}$  on  $S_\mathbf{u}^{\text{can}}(E)$  induced by the polarization of  $(\mathcal{E}^\Delta, \mathbb{D}^\Delta)$ . The following theorem is fundamental.

**Theorem 7.7** ([65]).  $(S_\mathbf{u}^{\text{can}}(E), W, \mathcal{N}, \mathcal{S}_\mathbf{u})$  is a polarized mixed twistor structure of weight 0.

It means that  $(S_\mathbf{u}^{\text{can}}(E), W)$  is a mixed twistor structure, and the pairing induced by  $\mathcal{S}_\mathbf{u}$  and  $\mathcal{N}$  satisfies some positivity condition. Theorem 7.7 is an analogue of the existence of the limit mixed Hodge structure, which is fundamental in the study of the asymptotic behaviour of polarized variation of Hodge structure due to E. Cattani, A. Kaplan, W. Schmid [9–11, 92] and M. Kashiwara, T. Kawai [42, 46].

**Reduction to polarized mixed Hodge structure.** We have one more reduction. The associated bundle  $\text{Gr}^W S_\mathbf{u}^{\text{can}}(E)$  is naturally equipped with the weight filtration  $W^{(0)}$ , a commuting tuple of morphisms  $\mathcal{N}^{(0)}$ , and a symmetric pairing  $\mathcal{S}^{(0)}$ . Moreover, we can take a  $G_m$ -action on  $(\text{Gr}^W S_\mathbf{u}^{\text{can}}(E), W^{(0)}, \mathcal{N}^{(0)}, \mathcal{S}^{(0)})$ . Under the equivalence between  $G_m$ -equivariant twistor structures and Hodge structures (Theorem 4.2) it is a polarized mixed Hodge structure of weight 0.

Many important properties are preserved by the grading with respect to the weight filtration of mixed twistor structure. Hence, we can deduce many important properties of polarized mixed twistor structure from the classical results for polarized mixed Hodge structure. For example, it permits us to prove that the nilpotent parts of  $\text{Res}_i(\mathbb{D})$  ( $i = 1, \dots, \ell$ ) have a strong constraint [65]. We can use it to obtain the norm estimate of flat sections and

holomorphic sections in terms of the parabolic structure and the monodromy, analogue to the Hodge case [65, 72].

**7.3. The Kobayashi-Hitchin correspondence.**

**Global prolongation.** Let  $X$  be any complex manifold with a simply normal crossing hypersurface  $D = \bigcup_{i \in \Lambda} D_i$ . Let  $(E, \bar{\partial}_E, \theta, h)$  be a good wild harmonic bundle on  $(X, D)$ . The  $\lambda$ -flat bundle  $(\mathcal{E}^\lambda, \mathbb{D}^\lambda)$  on  $X \setminus D$  can be naturally extended to a good filtered  $\lambda$ -flat bundle  $(\mathcal{P}_* \mathcal{E}^\lambda, \mathbb{D}^\lambda)$  on  $(X, D)$ . It is locally given as follows. For any  $P \in D$ , take a small neighbourhood  $U_P$  and a covering  $\varphi_P : U'_P \rightarrow U_P$  ramified over  $U_P \cap D$  such that  $\varphi_P^{-1}(E, \bar{\partial}_E, \theta, h)$  is unramifiedly good. By applying the procedure in §7.1, and taking the descent with respect to the natural action of the Galois group of  $\varphi_P$ , we obtain a good filtered  $\lambda$ -flat bundle on  $(U_P, U_P \cap D)$ . By gluing them, we obtain a global filtered  $\lambda$ -flat bundle on  $(X, D)$ .

**The Kobayashi-Hitchin correspondence.** Suppose that  $X$  is connected and projective with an ample line bundle  $L$ . Set  $n := \dim X$ .

**Theorem 7.8** ([64, 66, 72]). *For any good wild harmonic bundle  $(E, \bar{\partial}_E, \theta, h)$  on  $(X, D)$ , the associated good filtered  $\lambda$ -flat bundle  $(\mathcal{P}_* \mathcal{E}^\lambda, \mathbb{D}^\lambda)$  is poly-stable with  $\mu_L(\mathcal{P}_* \mathcal{E}^\lambda) = 0$  and  $\text{ch}_2(\mathcal{P}_* \mathcal{E}^\lambda) c_1(L)^{n-2} = 0$ .*

After the work of O. Biquard [5], Biquard-P. Boalch, [6], J. Li [54], B. Steer-A. Wren [102], C. Sabbah [79], C. Simpson [94, 95], and the author [64, 66, 72], we also have the converse.

**Theorem 7.9.** *Let  $(\mathcal{P}_* \mathcal{V}, \mathbb{D}^\lambda)$  be a stable good filtered  $\lambda$ -flat bundle on  $(X, D)$  with  $\mu_L(\mathcal{P}_* \mathcal{V}) = 0$  and  $\text{ch}_2(\mathcal{P}_* \mathcal{V}) c_1(L)^{n-2} = 0$ . Then, there exists a pluri-harmonic metric  $h$  of  $(V, \mathbb{D}^\lambda) := (\mathcal{V}, \mathbb{D}^\lambda)|_{X \setminus D}$  such that the filtered bundle associated to  $(V, h)$  is equal to  $\mathcal{P}_* \mathcal{V}$ .*

Strictly, a proof of Theorem 7.9 has not yet been written in the above generality. However, it is not difficult to prove it with the method of [64, 66] (the regular case) and [72] (the case  $\lambda = 1$ ).

**Characterization of semisimplicity.** Let  $X$  be any complex manifold with a normal crossing hypersurface  $D$ . An unramifiedly good meromorphic flat bundle  $(E, \bar{\partial}_E, \theta, h)$  is called  $\sqrt{-1}\mathbb{R}$ -wild if the following holds around any  $P \in D$ :

- Take any small holomorphic coordinate neighbourhood  $(U_P, z_1, \dots, z_n)$  of  $P$  such that  $U_P \cap D = \bigcup_{i=1}^\ell \{z_i = 0\}$ . For the decomposition of  $(E, \bar{\partial}_E, \theta, h)$  around  $P$  as in (7.1), the roots of  $\det(t \text{id}_{E_a} - f_{a,i})|_{z_i=0}$  are purely imaginary.

A good wild harmonic bundle is called  $\sqrt{-1}\mathbb{R}$ -wild if it is locally the descent of unramifiedly good  $\sqrt{-1}\mathbb{R}$ -wild harmonic bundle. (The terminology is slightly changed from that in [72].)

If  $(E, \bar{\partial}_E, \theta, h)$  is good  $\sqrt{-1}\mathbb{R}$ -wild, the good filtered flat bundle  $(\mathcal{P}_* \mathcal{E}^1, \mathbb{D}^1)$  equals the good Deligne-Malgrange [58], filtered bundle over  $(\mathcal{P} \mathcal{E}^1, \mathbb{D}^1)$  on  $(X, D)$ , which is a good filtered flat bundle canonically associated to the good meromorphic flat bundle. (See [70] for a review.)

If  $X$  is projective with an ample line bundle  $L$ , a good Deligne-Malgrange filtered bundle is always  $\mu_L$ -semistable. Moreover, it is  $\mu_L$ -polystable if and only if the underlying

meromorphic flat bundle is semisimple. Hence, we obtain the following characterization of semisimplicity of good meromorphic flat bundles.

**Theorem 7.10** ([72]). *Let  $X$  be a smooth complex projective variety with a simply normal crossing hypersurface  $D$ . A good meromorphic flat bundle  $(\mathcal{V}, \nabla)$  is semisimple if and only if  $(V, \nabla) := (\mathcal{V}, \nabla)|_{X \setminus D}$  has a pluri-harmonic metric  $h$  such that (i) the harmonic bundle  $(V, \nabla, h)$  is good  $\sqrt{-1}\mathbb{R}$ -wild, (ii) the meromorphic bundle  $\mathcal{P}V$  associated to  $(V, h)$  is  $\mathcal{V}$ .*

The case  $\dim X = 1$  is due to Simpson [95] and Sabbah [79]. As for the higher dimensional case, Jost-Zuo [40] constructed a pluri-harmonic metric  $h_{CJZ}$  for any semisimple flat bundle  $(V, \nabla)$  on a quasi-projective variety, i.e., the regular singular case. Their result was refined by the author, proving that  $(V, \nabla, h_{CJZ})$  is tame and  $\sqrt{-1}\mathbb{R}$ -good, and that the existence of such  $h_{CJZ}$  characterizes the semisimplicity. Later, another proof based on the Kobayashi-Hitchin correspondence was given in [66]. The possibly irregular case is settled in [72].

By using Theorem 7.10 and the existence of resolution of turning points (Theorem 5.1), we obtain a characterization of semisimplicity of algebraic meromorphic flat bundles. Suppose that  $X$  is projective. Let  $(\mathcal{V}, \nabla)$  be a meromorphic flat bundle on  $(X, D)$ . Take a birational morphism of complex projective manifolds  $\varphi : X' \rightarrow X$  such that (i)  $D' := \varphi^{-1}(D)$  is simply normal crossing, (ii)  $X' \setminus D' \simeq X \setminus D$ , (iii)  $\varphi^*(\mathcal{V}, \nabla)$  is good.

**Theorem 7.11.**  *$(\mathcal{V}, \nabla)$  is semisimple if and only if  $(V', \nabla') := \varphi^*(\mathcal{V}, \nabla)|_{X' \setminus D'}$  has a pluri-harmonic metric  $h'$  such that (i)  $(V', \nabla', h')$  is good  $\sqrt{-1}\mathbb{R}$ -wild, (ii)  $\mathcal{P}V' = \varphi^*\mathcal{V}$ .*

The author would like to mention that the resolution of turning points is required only in the case  $\dim X = 2$ . Using the Mehta-Ramanathan type theorem saying that the restriction of  $\mu_L$ -stable good filtered  $\lambda$ -flat bundle to arbitrarily ample and general hypersurface is also  $\mu_L$ -stable, we can reduce the characterization in the case  $\dim X \geq 3$  to that in the case  $\dim X = 2$ , although the statement for the characterization should be appropriately given in terms of  $V$  not  $V'$ . (See [72].)

**Semi-infinite variation of Hodge structure.** We also have an interesting generalization of Theorem 3.2. Suppose that  $X$  is an  $n$ -dimensional projective manifold with an ample line bundle  $L$ . Suppose that  $X, D$  and  $L$  are equipped with an algebraic action of  $\mathbb{C}^*$ . Note that the action of  $\mathbb{C}^*$  induces a holomorphic vector field  $\mathfrak{v}$  on  $X$  which is logarithmic to  $D$ .

**Theorem 7.12.** *Let  $(\mathcal{P}_*\mathcal{V}, \theta)$  be a good filtered Higgs bundle such that (i)  $(\mathcal{P}_*\mathcal{V}, \theta)$  is  $\mu_L$ -stable with  $\mu_L(\mathcal{P}_*\mathcal{V}) = 0$  and  $\text{ch}_2(\mathcal{P}_*\mathcal{V})c_1(L)^{n-2} = 0$ , (ii)  $\mathcal{P}_*\mathcal{V}$  is  $\mathbb{C}^*$ -equivariant for which  $t^*\theta = t^m\theta$  for some  $m \neq 0$ . Then, we obtain a complex polarized pure semi-infinite variation of Hodge structure on  $X \setminus D$  with the Euler field  $\mathfrak{v}$ .*

See [2, 38] for the definition of semi-infinite variation of Hodge structure. Indeed, by Theorem 7.11, we obtain a pluri-harmonic metric for  $(\mathcal{P}_*\mathcal{V}, \theta)$  which is  $S^1$ -equivariant. It induces a complex polarized pure semi-infinite variation of Hodge structure as explained in [74].

### 8. Twistor $\mathcal{D}$ -modules

**Holonomic  $\mathcal{D}$ -modules.** Let us recall the basic matters for  $\mathcal{D}$ -modules. See excellent textbooks [37] and [45] for more details. Let  $X$  be any complex manifold. Let  $\mathcal{D}_X$  denote the sheaf of holomorphic linear differential operators on  $X$ . A sheaf of modules over  $\mathcal{D}_X$  is called a  $\mathcal{D}_X$ -module. In this text, we prefer left  $\mathcal{D}_X$ -modules, although right  $\mathcal{D}_X$ -modules are more convenient in several aspects.

Let  $F_j\mathcal{D}_X \subset \mathcal{D}_X$  be the subsheaf of differential operators whose order is less than  $j$ . We have  $F_j\mathcal{D}_X \cdot F_k\mathcal{D}_X \subset F_{j+k}\mathcal{D}_X$ , and the associated graded algebra is isomorphic to the symmetric tensor product ring of the tangent sheaf  $\Theta_X$  of  $X$ .

For any coherent  $\mathcal{D}_X$ -module  $M$ , we can locally consider a filtration  $F$  on  $M$  by coherent  $\mathcal{O}_X$ -submodules  $F_j(M)$  satisfying (i)  $F_j(M) = 0$  if  $j \ll 0$  and  $\bigcup F_j(M) = M$ , (ii)  $F_j(\mathcal{D}_X) \cdot F_k(M) \subset F_{j+k}(M)$ , (iii)  $\bigoplus F_j(M)$  is finitely generated over  $\bigoplus F_j(\mathcal{D}_X)$ . Such a filtration is called a coherent filtration. The associated graded module  $\text{Gr}^F(M)$  is naturally a module over  $\text{Gr}^F(\mathcal{D}_X)$ , and induces a coherent sheaf on the cotangent bundle of  $X$ . Its support is independent of the choice of a coherent filtration, and hence makes sense globally. It is called the characteristic variety of  $M$  and denoted by  $Ch(M)$ . It is a classical theorem that  $\dim Ch(M) \geq \dim X$ . When  $\dim Ch(M) = \dim X$  holds,  $M$  is called holonomic. Flat bundles are naturally holonomic  $\mathcal{D}_X$ -modules. Any holonomic  $\mathcal{D}_X$ -module is locally obtained by gluing meromorphic flat bundles on subvarieties. (See [3].)

We have the standard functors such as pull back, push-forward, duality, tensor products, inner homomorphisms, etc., on the derived categories of  $\mathcal{D}$ -modules as in the case of  $\mathcal{O}$ -modules, although they are not straightforward generalizations. For example, the push-forward is given as follows. Let  $f : X \rightarrow Y$  be a morphism of complex manifolds. Let  $\Omega_X$  and  $\Omega_Y$  denote the canonical line bundles of  $X$  and  $Y$ , respectively. The push-forward of a left  $\mathcal{D}_X$ -module  $M$  is given by  $f_+(M) := Rf_* \left( (\Omega_X \otimes_{\mathcal{O}_X} M) \otimes_{\mathcal{D}_X}^L (\mathcal{O}_X \otimes_{f^{-1}\mathcal{O}_Y} f^{-1}\mathcal{D}_Y) \right) \otimes (\Omega_Y)^{-1}$ . If  $Y$  is a point, it is the de Rham cohomology of  $M$  up to the shift by the degree.

We also have the interesting functors called nearby cycle functors and vanishing cycle functors for holonomic  $\mathcal{D}$ -modules. Let us consider the case where  $X$  is an open subset of the product  $\mathbb{C} \times X_0$ . Let  $t$  be the standard coordinate function of  $\mathbb{C}$ . Let  $V_0\mathcal{D}_X$  denote the sheaf of subalgebras of  $\mathcal{D}_X$  generated by the pull back of  $\mathcal{D}_{X_0}$  and  $t\partial_t$  over  $\mathcal{O}_X$ . Let  $M$  be a holonomic  $\mathcal{D}_X$ -module. We fix a subset  $S \subset \mathbb{C}$  which is mapped to  $\mathbb{C}/\mathbb{Z}$  bijectively. We fix a total order  $\leq_S$  on  $S$ . We have a natural bijection  $\mathbb{Z} \times S \simeq \mathbb{C}$ . The lexicographic order on  $\mathbb{Z} \times S$  induces a total order on  $\mathbb{C}$ . Then, we have a unique filtration  $V(M)$  indexed by  $(\mathbb{C}, \leq)$  such that (i) each  $V_a(M)$  is  $V_0\mathcal{D}_X$ -coherent, (ii)  $tV_a(M) \subset V_{a-1}(M)$  and  $\partial_t V_a(M) \subset V_{a+1}(M)$ , (iii)  $-\partial_t t - a$  is locally nilpotent on  $\text{Gr}_a^V(M)$ . We set  $\psi_t(M) := \bigoplus_{-1 < a < 0} \text{Gr}_a^V(M)$  and  $\phi_t(M) := \bigoplus_{-1 < a \leq 0} \text{Gr}_a^V(M)$ . The functors  $\psi_t$  and  $\phi_t$  are called the nearby cycle functor and the vanishing cycle functor. Briefly said,  $\psi_t(M)$  and  $\phi_t(M)$  contain much information on the regular part of  $M$  along  $t$ . For any holomorphic function  $f$ , we can define  $\psi_f$  and  $\phi_f$  by using the graph of  $f$ .

The foundation having been well established,  $\mathcal{D}$ -modules have played an important role in various fields of mathematics. Holonomic  $\mathcal{D}$ -modules have several reincarnations, and are useful to relate several apparently different objects. One of the most important is the Riemann-Hilbert correspondence due to M. Kashiwara and Z. Mebkhout. For any  $\mathcal{D}_X$ -module  $M$ , we set  $\text{DR}_X(M) := \Omega_X \otimes_{\mathcal{D}_X}^L M$  as an object in the derived category of  $\mathbb{C}_X$ -

modules. Then,  $\mathrm{DR}_X$  induces an equivalence of the categories of regular holonomic  $\mathcal{D}_X$ -modules and perverse sheaves, i.e., the heart of the  $t$ -structure on the derived category of bounded cohomologically constructible complexes on  $X$  induced by the middle perversity.

In the irregular case, the complete generalization of the Riemann-Hilbert correspondence has not yet been established to the best of the author's knowledge. However, inspired by the work of D. Tamarkin [103], A. D'Agnolo and M. Kashiwara [18] obtained a kind of topological description of holonomic  $\mathcal{D}$ -modules, on the basis of the theory of ind-sheaves due to Kashiwara and P. Schapira [47] and the results on the local structure of meromorphic flat bundles mentioned in §5. This result is expected to lead us to thorough understanding of the generalized Riemann-Hilbert correspondence for holonomic  $\mathcal{D}$ -modules.

**Hodge modules.** Recall that a mixed Hodge structure is a  $\mathbb{Q}$ -vector space  $H_{\mathbb{Q}}$  equipped with a decreasing filtration  $F = (F^p \mid p \in \mathbb{Z})$  of  $H_{\mathbb{Q}} \otimes \mathbb{C}$  and an increasing filtration  $W = (W_m \mid m \in \mathbb{Z})$  of  $H_{\mathbb{Q}}$ , such that  $\mathrm{Gr}_n^W(H_{\mathbb{Q}}) \otimes \mathbb{C}$  with the induced filtrations  $F$  and  $\overline{F}$  is pure Hodge structure of weight  $n$ . Here,  $\overline{F}$  is the conjugate of  $F$ . It was introduced by P. Deligne as a nice structure on the cohomology groups of general complex algebraic varieties.

A. Beilinson, J. Bernstein, Deligne and O. Gabber [4] developed the theory of weights for perverse sheaves with Frobenius actions on algebraic varieties over finite fields. The dictionary of Deligne [20] between Frobenius action and mixed Hodge structure predicts a nice theory of regular holonomic  $\mathcal{D}$ -modules with mixed Hodge structure. It is realized by M. Saito as the theory of mixed Hodge modules [87, 88].

Roughly, a Hodge module on a complex manifold  $X$  consists of a regular holonomic  $\mathcal{D}$ -module  $\mathcal{M}$  with a coherent filtration  $F$  and a  $\mathbb{Q}$ -perverse sheaf  $P$  with an isomorphism  $\mathrm{DR}_X \mathcal{M} \simeq P \otimes_{\mathbb{Q}} \mathbb{C}$ . One of the important inventions of Saito is a nice method to impose conditions on such  $(\mathcal{M}, F, P)$ . In the case of variation of Hodge structure, we impose conditions by considering the restriction to points. For example, a variation of Hodge structure is pure if its restriction to each point is pure. Instead, Saito introduced the idea to impose the conditions by using the nearby and vanishing cycle functors along functions. It is particularly useful in his inductive argument to prove the Hard Lefschetz Theorem. See his papers [87, 88] and expositions [90, 91] for more details. See [37] for applications of Hodge modules to the representation theory.

**Twistor  $\mathcal{D}$ -modules.** Simpson's Meta-Theorem predicts a twistor version of Hodge modules. The theory of pure twistor  $\mathcal{D}$ -module was introduced by Sabbah [81, 83] and developed by him and the author [65, 72]. The mixed case was studied in [73].

The theory of Hodge modules is quite satisfactory. Most important  $\mathcal{D}$ -modules in algebraic geometry and representation theory naturally underlie mixed Hodge modules. Moreover, the Hodge filtrations play important roles in significant applications, but mixed twistor  $\mathcal{D}$ -modules do not have such filtration. (However, see the recent studies on irregular Hodge filtrations [30], inspired by the note of Deligne [21].) It might be asked why twistor  $\mathcal{D}$ -modules are interesting.

One answer is that there can still exist basic holonomic  $\mathcal{D}$ -modules which cannot underlie mixed Hodge modules. For example,  $(\mathcal{O}_{\mathbb{P}^1}(*\{0, \infty\}), d + \alpha dt/t)$  ( $\alpha \in \mathbb{C} \setminus \mathbb{R}$ ) and  $(\mathcal{O}_{\mathbb{P}^1}(*\infty), d + dt)$  cannot underlie any mixed Hodge modules. Indeed, if a holonomic  $\mathcal{D}$ -module  $M$  underlies a mixed Hodge module,  $M$  is regular, and the absolute value of the eigenvalues of the local monodromies of  $M$  is 1.

Relatedly, there has been a growing interest in a generalized Hodge theory for holonomic

$\mathcal{D}$ -modules in the context of irregular singularities. On one hand, Deligne discovered interesting resemblance between wildly ramified perverse sheaves in the positive characteristic case and irregular holonomic  $\mathcal{D}$ -modules. The dictionary of Deligne would predict a theory of weights for holonomic  $\mathcal{D}$ -modules, not only for regular holonomic  $\mathcal{D}$ -modules. On the other hand, it turns out that some natural objects in various fields of mathematics, including mathematical physics and singularity theory, are (expected to be) naturally equipped with variation of generalized Hodge structure such as semi-infinite variation of Hodge structure, TERP structure, non-commutative Hodge structure [2, 33, 48] etc. (Indeed, the equation for pluri-harmonic metrics also appeared in [13].) The construction of such variations of generalized Hodge structure, together with the method of M. Saito [89], is sometimes an important step to obtain flat structures of K. Saito (see a nice survey [86]), or equivalently Frobenius manifolds of B. Dubrovin [28]. (See [26], [82]. See also [78].) A theory of holonomic  $\mathcal{D}$ -modules with generalized Hodge structure would provide us, in some cases, a systematic way to understand such natural generalized Hodge structure. Because twistor structure is the most basic among generalized Hodge structures, the theory of twistor  $\mathcal{D}$ -modules should be fundamental.

Let us briefly describe the ingredients for twistor  $\mathcal{D}$ -modules. Sabbah [81] introduced the concept of  $\mathcal{R}$ -triples which is suitable to deal with irregular singularity. Let  $X$  be any complex manifold. Let  $\Theta_{\mathbb{C}_\lambda \times X / \mathbb{C}_\lambda}$  denote the relative tangent sheaf. Let  $\mathcal{R}_X$  denote the sheaf of subalgebras in  $\mathcal{D}_{\mathbb{C}_\lambda \times X}$  generated by  $\lambda \Theta_{\mathbb{C}_\lambda \times X / \mathbb{C}_\lambda}$  over  $\mathcal{O}_{\mathbb{C}_\lambda \times X}$ . As in the case of  $\mathcal{D}$ -modules, for any coherent  $\mathcal{R}_X$ -module  $\mathcal{M}$ , we have its characteristic variety  $Ch(\mathcal{M})$  in  $\mathbb{C}_\lambda \times T^*X$ . It is called holonomic if  $Ch(\mathcal{M})$  is contained in the product of  $\mathbb{C}_\lambda$  and a Lagrangian subvariety in  $T^*X$ . An  $\mathcal{R}_X$ -module is called strict if it is flat over  $p_1^{-1} \mathcal{O}_{\mathbb{C}_\lambda}$  where  $p_1 : \mathbb{C}_\lambda \times X \rightarrow \mathbb{C}_\lambda$  denotes the projection. Note that a filtered holonomic  $\mathcal{D}_X$ -module  $(M, F)$  induces a strict holonomic  $\mathcal{R}_X$ -module by the Rees construction and its analytification.

Set  $\mathcal{S} := \{\lambda \in \mathbb{C} \mid |\lambda| = 1\}$ . Let  $\mathfrak{D}\mathfrak{b}_{\mathcal{S} \times X / \mathcal{S}}$  denote the sheaf of distributions on  $\mathcal{S} \times X$  which are  $C^0$  with respect to  $\mathcal{S}$  (see [81]). Then,  $\mathfrak{D}\mathfrak{b}_{\mathcal{S} \times X / \mathcal{S}}$  is naturally an  $\mathcal{R}_{X|\mathcal{S} \times X}$ -module. Let  $\sigma : \mathcal{S} \times X \rightarrow \mathcal{S} \times X$  be given by  $\sigma(\lambda, P) = (-\lambda, P)$ . Then,  $\mathfrak{D}\mathfrak{b}_{\mathcal{S} \times X / \mathcal{S}}$  is also a  $\sigma^* \mathcal{R}_{X|\mathcal{S} \times X}$ -module by  $\sigma^*(f) \cdot \Phi = \overline{\sigma^*(f)} \Phi$ .

A sesqui-linear pairing  $C$  between  $\mathcal{R}_X$ -modules  $\mathcal{M}_i$  ( $i = 1, 2$ ) is an

$$\mathcal{R}_{X|\mathcal{S} \times X} \times \sigma^* \mathcal{R}_{X|\mathcal{S} \times X}\text{-homomorphism } C : \mathcal{M}_1|_{\mathcal{S} \times X} \times \sigma^* \mathcal{M}_2|_{\mathcal{S} \times X} \rightarrow \mathfrak{D}\mathfrak{b}_{\mathcal{S} \times X / \mathcal{S}}$$

Such a tuple  $\mathcal{T} = (\mathcal{M}_1, \mathcal{M}_2, C)$  is called an  $\mathcal{R}_X$ -triple. When  $\mathcal{M}_i$  are strict and holonomic,  $\mathcal{T}$  is called strict and holonomic.

If  $X$  is a point, we have  $\mathcal{R}_X = \mathcal{O}_{\mathbb{C}_\lambda}$ . A twistor structure  $V$  induces an  $\mathcal{R}_X$ -triple as follows. We regard  $\mathbb{P}^1$  as the gluing of  $\mathbb{C}_\lambda$  and  $\mathbb{C}_\mu$  with  $\lambda\mu = 1$ . Let  $\sigma : \mathbb{C}_\lambda \rightarrow \mathbb{C}_\mu$  be given by  $\sigma(\lambda) = -\bar{\lambda}$ . We set  $V_0 := V|_{\mathbb{C}_\lambda}$  and  $V_\infty := V|_{\mathbb{C}_\mu}$ . Then, we have  $\mathcal{O}_{\mathbb{C}_\lambda}$ -modules  $V_0$  and  $\sigma^* V_\infty^\vee$ . The gluing of  $V_0$  and  $V_\infty$  naturally induces a sesqui-linear pairing  $C_V$  of  $\sigma^* V_\infty^\vee$  and  $V_0$ . Thus, we obtain an  $\mathcal{R}_X$ -triple  $(\sigma^* V_\infty^\vee, V_0, C_V)$  from a twistor structure in the case where  $X$  is a point. For any complex manifold  $X$ , we can obtain a strict holonomic  $\mathcal{R}_X$ -triple from a variation of twistor structure on  $X$  in a similar way.

Roughly, we regard a strict holonomic  $\mathcal{R}_X$ -triple  $\mathcal{T} = (\mathcal{M}_1, \mathcal{M}_2, C)$  as a twistor structure on the  $\mathcal{D}_X$ -module  $\Xi_{DR}(\mathcal{T})$ , where  $\Xi_{DR}(\mathcal{T})$  is obtained as the specialization of  $\mathcal{M}_2$  along  $\{1\} \times X$ . By adapting Saito's strategy to use the nearby and vanishing cycle functors, Sabbah [81, 83] introduced the conditions for  $\mathcal{T}$  to be pure and polarizable. He made various innovations to make the strategy work appropriately, partially because of the difference of

ingredients. By using the strategy, the author [73] introduced the conditions for a filtered  $\mathcal{R}_X$ -triple to be a mixed twistor  $\mathcal{D}$ -module

**Remark 8.1.** *In the following, to simplify the statements, for any mixed twistor  $\mathcal{D}$ -modules their KMS-spectrum along any holomorphic functions are assumed to be contained in  $\mathbb{R} \times \sqrt{-1}\mathbb{R}$ . (It corresponds to  $\sqrt{-1}\mathbb{R}$ -wildness in §7.3.)*

One of the main results is the functoriality of twistor  $\mathcal{D}$ -modules as in the Hodge case. We have the Hard Lefschetz Theorem for polarizable pure twistor  $\mathcal{D}$ -modules [65, 72, 81]. It implies the following: Let  $f : X \rightarrow Y$  be a projective morphism of complex manifolds. Let  $\mathcal{T}$  be a polarizable pure twistor  $\mathcal{D}_X$ -module of weight  $w$ . Then, we have a naturally induced polarizable pure twistor  $\mathcal{D}_Y$ -modules  $f_+^i \mathcal{T}$  of weight  $w + i$  over the  $\mathcal{D}_Y$ -modules  $f_+^i \Xi_{DR}(\mathcal{T})$ . Moreover, for the morphism  $L : f_+^i(\mathcal{T}) \rightarrow f_+^{i+2}(\mathcal{T})$  induced by the first Chern class of a relatively ample line bundle, the morphisms  $L^i : f_+^{-i}(\mathcal{T}) \rightarrow f_+^i(\mathcal{T})$  are isomorphisms. As for the mixed case [73], we obtain the standard functors together with the nearby and vanishing cycle functors on the derived category of mixed twistor  $\mathcal{D}$ -modules on algebraic varieties, which are compatible with those on the derived category of holonomic  $\mathcal{D}$ -modules through  $\Xi_{DR}$ .

Because the conditions for pure and mixed twistor  $\mathcal{D}$ -modules are difficult to check, it is important to know which objects are really mixed twistor  $\mathcal{D}$ -modules, as in the Hodge case. It is the other main result in the theory. In the pure case, we have the following correspondence [72] between wild harmonic bundles and polarized pure twistor  $\mathcal{D}$ -modules:

- Let  $Z \subset X$  be a closed irreducible complex analytic subvariety with a resolution of singularity  $\varphi : Z' \rightarrow Z$ . Let  $D \subset Z'$  be a normal crossing hypersurface. Let  $(E, \bar{\partial}_E, \theta, h)$  be a good  $\sqrt{-1}\mathbb{R}$ -wild harmonic bundle on  $(Z', D)$ . Then, there exists a naturally defined polarized pure twistor  $\mathcal{D}$ -module over the holonomic  $\mathcal{D}_X$ -module  $\varphi_{!*} \mathcal{Q}\mathcal{E}^1$ . Here,  $\varphi_{!*} \mathcal{Q}\mathcal{E}^1$  is the minimal extension of the meromorphic flat bundle  $\mathcal{Q}\mathcal{E}^1$ , i.e., the image of  $\varphi_! \mathcal{Q}\mathcal{E}^1 \rightarrow \varphi_* \mathcal{Q}\mathcal{E}^1$ . Conversely, any polarized pure twistor  $\mathcal{D}$ -module is the direct sum of such minimal extensions of good  $\sqrt{-1}\mathbb{R}$ -wild harmonic bundles up to the shift of weights. We crucially use the meromorphic object with the nice property associated to any wild harmonic bundle described in §7.1 and §7.2.

In the mixed case, we have the following [73]:

- We continue to use the above notation. Let  $(\mathcal{V}, W)$  be an admissible variation of mixed twistor structure. Then, there exist mixed twistor  $\mathcal{D}_X$ -modules  $\varphi_*(\mathcal{V}, W)$  ( $\star = *, !$ ) over the holonomic  $\mathcal{D}$ -modules  $\varphi_* \Xi_{DR}(\mathcal{V})$ . Conversely, any mixed twistor  $\mathcal{D}_X$ -modules are locally described as the gluing of admissible variations of mixed twistor structures.

Admissible variation of mixed twistor structure is a twistor version of admissible variation of mixed Hodge structure in [43] and [101].

**Kashiwara's conjecture.** The most interesting result is obtained by the mixture with global analysis. By Theorem 7.11 and the above description of pure twistor  $\mathcal{D}$ -modules, we obtain a nice characterization of semisimplicity of algebraic holonomic  $\mathcal{D}$ -modules.

**Theorem 8.2** ([65, 72]). *For any holonomic  $\mathcal{D}$ -module  $M$  on any smooth projective variety,  $M$  is semisimple if and only if there exists a polarized pure twistor  $\mathcal{D}$ -module of weight 0 over  $M$ . ("0" can be replaced with any integer.)*

As an application, we can prove the following conjecture of Kashiwara.

**Theorem 8.3.** *Let  $X$  be a smooth complex algebraic variety. Let  $M$  be a semisimple holonomic  $\mathcal{D}_X$ -module.*

- *Let  $f : X \rightarrow Y$  be any projective morphism. We obtain holonomic  $\mathcal{D}_Y$ -modules  $f_+^i M$ . Then, each  $f_+^i M$  is semisimple. Moreover, for the morphism  $f_+^i M \rightarrow f_+^{i+2} M$  induced by the first Chern class of any relatively ample line bundle, the morphisms  $L^i : f_+^{-i} M \rightarrow f_+^i M$  ( $i \geq 0$ ) are isomorphisms. In particular,  $f_+ M$  is isomorphic to  $\bigoplus f_+^i M[-i]$  in the derived category of holonomic  $\mathcal{D}_Y$ -modules.*
- *Let  $g : X \rightarrow \mathbb{C}$  be any algebraic function. Let  $W$  denote the weight filtration on  $\psi_g(M)$  and  $\phi_g(M)$ . Then,  $\mathrm{Gr}^W \psi_g(M)$  and  $\mathrm{Gr}^W \phi_g(M)$  are also semisimple.*

This type of theorem was invented by Beilinson, Bernstein, Deligne and Gabber [4]. They proved such properties for perverse sheaves of geometric origin, by using their theory of weights for perverse sheaves together with the reduction to positive characteristic. It was vastly generalized by Saito who established such property for the regular holonomic  $\mathcal{D}$ -modules underlying his polarizable pure Hodge modules. Later, another Hodge theoretic proof in the case of geometric origin was given by M. de Cataldo and L. Migliorini [12]. The generalization to the case of semisimple regular holonomic  $\mathcal{D}$ -modules was obtained by two methods. One goes through the method of arithmetic geometry; V. Drinfeld [27] proved that a conjecture of A. J. de Jong implies the regular singular case of Kashiwara's conjecture, and the conjecture of de Jong was later proved by G. Böckle and C. Khare [7] and D. Gaitsgory [31]. The other is given by Sabbah and the author as a consequence of the theory of regular pure twistor  $\mathcal{D}$ -modules and tame harmonic bundles [65, 81], which the author calls Sabbah's program. It should be mentioned that Simpson [98] already suggested to apply the method of harmonic bundles to obtain such results. Finally, the general case is given in [72].

**Acknowledgement.** Research partially supported by the Grant-in-Aid for Scientific Research (C) (No. 22540078), the Grant-in-Aid for Scientific Research (A) (No. 22244003) and the Grant-in-Aid for Scientific Research (S) (No. 24224001), Japan Society for the Promotion of Science.

I thank the mentors in this study; Carlos Simpson, Claude Sabbah, Morihiko Saito. I appreciate the comments to this text by H ele ne Esnault, Simpson, Sabbah, Kari Vilonen and Ting Xue. I thank Pierre Deligne, Mark de Cataldo, Kenji Fukaya, William Fulton, David Gieseker, Akira Ishii, Masaki Kashiwara, Akira Kono, Mikiya Masuda, Tomohide Terasoma, Michael Thaddeus, Yoshifumi Tsuchimoto.

## References

- [1] Y. Andr e, *Structure des connexions m eromorphes formelles de plusieurs variables et semi-continuit e de l'irr egularit e*, Invent. Math. **170** (2007), 147–198.
- [2] S. Barannikov, *Quantum periods. I. Semi-infinite variations of Hodge structures*, Internat. Math. Res. Notices **23** (2001), 1243–1264.
- [3] A. Beilinson, *How to glue perverse sheaves*, IN *K-theory, arithmetic and geometry (Moscow, 1984–1986)*, Lect. Notes in Math. **1289**, Springer, Berlin, (1987), 42–51.

- [4] A. Beilinson, J. Bernstein, and P. Deligne, *Faisceaux pervers*, Analysis and topology on singular spaces, I (Luminy, 1981), Astérisque **100** (1982), 5–171.
- [5] O. Biquard, *Fibrés de Higgs et connexions intégrables: le cas logarithmique (diviseur lisse)*, Ann. Sci. École Norm. Sup. **30** (1997), 41–96.
- [6] O. Biquard and P. Boalch, *Wild non-abelian Hodge theory on curves*, Compos. Math. **140** (2004), 179–204.
- [7] G. Böckle and C. Khare, *Mod  $\ell$  representations of arithmetic fundamental groups II (A conjecture of A. J. de Jong)*, Compos. Math. **142** (2006), 271–294.
- [8] N. Borne, *Sur les représentations du groupe fondamental d'une variété privée d'un diviseur à croisements normaux simples*, Indiana Univ. Math. J. **58** (2009), 137–180.
- [9] E. Cattani and A. Kaplan, *Polarized mixed Hodge structures and the local monodromy of variation of Hodge structure*, Invent. Math. **67** (1982), 101–115.
- [10] E. Cattani, A. Kaplan, and W. Schmid, *Degeneration of Hodge structures*, Ann. of Math. **123** (1986), 457–535.
- [11] E. Cattani, A. Kaplan, and W. Schmid,  *$L^2$  and intersection cohomologies for a polarized variation of Hodge structure*, Invent. Math. **87** (1987), 217–252.
- [12] M. A. de Cataldo and L. Migliorini, *The Hodge theory of algebraic maps*, Ann. Sci. École Norm. Sup. (4) **38** (2005), 693–750.
- [13] S. Cecotti and C. Vafa, *Topological–anti-topological fusion*, Nuclear Phys. B, **367** (1991), 359–461.
- [14] B. Charbonneau and J. Hurtubise, *Singular Hermitian-Einstein monopoles on the product of a circle and a Riemann surface*, Int. Math. Res. Not. IMRN 2011, 175–216.
- [15] K. Corlette, *Flat  $G$ -bundles with canonical metrics*, J. Differential Geom. **28** (1988), 361–382.
- [16] ———, *Harmonic maps, rigidity, and Hodge theory*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994), 465–471, Birkhäuser, Basel, 1995.
- [17] M. Cornalba and P. Griffiths, *Analytic cycles and vector bundles on noncompact algebraic varieties*, Invent. Math. **28** (1975), 1–106.
- [18] A. D’Agnolo and M. Kashiwara, *Riemann-Hilbert correspondence for holonomic  $D$ -modules*, arXiv:1311.2374.
- [19] P. Deligne, *Équation différentielles à points singuliers réguliers*, Lectures Notes in Maths., vol. 163, Springer, 1970.
- [20] ———, *Théorie de Hodge I*, Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 1. Gauthier-Villars, Paris, 1971, pp. 425–430.
- [21] P. Deligne, B. Malgrange, and J-P. Ramis, *Singularités Irrégulières*, Documents Mathématiques **5**, Société Mathématique de France (2007), xii+188 pp.
- [22] K. Diederich and T. Ohsawa, *Harmonic mappings and disc bundles over compact Kähler manifolds*, Publ. Res. Inst. Math. Sci. **21** (1985), 819–833.
- [23] S. K. Donaldson, *Anti self-dual Yang-Mills connections over complex algebraic sur-*

- faces and stable vector bundles*, Proc. London Math. Soc. **50** (1985), 1–26.
- [24] S. K. Donaldson, *Infinite determinants, stable bundles and curvature*, Duke Math. J. **54** (1987), 231–247.
- [25] ———, *Twisted harmonic maps and the self-duality equations*, Proc. London Math. Soc. **55** (1987), 127–131.
- [26] A. Douai and C. Sabbah, *Gauss-Manin systems, Brieskorn lattices and Frobenius structures. I.*, Ann. Inst. Fourier (Grenoble) **53** (2003), 1055–1116.
- [27] V. Drinfeld, *On a conjecture of Kashiwara*, Math. Res. Lett. **8** (2001), 713–728.
- [28] B. Dubrovin, *Geometry and integrability of topological-antitopological fusion*, Comm. Math. Phys. **152** (1993), 539–564.
- [29] J. Eelles and J. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math. **86** (1964), 109–160.
- [30] H. Esnault, C. Sabbah, and J-D. Yu, with an appendix by M. Saito,  *$E_1$ -Degeneration of the irregular Hodge filtration*, arXiv:1302.4537.
- [31] D. Gaitsgory, *On de Jong’s conjecture*, Israel J. Math. **157** (2007), 155–191.
- [32] P. Griffiths, *Hodge theory and geometry*, Bull. London Math. Soc. **36** (2004), 721–757.
- [33] C. Hertling,  *$tt^*$  geometry, Frobenius manifolds, their connections, and the construction for singularities*, J. Reine Angew. Math. **555** (2003), 77–161.
- [34] C. Hertling and Ch. Sevenheck, *Nilpotent orbits of a generalization of Hodge structures*, J. Reine Angew. Math. **609** (2007), 23–80.
- [35] C. Hertling and C. Sevenheck, *Limits of families of Brieskorn lattices and compactified classifying spaces*, Adv. Math. **223** (2010), 1155–1224.
- [36] N. Hitchin, *The self-duality equations on a Riemann surface*, Proc. London Math. Soc. **55** (1987), 59–126.
- [37] R. Hotta, K. Takeuchi, and T. Tanisaki,  *$\mathcal{D}$ -modules, perverse sheaves, and representation theory*, Birkhäuser Boston, Inc., Boston, MA, 2008. xii+407 pp.
- [38] H. Iritani, *An integral structure in quantum cohomology and mirror symmetry for toric orbifolds*, Adv. Math. **222** (2009), 1016–1079.
- [39] J. Iyer and C. Simpson, *A relation between the parabolic Chern characters of the de Rham bundles*, Math. Ann. **338** (2007), 347–383.
- [40] J. Jost and K. Zuo, *Harmonic maps of infinite energy and rigidity results for representations of fundamental groups of quasiprojective varieties*, J. Differential Geom. **47** (1997), 469–503.
- [41] M. Kashiwara, *The Riemann-Hilbert problem for holonomic systems*, Publ. Res. Inst. Math. Sci. **20** (1984), 319–365.
- [42] ———, *The asymptotic behavior of a variation of polarized Hodge str.* Publ. Res. Inst. Math. Sci. **21** (1985), 853–875.
- [43] ———, *A study of variation of mixed Hodge structure.* Publ. Res. Inst. Math. Sci. **22** (1986), 991–1024.

- [44] ———, *Semisimple holonomic  $D$ -modules*, in *Topological Field Theory, Primitive Forms and Related Topics*, Prog. in Math., **160**, Birkhäuser, (1998), 267–271.
- [45] ———,  *$D$ -modules and microlocal calculus*, Translations of Mathematical Monographs, **217**. American Mathematical Society, Providence, 2003. xvi+254 pp.
- [46] M. Kashiwara and T. Kawai, *The Poincaré lemma for variations of polarized Hodge structure*, Publ. Res. Inst. Math. Sci. **23** (1987), 345–407.
- [47] M. Kashiwara and P. Schapira, *Ind-sheaves*, Astérisque **271** (2001), 136 pp.
- [48] L. Katzarkov, M. Kontsevich, and T. Pantev, *Hodge theoretic aspects of mirror symmetry*, in *From Hodge theory to integrability and TQFT  $tt^*$ -geometry*, Proc. Sympos. Pure Math., **78**, Amer. Math. Soc., Providence, RI, (2008), 87–174, math:0806.0107.
- [49] K. Kedlaya, *Good formal structures for flat meromorphic connections, I; Surfaces*, Duke Math. J., **154** (2010), 343–418. math:0811.0190.
- [50] ———, *Good formal structures for flat meromorphic connections, II: Excellent schemes*, J. Amer. Math. Soc. **24** (2011), 183–229. arXiv:1001.0544.
- [51] S. Kobayashi, *Curvature and stability of vector bundles*, Proc. Japan Acad. Ser. A Math. Sci. **58** (1982), 158–162.
- [52] L. Lafforgue, *Chtoucas de Drinfeld et correspondance de Langlands* Invent. Math. **147** (2002), 1–241.
- [53] A. Levelt, *Jordan decomposition for a class of singular differential operators*, Ark. Math. **13** (1975), 1–27.
- [54] J. Li, *Hermitian-Einstein metrics and Chern number inequalities on parabolic stable bundles over Kähler manifolds*, Comm. Anal. Geom. **8** (2000), 445–475.
- [55] M. Lübke and A. Teleman, *The Kobayashi-Hitchin correspondence*, World Scientific Publishing Co., Inc., River Edge, NJ, 1995.
- [56] H. Majima, *Asymptotic analysis for integrable connections with irregular singular points*, Lect. Notes in Math. **1075**, Springer-Verlag, Berlin, 1984. x+159 pp.
- [57] B. Malgrange, *La classification des connexions irrégulières à une variable*, In *Mathematics and physics (Paris, 1979/1982)*, Birkhäuser Boston, Boston, MA, (1983), 381–399.
- [58] ———, *Connexions méromorphes 2, Le réseau canonique*, Invent. Math. **124** (1996) 367–387.
- [59] M. Maruyama and K. Yokogawa, *Moduli of parabolic stable sheaves*, Math. Ann. **293** (1992), 77–99.
- [60] Z. Mebkhout, *Le théorème de comparaison entre cohomologies de de Rham sur le corps des nombres complexes*, C. R. Acad. Sci. Paris Sér. I Math. **305** (1987), 549–552.
- [61] ———, *Le théorème de comparaison entre cohomologies de de Rham d’une variété algébrique complexe et le théorème d’existence de Riemann*, Inst. Hautes Études Sci. Publ. Math. **69** (1989), 47–89.
- [62] V. Mehta and A. Ramanathan, *Restriction of stable sheaves and representations of the fundamental group*, Invent. Math. **77** (1984), 163–172.

- [63] V. Mehta and C. S. Seshadri, *Moduli of vector bundles on curves with parabolic structures*, Math. Ann. **248** (1980), 205–239.
- [64] T. Mochizuki, *Kobayashi-Hitchin correspondence for tame harmonic bundles and an application*, Astérisque **309** (2006).
- [65] ———, *Asymptotic behaviour of tame harmonic bundles and an application to pure twistor  $\mathcal{D}$ -modules I, II*, Mem. AMS. **185** (2007)
- [66] ———, *Kobayashi-Hitchin correspondence for tame harmonic bundles II*, Geometry&Topology **13** (2009), 359–455. math.DG/0602266.
- [67] ———, *Good formal structure for meromorphic flat connections on smooth projective surfaces*, in *Algebraic Analysis and Around*, Adv. Stud. in Pure Math. **54** (2009), 223–253, math:0803.1346.
- [68] ———, *Tame harmonic bundles and the application to pure twistor  $\mathcal{D}$ -modules*, Sugaku expositions. Sugaku Expositions **23** (2010), 105–131.
- [69] ———, *Asymptotic behaviour of variation of pure polarized TERP structures*, Publ. Res. Inst. Math. Sci. **47** (2011), 419–534, math:0811.1384.
- [70] T. Mochizuki, *On Deligne-Malgrange lattices, resolution of turning points and harmonic bundles*, Ann. Inst. Fourier (Grenoble) **59** (2009), 2819–2837.
- [71] ———, *The Stokes structure of good meromorphic flat bundle*, J. Inst. Math. Jussieu **10** (2011), 675–712.
- [72] ———, *Wild harmonic bundles and wild pure twistor  $\mathcal{D}$ -modules*, Astérisque **340**, 2011, x+607 pp.
- [73] ———, *Mixed twistor  $\mathcal{D}$ -modules*, arXiv:1104.3366.
- [74] ———, *Harmonic bundles and Toda lattices with opposite sign II*, Comm. Math. Phys. DOI:10.1007/s00220-014-1994-0 the second part of arXiv:1301.1718.
- [75] M. S. Narasimhan and C. S. Seshadri, *Stable and unitary vector bundles on a compact Riemann surface*, Ann. of Math. (2) **82** (1965), 540–567.
- [76] D. Panov, *Polyhedral Kähler manifolds*, Geom. Topol. **13** (2009), 2205–2252.
- [77] A. Reznikov, *All regulators of flat bundles are torsion*, Ann. of Math. (2) **141** (1995), 373–386.
- [78] T. Reichelt and C. Sevenheck, *Logarithmic Frobenius manifolds, hypergeometric systems and quantum  $\mathcal{D}$ -modules*, arXiv:1010.2118, to appear in J. Alg. Geom.
- [79] C. Sabbah, *Harmonic metrics and connections with irregular singularities*, Ann. Inst. Fourier (Grenoble) **49** (1999), 1265–1291.
- [80] ———, *Équations différentielles à points singuliers irréguliers et phénomène de Stokes en dimension 2*, Astérisque, **263** (2000).
- [81] ———, *Polarizable twistor  $\mathcal{D}$ -modules*, Astérisque, **300** (2005)
- [82] ———, *Fourier-Laplace transform of a variation of polarized complex Hodge structure*, J. Reine Angew. Math. **621** (2008), 123–158.
- [83] ———, *Wild twistor  $\mathcal{D}$ -modules*, in *Algebraic analysis and around*, Adv. Stud. Pure Math. **54** (2009), 293–353.

- [84] C. Sabbah, *Introduction to Stokes structures*, Lecture Notes in Mathematics, **2060**, Springer, Heidelberg, 2013. xiv+249 pp
- [85] ———, *Théorie de Hodge et correspondance de Hitchin-Kobayashi sauvages*, Séminaire Bourbaki 64<sup>ème</sup> année, 2011–2012.
- [86] K. Saito and A. Takahashi, “*From primitive forms to Frobenius manifolds*”, In “*From Hodge theory to integrability and TQFT  $tt^*$ -geometry*”, Proc. Sympos. Pure Math., **78**, Amer. Math. Soc., Providence, RI, (2008), 31–48.
- [87] M. Saito, *Modules de Hodge polarisables*, Publ. Res. Inst. Math. Sci. **24** (1988), 849–995.
- [88] ———, *Mixed Hodge modules*, Publ. Res. Inst. Math. Sci. **26** (1990), 221–333.
- [89] ———, *On the structure of Brieskorn lattice*, Ann. Inst. Fourier (Grenoble) **39** (1989), 27–72.
- [90] ———, *Mixed Hodge modules and applications*, Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990), 725–734, Math. Soc. Japan, Tokyo, 1991.
- [91] ———, *On the theory of mixed Hodge modules*, in Selected papers on number theory, algebraic geometry, and differential geometry, 47–61, Amer. Math. Soc. Transl. Ser. 2, **160**, Amer. Math. Soc., Providence, RI, 1994.
- [92] W. Schmid, *Variation of Hodge structure: the singularities of the period mapping*, Invent. Math. **22** (1973), 211–319.
- [93] Y. Sibuya, *Linear differential equations in the complex domain: problems of analytic continuation*, Kinokuniya, Tokyo, (1976) (in Japanese), Translations of Mathematical Monographs, **82**, American Mathematical Society, Providence, RI, (1990).
- [94] C. Simpson, *Constructing variations of Hodge structure using Yang-Mills theory and application to uniformization*, J. Amer. Math. Soc. **1** (1988), 867–918.
- [95] ———, *Harmonic bundles on non-compact curves*, J. Amer. Math. Soc. **3** (1990), 713–770.
- [96] ———, *The ubiquity of variations of Hodge structure*, Complex geometry and Lie theory (Sundance, UT, 1989), 329–348, Proc. Sympos. Pure Math., **53**, Amer. Math. Soc., Providence, RI, 1991.
- [97] ———, *Higgs bundles and local systems*, Publ. I.H.E.S., **75** (1992), 5–95.
- [98] ———, *Some families of local systems over smooth projective varieties*, Ann. of Math. (2) **138** (1993), 337–425.
- [99] ———, *Mixed twistor structures*, math.AG/9705006.
- [100] ———, *The Hodge filtration on nonabelian cohomology*, Proc. Sympos. Pure Math., **62**, Part 2, Amer. Math. Soc., Providence, RI, (1997), 217–281.
- [101] J. Steenbrink and S. Zucker, *Variation of mixed Hodge structure. I*, Invent. Math. **80** (1985), 489–542.
- [102] B. Steer and A. Wren, *The Donaldson-Hitchin-Kobayashi correspondence for parabolic bundles over orbifold surfaces*, Canad. J. Math. **53** (2001), 1309–1339.
- [103] D. Tamarkin, *Microlocal condition for non-displaceability*, arXiv:0809.1584v1.

- [104] K. Uhlenbeck and S. T. Yau, *On the existence of Hermitian Yang-Mills connections in stable bundles*, Comm. Pure Appl. Math., **39-S** (1986), 257–293.
- [105] K. Yokogawa, *Compactification of moduli of parabolic sheaves and moduli of parabolic Higgs sheaves*, J. Math. Kyoto Univ. **33** (1993) 451–504.
- [106] S. Zucker, *Hodge theory with degenerating coefficients:  $L^2$  cohomology in the Poincaré metric*, Ann of Math. (2) **109** (1979), 415–476.
- [107] K. Zuo, *Factorizations of nonrigid Zariski dense representations of  $\pi_1$  of projective algebraic manifolds*, Invent. Math. **118** (1994), 37–46.

RIMS, Kyoto University, Kyoto, Japan

E-mail: takuro@kurims.kyoto-u.ac.jp



# Some mathematical aspects of tumor growth and therapy

Benoît Perthame

**Abstract.** Mathematical models of tumor growth, written as partial differential equations or free boundary problems, are now in the toolbox for predicting the evolution of some cancers, using model based image analysis for example. These models serve not only to predict the evolution of cancers in medical treatments but also to understand the biological and mechanical effects that are involved in the tissue growth, the optimal therapy and, in some cases, in their implication in therapeutic failures. The models under consideration contain several levels of complexity, both in terms of the biological and mechanical effects, and therefore in their mathematical description. The number of scales, from the molecules, to the cell, to the organ and the entire body, explains partly the complexity of the problem. This paper focusses on two aspects of the problem which can be described with mathematical models keeping some simplicity. They have been chosen so as to cover mathematical questions which stem from both mechanical laws and biological considerations. I shall first present an asymptotic problem describing some mechanical properties of tumor growth and secondly, models of resistance to therapy and cell adaptation again using asymptotic analysis.

**Mathematics Subject Classification (2010).** 35K55, 35B25, 76D27, 92C50, 92D25.

**Keywords.** Tumor growth, Hele-Shaw equation, free boundary problems, structured population dynamics, resistance to therapy.

## 1. Introduction

Since the paper of H. P. Greenspan [33] in 1972, an increasing mathematical activity has been developing, that creates new models, new numerical methods, new analysis of partial differential equations representing various aspects of tumor growth and therapy. This activity follows the National Cancer Act, usually called ‘war on cancer’, signed in 1971 and the awareness that the disease becomes a major health problem in many countries. Despite several decisive progresses, the many faces of the problem, and their complexity in terms of scales, agents and scientific background, explain that cancer remains a challenge for XXIst century medicine.

Interestingly enough, many aspects have lead to mathematical modeling and I would like to mention some of them. The molecular basis of tumors are mutations of cells, which are modeled by random processes [29], and which opens the route of molecular targets for drug design. The number of scales, from the molecule, to the cell and to the organ, also explains the complexity of the phenomena. Considering an assembly of cells, bridging gene activity to cell behavior, are possible with agent based methods [3, 13, 34] giving a detailed account

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

of tissue growth and organization. However, in these notes, I will consider continuous models, used for large populations of cells, and only solid tumors even if blood cancers have also led to a core of mathematical literature, see [2, 21] and the references therein.

Continuous models allow for numerical simulations at the scale of the organ and are used for predicting tumor progression in combination with medical imaging [23, 24, 43, 56]. These models can incorporate several features as angiogenesis, the process by which necrotic cells in the core of the tumor emit molecular signals attracting new vasculature, adhesion to the extracellular matrix and its degradation, metastasis, proliferative or quiescent or necrotic states of the cells; these features and many others are described in the several surveys available in the literature [4, 6, 7, 32, 33, 43, 54]. Therapy is also a main concern in this field and has been considered in the field of optimal control with, for instance, the questions of drug optimization, interactions between cell cycle and circadian cycle, [18, 19, 39].

In order to present both the impact of physical laws and biological aspects, these notes address two different aspects of tumor growth. Considering fluid mechanical aspects, section 2 describes one of the simplest models in the area and is followed, in section 3 by the derivation of a free boundary problem in the ‘stiff law-of-state’ limit. Then, we turn to an approach, based on asymptotic analysis, to a question related to therapy and resistance to drugs; this is section 4.

## 2. Mechanical aspects of tissue growth

Solid tumors grow under the effect of cell proliferation limited by several factors. Space availability, and the pressure induced by higher cell population, appears to be the first cause of growth limitation by contact inhibition [9, 10, 13, 53]. This can be included in the simplest models for a cell population density  $n(x, t)$  where pressure generates both movement and growth limitation, leading to write

$$\begin{cases} \frac{\partial}{\partial t} n + \operatorname{div}(nv) = nG(p), & x \in \mathbb{R}^d, t \geq 0, \\ n(x, t = 0) = n^0(x) \geq 0, \\ v(x, t) = -\nabla p(x, t), & p(x, t) \equiv \Pi_\gamma(n(x, t)) := n(x, t)^\gamma, \quad \gamma > 1. \end{cases} \quad (2.1)$$

The rule  $v(x, t) = -\nabla p(x, t)$  is a simplified version of Darcy’s law expressing isotropic and homogeneous friction with the surrounding environment. This expression for the velocity field means that cells are only pushed by mechanical forces (variants are mentioned later). The particular choice for the law-of-state  $\Pi_\gamma(n) := n^\gamma$  is made for simplicity, see considerations on this issue in [17]. Finally the growth term, the right hand side in (2.1), is of Lotka-Volterra type, and takes into birth and death of cells. Because pressure generates contact inhibition, we assume that the  $C^1$  function  $G(\cdot)$  satisfies

$$G(0) = G_M > 0, \quad G'(\cdot) < 0, \quad G(P_h) = 0, \quad \text{for some } G_M > 0, P_h > 0. \quad (2.2)$$

The name ‘homeostatic pressure’ has been proposed for  $P_h$  ([53]). At this stage it might also be useful to mention that dimensions  $d = 2$  is relevant for *in vitro* experiments on a dish and  $d = 3$  is relevant both *in vitro* and *in vivo*. As well known for the porous medium equation, one property of such partial differential equations is to describe solutions with

compact support than expand [58]. For our purpose here, this is enough and we do not bother with a bounded domain and associated boundary conditions. This feature is however relevant both for realistic models and numerics.

As far as existence is concerned, this equation is standard and is a semi-linear version of the ‘porous medium equation’, [58]. Therefore, several bounds are known under some assumptions on the initial data.

Now, we follow closely [49]. Because we are interested in the dependence on the parameter  $\gamma$  (and large values of it), we consider a family of initial data  $n_\gamma^0$  such that for some constant  $K^0$ ,

$$\int_{\mathbb{R}^d} n_\gamma^0 dx \leq K^0, \quad p_\gamma^0 := \Pi_\gamma(n_\gamma^0) \leq P_h, \quad \int_{\mathbb{R}^d} |\nabla n_\gamma^0| dx \leq K^0. \quad (2.3)$$

**Proposition 2.1.** *With assumptions (2.2)–(2.3), the solution of equation (2.1) satisfies  $n(x, t) \geq 0$  and*

$$\begin{aligned} \int_{\mathbb{R}^d} n(x, t) dx &\leq K^0 e^{G_M t}, \quad \int_{\mathbb{R}^d} |\nabla n(x, t)| dx \leq K^0 e^{G_M t}, \\ \int_0^T \int_{\mathbb{R}^d} |\nabla p(x, t)| dx dt &\leq C(T, P_h, K^0), \\ p(x, t) &\leq P_h, \quad \int_{\mathbb{R}^d} p(x, t) dx \leq P_h^{(\gamma-1)/\gamma} K^0, \\ \int_0^T \int_{\mathbb{R}^d} |\nabla p(x, t)|^2 dx dt &\leq \frac{1 + \gamma G_M T}{\gamma - 1} P_h^{(\gamma-1)/\gamma} K^0. \end{aligned}$$

*Proof.* The estimates for  $n$  are straightforward. For the TV bound, we just notice that, the equation for  $n$  can also be written

$$\frac{\partial}{\partial t} n - \Delta \Phi(n) = nG(p(x, t)), \quad \text{with } \Phi'(n) = n\Pi'_\gamma(n).$$

Therefore, the equation for  $w_i = \frac{\partial n(x, t)}{\partial x_i}$  is

$$\frac{\partial}{\partial t} w_i - \text{div}[\Phi'(n)\nabla w_i] = w_i G(p(x, t)) + nG'(p(x, t)) \frac{\partial p(x, t)}{\partial x_i},$$

and finally

$$\frac{\partial}{\partial t} |w_i| - \text{div}[\Phi'(n)\nabla |w_i|] = |w_i|G(p(x, t)) - n|G'(p(x, t))| \left| \frac{\partial p(x, t)}{\partial x_i} \right| \leq |w_i|G_M.$$

After integration and use of the Gronwall lemma, this gives the  $L^1$  estimate on the gradient of  $n$  and keeping the term with  $\left| \frac{\partial p}{\partial x_i} \right|$  gives the bound on the gradient of  $p$  (see [49] for details).

The second line of bounds in Proposition 2.1 follows from the equation on the pressure. Namely, we compute

$$\frac{\partial}{\partial t} p - n\Pi'(n)\Delta p - |\nabla p|^2 = n\Pi'(n)G(p(x, t)). \quad (2.4)$$

This equation is in the strong form, the maximum principle applies and gives the bound  $p \leq P_h$ . It gives the  $L^1$  control on  $p$  because

$$p = n^\gamma = nn^{\gamma-1} = np^{(\gamma-1)/\gamma} \leq nP_h^{(\gamma-1)/\gamma},$$

and it remains to apply the  $L^1$  control on  $n$ .

The  $L^2$  estimate on the gradient is better seen when identifying the pressure, as  $p = n^\gamma$  in (2.4), to find

$$\frac{\partial}{\partial t} p - \gamma p \Delta p - |\nabla p|^2 = \gamma p G(p). \tag{2.5}$$

Integrating by parts, we obtain, for  $T > 0$ ,

$$\int_{\mathbb{R}^d} [p(x, T) - p^0(x)] dx + (\gamma - 1) \int_0^T \int_{\mathbb{R}^d} |\nabla p|^2 dx dt \leq \gamma G_M \int_0^T \int_{\mathbb{R}^d} p(x, t) dx dt.$$

which, combined with the  $L^1$  estimate for  $p$  gives the last inequality. □

The bounds in Proposition 2.1 are fine to ensure compactness in space. It remains to prove estimates implying time compactness. An easy way is to notice that under the assumption that  $n^0$  is a subsolution, that is

$$-\operatorname{div}(n^0 \nabla \Pi(n^0)) \leq n^0 G(p^0(x)),$$

we have  $\frac{\partial}{\partial t} n^0 \geq 0$ . We may apply the same argument as for space derivatives and  $w = \frac{\partial}{\partial t} n$  satisfies

$$\frac{\partial}{\partial t} w - \operatorname{div}[\Phi'(n) \nabla w] = w G(p(x, t)) + n G'(p(x, t)) \gamma n^{\gamma-1} w,$$

an equation which gives us the property

$$\frac{\partial}{\partial t} n^0 \geq 0 \implies \frac{\partial}{\partial t} n \geq 0. \tag{2.6}$$

This property is very strong and shows one limitation of the model at hand. It is incompatible with the observations that the cell population decreases in the center of the tumor, the necrotic core. This effect, which typically occurs at the size of  $1mm^3$ , can be obtained when the effects of nutrients are included in the equation, see (2.8) below.

In this situation, which we call ‘well prepared initial data’, we conclude

$$\frac{d}{dt} \int_{\mathbb{R}^d} |w(x, t)| dx \leq G_M \int_{\mathbb{R}^d} |w(x, t)| dx,$$

and thus

$$\int_{\mathbb{R}^d} \left| \frac{\partial}{\partial t} n(x, t) \right| dx \leq \int_{\mathbb{R}^d} |\operatorname{div}(n^0 \nabla \Pi(n^0)) + n^0 G(p^0(x))| dx. \tag{2.7}$$

It is possible to improve these estimates and avoid the restrictive assumption that the initial data is a subsolution. We recall from [49] the

**Proposition 2.2.** *For a constant  $r_G$  depending only on  $G(\cdot)$ , the estimates hold, for all  $t > 0$ ,*

$$\frac{\partial}{\partial t} p(x, t) \geq -\gamma r_G p(x, t) \frac{e^{-\gamma r_G t}}{1 - e^{-\gamma r_G t}}, \quad \frac{\partial}{\partial t} n(x, t) \geq -r_G n(x, t) \frac{e^{-r_G t}}{1 - e^{-r_G t}}.$$

These inequalities allow for a fast transition at  $t = 0$  (the right hand side is singular then). They were initiated in [25] and are stronger than those in (2.6) because they do not assume any further assumption on the initial data than those in Proposition 2.1. A remarkable feature here, is that the semi-linear source term improves the usual inequalities for porous medium equations, which are recovered for  $r_G \rightarrow 0$ .

To conclude this section, we present some additional effects which are used in more realistic models of tumor growth. A possible additional ingredient is to take into account nutrients. Then we arrive to the model, also treated in details in [49]

$$\begin{cases} \partial_t n - \operatorname{div}(n \nabla p) = n \Phi(p, c), \\ \partial_t c - \Delta c = -n \Psi(p, c), \\ c(x, t) = c_B > 0 \quad \text{as } |x| \rightarrow \infty, \end{cases} \tag{2.8}$$

where  $c$  denotes the density of nutrients, and  $c_B$  the far field supply of nutrients (from blood vessels). The coupling functions  $\Phi, \Psi$  are assumed to be smooth and to satisfy the intuitive hypotheses

$$\begin{cases} \partial_p \Phi < 0, & \partial_c \Phi \geq 0, & \Phi(P_h, c_B) = 0, \\ \partial_p \Psi \leq 0, & \partial_c \Psi \geq 0, & \Psi(p, 0) = 0. \end{cases} \tag{2.9}$$

Variants are possible; for instance, we could assume that nutrients are released continuously from a vasculature or an other source [17], several nutrients (oxygen, glucose) can be considered. More generally, the formalism of multiphase fluids can be used in the present context [14, 52] in order to represent the complexity of cell surrounding.

Another ingredient is to take into account active movement of cells and not only their passive movement under pressure forces. This leads to write the model, which is analyzed in [50],

$$\partial_t n - \operatorname{div}(n \nabla p) - \nu \Delta n = nG(p). \tag{2.10}$$

As mentioned in the introduction, many other biological aspects are possible, which have led to mathematical models and questions, and which we do not mention here.

### 3. The Hele-Shaw asymptotic and free boundary formulation

As long as cells are well separated, the pressure forces are negligible. When the population density increases, there is a maximum possible compaction which cannot be exceeded. To represent this effect with a fast transition, the simplest formalism is to consider the limit as  $\gamma \rightarrow \infty$  in the equation of state, see (2.1), and which we call the *stiff pressure asymptotic*. This type of modeling is mostly used in practical use of cancer models and software development [22, 23, 26, 32, 43, 54].

**3.1. Free boundary problem.** This limit results in a model that generalizes the Hele-Shaw equation of fluid mechanics and which is usually seen as a free boundary problem. The tumor occupies a domain  $\Omega(t)$ , healthy cells fill the space outside  $\Omega(t)$ . The boundary  $\partial\Omega(t)$  of the domain  $\Omega(t)$  is moving with the velocity

$$v_\infty(x, t) = -\nabla p_\infty(x, t) \tag{3.1}$$

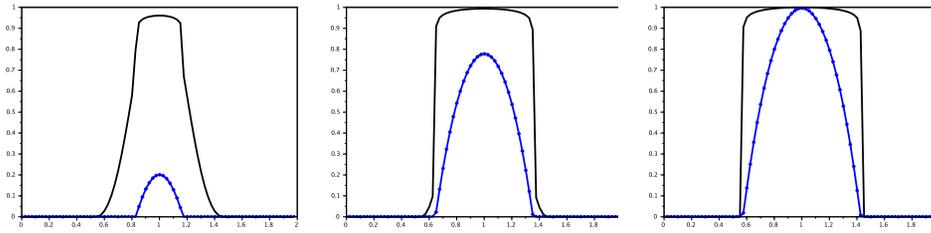


Figure 3.1. Numerical simulations at three different times of the system (2.1) with  $\gamma = 40$  and  $G(p) = 5(2 - p)$ . The density  $n$  is plotted in solid line whereas the pressure  $p$  is represented with  $-+-$  (smoother curve).

where the pressure field is computed thanks to the equation

$$\begin{cases} -\Delta p_\infty = G(p_\infty) & x \in \Omega(t), \\ p_\infty = 0 & \text{on } \partial\Omega(t). \end{cases} \tag{3.2}$$

In order to define this dynamic, some smoothness of the free boundary is necessary. Such a property has been widely studied, see [30, 32] and the references therein. An alternative is to set this problem in the general framework of viscosity solutions with a correct viscosity condition on the interface, see [36, 37]. Surface tension may also be included [1, 30, 32], then the Dirichlet boundary condition has to be changed to  $p_\infty = a\kappa(x, t)$  with  $a$  a parameter and  $\kappa$  the mean curvature.

As we mentioned earlier, the biophysical modeling gives growth terms  $G$  that depend on  $p$ , and not on  $n$  as in [8] for instance. Remarkably, this property allows us to extend nicely the usual Hele-Shaw theory and recover the semi-linear elliptic equation (3.2). A recent interest for the Hele-Shaw equation also arises in other fields of mathematics with the stochastic Loewner evolutions, Laplacian growth, diffusion limited aggregation, etc

**3.2. Weak formulation.** Besides the free boundary formulation, there is also a weak formulation of the limit  $\gamma \rightarrow \infty$  in the equation (2.1). This limit gives a more general setting allowing a ‘pretumor zone’ where healthy and tumor cells are present in a mixed state. This weak formulation was derived in [49] and leads to the equation

$$\begin{cases} \frac{\partial}{\partial t} n_\infty - \operatorname{div}(n_\infty \nabla p_\infty) = n_\infty G(p_\infty(x, t)), & x \in \mathbb{R}^d, t \geq 0, \\ n_\infty(x, t = 0) = n_\infty^0(x) \geq 0, \\ p_\infty(1 - n_\infty) = 0, & 0 \leq n_\infty \leq 1. \end{cases} \tag{3.3}$$

In other words, when  $n_\infty < 1$  then  $p_\infty = 0$ . Consequently,  $n_\infty$  and  $p_\infty$  are so weakly related that their dynamics can be somewhat independent. Nevertheless, a remarkable property is that the weak solution of (3.3) is unique (see [49]).

To present the result, we now insert the index  $\gamma$  to the notations  $n$  and  $p$  for the solutions of (2.1). The following result holds

**Theorem 3.1** (Hele-Shaw limit, [49]). *With the assumptions of Proposition 2.1, as  $\gamma \rightarrow \infty$ , we have*

$$n_\gamma \rightarrow n_\infty \leq 1, \quad p \rightarrow p_\infty \leq P_h \quad \text{a.e. in } \mathbb{R}^d \times (0, \infty),$$

$$\begin{aligned} \nabla p_\gamma \rightharpoonup \nabla p_\infty \quad & \text{in } L^2(\mathbb{R}^d \times (0, T))\text{-weak}, \quad \forall T > 0, \\ \frac{\partial}{\partial t} n_\infty \geq 0, \quad & \frac{\partial}{\partial t} p_\infty \geq 0. \end{aligned}$$

The limit of equation (2.1) is equation (3.3).

Notice that, from the BV (bounded variation) properties of  $n_\gamma$  and  $p_\gamma$  in Proposition 2.1, we derive strong compactness. We also conclude that

$$n_\infty \in L^\infty((0, T); L^1 \cap L^\infty(\mathbb{R}^d)), \quad p_\infty \in L^\infty((0, T) \times \mathbb{R}^d) \cap L^1((0, T) \times \mathbb{R}^d)$$

and that, as measures although we use the notation of  $L^1$  functions,

$$|\nabla n_\infty(x, t)| \text{ and } |\nabla p_\infty(x, t)|$$

are bounded with

$$\int_{\mathbb{R}^d} |\nabla n_\infty(x, t)| dx \leq K^0 e^{G_M t}, \quad \int_0^T \int_{\mathbb{R}^d} |\nabla p_\infty(x, t)| dx dt \leq C(T, P_h, K^0).$$

The other results follow immediately. For example, because

$$n_\gamma p_\gamma = n^{\gamma+1} = p_\gamma^{\frac{\gamma+1}{\gamma}},$$

and passing to the strong limits, we find in the limit the relation  $p_\infty(1 - n_\infty) = 0$ . Another property follows immediately from the same argument; because  $n_\gamma \nabla p_\gamma = \nabla p_\gamma^{\frac{\gamma+1}{\gamma}}$ , we find the relation

$$n_\infty \nabla p_\infty = p_\infty.$$

In other words, the equation on  $n_\infty$ , in (3.3), can also be written

$$\frac{\partial}{\partial t} n_\infty - \Delta p_\infty = n_\infty G(p_\infty(x, t)).$$

This is the form used in [49] to prove uniqueness of weak solutions.

A more difficult result is the derivation of the ‘complementary relation’, (3.4) below, which is equivalent to the strong convergence of  $\nabla p_\gamma$ .

**Theorem 3.2** (Complementary relation). *Additionally to Theorem 3.1, one also has*

$$\nabla p_\gamma \rightarrow \nabla p_\infty \quad \text{in } L^2_{\text{loc}}(\mathbb{R}^d \times (0, \infty))\text{-strong},$$

The ‘complementary relation’ also holds

$$p_\infty (\Delta p_\infty + G(p_\infty)) = 0 \quad \text{in } \mathcal{D}(\mathbb{R}^d \times (0, \infty)). \tag{3.4}$$

The complementary relation (3.4) is not an obstacle problem (a sign is incompatible) and the solution is not unique. It is a weak version of the equation (3.2) with

$$\Omega(t) = \{ p_\infty(x, t) > 0 \}, \tag{3.5}$$

as set which evolution cannot be deduced from (3.4), but from the weak formulation (3.3).

the meaning, in distributions, of (3.4) is that for all smooth test functions  $\varphi$  with compact support, it holds

$$\int_{\mathbb{R}^d \times (0, \infty)} \varphi(x, t) [-|\nabla p_\infty|^2 + p_\infty G(p_\infty)] - \int_{\mathbb{R}^d \times (0, \infty)} p_\infty \nabla \varphi \cdot \nabla p_\infty = 0$$

which makes sense with the available regularity for  $p_\infty$  in Proposition 2.1.

The proof of Theorem 3.2 relies on a functional analysis argument which uses the  $L^\infty$  control from below for  $\frac{\partial}{\partial t} n_\gamma \geq 0$  as given in Proposition 2.2.

**3.3. From the weak formulation to the free boundary statement.** To begin with, notice that  $\mathbf{I}_{\{\Omega(t)\}} = \mathbf{I}_{\{n_\infty(x,t)=1\}}$ . Indeed, on the one hand,  $\mathbf{I}_{\{\Omega(t)\}} \subset \mathbf{I}_{\{n_\infty(x,t)=1\}}$ . On the other hand, when  $p_\infty = 0$ , then from (3.3), we conclude that  $\frac{\partial}{\partial t} n_\infty = n_\infty G_M$ , which means that we cannot have  $n_\infty(x, t) = 1$  otherwise  $n_\infty$  would continue to grow thus contradicting the bound  $n_\infty(x, t) \leq 1$ .

Therefore, when  $n_\infty(x, t)$  takes the values 0 or 1 only, then we have

$$n_\infty(x, t) = \mathbf{I}_{\{\Omega(t)\}}. \tag{3.6}$$

In this situation and assuming some smoothness for  $\Omega(t)$ , it is easy to derive the Hele-Shaw free boundary formulation mentioned in Section 3.1. This is written in details (and in more generality in the sense below) when  $\Omega(t)$  is a ball in [49], then one can establish precisely the speed of the free boundary given by (3.1).

However, the weak formulation contains more than the free boundary statements (3.1), (3.2) which only holds true when initially  $n^0 = \mathbf{I}_{\{\Omega(t=0)\}}$  so as to ensure (3.6). One can formally see this, because in the interior of  $\Omega(t)$ , we can write  $\frac{\partial}{\partial t} n_\infty = 0$  and thus the weak formulation (3.3) gives immediately the elliptic equation (3.2). But, if there is a zone where  $n^0 < 1$ , then we still have  $n_\infty(x, t) < 1$  for some time. In this space-time zone, we have  $p_\infty = 0$  and (3.3) is reduced to the simple differential equation

$$\frac{\partial}{\partial t} n_\infty = n_\infty G_M.$$

A numerical simulation, illustrating this interpretation is displayed in Figure 3.1.

A similar, but less complete, theory can be carried out for the case with active motion (2.10), see [50], and for the system with nutrient (2.8) and furthermore, the permanent shape, given by a traveling wave can be written exactly [51].

### 4. Adaptation and resistance to drugs

Besides mechanical aspects which we have presented so far, mathematical models of tumor growth also deal with questions which are more connected to biology than mechanics, and resistance to treatment is a typical example. The subject of resistance is considered presently as one of the challenges is medical treatment (see [38, 40, 41, 57] and the references therein).

A possible modeling of this phenomena is related to Darwinian evolution and to selection of the fittest traits. A subject that bridges probability [16] for finite populations, game theory as introduced by J. Maynard Smith and PDEs, the formalism we use below.

**4.1. Population adaptive dynamic.** In the view of [40, 41], cells are assumed to carry a resistance phenotype  $y \in [0, 1]$ . In the simplest description, one considers the population density  $n(y, t)$ , this is usually called a *structured population*, [47]. One can postulate an equation for the dynamic of  $n(y, t)$ , expressing birth and death of cells. A general, yet simple, formalism is, following [42, 44, 48], to write a type of Lotka-Volterra equation

$$\frac{\partial}{\partial t} n(y, t) = n(y, t)R(y, \rho(t)) + \mu \Delta n(y, t), \quad \rho(t) = \int_0^1 n(y, t) dy,$$

with Neuman boundary conditions (these are somewhat artificial but simplify the presentation). The diffusion term stands for mutations; several other forms are possible as integral operators [5] and, as well as diffusion, can be derived from stochastic individual models [16]. Again the choice of diffusion is made for simplicity. The term  $R(y, \rho)$  represents the growth rate (death and birth), an example being

$$R(y, \rho) = b(y) - \rho k(y) - d(y) c_{th}, \tag{4.1}$$

with  $b(\cdot)$  the intrinsic division rate,  $d(\cdot)$  the death rate induced by the therapeutic drug given with the concentration  $c_{th}$ . Finally,  $k(\cdot)$  represents the death rate due to competition, for space and nutrients, with all the cells whatever is their resistance level. Therefore, in the general setting, we assume that, for some constant  $\alpha > 0$ ,

$$\frac{\partial}{\partial \rho} R(y, \rho) \leq -\alpha < 0.$$

Then, according to the interpretation of  $y$  as a resistance gene expression, we can assume some kind of resource allocation. When a cell uses energy to generate resistance, there is less energy for the cell division cycle, therefore we have

$$b'(\cdot) < 0, \quad d'(\cdot) < 0, \quad k'(y) < 0,$$

the last assumption means that resistant cells are also better competitors (an assumption that could be released by introducing another phenotypic trait).

The main qualitative property of solutions is better stated with a renormalization of time according to the scale  $\mu = \varepsilon^2 \mu_0$ ,  $t_{new} = \varepsilon t_{old}$ , with  $t_{old}$  the generation time,  $t_{new}$  the evolution time. This renormalization leads to re-write the equation on  $n(y, t)$  as

$$\varepsilon \frac{\partial}{\partial t} n_\varepsilon(y, t) = n_\varepsilon(y, t)R(y, \rho_\varepsilon(t)) + \varepsilon^2 \mu_0 \Delta n_\varepsilon(y, t), \quad \rho_\varepsilon(t) = \int_0^1 n_\varepsilon(y, t) dy. \tag{4.2}$$

This rescaling is standard in parabolic equation, in particular because it is the basis for deriving various front motions, see [31, 55] for instance.

The analysis carried out in [28, 42, 44, 47, 48] leads to use two main tools. The first one is a uniform Total Variation bound (TV in short) on  $\rho_\varepsilon(t)$

$$0 < c \leq \rho_\varepsilon(t) \leq C, \quad \int_0^T |\dot{\rho}_\varepsilon(t)| dt \leq C. \tag{4.3}$$

The lower bound expresses non-extinction and can be recovered a posteriori, it is however convenient to have it proved directly when this is possible. The BV bound is needed for nonlinear dependence on  $\rho$  in  $R(x, \rho)$ ; it is not fundamental for the case (4.1) for instance.

The second tool is the WKB change of unknown

$$u_\varepsilon(y, t) = \varepsilon \ln (n_\varepsilon(y, t))$$

and according to the observation of natural selection, as in standard in adaptive dynamics [27], the population should be highly concentrated around the fittest trait (think of a Gaussian). Then, initially one assumes that for some  $\bar{y}^0 \in (0, 1)$ ,

$$\begin{cases} n_\varepsilon^0(y) \xrightarrow{\varepsilon \rightarrow 0} \delta(y - \bar{y}^0) \text{ (weakly),} & u_\varepsilon^0 \text{ is bounded in } \text{Lip}(0, 1), \\ u_\varepsilon^0 \xrightarrow{\varepsilon \rightarrow 0} u^0, & \max_{0 < y < 1} u^0(y) = u^0(\bar{y}^0) \text{ (strict maximum).} \end{cases} \tag{4.4}$$

This initial concentration effect remains true for all times under structural assumptions on  $R$  (e.g. assuming that  $R$  is monotonic in  $y$  as in [48], or that  $R$  is concave in  $y$  as in [42]). Then, it is established that

$$\begin{cases} \rho_\varepsilon(t) \xrightarrow{\varepsilon \rightarrow 0} \bar{\rho}(t) \in L^\infty \cap TV(0, +\infty), & a.e. \\ n_\varepsilon(y, t) \xrightarrow{\varepsilon \rightarrow 0} \bar{\rho}(t)\delta(y - \bar{y}(t)). \end{cases} \tag{4.5}$$

The next question is to characterize the dynamic of the two unknowns  $\bar{\rho}(t)$  and  $\bar{y}(t)$ . The answer is expressed through the limiting behavior of  $u_\varepsilon(t)$ . Still under technical assumptions depending on the case at hand (monotonic or concave function  $R$ ), one has

$$u_\varepsilon(y, t) \xrightarrow{\varepsilon \rightarrow 0} u(y, t) \quad \text{uniformly, locally in time,}$$

and the functions  $u(y, t)$  and  $\bar{\rho}(t)$  satisfy the *constrained Hamilton-Jacobi equation*

$$\begin{cases} \frac{\partial}{\partial t} u(y, t) = R(y, \bar{\rho}(t)) + \mu_0 |\nabla u|^2, & 0 < x < 1, t \geq 0, \\ \max_{0 \leq y \leq 1} u(y, t) = 0 = u(\bar{y}(t), t), \\ u(y, t = 0) = u^0(y), \end{cases} \tag{4.6}$$

with Neuman boundary conditions (note that only cases in the full line have been studied so far). The interpretation is as follows:  $\bar{\rho}(t)$  is a Lagrange multiplier associated with the algebraic constraint that  $\max_y u(y, t) = 0$ . For this reason, the usual property of contraction in  $L^\infty$  of Hamilton-Jacobi equations is lost in the case with a constraint. However Lipschitz bounds for  $u(y, t)$  are still available (and motivate the corresponding assumption in (4.4)) and are enough to prove existence of a viscosity solution. Uniqueness is only known in the particular case when  $R$  has a specific form as in (4.1), see [48].

**4.2. Canonical equation and evolutionary stable distribution.** One can go further (to the expense of more regularity on  $u$ , a condition that can be proved with concavity assumptions on  $R$ ) and establish the form of canonical equation ([27]) as follows:

$$\begin{cases} R(\bar{y}(t), \bar{\rho}(t)) = 0, \\ \dot{\bar{y}}(t) = (-D^2 u(\bar{y}(t), t))^{-1} \cdot D_y R(\bar{y}(t), \bar{\rho}(t)). \end{cases}$$

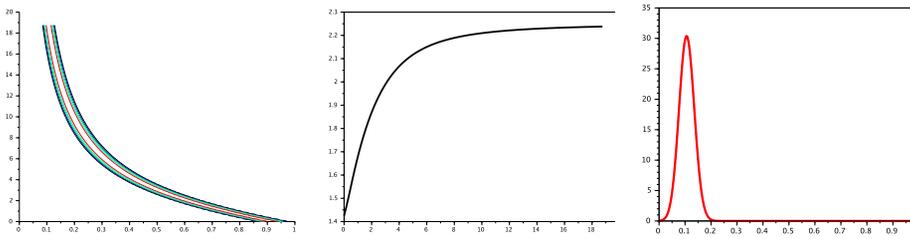


Figure 4.1. **No therapy.** Solution of system (4.2) with  $\mu_0 = 0$  for  $R(y, \rho) = \frac{3}{2} - y + \frac{\rho}{1.5+y}$ , and departing from a distribution concentrated near  $y = .95$  as a Gaussian with parameter  $\varepsilon = 0.02$ . Left: the isovalues of  $n(y, t)$ , abscissae are  $y$  and ordinates are  $t$ . Center: the function  $t \mapsto \rho(t)$ . Right: the distribution  $n(y, t_{\text{final}})$  at  $t_{\text{final}} = 20$ , which concentrates at the point  $y = \bar{y}_\infty = 0$ .

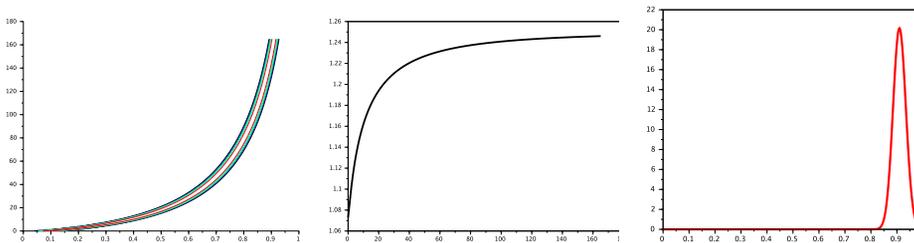


Figure 4.2. **With therapy.** Solution of system (4.2) with  $\mu_0 = 0$  for  $R(y, \rho) = \frac{3}{2} - y + \frac{\rho}{1.5+y} - C_{\text{Tox}}(1-x)$ , and departing from a distribution concentrated near  $y = .05$  as a Gaussian with parameter  $\varepsilon = 0.02$ . Left: the isovalues of  $n(y, t)$ , abscissae are  $y$  and ordinates are  $t$ . Center: the function  $t \mapsto \rho(t)$ . Right: the distribution  $n(y, t_{\text{final}})$  at  $t_{\text{final}} = 180$ , which concentrates at the point  $y = \bar{y}_\infty = 1$ .

Because  $R$  is invertible in  $\rho$ , the first equation gives  $\bar{\rho}(t)$  as a function of  $\bar{y}(t)$ , and then the ordinary differential equation for  $\bar{y}(t)$  is in closed form when  $u$  is known.

This form of a canonical equation is not explicit as long as  $u$  is not computed, but can however give some information on the sign of  $\dot{\bar{y}}(t)$  and on the long term dynamics of  $\bar{y}(t)$ . When a steady state is attained, it is called the evolutionary stable distribution [35] (ESD in short), a notion closely related the evolutionary stable strategy in adaptive dynamics [27].

For instance, for the case of (4.1) with no therapy,  $c_{\text{th}} = 0$ , and weak competition compared to proliferation ( $|b'|$  large compared to  $|k'|$ ), then we find

$$\dot{\bar{y}}(t) \leq 0 \quad \text{as } b' < 0,$$

because  $D^2u(\bar{y}(t), t) \leq 0$  at a maximum point. And we conclude that less resistant cells are selected. The dynamics will stop at an ESD  $(\bar{\rho}_\infty, \bar{y}_\infty)$  which achieves both conditions

$$R(\bar{y}_\infty, \bar{\rho}_\infty) = 0, \quad D_y R(\bar{y}_\infty, \bar{\rho}_\infty) = 0, \tag{4.7}$$

and this value  $\bar{y}_\infty$  corresponds to a maximum of  $R$  in  $y$ . Here we should emphasize that because  $\mu_0 = 0$ , the restriction that  $y \in (0, 1)$  is only useful for the biological interpretation. Mathematically, the dynamics might lead  $y(t)$  to become negative. This ESD is illustrated by Figure 4.1; with the rate function  $R(y, \rho)$  used in this figure, one can readily check that  $\bar{y}_\infty = 0$ .

For a strong drug concentration,  $c_{th}$  large, then one finds on the contrary

$$\dot{\bar{y}}(t) \geq 0 \quad \text{as} \quad -d' > 0$$

and thus resistant cells are selected which escape therapy if  $\rho(t)$  does not vanish. Here, it might occur that, for  $c_{th}$  large enough, then  $\rho(t)$  vanishes and the lower bound in (4.3) fails; this can be interpreted as recovery. In such a case, the constraint in the constrained Hamilton-Jacobi equation (4.6) does not hold because the Lagrange multiplier is fixed at  $\rho(t) = 0$ . The Figure 4.2 illustrates a case where resistance occurs and the ESD, characterized by (4.7), is for  $\bar{y}_\infty = 1$ .

The effect of a multi-therapy to prevent resistance can be included in the model and gives rise to the following extension

$$\frac{\partial}{\partial t} n(y, t) = n(y, t) \left[ \frac{b(y)}{1 + c_{Stat}} - \rho(t) k(y) - d(y) c_{Tox} \right], \quad \rho(t) = \int_0^1 n(y, t) dy. \tag{4.8}$$

Here two types of effects are taken into account;  $c_{Tox}$  represents the cytotoxic drugs which induce apoptosis (usually by DNA damage during the Synthesis phase of the cell cycle), and  $c_{Stat}$  represents cytostatic effects which slow down the cell cycle (for instance using molecules that inhibit cyclines). Then the question is to determine the optimal scheduling  $c_{Stat}(t)$ ,  $c_{Tox}(t)$  with constraints on total toxicity. This is studied in [20] with the constraints to keep a high enough population of healthy cells.

**4.3. Space structure and heterogeneity.** Selection of a monomorphic population (that means a single Dirac mass) as derived before is compatible with the Gause competitive exclusion principle which is used in ecology; with  $N$  environmental variables, a bioreactor can sustain  $N$  interacting species, [15]. Here  $N = 1$  and the environmental variable is just measured by the total population. However, genetic tests show a wide heterogeneity in tumor cells. Several explanations are possible as random mutations which generate a peaked distribution  $n(y, t)$  but not exactly a Dirac mass; the parameter  $\varepsilon$  is small but not zero. Another possible explanation is spacial heterogeneity within tumor environment due to local availability of nutrients. In order to write corresponding equations, one should describe population densities  $n(x, y, t)$  where  $x$  stands for the position and  $y$  for the phenotypical trait. With space and trait, to derive statements similar to those in (4.5)-(4.6) is a much more recent topic, see [11, 12, 46], with unexpected outcomes and difficulties. New phenomena as accelerating waves occur and mathematically, a priori bounds are more complicated because they should reflect the  $L^1$  theory in the trait and the  $L^\infty$  theory in space.

The model proposed in [41] contains aspects coming both from the spatial model with nutrient (2.8) and the evolutionary aspects introduced in Section 4.1. To begin with, we present a simpler version, taken from [45], which explains the expected behavior of the solutions. We denote by  $n_\varepsilon(y, x, t)$  the population density of cells which are located at the position  $x$ , with the trait  $y$ . To simplify we choose  $y \in (0, 1)$  and  $x \in \mathbb{R}$  to simplify the statements. The spatial dependence determines local conditions for trait adaptation, according to available nutrient concentration  $c(x, t)$ . Following the rescaling proposed in (4.2), we write

$$\varepsilon \partial_t n_\varepsilon(y, x, t) = [r(y)c_\varepsilon(t, x) - d(y)(1 + \varrho_\varepsilon(x, t))] n_\varepsilon(y, x, t), \quad x \in \mathbb{R}, \quad 0 < y < 1, \quad t \geq 0, \tag{4.9}$$

$$\frac{\partial}{\partial t} c_\varepsilon - \Delta_x c_\varepsilon(x, t) + [\varrho_\varepsilon(x, t) + \lambda] c_\varepsilon(x, t) = \lambda c_B, \quad x \in \mathbb{R}, t \geq 0, \quad (4.10)$$

$$\varrho_\varepsilon(x, t) = \int n_\varepsilon(y, x, t) dx, \quad x \in \mathbb{R}, t \geq 0. \quad (4.11)$$

In other words, we have chosen  $k(y) = d(y)$  in (4.1), neglected mutations and added a parameter  $x$  which dependency is ruled by a parabolic PDE. To handle the asymptotic behavior in (4.9), the main difficulty is to find strong estimates for  $\varrho_\varepsilon(x, t)$ . Uniform  $L^\infty$  bounds are immediate but strong compactness, as derived from (4.3) in the  $x$ -independent case, are not available. With technical assumptions that we skip here, it is proved in [45], that there is  $\varrho(y, t)$ ,  $X(y, t)$  such that

$$c_\varepsilon(x, t) \xrightarrow{\varepsilon \rightarrow 0} c(x, t) \quad \text{locally uniformly,}$$

$$\varrho_\varepsilon(x, t) \xrightarrow{\varepsilon \rightarrow 0} \varrho(x, t) \quad \text{pointwise,}$$

$$n_\varepsilon(y, x, t) \xrightarrow{\varepsilon \rightarrow 0} \varrho(x, t) \delta(y - Y(x, t)) \quad \text{weakly in measures.}$$

A qualitative consequence is heterogeneity which is expressed by the phenotypes  $Y(x, t)$ ,  $x \in \mathbb{R}$ , which are represented at a time  $t$ .

To be closer to the case of tumor treatment, the system used in [41] includes additional matter. The space variable represents the distance to the center, effect of therapeutic drugs are included (following the ideas leading to the equation (4.8)), and both nutrients and therapy are delivered from a vasculature on the boundary of the tumor.

## 5. Conclusion

One should keep in mind that mathematical biology is not a recent subject. It has a long record of success as the Lotka-Volterra equations in ecology, statistics and random processes in genetics, the Turing instability for pattern formation and developmental biology, the Hodgkin-Huxley system for electric pulse propagation along nerves, the Keller-Segel system for cell chemotaxis, and many others. Subjects as epidemiology, population genetics, neuroscience use mathematical models for a long time. Biofluids, biomechanics are now well established subjects with applications to medicine and medical industry. Even though more recent, mathematics motivated by questions around tumor growth are now numerous and a search on publications data basis shows a fast growing activity in the field. This fast development, can be observed in many other fields of life sciences under two effects. Biologists have now access to new experimental devices giving enormous quantities of data as images; data analysis is needed to handle them and mathematical modeling is needed to give sense to them. Physicists have entered the field massively and have now access to simplified living systems; it might be simpler for mathematicians to speak with them. However, because of the specificities of the living matter, classical models must be revisited with new variants. But new questions, on new models, also appear which require to develop new mathematical tools. This paper is an attempt to show these two faces.

**Acknowledgements.** Sorbonne Universités, CNRS, INRIA-Paris-Rocquencourt, Institut Universitaire de France.

## References

- [1] Nicholas D. Alikakos, Peter W. Bates, and Xinfu Chen, *Convergence of the Cahn-Hilliard Equation to the Hele-Shaw Model*, Arch. Rational Mech. Anal. **128** (1994), 165–205
- [2] M. Adimy, F. Crauste, and L. Pujo-Menjouet, *On the stability of a maturity structured model of cellular proliferation*, Discret. Cont. Dyn. Sys. Ser. A **12**(3) (2005), 501–522.
- [3] A. Anderson, M. A. J. Chaplain, and K. Rejniak. *Single-cell-based models in biology and medicine*, Birkhauser, Basel, 2007.
- [4] R. Araujo, D. McElwain, *A history of the study of solid tumour growth: the contribution of mathematical models*, Bull Math Biol **66** (2004), 1039–1091.
- [5] G. Barles, S. Mirrahimi and B. Perthame, *Concentration in Lotka-Volterra parabolic or integral equations: a general convergence result*, Methods Appl. Anal. **16**(3) (2009), 321–340.
- [6] N. Bellomo, N. K. Li., and P. K. Maini, *On the foundations of cancer modelling: selected topics, speculations, and perspectives*, Math. Models Methods Appl. Sci. **4** (2008), 593–646.
- [7] N. Bellomo and L. Preziosi, *Modelling and mathematical problems related to tumor evolution and its interaction with the immune system*, Math. Comput. Model. **32** (2000), 413–452.
- [8] Philippe Bénilan and Noureddine Igbida, *The mesa problem for Neumann boundary value problem*, J. Differential Equations **196** (2004), 301–315.
- [9] M. Bertsch, D. Hilhorst, H. Izuhara and M. Mimura, *A nonlinear parabolic-hyperbolic system for contact inhibition of cell-growth*, Diff. Eqs. Appl. **4** (2012), 137–157.
- [10] M. Bertsch, M. Mimura and T. Wakasa, *Modeling contact inhibition of growth: Traveling waves*, Networks and Heterogeneous Media **8** (2013), 131–147.
- [11] E. Bouin, V. Calvez, N. Meunier, S. Mirrahimi, B. Perthame, G. Raoul, and R. Voituriez, *Invasion fronts with variable motility: phenotype selection, spatial sorting and wave acceleration*, C. R. Math. Acad. Sci., Paris, **350**(15-16) (2012), 761–766.
- [12] E. Bouin and S. Mirrahimi, *A Hamilton-Jacobi approach for a model of population structured by space and trait*, Preprint hal-00849406, arXiv:1307.8332, 2013.
- [13] H. Byrne and D. Drasdo, *Individual-based and continuum models of growing cell populations: a comparison*, J. Math. Biol. **58** (2009), 657–687.
- [14] H. M. Byrne, J. R. King, D. L. S. McElwain, and L. Preziosi, *A two-phase model of solid tumor growth*, Appl. Math. Lett. **16** (2003), 567–573.
- [15] N. Champagnat and P.-E. Jabin, *The evolutionary limit for models of populations interacting competitively with many resources*, J. Differential Equations **251** (2011), 176–195.
- [16] N. Champagnat, R. Ferrière, and S. Méléard, *From Individual Stochastic Processes to Macroscopic Models in Adaptive Evolution*, Stochastic Models **24**, S1, (2008), 2–44.
- [17] P. Ciarletta, L. Foret, and M. Ben Amar, *The radial growth phase of malignant melanoma: multi-phase modelling, numerical simulations and linear stability analysis*, J. R. Soc., Interface **8** (2011), no 56, 345–368.

- [18] J. Clairambault, S. Gaubert, and T. Lepoutre, *Circadian rhythm and cell population growth*, *Mathematical and Computer Modelling* **53**(7-8) (2011), 1558–1567.
- [19] J. Clairambault, S. Gaubert, and B. Perthame, *Comparison of the Perron and Floquet eigenvalues in monotone differential systems and age structured equations*, *C. R. Acad. Sc., Paris*, **345**(10) (2007), 549–555.
- [20] J. Clairambault, A. Lorz, and E. Trélat, *Optimal control of cancer chemotherapies to overcome drug resistance*, Work in progress.
- [21] C. Colijn and M. C. Mackey, *Bifurcation and bistability in a model of hematopoietic regulation*, *SIAM J. App. Dynam. Sys.* **6**(2) (2007), 378–394.
- [22] T. Colin, D. Bresch, E. Grenier, B. Ribba, and O. Saut, *Computational modeling of solid tumor growth: the avascular stage*, *SIAM Journal of Scientific Computing* **32**(4) (2010), 2321–2344.
- [23] T. Colin, A. Iollo, D. Lombardi and O. Saut, *System identification in tumor growth modeling using semi-empirical eigenfunctions*, *Mathematical Models and Methods in Applied Sciences*, Vol. 22, No. 6, (2012), 1250003, (30 pp).
- [24] F. Cornelis, O. Saut, P. Cumsille, D. Lombardi, A. Iollo, J. Palussière, and T. Colin, *In vivo mathematical modeling of tumor growth from imaging date: Soon to come in the future?*, *Diagnostic and Interventional Imaging*, **94**(6), (2013), 593–600.
- [25] M. G. Crandall and M. Pierre, *Regularizing effects for  $u_t = \Delta\phi(u)$* , *Trans. Am. Math. Soc.*, Vol. 274, no 1, (1982), 159–168.
- [26] V. Cristini, J. Lowengrub, and Q. Nie, *Nonlinear simulations of tumor growth*, *J. Math. Biol.* **46** (2003), 191–224.
- [27] O. Diekmann, *A beginner's guide to adaptive dynamics*, In Rudnicki, R. (Ed.), *Mathematical modeling of population dynamics*, Banach Center Publications Vol. 63, (2004), 47–86.
- [28] O. Diekmann, P.-E. Jabin, S. Mischler, and B. Perthame, *The dynamics of adaptation: an illuminating example and a Hamilton-Jacobi approach*, *Th. Pop. Biol.* **67**(4) (2005), 257–271.
- [29] Richard Durrett, *Probability Models for DNA Sequence Evolution*, Springer, 2nd edition, 2008.
- [30] Joachim Escher and Gieri Simonett, *Classical solutions for Hele-Shaw models with surface tension*, *Adv. Differential Equations* **2**(4) (1997), 619–642.
- [31] L. C. Evans and P. E. Souganidis, *A PDE approach to geometric optics for certain semilinear parabolic equations*, *Indiana Univ. Math. J.* **38**(1) (1989), 141–172.
- [32] A. Friedman, *A hierarchy of cancer models and their mathematical challenges*, *DCDS(B)* **4**(1) (2004), 147–159.
- [33] H. P. Greenspan, *Models for the growth of a solid tumor by diffusion*, *Stud. Appl. Math.* **51** (1972), no. 4, 317–340.
- [34] S. Hoehme and D. Drasdo, *A cell-based simulation software for multi-cellular systems*, *Bioinformatics*, **26**(20), (2010), 2641–2642.
- [35] P.-E. Jabin and G. Raoul, *On selection dynamics for competitive interactions*, *J. Math. Biol.* **63**(3) (2011), 493–517.
- [36] I. C. Kim, *Uniqueness and existence results on viscosity solutions of the Hele-Shaw*

- and the Stefan problems, *Arch. Rat. Mech. Anal.* **168** (2003), no. 4, 299–328.
- [37] I. C Kim and A. Mellet, *Homogenization of a Hele-Shaw problem in periodic and random media*, *Arch. Ration. Mech. Anal.* **194** (2009), no. 2, 507–530.
- [38] M. Kimmel and A. Świerniak, *Control theory approach to cancer chemotherapy: benefiting from phase dependence and overcoming drug resistance*, pp. 185–221, in *Tutorials in Mathematical Biosciences III*, Lect. Notes Math. 1872, A. Friedman ed., Springer, 2006.
- [39] U. Ledzewicz and H. Schaettler, *Singular controls and chattering arcs in optimal control problems arising in biomedicine*, *Control and Cybernetics* **38** (2009), 1501–1523.
- [40] A. Lorz, T. Lorenzi, M. E. Hochberg, J. Clairambault, and B. Perthame, *Populational adaptive evolution, chemotherapeutic resistance and multiple anti-cancer therapies*, *ESAIM: Mathematical Modelling and Numerical Analysis*, Volume 47 (2), 2013, 377–399, DOI 10.1051/m2an/2012031.
- [41] A. Lorz, T. Lorenzi, J. Clairambault, A. Escargueil, and B. Perthame, *Effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors*, Preprint hal-00921266, (dec. 2013).
- [42] A. Lorz A., S. Mirrahimi, and B. Perthame, *Dirac mass dynamics in multidimensional nonlocal parabolic equations*, *Comm. in P. D. E.*, **36**(6) (2011), 1071–1098.
- [43] J. S. Lowengrub, H. B. Frieboes, F. Jin, Y.-L. Chuang, X. Li, P. Macklin, S. M. Wise, and V. Cristini, *Nonlinear modelling of cancer: bridging the gap between cells and tumours*, *Nonlinearity* **23** (2010), R1–R91.
- [44] S. Mirrahimi, G. Barles, B. Perthame, and P. E. Souganidis, *A singular Hamilton-Jacobi equation modeling the tail problem*, *SIAM J. Math. Anal.*, **44** No. 6 (2012), 4297–4319.
- [45] S. Mirrahimi and B. Perthame, *Concentration effect in a population adaptive evolution problem with space*, Work in preparation.
- [46] S. Mirrahimi and G. Raoul, *Dynamics of sexual populations structured by a space variable and a phenotypical trait*, *Theoretical Population Biology* **84** (2013), 87–103.
- [47] B. Perthame, *Transport equations in biology*, Series ‘Frontiers in Mathematics’, Birkhauser, 2007.
- [48] B. Perthame and G. Barles, *Dirac concentrations in Lotka-Volterra parabolic PDEs*, *Indiana Univ. Math. J.* **57**(7) (2008), 3275–3301.
- [49] B. Perthame, F. Quiròs, and J.-L. Vázquez, *The Hele-Shaw asymptotics for mechanical models of tumor growth*, *Arch. Ration. Mech. Anal.* **212**, No 1 (2014), 93–127.
- [50] ———, *Derivation of a Hele-Shaw type system from a cell model with active motion*, HAL-UPMC : hal-00906168.
- [51] B. Perthame, F. Quiròs, M. Tang, and N. Vauchelet, *Traveling wave solution of the Hele-Shaw model of tumor growth with nutrient*, *Mathematical Models and Methods in Applied Sciences*, in press.
- [52] L. Preziosi and A. Tosin, *Multiphase modeling of tumor growth and extracellular matrix interaction: mathematical tools and applications*, *J. Math. Biol.* **58** (2009), 625–656.

- [53] J. Ranft, M. Basan, J. Elgeti, J.-F. Joanny, J. Prost, and F. Jülicher, *Fluidization of tissues by cell division and apoptosis*, PNAS **107**, No 49, (2010), 20863–20868.
- [54] T. Roose, S. Chapman, and P. Maini, *Mathematical models of avascular tumour growth: a review*, SIAM Rev. **49**(2) (2007), 179–208.
- [55] P. E. Souganidis, *Front propagation: theory and applications*, CIME course on ‘viscosity solutions’. Lecture Notes in Math., Springer-Verlag, Berlin, 1998.
- [56] K. R. Swanson, R. C. Rockne, J. Claridge, M. A. J. Chaplain, E.C. Alvord Jr, and A.R.A. Anderson, *Quantifying the role of angiogenesis in malignant progression of gliomas: in silico modeling integrates imaging and histology*, Cancer Res. **71** (2011), 7366–7375.
- [57] C. Tomasetti and D. Levy, *An elementary approach to modeling drug resistance in cancer*, Math. Biosci. Eng. **7**, No 4, (2010), 905–18. [doi: 10.3934/mbe.2010.7.905.]
- [58] J.-L. Vázquez, *The porous medium equation. Mathematical theory*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, Oxford, 2007. ISBN: 978-0-19-856903-9.

Sorbonne Universités, UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions,  
F-75005, Paris, France

E-mail: Benoit.Perthame@upmc.fr



# O-minimality and Diophantine geometry

Jonathan Pila

**Abstract.** This lecture is concerned with some recent applications of mathematical logic to Diophantine geometry. More precisely it concerns applications of o-minimality, a branch of model theory which treats tame structures in real geometry, to certain finiteness problems descending from the classical conjecture of Mordell.

**Mathematics Subject Classification (2010).** Primary 03C64, 11G18.

**Keywords.** O-minimal structure, André-Oort conjecture, Zilber-Pink conjecture.

## 1. Introduction

This is a somewhat expanded version of my lecture at ICM 2014 in Seoul. It surveys some recent interactions between model theory and Diophantine geometry.

The Diophantine problems to be considered are of a type descending from the classical Mordell conjecture (theorem of Faltings). I will describe the passage from Mordell's conjecture to the far-reaching Zilber-Pink conjecture, which is very much open and the subject of lively study by a variety of methods on several fronts. The model theory is "o-minimality", which studies tame structures in real geometry, and offers powerful tools applicable to certain "definable" sets. In combination with an elementary analytic method for "counting rational points" it leads to a general result about the height distribution of rational points on definable sets. This result can be successfully applied to Zilber-Pink problems in the presence of certain functional transcendence and arithmetic ingredients which are known in many cases but seemingly quite difficult in general.

Both the methods and problems have connections with transcendental number theory. My further objective is to explain these connections and to bring out the pervasive presence of Schanuel's conjecture.

Though the broad family of Diophantine problems is the same, o-minimality is a rather different flavour of model theory to that employed in the Diophantine results of Hrushovski [62, 64] and the subsequent developments (e.g. [24, 129]; for more on "stability" and its applications see [63, 65]). However, both flavours involve fields with extra structure and hinge on suitable tame behaviour of the definable sets.

As there are excellent survey papers on these developments (e.g. [130, 131]), this exposition will stay at a broader level and keep technicalities to a minimum.

## 2. From Mordell to Zilber-Pink

**The Mordell conjecture.** Diophantine geometry deals in the first instance with the solution of systems of algebraic equations in integers and in rational numbers. It is a broad subject with a central place in number theory going back to antiquity. The problems we will consider are finiteness questions. One seeks to show that certain forms of Diophantine problems have only finitely many solutions, or a solution set that has a finite description in certain specific terms.

The ur-conjecture here is the Mordell conjecture asserting the finiteness of the number of rational points on curves of genus at least 2. For example, a non-singular plane quartic curve. This conjecture was proposed by Mordell [94] in 1922, and proved by Faltings [46] in 1983. In the meantime it evolved into the Mordell-Lang conjecture (ML; see Lang [77], I, 6.3) proved in the work of Faltings, Hindry, Laurent, McQuillan, Raynaud, Vojta, and others; see e.g. [18, 90, 97].

This was the first of three crucial steps in the evolution of the Mordell conjecture into what is known as the Zilber-Pink conjecture (ZP).

**The Mordell-Lang conjecture.** The first step, due to Lang (see e.g. [76]), recasts the conjecture in terms of a subvariety (i.e. irreducible closed algebraic subset defined over  $\mathbb{C}$ ; we identify varieties with their sets of complex points)  $V$  of a (semi-abelian) group variety  $X$ . The conjecture concerns the interaction of  $V$  with certain “special” subvarieties of  $X$  distinguished in terms of its group structure.

The simplest result of “Mordell-Lang” type concerns a curve  $V \subset X = \mathbb{G}_m^2$ . Here  $\mathbb{G}_m = \mathbb{G}_m(\mathbb{C}) = \mathbb{C}^\times$  is the multiplicative group of non-zero complex numbers, so  $V$  is the set of solutions in  $(\mathbb{C}^\times)^2$  of some irreducible (over  $\mathbb{C}$ ) polynomial  $F(x, y) = 0$ . The result, which appears in Lang [76] is the following. *If there are infinitely many points  $(\xi, \eta) \in V$  such that  $(\xi, \eta)$  is a torsion point of  $(\mathbb{C}^\times)^2$ , then  $F$  has either the form  $x^n y^m = \zeta$  or  $x^n = \zeta y^m$  for some non-negative integers  $n, m$  (not both zero) and root of unity  $\zeta$ . In the exceptional case  $V$  is a torsion coset: a coset of an irreducible algebraic subgroup (subtorus) of  $X$  by a torsion point, and, being positive-dimensional, contains infinitely many torsion points. Observe that a torsion point is a torsion coset (of the trivial group).*

With a view to generalisations, torsion cosets of  $X = \mathbb{G}_m^n$  will be called “special subvarieties” and torsion points “special points”. The (countable) collection of special subvarieties will be denoted  $\mathcal{S} = \mathcal{S}_X$ . For later use, general cosets of subtori will be called “weakly special subvarieties”. We observe that special points are Zariski dense in any special subvariety.

The Multiplicative Manin-Mumford conjecture, which is a special case of a theorem of Laurent [79] (for  $V$  defined over  $\mathbb{Q}$  it may be deduced from results of Mann [82]; see Dvornicich-Zannier [42] for generalisations, see also an independent proof by Sarnak [126]), asserts the converse. Consider a subvariety  $V \subset X$ .

(\*) *If special points are Zariski-dense in  $V$  then  $V$  is a special subvariety.*

Since the Zariski-closure of any set of points consists of finitely many irreducible components (\*) may be equivalently formulated as (\*)' or (\*)'' as follows.

(\*') *A component of the Zariski closure of a set of special points is special.*

(\*')'  *$V$  contains only finitely many maximal special subvarieties.*

If one replaces the group of torsion points by the division group  $\Gamma$  of a finitely generated subgroup of  $\mathbb{G}_m^n$ , and takes special subvarieties to be cosets of subtori by elements of  $\Gamma$ , then (\*) is the “Multiplicative Mordell-Lang conjecture”, a theorem of Laurent [79].

The Manin-Mumford conjecture (MM; proved by Raynaud [122, 123]) is the statement (\*) for a subvariety of an abelian variety with its torsion cosets as “special subvarieties” (the original formulations of Manin and Mumford concerned a curve of genus at least two embedded in its Jacobian); for the division group of a finitely generated subgroup it is ML. Note that ML, in both the multiplicative and abelian settings, is ineffective (one cannot bound the height of points, though one can bound their number).

While there is no explicit mention of rational points in the formulation of ML, implications for these, including the original Mordell conjecture, are recovered via the Mordell-Weil theorem: the group of rational points on an abelian variety over a number field is finitely generated ([78], I.4.1, [18]).

**The André-Oort conjecture.** André [1] and Oort [98] made conjectures analogous to the Manin-Mumford conjecture where the ambient variety  $X$  is a *Shimura variety* (the latter partially motivated by a conjecture of Coleman [33]). A combination of these has become known as the André-Oort conjecture (AO).

Shimura varieties have a central role in arithmetic geometry, in particular in the theory of automorphic forms see e.g. [91]. As the formal definition (see e.g. [91, 92]) is rather involved, I will just give some examples. The simplest examples are modular curves, for example the curve  $Y(1) = \mathrm{SL}_2(\mathbb{Z}) \backslash \mathbb{H}$  whose set of complex points is just the affine line  $\mathbb{C}$ , parameterising isomorphism classes (over  $\mathbb{C}$ ) of elliptic curves by their  $j$ -invariant (see e.g. [155]). However, the André-Oort conjecture is trivial for one-dimensional ambient varieties; the simplest non-trivial cases concern cartesian products of modular curves. The paradigm examples of Shimura varieties are the Siegel modular varieties  $\mathcal{A}_g$  parameterising principally polarised abelian varieties of dimension  $g$  [17].

Associated with a Shimura variety  $X$  is a countable collection  $\mathcal{S} = \mathcal{S}_X$  of *special subvarieties*, the zero-dimensional ones being called *special points*. For example, in  $Y(1)^2$ , a special subvariety of dimension 1 is: a “vertical line”  $x = j_0$  or “horizontal line”  $y = j_0$  where  $j_0$  is the  $j$ -invariant of an elliptic curve with *complex multiplication* (“CM”; i.e. a “singular modulus” see e.g. [155], §6); or the zero set of a *modular polynomial*  $\Phi_N(x, y)$  (see e.g. [155]). The other special subvarieties are  $Y(1)^2$  itself and special points, being the points for which both coordinates are singular moduli. There is also a larger (uncountable) collection of *weakly special subvarieties* which includes, in addition, all vertical and horizontal lines. In  $\mathcal{A}_g$  the special subvarieties become rather complicated to describe, but special points are again those  $x \in \mathcal{A}_g$  for which the corresponding abelian variety  $A_x$  is CM (see e.g. [98]). Special points are Zariski dense in any special subvariety, and AO asserts the converse:

*Let  $X$  be a Shimura variety and  $V \subset X$  a subvariety. Then (\*) holds.*

Equivalently, AO may be formulated as (\*'), or (\*'') which we take as the “official” version.

**Conjecture 2.1 (AO).** *Let  $X$  be a Shimura variety and  $V \subset X$ . Then  $V$  contains only finitely many maximal special subvarieties.*

The simplest non-trivial case of AO, for  $Y(1)^2$  (and more generally products of two modular curves), was established unconditionally by André [2]. AO is open in general, though

it is known to be true under the Generalised Riemann Hypothesis for CM fields (by work of Edixhoven, Klingler, Ullmo, and Yafaev [43, 45, 71, 144]) and it is known unconditionally in several cases and under various additional hypotheses on the special points in question (see [152]). In particular, AO for arbitrary products of modular curves was affirmed using  $o$ -minimality and point-counting in [108]. We will describe this approach below as well as further results which have been established by the same methods. Though unconditional, these results are ineffective in that they do not produce a bound on the height of the special points. The only effective result known is for products of two modular curves, due recently to Kühne [74] and Bilu-Masser-Zannier [16].

The broader class of *mixed Shimura varieties* (see e.g. [120]) includes for example the “mixed” variety  $\mathcal{X}_g$  associated with  $\mathcal{A}_g$ , namely  $\mathcal{A}_g$  fibered at each point by the abelian variety parameterised by that point, and analogous varieties with additional level structure (see e.g. [17]). These include elliptic modular surfaces. More exotic examples, like the Poincaré bi-extension ([13]), include copies of  $\mathbb{G}_m$  as special subvarieties. The second step in the evolution of ZP is to enlarge the category of “ambient” varieties to that of mixed Shimura varieties, which also have a geometrically defined collection of “special subvarieties” [120].

This gives a class of varieties in which all the Diophantine problems so far considered can be comprehended. The “special point conjecture” (\*) in this setting was formulated by André [1]. It contains AO and MM for CM abelian varieties, but it does not include the full MM or ML statements.

**The Zilber-Pink conjecture.** One further extension, which significantly enlarges its scope and reach, gives the Zilber-Pink conjecture. The setting is again  $V \subset X$  where  $X$  is a mixed Shimura variety, but instead of special subvarieties  $T$  contained in  $V$ , we consider (components of) intersections  $V \cap T$ , with  $T$  a special subvariety, which are *atypical in dimension* (see below).

This idea has three independent sources, expressed in different formulations. Zilber [159] formulated a version (“CIT”) in the setting of semi-abelian varieties, motivated by his work on the model theory of complex exponentiation (see below). Bombieri-Masser-Zannier [20] proved a theorem and formulated a conjecture about curves in  $\mathbb{G}_m^n$  originating with a question of Schinzel [132], leading to a result for intersections of subvarieties with one-dimensional tori and the formulation of a general conjecture in [21]. Pink [121] formulated a conjecture encompassing MM, ML, and AO by the same device of “unlikely intersections”, meaning intersections of a variety  $V \subset X$  with special subvarieties of codimension exceeding  $\dim V$ .

The irreducible components of the intersection of two subvarieties  $V, W \subset X$  typically have dimension

$$\dim V + \dim W - \dim X,$$

as one would expect by “counting conditions” (and never less if  $X$  is smooth [95]).

**Definition 2.2.** Let  $X$  be a mixed Shimura variety with collection  $\mathcal{S}$  of special subvarieties, and let  $V \subset X$ . An irreducible component  $A \subset V \cap T$ , where  $T \in \mathcal{S}$  is called an *atypical subvariety (of  $V$  in  $X$ )* if

$$\dim A > \dim V + \dim T - \dim X.$$

**Conjecture 2.3 (ZP).** *Let  $X$  be a mixed Shimura variety and  $V \subset X$ . Then  $V$  contains only finitely many maximal atypical subvarieties.*

This is essentially the formulation of Zilber and Bombieri-Masser-Zannier in Pink’s setting. There are several alternative formulations; see [22] for a proof that they are equivalent in the multiplicative setting. As it is always atypical for a proper subvariety of  $X$  to contain a special subvariety, ZP for  $X$  and all its special subvarieties implies the assertion  $(\ast'')$ , the “special point” or “generalised André-Oort” conjecture for  $X$ , via an inductive argument.

There has been a lot of work on problems subsumed within ZP. Nevertheless it is open even in the multiplicative case. I will describe a theorem established in work of Bombieri, Masser, Zannier, and Maurin that affirms ZP for a curve in  $\mathbb{G}_m^n$ .

An atypical subvariety of a curve in  $\mathbb{G}_m^n$  is either a point in its intersection with a subgroup of codimension at least 2, or the curve itself if it is contained in a subgroup of codimension 1. The following definition is convenient.

**Definition 2.4.** For a mixed Shimura variety  $X$  with its collection  $\mathcal{S}$  of special subvarieties and a non-negative integer  $k$  we let  $\mathcal{S}^{[k]}$  denote the (countable) union of all special subvarieties of  $X$  of codimension  $\leq k$ .

**Theorem 2.5** ([20, 22, 89]). *Let  $V \subset \mathbb{G}_m^n$  be a curve defined over  $\mathbb{C}$ . If  $V$  is not contained in a proper special subvariety then  $V \cap \mathcal{S}^{[2]}$  is a finite set. .*

An alternative proof (for  $V$  defined over  $\overline{\mathbb{Q}}$ , the algebraic closure of  $\mathbb{Q}$  in  $\mathbb{C}$ ) is given in [19] and, in conjunction with the “Bounded height theorem” of Habegger [52] leads to an effective result [53]. A proof of the main result of [20] using o-minimality and point-counting has been developed by Capuano [28].

ZP formally implies ML ([121, 159]), which may be seen in the multiplicative setting for curves as follows. Let  $V \subset \mathbb{G}_m^2$  be a curve and suppose that  $c_1, \dots, c_k \in \mathbb{C}^\times$  are multiplicatively independent (no nontrivial monomial on them gives unity). Define

$$V^* = \{(x, y, z_1, \dots, z_k) \in \mathbb{G}_m^{2+k} : (x, y) \in V, z_i = c_i, i = 1, \dots, k\}.$$

Two multiplicative conditions on  $(x, y, \bar{z}) \in V^*$  will in general mean that  $x$  and  $y$  belong to the division closure of the multiplicative group  $\langle c_1, \dots, c_k \rangle$  generated by  $c_1, \dots, c_k$ . Thus ZP for all  $\mathbb{G}_m^n$  implies ML for all  $\mathbb{G}_m^n$ .

I do not give a survey of results. The known results for abelian varieties are less complete than those for  $\mathbb{G}_m^n$ , and in the Shimura setting less complete still. Below I will discuss various specific problems that have been tackled using o-minimality and point-counting. See Zannier [156] for further discussion and references on ZP as well as more general problems under the rubric of “unlikely intersections”, and [157] for some specific problems and applications. See also Chambert-Loir [29]. For analogous results in other settings see [30, 85].

### 3. Transcendental Number Theory

**Classical results.** Transcendental number theory is concerned primarily with the algebraic nature of the values of special functions, especially the exponential function. I want to mention two famous results: Lindemann’s theorem (also known as the Lindemann-Weierstrass theorem) and Baker’s theorem (see e.g. [8]). Here  $\log x$  means any determination of the logarithm of  $x \in \mathbb{C}^\times$ .

**Theorem 3.1** (Lindemann-Weierstrass). *Let  $x_1, \dots, x_n \in \overline{\mathbb{Q}}$  be linearly independent over  $\mathbb{Q}$ . Then  $e^{x_1}, \dots, e^{x_n}$  are algebraically independent over  $\mathbb{Q}$ .*

**Theorem 3.2** (Baker). *Suppose that  $x_1, \dots, x_n \in \overline{\mathbb{Q}}$ . If  $\log x_1, \dots, \log x_n$  are linearly independent over  $\mathbb{Q}$  then they are linearly independent over  $\overline{\mathbb{Q}}$ .*

Baker's theorem has been partially extended to elliptic and abelian functions in work of Baker, Bertrand, Masser, Philippon, Wüstholz and others (see e.g. [9]). These developments also impacted substantially on Diophantine problems, but I want to note in particular that the Masser-Wüstholz isogeny estimates led to a new proof [86] of the Mordell conjecture. More recently, Kühne [74] uses quantitative results for linear forms in (elliptic and classical) logarithms in his unconditional proof of AO for products of two modular curves.

So the methods of Diophantine geometry and transcendence theory are cognate; but the underlying conjectures are also cognate in the work of Zilber on the model theory of exponentiation described below.

**Schanuel's conjecture.** Schanuel's conjecture (SC; see Lang [77], p.31) seems to encapsulate all reasonable transcendence properties of the exponential function.

**Conjecture 3.3** (SC). *Let  $z_1, \dots, z_n \in \mathbb{C}$  be linearly independent over  $\mathbb{Q}$ . Then*

$$\text{tr. deg.}_{\mathbb{Q}} \mathbb{Q}(z_1, \dots, z_n, e^{z_1}, \dots, e^{z_n}) \geq n.$$

The special case with all the  $z_i$  algebraic recovers Lindemann's theorem. The special case with all the  $\exp z_i$  algebraic is open for  $n \geq 2$ , the best result known towards "algebraic independence of logarithms" is Baker's theorem.

**"Ax-Schanuel".** Ax [5] (see also [4]) established Schanuel's conjecture in the setting of a differential field (apparently also conjectured by Schanuel; see [5]); this theorem is known as "Ax-Schanuel".

Let  $\mathbb{Q} \subset C \subset K$  be a tower of fields and  $\{D_1, \dots, D_m\}$  a set of commuting derivations of  $K$  with  $C = \bigcap_{\mu} \ker D_{\mu}$ . By "rank" below we mean rank over  $K$ .

**Definition 3.4.** Elements  $x_1, \dots, x_n \in K$  are called *linearly independent over  $\mathbb{Q}$  modulo  $C$*  if there is no nontrivial relation  $\sum_{\nu} q_{\nu} x_{\nu} = c$  where  $q_{\nu} \in \mathbb{Q}$ ,  $c \in C$ .

Ax's theorem is then the following. Condition (a) encapsulates " $y_{\nu} = e^{x_{\nu}}$ " in a general differential field. However, by the Seidenberg embedding theorem [133, 134], a finitely generated differential field may be embedded into a field of meromorphic functions.

**Theorem 3.5** ("Ax-Schanuel"). *Let  $x_{\nu}, y_{\nu} \in K^{\times}$ ,  $\nu = 1, \dots, n$ , with*

- (a) *for all  $\mu, \nu$ ,  $D_{\mu} y_{\nu} = y_{\nu} D_{\mu} x_{\nu}$ ;*
- (b) *the  $x_{\nu}$  are linearly independent over  $\mathbb{Q}$  modulo  $C$  [or (b'), the  $y_{\nu}$  are multiplicatively independent over  $C$ ].*

*Then*

$$\text{tr. deg.}_C C(x_1, \dots, x_n, y_1, \dots, y_n) \geq n + \text{rank}(D_{\mu} x_{\nu})_{\mu=1, \dots, m, \nu=1, \dots, n}.$$

This implies a (weaker) variant in the complex setting that will be important in the sequel. A statement along these lines was established by Ax [6] in the semiabelian setting.

We consider  $\pi : \mathbb{C}^n \rightarrow \mathbb{G}_m^n$  given by  $z \mapsto e(z) = \exp(2\pi iz)$  on each coordinate. Fix  $V \subset \mathbb{G}_m^n$ . Ax-Schanuel implies that the "best" intersections of  $\pi^{-1}(V)$  with algebraic subvarieties  $W \subset \mathbb{C}^n$  are achieved by weakly special  $W$ . We formulate a precise statement as follows.

**Definition 3.6.**

- (1) A *component with respect to  $V$*  is a complex analytically irreducible component  $A$  of  $W \cap \pi^{-1}(V)$  for some irreducible algebraic  $W \subset \mathbb{C}^n$ .
- (2) If  $A$  is a component w.r.t.  $V$  we define its *defect*  $\delta(A)$  to be  $\dim \text{Zcl}(A) - \dim A$  where  $\text{Zcl}(A)$  is the Zariski closure of  $A$ .
- (3) A component  $A$  w.r.t.  $V$  is called *optimal* for  $V$  if there is no component  $B$  w.r.t.  $V$  with  $A \subset B$ ,  $A \neq B$ , and  $\delta(B) \leq \delta(A)$ . Note that if  $A$  is optimal it must be a component of  $\text{Zcl}(A) \cap \pi^{-1}(V)$ .
- (4) A component  $A$  w.r.t.  $V$  is called *weakly special* if it is a component of  $W \cap \pi^{-1}(V)$  for some weakly special  $W = \text{Zcl}(A)$ .

Then Ax-Schanuel implies the following statement (see [111]).

**Theorem 3.7.** *An optimal component w.r.t.  $V \subset \mathbb{G}_m^n$  is weakly special.*

In particular, we have  $\delta(W) = 0$  just if  $W \subset \pi^{-1}(V)$ . An optimal component with defect zero is then a maximal irreducible algebraic subvariety contained in  $\pi^{-1}(V)$ .

**Corollary 3.8.** *A maximal algebraic subvariety  $W \subset \pi^{-1}(V)$  is weakly special.*

Another way to formulate the corollary is that if algebraic functions  $z_1, \dots, z_m$  (say elements of the function field  $\mathbb{C}(W)$ ) are linearly independent modulo constants (i.e. the locus  $z_1, \dots, z_m$  is not contained in any weakly special subvariety) then the exponentials  $\exp z_1, \dots, \exp z_n$  are algebraically independent over  $\mathbb{C}$ , which is a functional analogue of Lindemann's theorem. Accordingly I call the assertion of the corollary and its various analogues "Ax-Lindemann"; see also [15].

Tsimerman [141] has recently given a new proof of Ax-Schanuel via o-minimality and point-counting.

## 4. Model Theory

**Model theory of  $\mathbb{C}$  and  $\mathbb{R}$ .** See e.g. [160]. The first-order theory of the complex field  $(\mathbb{C}, +, \times, 0, 1)$  is just the theory of algebraically closed fields of characteristic zero and is *categorical* (has a unique model up to isomorphism) in every uncountable power. This is a very strong property of a theory. Algebraically closed fields are also "strongly minimal": the definable subsets of  $\mathbb{C}$  are either finite or cofinite. Indeed, by quantifier elimination, the definable (with parameters from  $\mathbb{C}$ ) subsets of  $\mathbb{C}^n$  are precisely the constructible sets: the Boolean algebra generated by the zero-sets of polynomials (with coefficients in  $\mathbb{C}$ ) in  $\mathbb{C}^n$ . The theory is also decidable.

Strong minimality fails for the real field as the order is definable, whence intervals are definable; however (Tarski-Seidenberg theorem) the definable sets are just the semi-algebraic sets: finite boolean combinations of sets defined by finitely many polynomial equalities and inequalities. The theory is again decidable (Tarski [139]). A definable subset of  $\mathbb{R}$  is still relatively simple, being a finite union of points and (possibly unbounded) intervals.

**Model theory of complex exponentiation.** The integers are definable in the complex numbers with exponentiation:

$$\mathbb{Z} = \{z \in \mathbb{C} : \forall w \in \mathbb{C} (\exp(w) = 1 \rightarrow \exp(zw) = 1)\}.$$

Therefore, by Gödel’s Theorem, the first-order theory of  $\mathbb{C}_{\text{exp}} = (\mathbb{C}, +, \times, 0, 1, \exp)$  is undecidable and the definable sets can be “wild”. The theory is very far from categorical. Nevertheless, Zilber showed that categoricity can be recovered if one works with a stronger infinitary logic. He used a Hrushovski-style construction in which Schanuel’s conjecture plays a fundamental role to construct [158] a candidate “logically perfect” algebraically closed field of power continuum with a “standard” (cyclic kernel) exponentiation,  $\mathbb{B}_{\text{exp}}$ , and conjectured that this field is isomorphic to  $\mathbb{C}_{\text{exp}}$  (entailing SC and more; see e.g. [34, 35]).

Considering the first-order theory of this structure led Zilber to his “CIT” conjecture [159] in the setting of  $\mathbb{G}_m^n$  and more generally semiabelian varieties: it is the “difference” between SC and a uniform version of SC that admits first-order axiomatization.

**Conjecture 4.1** (Uniform Schanuel Conjecture; USC). *Let  $V \subset \mathbb{C}^{2n}$  be a closed algebraic set defined over  $\mathbb{Q}$  with  $\dim V < n$ . There exists a finite set  $\mu(V)$  of proper  $\mathbb{Q}$ -linear subspaces of  $\mathbb{C}^n$  such that if*

$$(z_1, \dots, z_n, e^{z_1}, \dots, e^{z_n}) \in V$$

*then there is  $M \in \mu(V)$  and  $\bar{k} \in \mathbb{Z}^n$  and such that  $(z_1 + 2\pi i k_1, \dots, z_n + 2\pi i k_n) \in M$ . Moreover if  $M$  is codimension 1 (in  $\mathbb{C}^n$ ) then  $k = 0$ .*

Part of this program has been carried out for the  $j$ -function by Harris [59] and more generally for Shimura curves [36] by Daw-Harris. A very general picture of “special subvarieties” and generalised Schanuel conjectures is set out in [161].

**Model theory of real exponentiation.** O-minimality grew out of the attempt to understand the model theory of the real field with exponentiation. The real exponential has no overt periodic behaviour, thus no obvious source of “Gödelian problems”. Upon proving the decidability of the real field, Tarski [139] asked whether the theory of the real field with exponentiation, i.e. the structure  $\mathbb{R}_{\text{exp}} = (\mathbb{R}, +, \times, 0, 1, \exp)$ , is decidable.

In studying this question, van den Dries [38] noted the key role played by the above mentioned finiteness property of semi-algebraic sets and formulated the condition “a definable subset of  $\mathbb{R}$  is a finite union of points and intervals” that is the key defining property of a general theory of “o-minimal structures” subsequently undertaken by Pillay and Steinhorn [118] (see also [73, 119]). They prove the fundamental Cell Decomposition Theorem, from which the remarkable tameness and uniformity properties of o-minimal structures flow. For completeness I include a “model-theory free” definition of an o-minimal structure over the real field. Being a “structure” means that the sets in  $\bigcup_n \Sigma_n$  are precisely the definable sets (with parameters) in a suitable expansion of the real field.

**Definition 4.2.**

- (1) A *pre-structure* is a sequence  $\Sigma = (\Sigma_n)_{n=1,2,\dots}$  where each  $\Sigma_n$  is a collection of subsets of  $\mathbb{R}^n$ .
- (2) A pre-structure  $\Sigma$  is called a *structure (over the real field)* if, for all  $n, m = 1, 2, \dots$  with  $m \geq n$ , the following conditions are satisfied:

- (a)  $\Sigma_n$  is a Boolean algebra
- (b)  $\Sigma_n$  contains every semi-algebraic subset of  $\mathbb{R}^n$
- (c) if  $A \in \Sigma_n$  and  $B \in \Sigma_m$  then  $A \times B \in \Sigma_{n+m}$
- (d) if  $A \in \Sigma_n$  then  $\pi(A) \in \Sigma_m$  where  $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a coordinate projection.

If  $\Sigma$  is a structure and  $Z \subset \mathbb{R}^n$  we say that  $Z$  is *definable* in  $\Sigma$  if  $Z \in \Sigma_n$ .

- (3) A structure  $\Sigma$  is called *o-minimal* if the boundary of each set in  $\Sigma_1$  is a finite set of points.

A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is said to be *definable* in a structure  $\Sigma$  if its graph is. If  $A, \dots, f, \dots$  are sets or functions then  $\mathbb{R}_{A, \dots, f, \dots}$  denotes the smallest structure containing  $A, \dots, f, \dots$ . By a *definable family* of sets we mean a definable subset  $Z \subset \mathbb{R}^n \times \mathbb{R}^m$  which we view as a family of fibres  $Z_y \subset \mathbb{R}^n$  as  $y$  varies over the projection of  $Z$  onto  $\mathbb{R}^m$  (which is definable, along with all the fibres  $Z_y$ ). A family of functions is definable if the family of their graphs is.

In the sequel, a *definable set* will mean a definable set in some o-minimal structure over  $\mathbb{R}$ . The o-minimal condition has very strong consequences for definable sets and functions. For example, a definable function is continuous (and also differentiable) except at finitely many points. Moreover, in a definable family of functions, the number of points of discontinuity (or of non-differentiability) is bounded uniformly for all members of the family. Another example (relevant later): in a definable family, the set of parameters for which the fibre has a given dimension is definable. For these and other properties see van den Dries [39].

Of course the theory is only useful if there are non-trivial examples. van den Dries observed that the o-minimality of the structure  $\mathbb{R}_{\text{an}}$  generated by all *restricted analytic functions*, i.e. all  $f : T \rightarrow \mathbb{R}$  where  $T \subset \mathbb{R}^n$  is a compact box and  $f$  is analytic on an open neighbourhood of  $T$ , follows from a fundamental theorem of Gabrielov [49] on subanalytic functions.

This is a large and useful structure, but did not answer the question raised by van den Dries for the (unrestricted) exponential function. This was affirmed by Wilkie [148] (who in fact established the “model-completeness” of  $\mathbb{R}_{\text{exp}}$ , giving its o-minimality in view of the results of Khovanskii [68]).

**Theorem 4.3** (Wilkie [148]). *The structure  $\mathbb{R}_{\text{exp}}$  is o-minimal.*

The structure  $\mathbb{R}_{\text{an, exp}}$  generated by the union of  $\mathbb{R}_{\text{exp}}$  and  $\mathbb{R}_{\text{an}}$  is o-minimal ([41], see also [40]). Note that in general the structure generated by the union of o-minimal structures need not be o-minimal ([125]): there is no “largest” minimal structure over  $\mathbb{R}$ . Larger and stranger o-minimal structures followed ([125, 138]), but  $\mathbb{R}_{\text{an, exp}}$  suffices for all the applications we will consider ( $\mathbb{R}_{\text{an}}$  doesn’t).

Macintyre and Wilkie [80] affirmed Tarski’s original question assuming SC.

**Theorem 4.4** ([80]). *Assuming SC, the theory of  $\mathbb{R}_{\text{exp}}$  is decidable.*

## 5. Counting Points

**Counting rational points in algebraic varieties.** Counting solutions to a Diophantine equation up to a given height  $T$  and probing the behaviour of their number  $N(T)$  as  $T \rightarrow \infty$  is

a well-travelled path in Diophantine geometry, especially in connection with Waring's problem and, more recently, the Batyrev-Manin conjectures; see e.g. [61]. For example, it is believed (see [137]) that there is no positive integer  $n$  which can be written as a sum of two fifth powers in two essentially different ways. This amounts to saying that all solutions in non-negative integers to

$$X^5 + Y^5 = U^5 + V^5$$

are *trivial* in that  $\{X, Y\} = \{U, V\}$ . Hooley proved (see [60], improved in [26]) that there are at most  $O_\epsilon(T^{5/3+\epsilon})$  non-trivial solutions with  $0 \leq X, Y, U, V \leq T$ , which are thus dominated by the  $2T^2 + O(T)$  trivial ones. The conjectures of Bombieri and Lang (see e.g. [18], 14.3.7, [61], F.5.2) imply that all but finitely many rational points on a variety lie in the geometrically defined "special set".

Thus, conjecturally, general Diophantine problems, like the special ones of Mordell-Lang type, only have infinitely many solutions if there is a "reason".

**Counting rational points in definable sets.** Prompted by questions posed by Sarnak (motivated by his analytic proof of the multiplicative Manin-Mumford conjecture [126]; see also [127]), Bombieri-Pila [23] counted integer points up to a given height on plane curves in various categories (convex, transcendental real-analytic, algebraic) by an elementary real-variable method. The same idea was applied to rational points on a real analytic plane curve in [104]. Heath-Brown [60] introduced a variant  $p$ -adic "determinant method" applicable to rational points on algebraic varieties in any dimension, which prompted the idea of applying the "real" version to count rational points in higher-dimensional sets defined by analytic conditions.

We define the *height* of a rational number  $x = a/b$  in lowest terms (i.e.  $\gcd(a, b) = 1$ ) by  $H(x) = \max(|a|, |b|)$ , and the *height* of a tuple  $x = (x_1, \dots, x_n) \in \mathbb{Q}^n$  by  $H(x) = \max(H(x_i), i = 1, \dots, n)$ . For a set  $Z \subset \mathbb{R}^n$  we put

$$Z(\mathbb{Q}, T) = \{x \in Z : x \in \mathbb{Q}^n, H(x) \leq T\},$$

and define the *counting function* of  $Z$  by

$$N(Z, T) = \#Z(\mathbb{Q}, T).$$

We would like to have a result expressing that a "reasonable" set  $Z \subset \mathbb{R}^n$  has "few" rational points unless there is a "reason". We will take "reasonable" to mean definable. If  $Z$  contains positive dimensional semi-algebraic subsets (e.g. a piece of a line or circle) then these may contain quite a lot of algebraic points; thus we will exclude such subsets from the counting.

**Definition 5.1.** Let  $Z \subset \mathbb{R}^n$ . We define the *algebraic part*  $Z^{\text{alg}}$  of  $Z$  to be the union of positive-dimensional connected semi-algebraic subsets of  $Z$ .

The algebraic part is a coarse analogue of the "special set". But one cannot expect finiteness of rational points outside the algebraic part in view of curves like  $y = 2^x$ . The following theorem provides a sense in which there are "few" rational points outside the algebraic part of a definable set.

**Theorem 5.2** (Counting Theorem; Pila-Wilkie [116]). *Let  $Z \subset \mathbb{R}^n$  be a definable set and  $\epsilon > 0$ . Then there is a constant  $c(Z, \epsilon)$  such that, for all  $T$ ,*

$$N(Z - Z^{\text{alg}}, T) \leq c(Z, \epsilon)T^\epsilon.$$

Suppose  $Z$  is the image of a map  $\phi : (0, 1)^k \rightarrow \mathbb{R}^n$ . The underlying analytic idea of [23], extended to higher dimension in [105], is that  $Z(\mathbb{Q}, T)$  is contained in the intersection  $Z \cap V$  of  $Z$  with “few” hypersurfaces  $V$  of some suitable degree  $d = d(Z, \epsilon)$ , whose number depends on the maximum size of coordinate functions of  $\phi$  and some number (depending on  $\epsilon$ ) of their partial derivatives. The key to proving the Counting Theorem is a parameterisation theorem ([116], Thm 2.3) by means of which the intersections  $Z \cap V$  can be realised as images of finitely many maps whose derivatives up to a given order are bounded uniformly as  $V$  varies in the family of all hypersurfaces of given degree. This o-minimal version of the “Algebraic Lemma” of Yomdin-Gromov [50, 154] allows the analytic idea to be applied inductively; it yields a result which is uniform for definable families.

One can establish a bound of the same quality for algebraic points of bounded degree. For a definable set  $Z$  and  $k \geq 1$  put

$$Z(k, T) = \{x \in Z : [\mathbb{Q}(x_i) : \mathbb{Q}] \leq k, H(x_i) \leq T, i = 1, \dots, n\},$$

$$N_k(Z, T) = \#Z(k, T),$$

where  $H(x)$  here is the multiplicative height (see [18], 1.5.7) of an algebraic number.

Then for definable  $Z$ , positive  $k$  and  $\epsilon > 0$  we have

$$N_k(Z - Z^{\text{alg}}, T) \leq c(Z, k, \epsilon)T^\epsilon.$$

The result is again uniform for  $Z$  in definable families.

A further refinement, necessitated by applications, makes the result look more like a generalised “special point” statement. Namely, one shows that  $Z(k, T)$  is contained in “few” definable connected subsets which locally coincide with semialgebraic sets (“blocks”), and which come from finitely many (depending on  $Z, k, \epsilon$ ) definable families. For the precise statement I refer to [108].

To make the result effective for a particular o-minimal structure one would need an effective bound on the number of connected components of a definable set in that structure, as a function of the “complexity” of the formula which defines it. This is known in only special cases [11].

Non-archimedean analogues have been announced by Cluckers-Comte-Loeser [31]. For an earlier result about integer points on definable curves, including a much stronger bound for curves in  $\mathbb{R}_{\text{an}}$ , see Wilkie [151]. For a still earlier application of Khovanskii theory to “unlikely intersections” see Cohen-Zannier [32].

**Wilkie’s conjecture.** The Counting Theorem cannot be much improved in general (see [105]); in particular one cannot in general replace the  $\ll_\epsilon T^\epsilon$  bound with a power of  $\log T$ . However, Wilkie ([116], 1.11) has conjectured:

**Conjecture 5.3.** *Let  $Z$  be definable in  $\mathbb{R}_{\text{exp}}$ . Then there are constants  $C(Z), c(Z)$  such that*

$$N(Z - Z^{\text{alg}}, T) \leq C(\log T)^c.$$

Partial results are established in [27, 66, 67, 107], some for sets definable in the larger o-minimal structure  $\mathbb{R}_{\text{Pfa}}^{\text{aff}}$ . It would be nice to go further and establish such results for the structures in which (the restrictions of) the uniformising maps of mixed Shimura varieties are definable.

### 6. O-minimality and “special point” problems

**The setting.** The basic strategy is due to Zannier, who proposed using the Counting Theorem to give a new proof of MM. This was implemented in [117]. He and Masser saw that the same strategy could be applied to certain “relative Manin-Mumford” problems posed by Masser ([87], further described below). These turn out to be also special cases of ZP.

The generalisation of the Counting Theorem to algebraic points [106], and the analogies between MM and AO (as highlighted e.g. in [142, 153]) prompted the idea of applying the same idea to the latter problem: for products of  $Y(1)$ , for example, the “special points” correspond to tuples of  $j$ -invariants of elliptic curves with complex multiplication. These are precisely the points  $j(z)$  where  $z \in \mathbb{H} = \{\tau \in \mathbb{C} : \text{Im}(\tau) > 0\}$  is a quadratic irrationality, thus they correspond to algebraic points in  $\mathbb{H}^n$  of bounded degree.

Indeed all the “special point” problems take a similar form. The ambient variety  $X$  has a transcendental uniformisation

$$\pi : U \rightarrow X$$

by a complex domain  $U$  with certain properties. Examples (set  $e(u) = \exp(2\pi iu)$ ):

- (i)  $X = \mathbb{G}_m^n, U = \mathbb{C}^n, \pi(u_1, \dots, u_n) = (e(u_1), \dots, e(u_n))$ ;
- (ii)  $X$  an abelian variety,  $U = \mathbb{C}^{\dim X}$ ,  $\pi$  periodic under a suitable lattice  $\Lambda$ ;
- (iii)  $X = Y(1)^n, U = \mathbb{H}^n$ , and  $\pi(u_1, \dots, u_n) = (j(u_1), \dots, j(u_n))$ ;
- (iv)  $X = \mathcal{A}_g, U = \mathbb{H}_g$ , Siegel upper half-space,  $\pi$  is  $\text{Sp}_{2g}(\mathbb{Z})$ -invariant [17].

While a general abelian variety is not a mixed Shimura variety, it is a subvariety of one, and the “induced” ZP on its subvarieties is equivalent to the statement of ZP when  $X$  is endowed with its torsion cosets as “special subvarieties” [121].

This picture is essentially the same for any Shimura (or mixed Shimura) variety  $X$ , where  $U$  may be taken to be an open domain in some ambient complex affine space,  $\pi$  is invariant under a discrete arithmetic subgroup  $\Gamma$  of a real algebraic group  $G$  acting on  $U$  as biholomorphisms, and where  $U$  and the (graph of the)  $G$  action on it are semi-algebraic.

In each case, the pre-images of special points are algebraic points of bounded degree, when considered in suitable real coordinates on  $U$  (e.g. for an abelian variety we take a basis of  $\Lambda$  to define our real coordinates). Thus the search for special points in  $V \subset X$  can be translated into a search for their pre-images (which we will also just call special points) in  $\pi^{-1}(V)$ . This set is in general far from algebraic.

Moreover, components of pre-images of special subvarieties are algebraic (in a sense described below) and appear in definable families. For example in  $Y(1)^n$  they are the just the subvarieties of  $\mathbb{H}^n$  defined by some collection of equations of the form  $z_i = g_{ij}z_j$ , where  $(i, j) \in E \subset \{1, \dots, n\}^2$  and  $g_{ij} \in \text{GL}_2^+(\mathbb{Q})$  acting by Mobius transformations. Here  $E$  is any set, possibly empty, and we allow  $i = j$  in which case the fixed point  $z_i$  is quadratic. They sit in the family of subvarieties defined by relations from  $\text{SL}_2(\mathbb{R})$ , and this family is definable (indeed semi-algebraic).

As the map  $\pi$  is invariant under the group  $\Gamma$  acting on  $U$  (in the examples:  $\mathbb{Z}^n, \Lambda, \text{SL}_2(\mathbb{Z})^n, \text{Sp}_{2g}(\mathbb{Z})$ ), we may restrict our attention to a fundamental domain  $F$  for this action, which may also be taken to be semi-algebraic. We let

$$Z = \pi^{-1}(V) \cap F$$

and we are now interested in certain algebraic points of bounded degree in  $Z$ .

**Definability.** The map  $e : \mathbb{C} \rightarrow \mathbb{C}^\times$  (i.e. its graph in  $\mathbb{C} \times \mathbb{C}^\times$ ) is not definable in any o-minimal structure, due to the infinite discrete group acting, and the same holds for the map  $\pi : U \rightarrow X$  for every mixed Shimura variety  $X$  of positive dimension. But in all the examples given so far, the restriction of  $\pi$  to a suitable fundamental domain  $F$  for  $\Gamma$  is definable in  $\mathbb{R}_{\text{an}, \text{exp}}$ .

**Theorem 6.1** (Peterzil-Starchenko [103]). *The restriction of  $\pi : \mathbb{H}_g \rightarrow \mathcal{A}_g$  to the classical fundamental domain for the  $\text{Sp}_{2g}(\mathbb{Z})$  action is definable.*

Indeed the corresponding assertion holds for  $\mathcal{X}_g$ , generalising the earlier result by the same authors for Weierstrass  $\wp$ -functions [100] established in the course of a study of non-standard complex tori. The generalisation of this result to all Shimura varieties has been announced by Klingler-Ullmo-Yafaev [72], and to all mixed Shimura varieties by Gao [48]. With these results, o-minimal methods are available across the full breadth of the Zilber-Pink conjecture.

**The strategy.** The Counting Theorem tells us that  $Z(k, T)$  is contained in “few” blocks contained in  $Z^{\text{alg}}$ . In the arithmetic settings, one then has essentially two tasks to turn this statement into the Diophantine conclusion:

- (i) to characterise  $Z^{\text{alg}}$  as (essentially) coinciding with the exceptional locus in the Diophantine problem, i.e. weakly special subvarieties. This is a problem in *functional transcendence*.
- (ii) to reduce “few” (i.e.  $\ll_\epsilon T^\epsilon$ ) to finite. This is effected by playing off the upper bound against a *lower bound for the size of the Galois orbit of a special point*.

**Characterising the algebraic part.** We need to understand  $Z^{\text{alg}}$ , but it is more natural to consider first  $\pi^{-1}(V)^{\text{alg}}$ , which turns out to be a union of complex algebraic subvarieties (intersected with  $U$ ). These will generally not be fully contained in  $Z$  (or indeed in  $F$ ).

I have not defined weakly special varieties except in the case of exponentiation, but they may be characterised by the following result of Ullmo-Yafaev [145]. By an “algebraic subvariety of  $U$ ” we will mean a (complex analytically irreducible) component of  $W \cap U$  where  $W$  is an algebraic subvariety of the ambient space (we always assume that the uniformising space  $U$  is semialgebraic).

**Theorem 6.2** (Ullmo-Yafaev [145]). *Let  $X$  be a Shimura variety. A subvariety  $W \subset X$  is weakly special if and only if the components of its pre-image in  $U$  are algebraic subvarieties.*

Weakly special subvarieties are thus precisely the algebraic varieties preserved (as algebraic) by  $\pi$ . It turns out that the algebraic part of  $\pi^{-1}(V)$  is equal to the union of weakly special subvarieties of positive dimension it contains:

**Theorem 6.3** (“Ax-Lindemann”). *For  $X$  and  $\pi : U \rightarrow X$  as in our examples, a maximal algebraic  $W \subset \pi^{-1}(V)$  is weakly special.*

Thus, the algebraic part of  $\pi^{-1}(V)$ , a coarse analogue of the special set, turns out to be a close relative. For the exponential function this follows, as already observed, from Ax-Schanuel; for abelian varieties it is likewise due to Ax [6]; see also [25, 69]. For  $\mathcal{A}_g$  it is due to Pila-Tsimerman [114], building on [108, 113, 147]. Klingler-Ullmo-Yafaev [72] have announced the result for all Shimura varieties (and indeed a bit more generally), and a further

generalisation to all mixed Shimura varieties has been announced by Gao [48]. A version for the modular function “with derivatives” is in [109]. While Ax’s theorem is in the setting of differential fields and is proved by differential algebra, in all the Shimura variety settings mentioned “Ax-Lindemann” is proved directly in the complex setting using o-minimality and point-counting. This uses the fact that the group  $\Gamma$  gives rise to “many” integer points in suitable definable subsets of  $G$ . Mok has indicated how such results can be proved via complex differential geometry.

**From “few” to finite.** The definability of  $Z$  allows the Counting Theorem to be applied to the relevant algebraic points. This implies that there are “few” such points. How does one get from this to a finiteness statement?

The key here is that special points in  $X$  are algebraic, and are (at least conjecturally) of high degree, while their pre-images have small height, relative to a suitable measure of their “complexity”. For example, a root of unity  $\zeta$  of order (precisely)  $T$  has degree

$$[\mathbb{Q}(\zeta) : \mathbb{Q}] = \phi(T) \gg_{\epsilon} T^{1-\epsilon}$$

for every positive  $\epsilon$  (see e.g. [58], 18.4, Theorem 327), while its pre-image (under the map  $e(z) = \exp(2\pi iz)$ ) in the fundamental domain  $F = \{z \in \mathbb{C} : 0 \leq \operatorname{Re} z < 1\}$  has height  $T$ . (The “complexity measure” here is the order  $T$ .)

Lower bounds for the size of Galois orbits of torsion points in abelian varieties are much studied, e.g. in connection with isogeny estimates and Serre’s Open Image Theorem. Suitable results for [117] are due to Masser [83]. For products of elliptic modular surfaces (torsion points on CM curves) one has results of Silverberg [136].

Lower bounds for the size of Galois orbits of special points are essential in all current approaches to AO. The following was suggested by Edixhoven [44] for special points in  $\mathcal{A}_g$ , where an appropriate complexity measure for  $x \in \mathcal{A}_g$  is afforded by the discriminant  $\Delta(x)$  of the centre of the endomorphism ring of the corresponding abelian variety  $A_x$ .

**Conjecture 6.4.** *Let  $g \geq 1$ . There exist positive constants  $C_g, \delta_g$  such that, for a CM point  $x \in \mathcal{A}_g$ ,*

$$[\mathbb{Q}(x) : \mathbb{Q}] \geq C_g |\Delta(x)|^{\delta_g}.$$

For  $g = 1$  the conjecture is affirmed by the theory of complex multiplication of elliptic curves and the (ineffective) Landau-Siegel lower bound for class numbers [75, 135]. It has been affirmed unconditionally for  $g \leq 6$ , and for all  $g$  under GRH, by Tsimerman [140] (for the latter see also [146]).

A suitable upper bound for the height of the pre-image in a fundamental domain of a special point in  $\mathcal{A}_g$  is established in [113]; its generalisation to general Shimura varieties is expected. Then the upper and lower bounds for the number of points outside the algebraic part are incompatible for large height: there are only finitely many “isolated” special points.

**Concluding the proof.** Once the “Ax-Lindemann” result is established, a further property follows: that the maximal weakly special subvarieties contained in  $\pi^{-1}(V)$  come from a finite number of “families” (because being “optimal” is a definable condition on a larger semi-algebraic collection of subvarieties of  $U$  containing all the weakly special subvarieties, while by Ax-Lindemann the weakly special families in which optimal subvarieties lie are characterised by rational data: a definable subset of  $\mathbb{Q}$  must be finite). In the exponential case, this means that they are translates of finitely many rational linear spaces. In the other

cases, one can also view the weakly special subvarieties in a given family as “translates”, parameterised by points in a suitable “quotient”. The translate is special if and only if the corresponding parameter is a special point.

This finally enables the argument to be concluded by induction as follows. Given  $V$ , one has finitely many families  $U_i, i = 1 \dots, k$  of weakly special subvarieties, parameterised by points of some ambient varieties  $X_i$  of the same general type, but of lower dimension (except that points are weakly special and are parameterised by  $X$  itself). With each one has a subvariety  $V_i \subset X_i$  consisting of those parameters for which the corresponding weakly special subvariety is contained in  $V$ . One may suppose by induction that  $V$  contains only finitely many special subvarieties of positive dimension. Then apply the Counting Theorem directly to see that a special point of large complexity has “many” Galois conjugates over  $V$  and all its positive dimensional special subvarieties, and leads to a contradiction. To conclude one observes that there are only finitely many special points whose complexity is below a given bound.

**Theorem 6.5** ([37, 108, 113, 114, 143, 147]). *AO holds for  $\mathcal{A}_g^n, n \geq 1, g \leq 6$ .*

The efficacy of the Counting Theorem in these applications lies firstly in that it may be applied even when the Galois lower bounds are far from optimal: the problem then devolves to understanding the algebraic part, which is a question in functional transcendence. Secondly, it can be applied to this latter problem due to the arithmetic nature of  $\Gamma$ . Ullmo [143] has shown that, for a Shimura variety  $X$ , these ingredients (definability of  $\pi$  on  $F$ , Ax-Lindemann, lower bound for Galois orbits, upper bound for the height of a pre-image of a special point in  $F$ , the last two in terms of a suitable “complexity”) suffice to establish AO for  $X$ .

Special points results in mixed settings are obtained in [3, 48, 110]. A proof of semi-abelian MM along these lines is in [101].

### 7. O-minimality and atypical intersections

**Torsion anomalous points.** For  $\lambda \in \mathbb{P}^1 - \{0, 1, \infty\}$  we denote by  $E_\lambda$  the elliptic curve in Legendre form defined (in affine coordinates) by

$$y^2 = x(x - 1)(x - \lambda).$$

We let  $P_\lambda, Q_\lambda \in E_\lambda$  be the points

$$P_\lambda = (2, \sqrt{2(2 - \lambda)}), \quad Q_\lambda = (3, \sqrt{6(3 - \lambda)})$$

(with some fixed determination of  $\sqrt{\phantom{x}}$ ; whether the point is torsion is independent of the choice). Masser and Zannier [87, 88] prove the following theorem.

**Theorem 7.1.** *There are only finitely many complex numbers  $\lambda \neq 0, 1$  such that  $P_\lambda$  and  $Q_\lambda$  are both torsion point in  $E_\lambda$ .*

This is a “Relative Manin-Mumford” problem, in that it concerns a curve (the locus of  $(P_\lambda, Q_\lambda)$ ) in a family of abelian varieties (the squares of the  $E_\lambda$  for  $\lambda \in \mathbb{P}^1 - \{0, 1, \infty\}$ ). A general “Relative Manin-Mumford” conjecture is framed by Pink [121] where it is shown to

follow from his general conjecture (but note that it requires a slight correction: see Bertrand [13]).

The relative Manin-Mumford conjecture for a curve in the Poincaré bi-extension has been announced by Bertrand-Masser-Pillay-Zannier [14]. See Zannier [157] for further developments and applications.

**Atypical modular intersections.** It is natural then to apply a similar strategy to other problems of atypical intersections. Since special subvarieties are defined by rational (or bounded degree algebraic) data, and the dimension conditions characterising atypical intersections are detectable by definable sets, the methods are *prima facie* available once one has definability of  $\pi$  on  $F$ .

Habegger and Pila [56] establish a partial analogue of Theorem 2.5 concerning atypical intersections of a curve in  $Y(1)^n$ : i.e. points where the coordinates satisfy two independent “special” relationships (either the elliptic curves corresponding to two coordinates are isogenous, or the curve corresponding to one coordinate is CM). The result again depends on a functional transcendence statement (algebraic independence of “modular logarithms”) and a suitable lower bound for Galois orbits. The lower bound is obtained only under an additional hypothesis.

**Definition 7.2.** For a curve  $V \subset Y(1)^n$ , define  $\deg_i V$  to be the number of intersections of  $V$  with the hyperplane determined by a generic fixed value of the  $i$ th coordinate. The curve  $V$  is called *asymmetric* if, among the positive  $\deg_i V$ , there are no repetitions, save that one value may appear at most twice.

**Theorem 7.3** ([56]). *Let  $V \subset Y(1)^n$  be an asymmetric curve defined over  $\overline{\mathbb{Q}}$ . If  $V$  is not contained in a proper special subvariety then  $V \cap \mathcal{S}^{[2]}$  is a finite set.*

Looking to atypical intersection problems more generally, it seems reasonable to conjecture the following “complex Ax-Schanuel” statement.

**Conjecture 7.4** (Weak Complex Ax; WCA). *Let  $X$  be a mixed Shimura variety, with its uniformisation  $\pi : U \rightarrow X$ , and  $V \subset X$ . Then an optimal component for  $V$  is weakly special.*

Habegger and Pila [57] show that WCA for  $Y(1)^n$  together with a conjecture on the size of Galois orbits of certain “optimal” atypical intersections in  $Y(1)^n$  enable the point-counting strategy to be carried through to give ZP for  $Y(1)^n$ . (A proof of WCA for  $Y(1)^n$  has been announced in [115].) The same ideas yield an unconditional result for curves in abelian varieties. We give some definitions in order to formulate this conjecture.

Let  $X$  be a mixed Shimura variety and  $\mathcal{S}$  its collection of special subvarieties. Since  $\mathcal{S}$  is closed under taking irreducible components of intersections, for any subvariety  $A \subset X$  there is a smallest special subvariety containing  $A$  which we denote  $\langle A \rangle$ . We call  $\partial(A) = \dim \langle A \rangle - \dim A$  the *defect* of  $A$ . Fix  $V \subset X$ . A subvariety  $A \subset V$  is called *optimal* (for  $V$ ) if there is no subvariety  $B \subset V$  with  $A \subset B$ ,  $A \neq B$  and  $\partial(B) \leq \partial(A)$ .

Another formulation of ZP for  $X$  is then that for any  $V \subset X$  there are only finitely many optimal subvarieties. (Apart from  $V$  itself, which is optimal for its defect, any optimal proper subvariety of  $V$  must be atypical.)

**Definition 7.5.** The *complexity*  $\Delta(T)$  of  $T \in \mathcal{S}_{Y(1)^n}$  is the maximum of the absolute values of the discriminant of any fixed (quadratic) coordinates and the heights of any  $g \in \mathrm{GL}_2(\mathbb{Q})^+$  defining a pre-image of  $T$  in  $\mathbb{H}^n$  (see [108]).

**Conjecture 7.6** (Large Galois Orbits; LGO). *Let  $X = Y(1)^n$  and  $V \subset X$  defined over  $K$ , a field finitely generated over  $\mathbb{Q}$ . Then there are constants  $C(V), \delta(V) > 0$  such that for any optimal isolated point component  $\{x\}$  one has*

$$[K(x) : K] \geq C(V)\Delta(\{\{x\}\})^{\delta(V)}.$$

The following two results are announced in [57].

**Theorem 7.7.** *Assume WCA for  $Y(1)^n$  and LGO. Then ZP holds for  $Y(1)^n$ .*

The same blueprint works for abelian varieties (and I would expect a suitable formulation to apply to any mixed Shimura variety). WCA is known for abelian varieties (Ax [6]) while LGO may be affirmed unconditionally for curves when everything is defined over  $\overline{\mathbb{Q}}$ . This relies on a height inequality of Rémond [124].

**Theorem 7.8.** *Let  $X$  be an abelian variety,  $V \subset X$  a curve, both defined over  $\overline{\mathbb{Q}}$ . If  $V$  is not contained in a proper special subvariety then  $V \cap S^{[2]}$  is a finite set.*

**Analogue of Mordell-Lang.** Just as ZP for curves in  $\mathbb{G}_m^n$  entails ML for curves, ZP for curves in  $Y(1)^n$  entails an analogue of ML. The same circle of ideas (o-minimality, point counting, and lower bounds for Galois orbits coming from isogeny estimates for elliptic curves [86]) enable a proof of “modular ML” for a general subvariety of  $Y(1)^n$ ; see [56, 110]. The latter includes an extension to products of elliptic modular surfaces. Various partial results for subvarieties of  $\mathcal{A}_g$  have been obtained by Orr, including the following full result for curves.

**Theorem 7.9** (Orr [99]). *A curve  $V \subset \mathcal{A}_g$  having infinitely many points for which the corresponding abelian varieties are isogenous is weakly special.*

A further result in this general area (though not a special case of ZP) is an analogue of the Tate-Voloch conjecture for products of modular curves, proved by Habegger [55]. I will not state the result, but note that the proof makes use of the above mentioned modular Mordell-Lang, established via o-minimality, while results of Scanlon [128] in the original semi-abelian setting made use of the model theory of difference fields.

**Some further questions.** The section above contains many stated results but they are at the same time fragmentary. Functional transcendence questions and lower bounds for Galois orbits seem to pose significant (though fascinating) challenges. I would like to conclude with some further questions that seem to arise naturally from the considerations around the Ax-Schanuel theme.

Consider a mixed Shimura variety  $X$ , its uniformisation  $\pi : U \rightarrow X$ , and an algebraic subvariety  $W \subset U$  in the sense defined earlier. When  $W$  is an orbit of a suitable kind, results from ergodic theory (“Ragunathan conjecture”) govern when  $\pi(W)$  is dense (in the usual analytic topology) in a weakly special subvariety. Ax-Lindemann says that the Zariski closure of  $\pi(W)$  is always weakly special.

**Question 7.10.** *For  $\pi : U \rightarrow X$  and an algebraic  $W \subset U$  as above:*

- (1) *Are there natural conditions under which  $\pi(W)$  is dense in  $X$ ?*
- (2) *Are there natural conditions under which  $\pi(W)$  intersects every algebraic subvariety  $V \subset X$  of complementary dimension (cf Ax [7])?*

Ax-Schanuel is naturally stated (and proved) in the setting of a differential field. The function  $j(z)$  satisfies a certain nonlinear third order algebraic differential equation, and none of lower order [81]. Specifically (see e.g. [84]),

$$J(j, j', j'', j''') = Sj + \frac{j^2 - 1968j + 2654208}{2j^2(j - 1728)^2} (j')^2 = 0,$$

where  $Sf$  denotes the *Schwarzian derivative*  $Sf = \frac{f'''}{f'} - \frac{3}{2} \left(\frac{f''}{f'}\right)^2$  and  $'$  indicates differentiation with respect to  $z$ . The full solution set is  $\{j(gz) : g \in \text{SL}_2(\mathbb{C})\}$ .

**Definition 7.11.** Let  $\mathbb{Q} \subset C \subset K$  be a tower of fields and  $\{D_\mu\}$  a set of commuting derivations of  $K$  with  $C = \bigcap_\mu \ker D_\mu$ . Elements  $j_1, \dots, j_n \in K$  are called *modular-independent* if no  $j_\nu \in C$  and no relation  $\Phi_N(j_\nu, j_\mu) = 0$  holds with  $N \geq 1, \nu \neq \mu$ .

We can formulate a conjecture giving a modular analogue of ‘‘Ax-Schanuel’’ in a differential field setting. It implies WCA for  $Y(1)^n$  as well as the result of [109]. (A modular analogue of Schanuel’s conjecture may be deduced from the Grothendieck-Andr e period conjecture [1], as explicated by Bertolin [12]; see [111].) Condition (a) below stipulates that  $j'_\nu, j''_\nu$  are the derivatives of  $j_\nu$  with respect to  $z_\nu$  and that  $j_\nu$  satisfies the  $j$  equation with respect to  $z_\nu$  for each  $\nu$ . The modular independence (b) implies that the quantities which appear in the denominator in  $J$  are non-zero. A corresponding ‘‘modular ZP with derivatives’’ is framed in [112].

**Conjecture 7.12.** *With  $K$  as above let  $z_\nu, j_\nu, j'_\nu, j''_\nu, j'''_\nu \in K^\times, \nu = 1, \dots, n$ , with*

(a) *for all  $\nu, \mu, D_\mu j_\nu = j'_\nu D_\mu z_\nu, D_\mu j'_\nu = j''_\nu D_\mu z_\nu, D_\mu j''_\nu = j'''_\nu D_\mu z_\nu$ , and*

$$J(j_\nu, j'_\nu, j''_\nu, j'''_\nu) = 0;$$

(b) *the  $j_\nu$  are modular-independent.*

*Then*

$$\text{tr. deg.}_C C(z_1, \dots, z_n, j_1, \dots, j_n, j'_1, \dots, j'_n, j''_1, \dots, j''_n) \geq 3n + \text{rank}(D_\mu z_\nu).$$

Freitag and Scanlon [47] have shown that the set defined by the differential equation satisfied by the  $j$ -function in a differentially closed field of characteristic zero is ‘‘strongly minimal’’ and ‘‘geometrically trivial’’. This uses the ‘‘modular Ax-Lindemann-Weierstrass with derivatives’’ result in [109]. For an introduction to differential fields in a model-theoretic setting, including definitions of the above terms (and related results on Painlev e transcendents) see Nagloo-Pillay [96]. One would like to generalise these results appropriately to the uniformising functions of mixed Shimura varieties.

**Acknowledgements.** My thanks to Philipp Habegger, Jacob Tsimerman, Alex Wilkie, Umberto Zannier, and Boris Zilber for valuable comments on drafts of this article. I am further most grateful to these colleagues, as well as to Enrico Bombieri, Peter Sarnak, and Thomas Scanlon for our collaborations and discussions regarding the problems considered. Finally, I thank the EPSRC for partial support of some of my research described herein under grant EP/J019232/1.

## References

- [1] Y. André, *G-functions and Geometry*, Aspects of Mathematics E13, Vieweg, Braunschweig, 1989.
- [2] ———, *Finitude des couples d'invariants modulaires singuliers sur une courbe algébrique plane non modulaire*, *Crelle* **505** (1998), 203–208.
- [3] ———, *Shimura varieties, subvarieties, and CM points*, Six lectures at the University of Hsinchu, August–September 2001.
- [4] J. Ax, *Transcendence and differential algebraic geometry*, Proc. ICM, Nice, 1970.
- [5] ———, *On Schanuel's conjectures*, *Annals* **93** (1971), 252–268.
- [6] ———, *Some topics in differential algebraic geometry I: Analytic subgroups of algebraic groups*, *Amer. J. Math.* **94** (1972), 1195–1204.
- [7] ———, *Some topics in differential algebraic geometry II: On the zeros of theta functions*, *Amer. J. Math.* **94** (1972), 1205–1213.
- [8] A. Baker, *Transcendental Number Theory*, CUP, 1974, 1979.
- [9] A. Baker and G. Wüstholz, *Logarithmic Forms and Diophantine Geometry*, New Mathematical Monographs: **9**, CUP, 2007.
- [10] M. Bays and P. Habegger, *A note on divisible points on curves*, arXiv:1301.5674.
- [11] A. Berarducci and T. Servi, *An effective version of Wilkie's theorem of the complement and some effective o-minimality results*, *Ann. Pure Appl. Logic* **125** (2004), 43–74.
- [12] C. Bertolin, *Périodes de 1-motifs et transcendance*, *J. Number Th.* **97** (2002), 204–221.
- [13] D. Bertrand, *Unlikely intersections in Poincaré biextensions over elliptic schemes*, *Notre Dame J. Formal Logic* **54** (2013; Oléron proceedings), 365–375.
- [14] D. Bertrand, D. Masser, A. Pillay, and U. Zannier, *Relative Manin-Mumford for semi-abelian surfaces*, arXiv:1307.1008.
- [15] D. Bertrand and A. Pillay, *A Lindemann-Weierstrass theorem for semi-abelian varieties over function fields*, *J. Amer. Math. Soc.* **23** (2010), 491–533.
- [16] Y. Bilu, D. Masser, and U. Zannier, *An effective “theorem of André” for CM points on plane curves*, *Math. Proc. Camb. Phil. Soc.* **154** (2013), 145–152.
- [17] C. Birkenhake and H. Lange, *Complex Abelian Varieties*, second edition, *Grund. math. Wiss.* **302**, Springer, Berlin, 2004.
- [18] E. Bombieri and W. Gubler, *Heights in Diophantine Geometry*, New Mathematical Monographs: **4**, CUP, 2006.
- [19] E. Bombieri, P. Habegger, D. Masser, and U. Zannier, *A note on Maurin's theorem*, *Rend. Lincei. Mat. Appl.* **21** (2010), 251–260.
- [20] E. Bombieri, D. Masser, and U. Zannier, *Intersecting a curve with algebraic subgroups of multiplicative groups*, *IMRN* **20** (1999), 1119–1140.
- [21] ———, *Anomalous subvarieties – structure theorems and applications*, *IMRN* **19** (2007), 33 pages.
- [22] ———, *On unlikely intersections of complex varieties with tori*, *Acta Arithmetica* **133** (2008), 309–323.

- [23] E. Bombieri and J. Pila, *The number of integral points on arcs and ovals*, Duke Math. J. **59** (1989), 337–357.
- [24] E. Bouscaren, *Groups interpretable in theories of fields*, Proc. ICM 2002, Beijing, Volume II, 3–12.
- [25] W. D. Brownawell and K. K. Kubota, *Algebraic independence of Weierstrass functions and some related numbers*, Acta Arith. **33** (1977), 111–149.
- [26] T. D. Browning and R. Heath-Brown, *Plane curves in boxes and equal sums of two powers*, Math. Z. **251** (2005), 233–247.
- [27] L. Butler, *Some cases of Wilkie’s conjecture*, Bull. LMS **44** (2012), 642–660.
- [28] L. Capuano, *Unlikely Intersections and Applications to Diophantine Geometry*, Ph.D. Thesis, 2013.
- [29] A. Chambert-Loir, *Relations de dépendance et intersections exceptionnelles*, Séminaire Bourbaki (2010–2011), exposé 1032, Asterisque **348** (2012), 149–188.
- [30] Z. Chatzidakis, D. Ghioca, D. Masser, and G. Maurin, *Unlikely, likely and impossible intersections without algebraic groups*, Rend. Lincei Math. Appl., to appear.
- [31] R. Cluckers, G. Comte, F. Loeser, *Non-archimedean Yomdin-Gromov parametrizations and points of bounded height*, arXiv:1404.1952.
- [32] P. B. Cohen and U. Zannier, *Fewnomials and intersections of lines with real analytic subgroups in  $\mathbb{G}_m^n$* , Bull. London Math. Soc. **34** (2002), 21–32.
- [33] R. Coleman, *Torsion points on curves*, Galois representations and arithmetic algebraic geometry, (Kyoto, 1985/Tokyo, 1986), 235–247, Adv. Stud. Pure Math. **12**, North-Holland, Amsterdam.
- [34] P. D’Aquino, A. Macintyre, G. Terzo, *From Schanuel’s conjecture to Shapiro’s conjecture*, arXiv:1206.6747.
- [35] ———, *Comparing  $\mathbb{C}$  and Zilber exponential fields*, zero sets of exponential polynomials, arXiv:1310.6891.
- [36] C. Daw and A. Harris, *Categoricity of modular and Shimura curves*, arXiv:1304.4797.
- [37] C. Daw and A. Yafaev, *An unconditional proof of the André-Oort conjecture for Hilbert modular surfaces*, Manuscripta **135** (2011), 263–271.
- [38] L. van den Dries, *Remarks on Tarski’s problem concerning  $(\mathbb{R}, +, \cdot, \exp)$* , in Logic colloquium ’82, pp. 97–121, Lolli, Longo, and Marcja, editors, North Holland, 1984.
- [39] ———, *Tame Topology and O-minimal Structures*, LMS Lecture Note Series **248**, CUP, 1998.
- [40] L. van den Dries, A. Macintyre, and D. Marker, *The elementary theory of restricted analytic fields with exponentiation*, Annals **140** (1994), 183–205.
- [41] L. van den Dries and C. Miller, *On the real exponential field with restricted analytic functions*, Israel J. Math. **85** (1994), 19–56.
- [42] R. Dvornicich and U. Zannier, *Sums of roots of unity*, Monatsh. Math. **129** (2000), 97–108.
- [43] S. J. Edixhoven, *Special points on products of modular curves*, Duke Math. J. **126** (2005), 325–348.
- [44] S. J. Edixhoven, B. J. J. Moonen, and F. Oort, *Open problems in algebraic geometry*,

- Bull. Sci. Math. **125** (2001), 1–22.
- [45] S. J. Edixhoven and A. Yafaev, *Subvarieties of Shimura varieties*, Annals **157** (2003), 621–645.
- [46] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Inventiones **73** (1983), 349–366.
- [47] J. Freitag and T. Scanlon, *Strong minimality of the  $j$ -function*, arXiv:1402.4588.
- [48] Z. Gao, *Towards the generalised André-Oort conjecture: The Ax-Lindemann theorem and lower bounds for Galois orbits of special points of mixed Shimura varieties*, arXiv:1310.1302.
- [49] A. Gabrielov, *Projections of semi-analytic sets*, Funct. Anal. Appl. **2** (1968), 282–291.
- [50] M. Gromov, *Entropy, homology and semialgebraic geometry* [after Y. Yomdin], Séminaire Bourbaki, 1985–86, exposé 663, Astérisque **145-146** (1987), 225–240.
- [51] P. Habegger, *Intersecting subvarieties of  $\mathbf{G}_m^n$  with algebraic subgroups*, Math. Annalen **342** (2008), 449–466.
- [52] ———, *On the bounded height conjecture*, IMRN **2009**, 860–886.
- [53] ———, *Effective height upper bounds on algebraic tori*, arXiv:1201.3255.
- [54] ———, *Torsion points on elliptic curves in Weierstrass form*, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (5) **12** (2013), 687–715.
- [55] ———, *A Tate-Voloch conjecture in a product of modular curves*, IMRN, to appear.
- [56] P. Habegger and J. Pila, *Some unlikely intersections beyond André–Oort*, Compositio **148** (2012), 1–27.
- [57] ———, *O-minimality and certain atypical intersections*, preprint.
- [58] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Fifth edition, OUP, 1979.
- [59] A. Harris, *Categoricity of the two-sorted  $j$ -function*, arXiv:1304.4787.
- [60] R. Heath-Brown, *The density of rational points on curves and surfaces*, Annals **155** (2002), 553–595.
- [61] M. Hindry and J. H. Silverman, *Diophantine Geometry: An Introduction*, Graduate Texts in Mathematics **201**, Springer, New York, 2000.
- [62] E. Hrushovski, *The Mordell-Lang conjecture for function fields*, J. Amer. Math. Soc. **9** (1996), 667–690.
- [63] ———, *Geometric model theory*, Proc. ICM Berlin, 1998, Volume I, 281–302.
- [64] ———, *The Manin-Mumford conjecture and the model theory of difference fields*, Ann. Pure. Appl. Logic **112** (2001), 43–115.
- [65] ———, *Stable group theory and approximate subgroups*, J. Amer. Math. Soc. **25** (2012), 189–243.
- [66] G. Jones and M. Thomas, *The density of algebraic points on certain Pfaffian surfaces*, QJM **63** (2012), 637–651.
- [67] G. Jones, C. Miller, and M. Thomas, *Mildness and the density of rational points on certain transcendental curves*, Notre Dame J. Formal Logic **52** (2011), 67–74.
- [68] A. G. Khovanskii, *Fewnomials*, Translations of Math. Monographs **88**, AMS, Provi-

- dence, 1991.
- [69] J. Kirby, *The theory of the exponential differential equations of semiabelian varieties*, *Selecta Math.* **15** (2009), 445–486.
  - [70] J. Kirby, B. Zilber, *Exponential fields and atypical intersections*, arXiv:1108.1075.
  - [71] B. Klingler and A. Yafaev, *The André-Oort conjecture*, *Annals*, to appear.
  - [72] B. Klingler, E. Ullmo, and A. Yafaev, *The hyperbolic Ax-Lindemann-Weierstrass conjecture*, 2013 manuscript.
  - [73] J. Knight, A. Pillay, and C. Steinhorn, *Definable sets in ordered structures. II*, *Trans. AMS* **295** (1986), 593–605.
  - [74] L. Kühne, *An effective result of André-Oort type*, *Annals* **176** (2012), 651–671.
  - [75] E. Landau, *Bemerkung zum Heilbronnschen Satz*, *Acta Arithmetica* **1** (1935), 2–18.
  - [76] S. Lang, *Division points on curves*, *Ann. Mat. Pura Appl.* (4) **70** (1965), 229–234.
  - [77] ———, *Introduction to Transcendental Numbers*, Addison-Wesley, Reading, 1966.
  - [78] ———, *Number Theory III: Diophantine Geometry*, *Encyclopaedia of Mathematical Sciences* **60**, Springer, Berlin, 1991.
  - [79] M. Laurent, *Équations diophantiennes exponentielles*, *Inventiones* **78** (1984), 299–327.
  - [80] A. Macintyre and A. Wilkie, *On the decidability of the real exponential field*, *Kreiseliania*, 441–467, A K Peters, Wellesley, MA, 1996.
  - [81] K. Mahler, *On algebraic differential equations satisfied by automorphic functions*, *J. Austral. Math. Soc.* **10** (1969), 445–450.
  - [82] H. B. Mann, *On linear relations between roots of unity*, *Mathematika* **12** (1965), 107–117.
  - [83] D. Masser, *Small values of the quadratic part of the Néron-Tate height on an abelian variety*, *Compositio* **53** (1984), 153–170.
  - [84] ———, *Heights, Transcendence, and Linear Independence on Commutative Group Varieties*, *Lecture Notes in Mathematics* **1819**, Amoroso and Zannier, eds, 1–51, Springer-Verlag, Berlin, 2003.
  - [85] ———, *Unlikely intersections for curves in multiplicative groups over positive characteristic*, *QJM*, published online 2013.
  - [86] D. Masser and G. Wüstholz, *Isogeny estimates for abelian varieties, and finiteness theorems*, *Annals* **137** (1993), 459–472.
  - [87] D. Masser and U. Zannier, *Torsion anomalous points and families of elliptic curves*, *C. R. Math. Acad. Sci. Paris* **346** (2008), 491–494.
  - [88] ———, *Torsion anomalous points and families of elliptic curves*, *Amer. J. Math.* **132** (2010), 1677–1691.
  - [89] G. Maurin, *Courbes algébriques et équations multiplicatives*, *Math. Annalen* **341** (2008), 789–824.
  - [90] M. McQuillan, *Division points on semi-abelian varieties*, *Inventiones* **120** (1995), 143–159.
  - [91] J. S. Milne, *Introduction to Shimura varieties*, *Harmonic Analysis, the Trace Formula*,

- and Shimura Varieties, 265–378, Clay Mathematics Proceedings **4** AMS, Providence, 2005.
- [92] B. J. J. Moonen, *Linearity properties of Shimura varieties, I*, J. Alg. Geom. **7** (1998), 539–567.
- [93] B. Moonen and F. Oort, *The Torelli locus and special subvarieties*, Handbook of Moduli, Volume II, Farkas and Morrison, editors, 549–594, Advanced Lectures in Mathematics **25**, International Press, Boston, 2013.
- [94] L. J. Mordell, *On the rational solutions of the indeterminate equation of third and fourth degrees*, Proc. Camb. Phil. Soc. **21** (1922), 179–192.
- [95] D. Mumford, *Algebraic Geometry I: Complex Projective Varieties*, Grundlehren der math. Wiss. **221**, Springer, Berlin, 1976.
- [96] R. Nagloo and A. Pillay, *On algebraic relations between solutions of a generic Painlevé equation*, arXiv:1112.2916.
- [97] F. Oort, “*The*” general case of S. Lang’s conjecture (after Faltings), pp 117–122 in: Edixhoven and Evertse, editors, Diophantine Approximation and Abelian Varieties, Lecture Notes in Mathematics **1566**, Springer, Berlin, 1993.
- [98] ———, *Canonical lifts and dense sets of CM points*, Arithmetic Geometry, Cortona, 1994, 228–234, F. Catanese, editor, Symposia. Math., XXXVII, CUP, 1997.
- [99] M. Orr, *Families of abelian varieties with many isogenous fibres*, Crelle, DOI: 10.1515/crelle-2013-0058, published online July 2013.
- [100] Y. Peterzil and S. Starchenko, *Uniform definability of the Weierstrass  $\wp$  functions and generalized tori of dimension one*, Selecta Math. N. S. **10** (2004), 525–550.
- [101] ———, *Around Pila-Zannier: the semiabelian case*, 2009 preprint.
- [102] ———, *Tame complex geometry and o-minimality*, Proc. ICM Hyderabad, 2010.
- [103] ———, *Definability of restricted theta functions and families of abelian varieties*, Duke Math. J. **162** (2013), 731–765.
- [104] J. Pila, *Geometric postulation of a smooth curve and the number of rational points*, Duke Math. J. **63** (1991), 449–463.
- [105] ———, *Integer points on the dilation of a subanalytic surface*, QJM **55** (2004), 207–223.
- [106] ———, *On the algebraic points of a definable set*, Selecta Math. N. S. **15** (2009), 151–170.
- [107] ———, *Counting rational points on a certain exponential-algebraic surface*, Ann. Inst. Fourier **60** (2010), 489–514.
- [108] ———, *O-minimality and the André-Oort conjecture for  $\mathbb{C}^n$* , Annals **173** (2011), 1779–1840.
- [109] ———, *Modular Ax-Lindemann-Weierstrass with derivatives*, Notre Dame J. Formal Logic **54** (2013; Oléron proceedings), 553–565.
- [110] ———, *Special point problems with elliptic modular surfaces*, Mathematika **60** (2014), 1–31.
- [111] ———, *Functional transcendence via o-minimality*, lecture notes, LMS-EPSRC Minicourse, 2013, available from the author’s web page.

- [112] ———, Modular Zilber-Pink “with derivatives”, working paper.
- [113] J. Pila and J. Tsimerman, *The André-Oort conjecture for the moduli space of abelian surfaces*, *Compositio* **149** (2013), 204–216.
- [114] ———, *Ax-Lindemann for  $\mathcal{A}_g$* , *Annals* **179** (2014), 659–681.
- [115] ———, *Ax-Schanuel for the  $j$ -function*, in preparation.
- [116] J. Pila and A. J. Wilkie, *The rational points of a definable set*, *Duke Math. J.* **133** (2006), 591–616.
- [117] J. Pila and U. Zannier, *Rational points in periodic analytic sets and the Manin-Mumford conjecture*, *Rend. Lincei Mat. Appl.* **19** (2008), 149–162.
- [118] A. Pillay and C. Steinhorn, *Definable sets in ordered structures I*, *Trans. AMS* **295** (1986), 565–592.
- [119] ———, *Definable sets in ordered structures III*, *Trans. AMS* **309** (1988), 469–476.
- [120] R. Pink, *A combination of the conjectures of Mordell-Lang and André-Oort*, *Geometric methods in algebra and number theory*, F. Bogomolov, Y. Tschinkel, editors, pp 251–282, *Prog. Math.* **253**, Birkhauser, Boston MA, 2005.
- [121] ———, *A common generalization of the conjectures of André-Oort, Manin-Mumford, and Mordell-Lang*, 2005 preprint, available from the author’s webpage.
- [122] M. Raynaud, *Courbes sur une variété abélienne et points de torsion*, *Inventiones* **71** (1983), no. 1, 207–233.
- [123] ———, *Sous-variétés d’une variété abélienne et points de torsion*, in *Arithmetic and geometry*, Volume I, pp 327–352, *Progr. Math.* **35**, Birkhauser, Boston MA, 1983.
- [124] G. Rémond, *Intersection de sous-groupes et sous-variétés II*, *J. Inst. Math. Jussieu* **6** (2007), 317–348.
- [125] J.-P. Rolin, P. Speissegger and A.J. Wilkie, *Quasianalytic Denjoy-Carleman classes and  $o$ -minimality*, *J. Amer. M. Soc.* **16** (2003), 751–777.
- [126] P. Sarnak, *Torsion points on varieties and homology of abelian covers*, 1988 preprint.
- [127] P. Sarnak and S. Adams, *Betti numbers of congruence groups (with an appendix by Z. Rudnick)*, *Israel J. Math.* **88** (1994), 31–72.
- [128] T. Scanlon, *The conjecture of Tate and Voloch on  $p$ -adic proximity to torsion*, *IMRN* **1999**, 909–914.
- [129] ———, *Local André-Oort conjecture for the universal abelian variety*, *Inventiones* **163** (2006), 191–211.
- [130] ———, *Counting special points: logic*, *Diophantine geometry and transcendence theory*, *Current Events Bulletin*, AMS, 2011, also *Bull. AMS* **49** (2012), 51–71.
- [131] T. Scanlon,  *$O$ -minimality as an approach to the André-Oort conjecture*, *Panoramas et Synthèses*, to appear.
- [132] A. Schinzel, *Reducibility of lacunary polynomials, X*, *Acta Arith.* **53** (1989), 47–97.
- [133] A. Seidenberg, *Abstract differential algebra and the analytic case*, *Proc. A. M. S.* **9** (1958), 159–164.
- [134] ———, *Abstract differential algebra and the analytic case. II*, *Proc. A. M. S.* **23** (1969), 689–691.

- [135] C.-L. Siegel, *Über die Classenzahl quadratischer Zahlkörper*, Acta Arithmetica **1** (1935), 83–86.
- [136] A. Silverberg, *Torsion points on abelian varieties of CM-type*, Compositio **68** (1988), 241–249.
- [137] C. M. Skinner and T. D. Wooley, *Sums of two  $k$ th powers*, Crelle **462** (1995), 57–68.
- [138] P. Speissegger, *The Pfaffian closure of an  $o$ -minimal structure*, J. Reine Angew. Math. **508** (1999), 198–211.
- [139] A. Tarski, *A Decision Method for Elementary Algebra and Geometry*, RAND Corporation, Santa Monica, 1948.
- [140] J. Tsimerman, *Brauer-Siegel for arithmetic tori and lower bounds for Galois orbits of special points*, J. Amer. Math. Soc. **25** (2012), 1091–1117.
- [141] ———, *Ax-Schanuel and  $o$ -minimality*, preprint available from the author’s webpage.
- [142] E. Ullmo, *Manin-Mumford, André-Oort, the equidistribution point of view*, Equidistribution in Number Theory, an Introduction, 103–138, NATO Sci. Ser. II Math. Phys. Chem **237**, Springer, Dordrecht, 2007.
- [143] ———, *Quelques applications du théorème de Ax-Lindemann hyperbolique*, Compositio **150** (2014), 175–190.
- [144] E. Ullmo and A. Yafaev, *Galois orbits and equidistribution of special subvarieties: towards the André-Oort conjecture*, Annals, to appear.
- [145] ———, *A characterisation of special subvarieties*, Mathematika **57** (2011), 833–842.
- [146] ———, *Nombres de classes des tores de multiplication complexe et bornes inférieures pour orbites Galoisiennes de points spéciaux*, arXiv:1209.0942, and Bull. SMF, to appear.
- [147] ———, *Hyperbolic Ax-Lindemann in the Cocompact case*, Duke Math. J. **163** (2014), 433–463.
- [148] A. J. Wilkie, *Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function*, J. Amer. M. Soc. **9** (1996), 1051–1094.
- [149] ———,  *$O$ -minimality*, Proc. ICM Berlin, 1998, Volume I, 633–636.
- [150] ———, *A theorem of the complement and some new  $o$ -minimal structures*, Selecta Math. N.S. **5** (1999), 397–421.
- [151] ———, *Diophantine properties of sets definable in an  $o$ -minimal structure*, J. Symbolic Logic **69** (2004), 851–861.
- [152] A. Yafaev, *The André-Oort conjecture—a survey*,  $L$ -functions and Galois Representations, Burns, Buzzard, and Nekovar, editors, London Math. Soc. Lecture Notes **320**, pp 381–406, CUP, 2007.
- [153] ———, *Galois orbits and equidistribution: Manin-Mumford and André-Oort*, J. Th. Nombres Bordeaux **21** (2009), 491–500.
- [154] Y. Yomdin,  *$C^k$ -resolution of semialgebraic mappings*, Israel J. Math. **57** (1987), 301–317.
- [155] D. Zagier, *Elliptic modular functions and their applications*, in The 1-2-3 of Modular Forms, Universitext, Springer, Berlin, 2008.

- [156] U. Zannier, *Some Problems of Unlikely Intersections in Arithmetic and Geometry*, with appendices by D. Masser, *Annals of Mathematics Studies* **181**, Princeton University Press, 2012.
- [157] ———, *Elementary integration of differentials in families and conjectures of Pink*, these proceedings.
- [158] B. Zilber, *Pseudo-exponentiation on algebraically closed fields of characteristic zero*, *Ann. Pure Appl. Logic* **132** (2005), 67–95.0
- [159] ———, *Exponential sums equations and the Schanuel conjecture*, *J. London Math. Soc. (2)* **65** (2002), 27–44.
- [160] ———, *Model Theory, Chapter X* in *A Course in Mathematical Logic for Mathematicians*, by Yu. I. Manin, second edition, *GTM 53*, Springer, New York, 2010.
- [161] ———, *Model theory of special subvarieties and Schanuel-type conjectures*, preprint, 2013.

Mathematical Institute, Andrew Wiles Building, University of Oxford, Oxford, UK.

E-mail: pila@maths.ox.ac.uk

# Quasi-randomness and the regularity method in hypergraphs

Vojtěch Rödl

**Abstract.** The probabilistic method is one of the most successful techniques in combinatorics. It enables one to prove results about deterministic objects by immersing them into specially designed probability spaces. One of the more recent techniques employs the idea of quasi-randomness. A quasi-random object is a deterministic object which shares important properties with “typical” objects of the same kind. Szemerédi’s regularity lemma asserts, quite remarkably, that every graph can be decomposed into relatively few subgraphs that are quasi-random. In appropriate situations quasi-randomness enables one to find and to enumerate subgraphs of a given isomorphism type. This approach has led to many applications in extremal combinatorics. We discuss some developments and applications of this method and focus on its extensions to hypergraphs.

**Mathematics Subject Classification (2010).** Primary 05C35; Secondary 05C65.

**Keywords.** Szemerédi’s theorem, removal lemma, quasi-randomness, Ramsey theory.

## 1. Introduction

*Probabilistic combinatorics* is a very active area in modern discrete mathematics with its fast development over the last few decades leading to the solution of several important problems in combinatorics. The use of probabilistic techniques in discrete mathematics was pioneered by Paul Erdős [19] more than sixty years ago. In his first applications of probabilistic arguments, he proved results about deterministic objects by embedding them into specially designed probability spaces. Since then, many new tools and advanced techniques have been developed.

One such technique is based on the notion of *quasi-randomness*. About forty years ago, E. Szemerédi pioneered the use of quasi-randomness in combinatorics leading to the *regularity method* in graph theory, which turned out to be a powerful technique in *extremal combinatorics*. Roughly speaking, the regularity method involves the decomposition of large structures into a bounded number of quasi-random “blocks,” that is, substructures that share fundamental properties with genuinely random structures of the same “order” and “density”. This research has some of its roots in *Ramsey theory* and we shall discuss these connections below.

Ramsey theory, named after the seminal contribution of F. P. Ramsey [48], is an important branch of combinatorics. A typical result in Ramsey theory asserts that for some given discrete structure  $F$  and some integer  $r$  there exists a discrete structure  $H$  such that any partition (or coloring) of  $H$  into  $r$  classes has the property that a copy of  $F$  is completely

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

contained in one of the  $r$  partition classes. For example, one of the first results of this type can be found in the work of D. Hilbert [40], where it was shown that for every  $\ell$  and for every finite partition of the natural numbers  $\mathbb{N} = \{1, 2, 3, \dots\}$  there exists a partition class which contains an affine cube of dimension  $\ell$ , i.e., a set of the form  $\{x_0 + \sum_{i=1}^{\ell} \varepsilon_i x_i : \varepsilon_i \in \{0, 1\}\}$  for some  $x_0, x_1, \dots, x_{\ell} \in \mathbb{N}$ . Another classical Ramsey-type result for the integers is van der Waerden's theorem, which we discuss below.

**1.1. Arithmetic progressions in Ramsey theory.** In 1927 van der Waerden [66] proved that every partition of the integers into finitely many classes yields arithmetic progressions of every finite length that are completely contained in one of the classes. We state the following finite version of it.

**Theorem 1.1** (van der Waerden's theorem). *For all integers  $r \geq 2$  and  $k \geq 3$  there exists some integer  $W = W(k, r)$  such that for every  $n \geq W$  the following holds. For every partition  $C_1 \cup \dots \cup C_r = [n] = \{1, \dots, n\}$  of the first  $n$  integers, there exists some partition class  $C_i$  such that  $C_i$  contains an arithmetic progression with  $k$  elements ( $AP_k$ ).*

It is of interest to find reasonable estimates for the smallest integer  $W = W(k, r)$  in Theorem 1.1. The original proof of van der Waerden is based on a double induction and the resulting bounds, even in the special case  $r = 2$ , are of Ackermann type, in sharp contrast to the best known lower bounds, which are only exponential in  $k$ . The upper bound was dramatically improved by Shelah [58] and the dependency of  $W$  on  $k$  (for  $r = 2$ ) was reduced to an iterated tower function. More precisely, for  $x \in \mathbb{N}$  we denote by  $\text{tow}(x)$  a tower of twos of height  $x$  and we define  $w(x)$  recursively by  $w(1) = 2$  and  $w(x) = \text{tow}(w(x - 1))$ . Shelah's proof yields  $W(k, 2) \leq w(O(k))$ .

The lack of a good upper bound partially motivated Erdős and Turán [23] to propose a strengthening of van der Waerden's theorem. They conjectured that the largest partition class must always contain an arithmetic progression of the required length, which would imply that the appearance of the arithmetic progression is already forced if a given subset of the first  $n$  integers has positive density. For integers  $k$  and  $n$  they denoted by  $r_k(n)$  the cardinality of the largest subsets of  $[n]$  that contains no arithmetic progression of length  $k$  ( $AP_k$ ), i.e.,

$$r_k(n) = \max\{|A| : A \subseteq [n] \text{ and } A \text{ contains no } AP_k\}.$$

Erdős and Turán conjectured that  $r_k(n) = o(n)$  for all  $k \geq 3$ . A positive resolution clearly implies van der Waerden's theorem, since  $r_k(n) \leq n/r$  implies  $W(k, r) \leq n$ .

The first nontrivial case  $k = 3$  of this conjecture was addressed by Roth [55]. Szemerédi first proved the conjecture for  $k = 4$  in [60] and in 1975 he settled the conjecture for every  $k$  [61].

**Theorem 1.2** (Szemerédi's theorem). *For every integer  $k \geq 3$  and every  $\delta > 0$  there exists an  $n_0$  such that for every  $n \geq n_0$  we have  $r_k(n) \leq \delta n$ .*

Since Szemerédi's combinatorial proof used van der Waerden's theorem, it did not give new bounds for  $W(k, r)$ . Shortly after Szemerédi's proof appeared, Furstenberg [28] found a new proof of Theorem 1.2. This new proof was based on ergodic theory and on the axiom of choice and as a consequence it did not yield quantitative estimates without the addition of substantial further ideas (which were supplied in [63]). Eventually Gowers obtained another proof of Szemerédi's theorem [34], which used, among many other things, exponential-sum

estimates, which also played a crucial role in Roth’s proof [55] for  $k = 3$ . The work of Gowers implies the best known quantitative estimates for Szemerédi’s theorem for general  $k$  (see Sanders [57] for the best current bound for  $k = 3$  and for  $k = 4$  see Green and Tao [37]). In particular, the work of Gowers implies the following bound on  $W(k, r)$  in van der Waerden’s theorem

$$W(k, r) \leq 2^{2^{r \cdot 2^{k+9}}}.$$

**1.2. Triangle removal lemma.** Together with Frankl, we studied another, purely combinatorial, approach to Szemerédi’s theorem, which is based on an extremal hypergraph problem that can be traced back to the work of Ruzsa and Szemerédi. These authors observed that Roth’s theorem (Theorem 1.2 for  $k = 3$ ) follows from the so-called *triangle removal lemma* in graph theory. Below we state this lemma in its modern form. Note that the statement proved by Ruzsa and Szemerédi is slightly weaker, however, their proof can be adjusted to give the following result, which asserts that graphs  $G = (V, E)$  that contain  $o(|V|^3)$  triangles can be made triangle-free by the removal of  $o(|V|^2)$  edges.

**Theorem 1.3** (Triangle removal lemma). *For every  $\delta > 0$  there exist some  $c > 0$  and  $n_0$  such that every graph  $G = (V, E)$  on  $|V| = n \geq n_0$  vertices that contains at most  $cn^3$  triangles, i.e., at most  $cn^3$  complete subgraphs on three vertices, has a subgraph  $\hat{G} = (V, \hat{E})$  with no triangles at all and with  $|E \setminus \hat{E}| \leq \delta n^2$ .*

All known proofs of Theorem 1.3 are based on Szemerédi’s regularity lemma for graphs (see Theorem 3.1) or use some variant of this lemma and we will give such a proof in Section 3.1. Currently, the best known numerical dependency is given by a proof of Fox [24], which requires  $c^{-1}$  to be a tower of twos with height linear in the logarithm of  $\delta^{-1}$ , i.e.,  $c^{-1} = \text{tow}(C \log(\delta^{-1}))$  for some constant  $C > 0$ .

Ruzsa and Szemerédi [56] observed that Roth’s theorem (Theorem 1.2 for  $k = 3$ ) follows from Theorem 1.3 by a simple construction. Let  $A \subseteq [n]$  contain no  $AP_3$ . With  $A$  we associate a three-partite graph  $G = (V_1 \cup V_2 \cup V_3, E)$  whose vertex set consists of three (formally) disjoint sets  $V_1 = [n]$ ,  $V_2 = [2n]$ , and  $V_3 = [3n]$ . For every  $x \in [n]$  and  $a \in A$  we include the triangle on the vertices

$$x \in V_1, \quad x + a \in V_2, \quad \text{and} \quad x + 2a \in V_3.$$

By definition every edge of  $G$  is contained in at least one triangle. On the other hand, if  $x \in V_1, y \in V_2$ , and  $z \in V_3$  spans a triangle in  $G$ , then there exist  $a, a',$  and  $a'' \in A$  with

$$a = y - x, \quad a' = z - 2x, \quad \text{and} \quad a'' = 2y - z,$$

which yields  $a' + a'' = 2a$ . Since  $A$  contains no  $AP_3$ , this implies  $a = a' = a''$  and as a consequence,  $y = x + a$  and  $z = 2x + a$ . Therefore,  $G$  has the property that every edge of  $G$  is contained in precisely one triangle. In particular, we showed that the graph  $G$  defined by  $A$  has the property that every edge is contained in precisely one triangle and that  $|E| = 3n|A|$  and  $|V| = 6n$ .

Hence, the number of triangles is bounded by  $n|A| \leq n^2 = o(|V|^3)$  and the triangle removal lemma implies that we can remove  $o(n^2)$  edges from  $G$  to obtain a triangle-free graph. Since every removed edge destroys at most one triangle, the number of triangles  $n|A|$  is bounded by  $o(n^2)$  and Roth’s theorem follows.

With Frankl we observed that several possible generalizations of the triangle removal lemma imply Szemerédi's theorem by similar constructions (see, e.g., [22, 26, 49]). One of these generalization was first proved with Frankl [26] for 3-uniform hypergraphs. For higher uniformities it was proved independently by Gowers [35] and in joint work with Nagle, Schacht, and Skokan [45, 54]. Several other proofs of Szemerédi's theorem appeared over the last decade (see [38, page 272]).

## 2. Removal lemma and applications

In this section, we introduce an extension of Theorem 1.3 and state some of its applications. We recall that a  $k$ -uniform hypergraph with vertex set  $V$  is a pair  $H = (V, E)$ , where  $E \subseteq \binom{V}{k} = \{K \subseteq V : |K| = k\}$  is a system of  $k$ -tuples, called *hyperedges* of  $H$ . A  $k$ -uniform hypergraph  $H' = (V', E')$  is a *subhypergraph* of  $H$  if  $V' \subseteq V$  and  $E' \subseteq E$ . A *clique*  $K_\ell^{(k)}$  is a *complete*  $k$ -uniform hypergraph with  $\ell$  vertices and all  $\binom{\ell}{k}$  hyperedges.

**Theorem 2.1** (Removal lemma). *For every  $\delta > 0$  and every  $k \geq 2$ , there exist  $c > 0$  and  $n_0$  such that any  $k$ -uniform hypergraph  $H$  with  $n \geq n_0$  vertices that contains at most  $cn^{k+1}$  cliques  $K_{k+1}^{(k)}$  has a subhypergraph  $\widehat{H} = (V, \widehat{E})$  with no copy of  $K_{k+1}^{(k)}$  at all and with  $|E \setminus \widehat{E}| \leq \delta n^k$ .*

In other words, any hypergraph  $H$  containing  $o(n^{k+1})$  cliques  $K_{k+1}^{(k)}$  can be made  $K_{k+1}^{(k)}$ -free by the omission of  $o(n^k)$  hyperedges. We will also use the following immediate corollary of Theorem 2.1

**Corollary 2.2.** *Let  $k \geq 2$  and let  $H = (V, E)$  be a  $k$ -uniform hypergraph on  $n$  vertices with the property that every hyperedge  $e \in E$  is contained in precisely one clique  $K_{k+1}^{(k)}$  in  $H$ . Then  $|E| = o(n^k)$ .*

Corollary 2.2 follows easily from Theorem 2.1. Suppose every hyperedge of an  $n$ -vertex hypergraph  $H = (V, E)$  is contained in precisely one  $K_{k+1}^{(k)}$ , then the number of copies  $\#\{K_{k+1}^{(k)} \subseteq H\}$  of  $K_{k+1}^{(k)}$  in  $H$  satisfies

$$\#\{K_{k+1}^{(k)} \subseteq H\} = \frac{|E|}{k+1} \leq \binom{n}{k} = o(n^{k+1}). \quad (2.1)$$

Hence, it follows from Theorem 2.1 that we may remove  $o(n^k)$  hyperedges of  $H$  to be left with a  $K_{k+1}^{(k)}$ -free subhypergraph. Since every omitted hyperedge destroys at most one copy of  $K_{k+1}^{(k)}$  this means that the number of copies of  $K_{k+1}^{(k)}$  was bounded by  $o(n^k)$ . Therefore, from the first identity in (2.1), we obtain  $|E| = o(n^k)$ .

**2.1. Removal lemma and density theorems.** The removal lemma for hypergraphs (Theorem 2.1) implies Szemerédi's theorem (Theorem 1.2) by generalizing the construction following Theorem 1.3 (see, e.g., [26]). We omit this reduction here and discuss the connection of the removal lemma with density versions of Ramsey-type theorems, first obtained by Furstenberg and Katznelson.

In 1977 Furstenberg [28] gave an alternative proof of Szemerédi's theorem. Subsequently, Furstenberg and Katznelson [29, 30] refined this proof and were able to derive

several generalizations. The following result from [29] can be viewed as a density version of the Gallai-Witt theorem (see [47, p. 123] and [67]).

**Theorem 2.3.** *Let  $K$  be a finite subset of  $\mathbb{R}^m$  and let  $\delta > 0$ . Then there exists a finite subset  $W \subset \mathbb{R}^m$  such that any  $Y \subset W$  with  $|Y| > \delta|W|$  contains a translated, scaled copy of  $K$ , i.e., a set of the form  $y + \lambda K$  for some  $y \in \mathbb{R}^m$  and some  $\lambda \neq 0$ .*

*If in addition  $K \subset [k]^m$  for some positive integer  $k$ , then  $W = [N]^m$  has the above property for any sufficiently large  $N = N(k, m, \delta)$ .*

Note that for  $m = 1$  and  $K = [k]$  Theorem 2.3 reduces to Theorem 1.2. Below we will follow a construction of Solymosi [59] to deduce Theorem 2.3 from the removal lemma in the case when  $K$  is a particular simplex. The general result follows by a similar construction and appropriate projections (see, e.g., [53, 59]).

**Corollary 2.4.** *For every  $k \geq 2$  and  $\delta > 0$ , there exists  $N_0$  such that for every  $N \geq N_0$  every  $A \subset [N]^k$  with  $|A| > \delta N^k$  contains a translated, scaled copy of the simplex  $S_k$  consisting of the zero-vector and the standard basis  $e_1, \dots, e_k$  of  $\mathbb{R}^k$ .*

*Proof of Corollary 2.4 from Corollary 2.2.* Let  $k \geq 2$  be fixed and assume that for some  $\delta > 0$  and an infinite sequence of  $N \in \mathbb{N}$  there is a set  $A \subset [N]^d$  with  $|A| \geq \delta N^k$  that contains no translated, scaled copy of the simplex  $S_k$ .

We consider the  $(k + 1)$ -partite,  $k$ -uniform hypergraph  $H = (V, E)$  with vertex partition  $V = V_0 \cup \dots \cup V_k$ , where each  $V_i$  represents all affine hyperplanes parallel to one of the faces of the simplex  $S_k$  that intersect a point from  $[N]^k$ . Let

$$V_0 = \{D(t) : t = k, \dots, kN\} \text{ where } D(t) = \{(x_1, \dots, x_k) \in [N]^k : \sum_{i=1}^k x_i = t\}$$

consists of hyperplanes parallel to the face spanned by the standard basis and the collection of these planes covers all points of  $[N]^k$ . For  $i = 1, \dots, k$  the vertex class  $V_i$  consists of hyperplanes that are perpendicular to  $e_i$ , i.e.,

$$V_i = \{X_i(t) : t \in [N]\} \text{ where } X_i(t) = \{(x_1, \dots, x_k) \in [N]^k : x_i = t\}.$$

It follows directly from the definition that  $|V_0| = kN - k + 1$  and  $|V_i| = N$  for every  $i = 1, \dots, k$ . Note that any selection of  $k$  hyperplanes from  $k$  different vertex classes intersect in precisely one point in  $[N]^k$  and we include the hyperedge consisting of these  $k$  hyperplanes in  $H$  if and only if their intersection is in  $A$ .

Moreover, for every point  $a \in A$  and for every  $i = 0, \dots, k$  there exists a unique hyperplane in  $V_i$ , which contains  $a$ . Consequently,  $|E| = (k + 1)|A|$  and the  $k + 1$  hyperplanes containing a given  $a \in A$  form a clique  $K_{k+1}^{(k)}$  in  $H$ . On the other hand, the vertices of any clique that is *not* of that form would define a translated, scaled copy of  $S_k$  in  $A$ . Indeed the  $k + 1$  points in  $A$  obtained as intersections of any  $k$  of the  $k + 1$  corresponding hyperplanes define such a copy.

Therefore,  $H$  is a  $k$ -uniform hypergraph consisting of  $2kN - k + 1$  vertices and  $(k + 1)|A|$  hyperedges, and every hyperedge is contained in precisely one  $K_{k+1}^{(k)}$ . Corollary 2.2 implies  $|E| = o(N^k)$ , which yields the contradiction  $|A| = o(N^k)$ .  $\square$

We also state another result of Furstenberg and Katznelson [30], which addresses affine subspaces in high dimensional vector spaces over finite fields.

**Theorem 2.5.** *Let  $\mathbb{F}_q$  be the finite field with  $q$  elements. Then for every positive integer  $k$  and every  $\delta > 0$ , there exists  $n_0 = n_0(q, k, \delta)$  such that for  $n \geq n_0$  any subset  $A \subset \mathbb{F}_q^n$  with  $|A| > \delta|\mathbb{F}_q^n| = \delta q^n$  contains a  $k$ -dimensional affine subspace.*

We remark that Theorem 2.5 can be viewed as a density version of a special case of the Graham-Rothschild-Leeb theorem [36] in Ramsey theory. This latter theorem resolves a conjecture of Gian-Carlo Rota and asserts that for all integers  $k \geq \ell$  and  $r$ , and every finite field  $\mathbb{F}_q$  any partition into  $r$  classes of the  $\ell$ -dimensional affine subspaces of  $\mathbb{F}_q^n$  for sufficiently large  $n$  yields a  $k$ -dimensional affine subspace such that all its  $\ell$ -dimensional affine subspaces belong to the same partition class. Theorem 2.5 strengthens this result for  $\ell = 0$  in the same way as Szemerédi’s theorem strengthens van der Waerden’s theorem. We also note that for  $\ell > 0$  such a density version is known to be false.

For  $k = 1$  the reduction of Theorem 2.5 to the removal lemma appeared in joint work with Frankl [26] and we will give a similar reduction below. For general  $k$  such a reduction appeared in [53].

*Proof of Theorem 2.5 for  $k = 1$ .* Let  $\alpha_0 = 0, \alpha_1 = 1, \alpha_2, \dots, \alpha_{q-1}$  be the elements of  $\mathbb{F}_q$  and let  $A \subseteq \mathbb{F}_q^n$  with  $|A| \geq \delta q^n$  for some sufficiently large  $n$  be given. Suppose  $A$  contains no affine line from  $\mathbb{F}_q^n$ . We consider a  $q$ -partite,  $(q-1)$ -uniform hypergraph  $H = (V, E)$  with vertex classes  $V = V_0 \cup \dots \cup V_{q-1}$ . Here every set  $V_i$  is a copy of  $\mathbb{F}_q^n$ . For  $j = 0, \dots, q-1$  and  $\mathbf{v}_i \in V_i$  for  $i \neq j$ , we include the hyperedge  $\{\mathbf{v}_0, \dots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \dots, \mathbf{v}_{q-1}\}$  in  $H$  if

$$\sum_{i \neq j} (\alpha_j - \alpha_i) \mathbf{v}_i \in A \tag{2.2}$$

and again we will show that every hyperedge of  $H$  is contained in precisely one clique  $K_q^{(q-1)}$ . For any fixed  $j \in \{0, \dots, q-1\}$  we observe that a hyperedge  $\{\mathbf{v}_0, \dots, \mathbf{v}_{j-1}, \mathbf{v}_{j+1}, \dots, \mathbf{v}_{q-1}\}$  is contained in a clique spanned by  $\{\mathbf{v}_0, \dots, \mathbf{v}_{q-1}\}$  where

$$\mathbf{v}_j = - \sum_{i \neq j} \mathbf{v}_i.$$

Indeed, the choice of  $\mathbf{v}_j$  implies that  $\sum_{i=0}^{q-1} \mathbf{v}_i = \mathbf{0}$  and, hence,

$$\sum_{i \neq \ell} (\alpha_\ell - \alpha_i) \mathbf{v}_i = \sum_{i=0}^{q-1} (\alpha_\ell - \alpha_i) \mathbf{v}_i + (\alpha_j - \alpha_\ell) \sum_{i=0}^{q-1} \mathbf{v}_i \stackrel{(2.2)}{\in} A$$

for every  $\ell \in \{0, \dots, q-1\}$ . Hence, it follows from the construction of  $H$  that  $\{\mathbf{v}_0, \dots, \mathbf{v}_{q-1}\}$  spans a clique  $K_q^{(q-1)}$ .

On the other hand, if  $\{\mathbf{v}_0, \dots, \mathbf{v}_{q-1}\}$  spans a clique, then there exist vectors  $\mathbf{a}_0, \dots, \mathbf{a}_{q-1} \in A$  such that for every  $j = 0, \dots, q-1$  we have

$$\mathbf{a}_j = \sum_{i \neq j} (\alpha_j - \alpha_i) \mathbf{v}_i = \sum_{i=0}^{q-1} (\alpha_j - \alpha_i) \mathbf{v}_i.$$

Recalling that  $\alpha_0 = 0$ , we see that

$$\mathbf{a}_j - \mathbf{a}_0 = \sum_{i=0}^{q-1} (\alpha_j - \alpha_i - \alpha_0 + \alpha_i) \mathbf{v}_i = \alpha_j \sum_{i=0}^{q-1} \mathbf{v}_i$$

for every  $j = 0, \dots, q-1$ . Since  $A$  contains no affine line from  $\mathbb{F}_q^n$  this implies that  $\sum_{i=0}^{q-1} \mathbf{v}_i$  must be the zero-vector in  $\mathbb{F}_q^n$  and  $\mathbf{a}_0 = \dots = \mathbf{a}_{q-1}$ . Therefore, if  $\mathbf{v}_0, \dots, \mathbf{v}_{q-1}$  spans a clique, then  $\sum_{i=0}^{q-1} \mathbf{v}_i = \mathbf{0}$ .

Summarizing, we showed that every hyperedge of  $H$  is contained in precisely one clique  $K_q^{(q-1)}$ . For every given  $\mathbf{a} \in A$  and any choice of  $\mathbf{v}_0 \in V_0, \dots, \mathbf{v}_{q-3} \in V_{q-3}$  there exists a unique  $\mathbf{v}_{q-2} \in V_{q-2}$  such that the defining sum from (2.2) gives  $\mathbf{a}$ . Therefore, the vertex classes  $V_0, \dots, V_{q-2}$  span  $|A|q^{n(q-2)}$  hyperedges. Owing to the symmetry of this construction we see that  $H$  contains

$$|E| = q|A|q^{n(q-2)}$$

hyperedges. Moreover,  $|V| = q^{n+1}$  and, therefore, Corollary 2.2 applied for  $(q-1)$ -uniform hypergraphs implies

$$|A| = \frac{|E|}{q^{n(q-2)+1}} = \frac{o(|V|^{q-1})}{q^{n(q-2)+1}} = o(q^{(n+1)(q-1)-n(q-2)-1}) = o(q^{n+q-2}) = o(q^n),$$

since  $q$  is fixed and  $n \rightarrow \infty$ . □

The original proofs of Theorems 2.3 and 2.5, due to Furstenberg and Katznelson [29, 30], were based on ergodic theory and gave no quantitative bounds for the involved parameters. The proofs based on the removal lemma gave the first quantitative proofs, though the bounds are quite poor and are of Ackermann type.

The density version of the Hales-Jewett theorem [39], which was obtained by Furstenberg and Katznelson [31], provides a natural extension of Theorem 2.5. Roughly speaking, this abstract generalisation asserts that Theorem 2.5 holds even if one requires that the  $k$ -dimensional affine subspaces have a very special nature. The density version of the Hales-Jewett theorem not only extends Theorem 2.5, but also implies Szemerédi’s theorem and its multidimensional version (Theorem 2.3).

Owing to these connections it seemed to be natural to investigate whether the removal lemma itself, or some appropriate variant of it, could lead to a new proof of the density version of the Hales-Jewett theorem. In fact, the first Polymath project initiated by Gowers explored such a possibility, though in the end this project culminated with a different combinatorial proof [46] (see [6] for another proof obtained at that time and [16] for a recent new proof).

**2.2. A generalized removal lemma and property testing.** In Section 2 we stated a special case of the removal lemma concerning  $k$ -uniform hypergraphs with few copies of the clique  $K_{k+1}^{(k)}$  (Theorem 2.1). There is nothing special about the clique and all known proofs easily extend to arbitrary hypergraphs. For graphs such a statement first appeared in [27].

**Theorem 2.6.** *Let  $\delta > 0$  and let  $\ell \geq k \geq 2$  be integers. For every  $k$ -uniform hypergraph  $F$  on  $\ell$  vertices there exist constants  $c > 0$  and  $n_0$  such that the following holds for every  $k$ -uniform hypergraph  $H$  on  $n \geq n_0$  vertices. If  $H$  contains at most  $cn^\ell$  copies of  $F$ , then one can delete  $\delta n^k$  hyperedges from  $H$  so that the resulting subhypergraph  $\widehat{H}$  contains no copy of  $F$ .*

Here we discuss further generalizations of Theorem 2.6. One possible generalization of Theorem 2.6 is to replace the single hypergraph  $F$  by a possibly infinite family  $\mathcal{F}$  of  $k$ -uniform hypergraphs. Such a result was first proved for graphs by Alon and Shapira [4] in

the context of property testing. For a family of graphs  $\mathcal{F}$  consider the class  $\text{Forb}(\mathcal{F})$  of all graphs  $H$  containing no member of  $\mathcal{F}$  as a subgraph. Clearly  $\text{Forb}(\mathcal{F})$  is monotone, i.e., if  $H \in \text{Forb}(\mathcal{F})$  and  $H'$  is a subgraph of  $H$  (obtained from  $H$  by successive vertex and edge deletions), then  $H' \in \text{Forb}(\mathcal{F})$ . Moreover, it is easy to see that for every monotone family of graphs  $\mathcal{P}$  (a so-called *monotone property*  $\mathcal{P}$ ) there exists a family  $\mathcal{F}$  such that  $\mathcal{P} = \text{Forb}(\mathcal{F})$ . Alon and Shapira proved the following in [4].

**Theorem 2.7.** *For every  $\delta > 0$  and for (a possibly infinite) family of graphs  $\mathcal{F}$  there exist constants  $c > 0$ ,  $L > 0$ , and  $n_0$  such that the following holds for every graph  $G$  on  $n \geq n_0$  vertices. If for every  $\ell = 1, \dots, L$  and every  $F \in \mathcal{F}$  on  $\ell$  vertices,  $G$  contains at most  $cn^\ell$  copies of  $F$ , then one can delete  $\delta n^2$  edges from  $G$  so that the resulting subgraph  $\hat{G}$  contains no copy of any member of  $\mathcal{F}$ .*

Theorem 2.6 for  $k = 2$  is equivalent to Theorem 2.7 in the special case when  $\mathcal{F}$  consists of only one graph. While for finite families  $\mathcal{F}$  Theorem 2.7 can be proved along the lines of the proof of Theorem 2.6 (or be deduced from Theorem 2.6 directly), for infinite families  $\mathcal{F}$  the proof of Theorem 2.7 is more sophisticated.

Perhaps one of the earliest results of this nature was obtained by Bollobás, Erdős, Simonovits, and Szemerédi [8], who essentially proved Theorem 2.7 for the special family  $\mathcal{F}$  of blow-ups of odd cycles. In [17] answering a question of Erdős (see, e.g., [20]) with Duke we generalized the result from [8] and proved Theorem 2.7 for the families  $\mathcal{F}$  of  $(r + 1)$ -chromatic graphs  $r \geq 2$ .

The proof of Theorem 2.7 relies on a strengthened version of Szemerédi's regularity lemma (see Theorem 3.1), which was obtained by Alon, Fischer, Krivelevich, and M. Szegedy [2] by iterating Szemerédi's regularity lemma for graphs.

Another natural variant of Theorem 2.6 is an *induced* version. We say a graph  $G$  contains an *induced* copy of  $F$  if there is an injective map from  $V(F)$  to  $V(G)$  which preserves edges and non-edges. For graphs this was first considered by Alon, Fischer, Krivelevich, and M. Szegedy [2]. Note that, if one wishes to “destroy” *induced* copies of a given graph  $F$  in a graph  $G$  one may not only remove edges from  $G$  but one may also *add* edges to  $G$ .

**Theorem 2.8.** *For every  $\delta > 0$  and for every graph  $F$  on  $\ell$  vertices there exist constants  $c > 0$  and  $n_0$  such that the following holds for every graph  $G = (V, E)$  on  $n \geq n_0$  vertices. If  $G$  contains at most  $cn^\ell$  induced copies of  $F$ , then one can change  $\delta n^2$  pairs from  $V$  (deleting or adding an edge) so that the resulting graph  $\hat{G}$  contains no induced copy of  $F$ .*

In [3] Alon and Shapira proved a common generalization of Theorem 2.7 and Theorem 2.8, extending Theorem 2.8 from forbidding a single induced graph  $F$  to forbidding a family of graphs  $\mathcal{F}$  as induced subgraphs. With Schacht [52] we obtained the hypergraph generalisation of this result (see Theorem 2.9 below).

For a family of  $k$ -uniform hypergraphs  $\mathcal{F}$ , let  $\text{Forb}_{\text{ind}}(\mathcal{F})$  be the family of all hypergraphs  $H$  that contain no induced copy of any member of  $\mathcal{F}$ . Clearly,  $\text{Forb}_{\text{ind}}(\mathcal{F})$  is a *hereditary property* of hypergraphs, i.e., if  $H \in \text{Forb}_{\text{ind}}(\mathcal{F})$  and  $H[U]$  is the induced subhypergraph on some  $U \subseteq V(H)$ , then  $H[U] \in \text{Forb}_{\text{ind}}(\mathcal{F})$ .

We denote by  $H \Delta \hat{H}$  the symmetric difference of the sets of hyperedges of two hypergraphs  $H$  and  $\hat{H}$ . For a constant  $\delta > 0$  and a possibly infinite family of  $k$ -uniform hypergraphs  $\mathcal{P}$  we say a given hypergraph  $H$  is  $\delta$ -far from  $\mathcal{P}$  if every hypergraph  $G \in \mathcal{P}$  on the same vertex set satisfies  $|H \Delta G| > \delta n^k$ .

**Theorem 2.9.** *For every  $\delta > 0$  and for every (possibly infinite) family  $\mathcal{F}$  of  $k$ -uniform hypergraphs there exist constants  $c > 0$ ,  $L > 0$ , and  $n_0$  such that the following holds for every  $k$ -uniform hypergraph  $H$  on  $n \geq n_0$  vertices. If for every  $\ell = 1, \dots, L$  and every  $F \in \mathcal{F}$  on  $\ell$  vertices,  $H$  contains at most  $cn^\ell$  induced copies of  $F$ , then  $H$  is not  $\delta$ -far from  $\text{Forb}_{\text{ind}}(\mathcal{F})$ .*

Theorem 2.9 asserts that if  $H$  contains only “a few” induced copies of all forbidden hypergraphs  $F$  up to some bounded size, then one can alter at most  $o(n^k)$   $k$ -tuples so that the resulting hypergraph  $\hat{H}$  contains no induced copy of any member of  $\mathcal{F}$ , i.e., so that  $\hat{H} \in \text{Forb}_{\text{ind}}(\mathcal{F})$ .

As mentioned above, for graphs Theorem 2.9 was first obtained by Alon and Shapira [3], whose proof is again based on the strong version of Szemerédi’s regularity lemma from [2]. Another proof for graphs was found by Lovász and B. Szegedy [43] (see also [9]). Below we discuss a consequence of Theorem 2.9 in the area of *property testing* introduced by Goldreich, Goldwasser, and Ron [32].

Recall that for every hereditary property  $\mathcal{P}$  of  $k$ -uniform hypergraphs, there exists a family of  $k$ -uniform hypergraphs  $\mathcal{F}$  such that  $\mathcal{P} = \text{Forb}_{\text{ind}}(\mathcal{F})$ . Consequently, Theorem 2.9 states that if  $H$  is  $\delta$ -far from some hereditary property  $\mathcal{P} = \text{Forb}_{\text{ind}}(\mathcal{F})$ , then it contains many ( $cn^{|V(F)|}$ ) induced copies of some “forbidden” hypergraph  $F \in \mathcal{F}$  of size at most  $L$ , which “proves” that  $H$  is not in  $\mathcal{P}$ . In other words, if  $H$  is  $\delta$ -far from some given hereditary property  $\mathcal{P}$ , then it is “easy” to detect that  $H \notin \mathcal{P}$ . More precisely, we say a property  $\mathcal{P}$  of hypergraphs is *testable with one-sided error* if for every  $\delta > 0$  there exists a constant  $q = q(\mathcal{P}, \delta)$  and a randomized algorithm  $\mathcal{A}$  satisfying: *For a given hypergraph  $H = (V, E)$  the algorithm  $\mathcal{A}$  can query some oracle whether a  $k$ -tuple  $K$  of  $V$  spans a hyperedge in  $H$  or not. After at most  $q$  queries the algorithm outputs*

- $H \in \mathcal{P}$  with probability 1 if  $H \in \mathcal{P}$  and
- $H \notin \mathcal{P}$  with probability at least  $2/3$  if  $H$  is  $\delta$ -far from  $\mathcal{P}$ .

*If  $H \notin \mathcal{P}$  and not  $\delta$ -far from  $\mathcal{P}$ , then there are no guarantees for the output of  $\mathcal{A}$ .*

Furthermore, we say a property  $\mathcal{P}$  is *decidable* if there is an algorithm which for every hypergraph  $H$  distinguishes in finite time if  $H \in \mathcal{P}$  or  $H \notin \mathcal{P}$ . In this context Theorem 2.9 implies the following result, which generalizes to hypergraphs a result for graphs proved by Alon and Shapira [3].

**Corollary 2.10.** *Every decidable, hereditary property of hypergraphs is testable with one-sided error.*

*Proof.* Let a decidable, hereditary property  $\mathcal{P} = \text{Forb}_{\text{ind}}(\mathcal{F})$  and some  $\delta > 0$  be given. By Theorem 2.9, there exist constants  $c > 0$ ,  $L$  and  $n_0 \in \mathbb{N}$  such that any  $k$ -uniform hypergraph on  $n \geq n_0$  vertices which is  $\delta$ -far from exhibiting  $\mathcal{P}$  contains at least  $cn^{|V(F)|}$  copies of some  $F \in \mathcal{F}$  with  $|V(F)| \leq L$ .

Let  $s \in \mathbb{N}$  be such that  $(1 - c)^{\lfloor s/L \rfloor} < 1/3$  and set  $m = \max\{s, n_0\}$ . We claim that there is a one-sided tester with query complexity  $q = \binom{m}{k}$  for  $\mathcal{P}$ . Let  $H$  be a  $k$ -uniform hypergraph on  $n$  vertices. If  $n \leq m$ , then the tester simply queries all edges of  $H$  and since  $\mathcal{P}$  is decidable, there is an exact algorithm with running time only depending on the fixed  $m$  which determines correctly if  $H \in \mathcal{P}$  or not.

Hence we may assume that  $n > m$ . Then we choose uniformly at random a set  $S$  of  $s$  vertices from  $H$  and consider the hypergraph  $H[S] = H \cap \binom{S}{k}$  induced on  $S$ . If  $H[S]$

has  $\mathcal{P}$ , then the tester outputs “ $H \in \mathcal{P}$ ” and otherwise “ $H \notin \mathcal{P}$ .” Since  $\mathcal{P}$  is decidable and  $s$  is fixed the algorithm decides whether or not  $H[S]$  is in  $\mathcal{P}$  in constant time (constant only depending on  $s$  and  $\mathcal{P}$ ).

Clearly, if  $H \in \mathcal{P}$  or  $n \leq m$ , then this tester outputs correctly and hence it is one-sided. On the other hand, if  $H$  is  $\delta$ -far from  $\mathcal{P}$  and  $n > m$ , then because of Theorem 2.9 a random  $\ell$ -element set spans a copy of  $F$  for some  $F \in \mathcal{F}$  on  $\ell \leq L$  vertices with probability at least  $cn^\ell / \binom{n}{\ell} \geq c$ . Hence the probability that  $S$  does not span any copy of  $F$  is at most

$$(1 - c)^{\lfloor s/\ell \rfloor} \leq (1 - c)^{\lfloor s/L \rfloor} < 1/3.$$

In other words,  $S$  spans a copy of  $F$  with probability at least  $2/3$ . □

We remark that Corollary 2.10 only yields the existence of such a testing algorithm. In fact, the parameter  $L$  in Theorem 2.9 is not necessarily computable and there exist decidable, hereditary properties for which the query complexity  $q$  is not computable (see [5] for details). Theorem 2.9 was further extended by Austin and Tao [7] and a more thorough discussion of the numerical aspects of removal lemmas in graphs can be found in the recent survey of Conlon and Fox [14].

### 3. Regularity method for graphs

All known proofs of the removal lemma rely on the *regularity method* for graphs and hypergraphs or variants of it. Here we describe the proof of Theorem 1.3 based on Szemerédi’s regularity lemma for graphs from [62] and discuss the connection between the regularity lemma and theory of *quasi-random graphs*.

**3.1. Szemerédi’s regularity lemma.** Szemerédi’s regularity lemma [62] is an important tool in *extremal and probabilistic combinatorics* with applications in *graph theory, combinatorial number theory, discrete geometry, and theoretical computer science*. Roughly speaking, the regularity lemma asserts that any given graph  $G = (V, E)$  can be “approximated” by a bounded number of quasi-random bipartite graphs, i.e., by bipartite graphs that share important properties with random bipartite graphs of the same order and density. The quasi-randomness is made precise by a quantitative version of uniform edge distribution. More precisely, given a graph  $G = (V, E)$  and a pair of disjoint non-empty subsets  $A, B \subseteq V$  we define the *density* of the bipartite subgraph induced on  $A$  and  $B$  by

$$d(A, B) = \frac{e(A, B)}{|A||B|},$$

where  $e(A, B)$  denotes the number of edges  $\{a, b\} \in E$  with  $a \in A$  and  $b \in B$ . Such a pair  $(A, B)$  is  $\varepsilon$ -regular if

$$|d(A, B) - d(A', B')| < \varepsilon \tag{3.1}$$

for all subsets  $A' \subseteq A$  and  $B' \subseteq B$  with  $|A'| > \varepsilon|A|$  and  $|B'| > \varepsilon|B|$ . Moreover, we say  $(A, B)$  is  $(\varepsilon, d)$ -regular for some  $d \geq 0$  if we can replace  $d(A, B)$  in (3.1) by  $d$ .

Note that  $\varepsilon$ -regularity is a property shared by most bipartite graphs with a given number of edges, i.e., for any  $\varepsilon > 0$ , there are  $(1 - o(1)) \binom{|A||B|}{m}$  bipartite  $\varepsilon$ -regular graphs with bipartition  $(A, B)$  and  $m$  edges, where  $o(1) \rightarrow 0$  as  $|A|, |B| \rightarrow \infty$ .

The regularity lemma asserts that for any given  $\varepsilon > 0$ , the edge set of any sufficiently large graph can be decomposed into a bounded number of blocks, almost all of which are  $\varepsilon$ -regular.

**Theorem 3.1** (Szemerédi’s regularity lemma). *Let  $\varepsilon > 0$  and an integer  $t_0$  be given. There exists a positive integer  $T_0 = T_0(\varepsilon, t_0)$  such that any graph  $G = (V, E)$  admits a partition  $V = V_1 \cup V_2 \cup \dots \cup V_t$  satisfying*

- (i)  $t_0 \leq t \leq T_0$
- (ii)  $|V_1| \leq |V_2| \leq \dots \leq |V_t| \leq |V_1| + 1$ , and
- (iii) all but at most  $\varepsilon t^2$  pairs  $(V_i, V_j)$  are  $\varepsilon$ -regular.

One of the most important consequences, which is used in many applications of the regularity lemma, is that  $\varepsilon$ -regularity can be used to embed given graphs. The following observation, sometimes called the *counting lemma*, can be easily proved.

**Proposition 3.2** (Counting lemma). *For all  $\ell \geq 2$  and constants  $\gamma > 0$  and  $d > 0$  there exist  $\varepsilon = \varepsilon(\ell, d, \gamma) > 0$  and  $m_0 = m_0(\ell, d, \gamma)$  so that the following holds.*

*Let  $G = (V_1 \cup \dots \cup V_\ell, E)$  be an  $\ell$ -partite graph with vertex set  $V = \bigcup_{i=1}^\ell V_i$  and  $|V_1| = |V_2| = \dots = |V_\ell| = m \geq m_0$ , where all pairs  $(V_i, V_j)$  with  $1 \leq i < j \leq \ell$  are  $\varepsilon$ -regular with density at least  $d$ . Then the number of copies of  $K_\ell$  in  $G$  is at least  $(1 - \gamma)d^{\binom{\ell}{2}}m^\ell$ . Moreover, if every pair  $(V_i, V_j)$  is  $(\varepsilon, d)$ -regular, then the number of copies of  $K_\ell$  in  $G$  is in the interval  $(1 \pm \gamma)d^{\binom{\ell}{2}}m^\ell$ .*

We deduce Theorem 1.3 from Theorem 3.1 and Proposition 3.2. The argument follows the lines of [21], [27], and [1], whereas the original proof of Ruzsa and Szemerédi [56] used several iterations of an early version of the regularity lemma.

*Proof of Theorem 1.3.* We first address the promised constants. For given  $\delta > 0$  we have to define  $c > 0$ . We first set  $t_0 = \lceil 3/\delta \rceil$ ,  $\gamma = 1/2$ ,  $d_0 = \delta/3$ , and take  $\varepsilon = \varepsilon(3, d_0, \gamma) > 0$  to be the constant guaranteed by Proposition 3.2. Without loss of generality we may assume that  $\varepsilon < \delta/3$ . We apply Theorem 3.1 with  $\varepsilon$  and  $t_0$  defined above, and let  $T_0 = T_0(\varepsilon, t_0)$  be the guaranteed constant. We finally set

$$c = \frac{1}{4} \left( \frac{d_0}{T_0} \right)^3$$

and let  $n_0$  be sufficiently large.

Suppose  $G = (V, E)$  is a graph on  $n \geq n_0$  vertices that contains at most  $cn^3$  triangles. We will find a triangle-free subgraph  $\widehat{G} = (V, \widehat{E})$  of  $E$  with  $|E \setminus \widehat{E}| \leq \delta n^2$ .

We apply Theorem 3.1 to the graph  $G$  and obtain a partition  $V = V_1 \cup \dots \cup V_t$  satisfying properties (i)–(iii) of the conclusion of Theorem 3.1. To obtain the promised triangle-free subgraph  $\widehat{G}$ , delete from  $G$  any edge  $\{v_i, v_j\} \in E$  with  $v_i \in V_i$  and  $v_j \in V_j$  such that either  $i = j$ , or  $d_G(V_i, V_j) < d_0$  or the pair  $(V_i, V_j)$  is not  $\varepsilon$ -regular. It follows from the choice of  $\varepsilon, d_0$  and  $t_0$  that at most

$$t \binom{\lceil n/t \rceil}{2} + \binom{t}{2} d_0 \left\lceil \frac{n}{t} \right\rceil^2 + \varepsilon \binom{t}{2} \left\lceil \frac{n}{t} \right\rceil^2 \leq \delta n^2$$

edges of  $G$  were removed.

We will prove that the resulting graph  $\widehat{G}$  is triangle-free. Suppose to the contrary that  $\widehat{G}$  contains a triangle. By our construction of  $\widehat{G}$ , it follows that the vertex set  $\{v_i, v_j, v_k\}$  of this triangle satisfies  $v_i \in V_i, v_j \in V_j, v_k \in V_k$  for some  $1 \leq i < j < k \leq t$ . Consequently each of the pairs  $(V_i, V_j), (V_j, V_k)$  and  $(V_i, V_k)$  contains at least one edge and, hence, all these pairs are  $\varepsilon$ -regular with density at least  $d_0$ . Since the assumptions of Proposition 3.2 are met by the 3-partite graph  $\widehat{G}[V_i, V_j, V_k]$  induced on  $V_i \cup V_j \cup V_k$ , the subgraph  $\widehat{G}$  of  $G$  contains at least

$$\left(1 - \frac{1}{2}\right) d_0^3 \left\lfloor \frac{n}{t} \right\rfloor^3 > \frac{d_0^3}{4T_0^3} n^3 = cn^3$$

triangles, which contradicts our hypothesis that  $G$  has at most  $cn^3$  such copies. □

**3.2. Regular approximation lemma for graphs.** Szemerédi’s proof of the regularity lemma leads to a tower-type dependency of the involved parameters. More precisely, the proof yields  $T_0(\varepsilon, t_0)$  of the form

$$2^{2^{\dots^{2^{t_0}}}}$$

with a tower of twos of height  $c/\varepsilon^5$  for some absolute constant  $c > 0$ . Eventually Gowers [33] showed a similar lower bound. In fact, he showed that there are graphs with the property that every partition satisfying properties (i)–(iii) of Theorem 3.1 consists of at least  $T$  parts, where  $T$  is given by a tower of twos of height polynomial in  $1/\varepsilon$ . Very recently, Fox and L. M. Lovász [25] closed the gap and determined the exact degree of the polynomial in the height of the tower for a version of the regularity lemma that is closely related to the original version of Szemerédi.

The result of Gowers implies, in particular, that  $t \gg 1/\varepsilon$  is unavoidable in Theorem 3.1. Below we discuss a variant of the regularity lemma that allows this dependency to be avoided. The price for that is the necessity of changing the edge set of the given graph slightly (see property (b) in Theorem 3.3). This lemma appeared in the work on hypergraph generalizations of the regularity lemma [51] (see Section 5 below) and in the work of Lovász and B. Szegedy [44, Lemma 5.2]. Its origin can be traced back to the work of Alon, Fischer, Krivelevich, and M. Szegedy [2].

**Theorem 3.3.** *For every  $\mu > 0$ , every function  $\varepsilon: \mathbb{N} \rightarrow (0, 1]$ , and every  $t_0 \in \mathbb{N}$  there exist  $T_0 = T_0(\mu, \varepsilon(\cdot), t_0)$  and  $n_0$  such that for every graph  $H = (V, E)$  with at least  $|V| = n \geq n_0$  vertices the following holds. There exists a partition  $V = V_1 \cup \dots \cup V_t$  with  $t_0 \leq t \leq T_0$  and  $|V_1| \leq \dots \leq |V_t| \leq |V_1| + 1$ , and there exists a graph  $G = (V, E')$  on the same vertex set such that*

- (a) all pairs  $(V_i, V_j)$  are  $\varepsilon(t)$ -regular in  $G$ , and
- (b)  $|E \triangle E'| = |E \setminus E'| + |E' \setminus E| \leq \mu n^2$ .

The main difference between Theorem 3.3 and Theorem 3.1 is in the choice of  $\varepsilon$  being a function of  $t$ . In particular, we may choose  $\varepsilon = \varepsilon(t) \ll 1/t$ . In view of Gowers’s work discussed above, Theorem 3.3 must fail for such an  $\varepsilon$  and  $\mu = 0$ . Dealing with such a small choice of  $\varepsilon$  costs us a price, which we pay in the form of perturbing the graph slightly.

The following counting lemma is suited for applications of Theorem 3.3 and is a simple consequence of Proposition 3.2.

**Corollary 3.4.** *For all integers  $\ell$  and constants  $\gamma > 0$  and  $d > 0$  there exist  $\nu > 0$ ,  $\varepsilon > 0$  and  $m_0$  so that the following holds. Suppose*

- (i)  $G = (V_1 \cup \dots \cup V_\ell, E)$  is an  $\ell$ -partite graph with vertex set  $V = \bigcup_{i=1}^\ell V_i$  and  $|V_1| = \dots = |V_\ell| = m \geq m_0$ , where all the pairs  $(V_i, V_j)$  with  $1 \leq i < j \leq \ell$  are  $\varepsilon$ -regular with density  $d$  and
- (ii)  $H = (V_1 \cup \dots \cup V_\ell, E')$  is an  $\ell$ -partite graph on the same vertex partition with  $|E \Delta E'| \leq \nu m^2$ .

Then the number of copies of  $K_\ell$  in  $H$  is in the interval  $(1 \pm \gamma)d^{\binom{\ell}{2}}m^\ell$ .

It is easy to see that Corollary 3.4 follows from Proposition 3.2. In fact, with an appropriate choice of  $\varepsilon = \varepsilon(\ell, d, \gamma/2)$  Proposition 3.2 implies that the graph  $G$  given by Corollary 3.4 contains  $(1 \pm \gamma/2)d^{\binom{\ell}{2}}m^\ell$  copies of  $K_\ell$ . Moreover, every pair appearing in the symmetric difference of the edge sets of  $G$  and  $H$  may extend to at most  $m^{\ell-2}$  copies of  $K_\ell$  and thus the number of copies of  $K_\ell$  in  $G$  and  $H$  may differ by at most  $\nu m^\ell$ . Consequently,  $H$  contains

$$\left(1 \pm \frac{\gamma}{2}\right) d^{\binom{\ell}{2}} m^\ell \pm \nu m^\ell$$

copies of  $K_\ell$  and, hence, for  $\nu < \gamma d^{\binom{\ell}{2}}/2$  the result follows.

Note that just as Theorem 3.1 and Proposition 3.2 imply Theorem 1.3, one can give a proof based on Theorem 3.3 and Corollary 3.4. In fact, one can apply Theorem 3.3 with  $\mu = \nu\delta/10$ , where  $\delta$  is given by the removal lemma and  $\nu$  is given by Corollary 3.4 applied with  $\gamma = 1/2$  and  $d = \delta/10$ . Then one follows the proof of Theorem 1.3 line by line and instead of deleting the edges of pairs that are not  $\varepsilon$ -regular we may simply delete the edges contained in the pairs  $(V_i, V_j)$  where the edge sets of  $H$  and  $G$  differ by more than  $\nu|V_i||V_j|$  pairs.

### 4. Quasi-randomness in graphs and hypergraphs

Szemerédi’s regularity lemma decomposes the edge set of any given graph into bipartite graphs, most of which are  $\varepsilon$ -regular. Roughly speaking,  $\varepsilon$ -regularity resembles the uniform edge distribution of random bipartite graphs and is shared by most bipartite graphs of given size and density. Closely related to the concept of  $\varepsilon$ -regularity is the research concerning *quasi-random graphs*, where one focuses on graph properties that are shared by almost all graphs. As it turns out  $\varepsilon$ -regularity (or its analogue  $\text{Disc}_d$  defined below) is such a property. We will state some of these properties and discuss their relation. We then state its extensions to hypergraphs, which leads to the concepts the regularity method for hypergraphs is based on.

**4.1. Quasi-random graphs.** The systematic study of quasi-random graphs was pioneered by Thomason [65] and by Chung, Graham, and Wilson [13]. Here we restrict ourselves to the following properties: uniform edge distribution ( $\text{Disc}$ ), subgraph count ( $\text{Count}$ ), and small deviation ( $\text{Dev}$ ), all defined below.

**Definition 4.1.** Let  $d \in [0, 1]$ . We say a sequence  $\mathcal{G} = (G_n)_{n \in \mathbb{N}}$  of graphs satisfies property

$\text{Disc}_d$  if for every  $\varepsilon > 0$  all but finitely many graphs  $G_n = (V_n, E_n)$  of  $\mathcal{G}$  have the property that for every  $U \subseteq V_n$  we have

$$\left| e(U) - d \binom{|U|}{2} \right| \leq \varepsilon |V_n|^2, \tag{4.1}$$

where  $e(U)$  is the number of edges of  $G_n$  contained in  $U$ .

$\text{Count}_d$  if for every  $\varepsilon > 0$  and for every graph  $F$  all but finitely many graphs  $G_n = (V_n, E_n)$  of  $\mathcal{G}$  contain  $(d^{|E(F)|} \pm \varepsilon) |V_n|^{|V(F)|}$  labeled copies of  $F$ .

$\text{Dev}_d$  if for every  $\varepsilon > 0$  all but finitely many graphs  $G_n = (V_n, E_n)$  of  $\mathcal{G}$  satisfy

$$\sum_{u_0, u_1 \in V_n} \sum_{w_0, w_1 \in V_n} \prod_{i, j \in \{0, 1\}} (\mathbb{1}_{E_n}(u_i, w_j) - d) \leq \varepsilon |V_n|^4, \tag{4.2}$$

where  $\mathbb{1}_{E_n} : V_n^2 \rightarrow \{0, 1\}$  is the indicator function of  $E_n$  with  $\mathbb{1}_{E_n}(u, w) = 1$  if  $\{u, w\} \in E_n$  and 0 otherwise.

We say one such property  $\mathcal{P}_d$  implies another such property  $\mathcal{Q}_d$  if every graph sequence satisfying  $\mathcal{P}_d$  also satisfies  $\mathcal{Q}_d$ . Moreover, two such properties are equivalent if any sequence of graphs  $\mathcal{G}$  satisfies either both or none of the properties.

Chung, Graham, and Wilson considered a larger and slightly different set of properties and proved that they are equivalent in the sense defined in Definition 4.1. In fact, some of those equivalences had already appeared before in the literature. Since then, the list of equivalent properties has been extended by several researchers. For our discussion here we state the following partial version of the Chung-Graham-Wilson theorem.

**Theorem 4.2.** *For graphs  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  are equivalent for any  $d > 0$ .*

It is easy to show that random graphs with density  $d$  satisfy  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  with high probability, i.e., with probability tending to one when the number of vertices tend to infinity. Therefore, graphs satisfying one (and therefore in view of Theorem 4.2 all) of these properties are called *quasi-random graphs*. Strictly speaking, we should always consider sequences of graphs here. However, we will also say that a single graph  $G$  satisfies  $\text{Disc}_d(\varepsilon)$  if (4.1) holds. Similarly, we will consider graphs  $G$  satisfying  $\text{Count}_d(\varepsilon, F)$ , or  $\text{Dev}_d(\varepsilon)$ . We shall refer to such a graph  $G$  as a quasi-random graph.

We will prove Theorem 4.2 in the next section. The implication “ $\text{Disc}_d \Rightarrow \text{Count}_d$ ” is of particular interest. It states that for every graph  $F$  and every  $\gamma > 0$  there exists an  $\varepsilon > 0$  such that any sufficiently large graph  $G$  that satisfies  $\text{Disc}_d(\varepsilon)$  must satisfy  $\text{Count}(\gamma, F)$ . This is closely related to Proposition 3.2 with  $F = K_\ell$ .

In Section 5 we outline the regularity method for hypergraphs, which allows us to extend the proof of the triangle removal lemma (Theorem 1.3) based on Szemerédi’s regularity lemma (Theorem 3.1) and the counting lemma (Proposition 3.2) to hypergraphs. For that we define appropriate extensions of  $\text{Disc}_d$  and  $\text{Count}_d$ . We remark that an alternative approach to the regularity method for hypergraphs based on extensions of  $\text{Dev}_d$  was developed by Gowers [35].

**4.2. Equivalences of quasi-random properties.** In this section we sketch some of the ideas in the proof of Theorem 4.2 and we prove the equivalence of  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$ . For that we will verify the following three implications

$$\text{Disc}_d \Rightarrow \text{Count}_d \Rightarrow \text{Dev}_d \Rightarrow \text{Disc}_d. \tag{4.3}$$

We begin our discussion with the implication “ $\text{Disc}_d \Rightarrow \text{Count}_d$ ” and present a proof which borrows some ideas of Tao from [64].

*Proof of  $\text{Disc}_d \Rightarrow \text{Count}_d$ .* It suffices to show that for every graph  $F$  and  $\varepsilon > 0$  there exists  $\delta > 0$  such that every sufficiently large graph  $G = (V, E)$  satisfying  $\text{Disc}_d(\delta)$  also satisfies  $\text{Count}_d(\varepsilon, F)$ .

The proof consists of two parts. First we appeal to the identity

$$e(U, W) = e(U \cup W) - e(U \setminus W) - e(W \setminus U) + e(U \cap W),$$

valid for every  $U, W \subseteq V$ , where the edges contained in  $U \cap W$  are counted twice in  $e(U, W)$ . Consequently,  $\text{Disc}_d(\delta)$  implies that one can control the number  $e(U, W)$  up to an additive error of  $4\delta|V|^2$ . In other words, we obtain a “two-set-version” of  $\text{Disc}_d$ .

In the second part we continue by induction on  $|E(F)|$ . Since the statement  $\text{Count}_d(F, \varepsilon)$  is trivial for graphs without edges if  $n = |V|$  is sufficiently large, we may assume that for sufficiently small  $\delta$  any graph satisfying  $\text{Disc}_d(\delta)$  also satisfies  $\text{Count}_d(F', \varepsilon/2)$  for some subgraph  $F'$  of  $F$  with one fewer edge. Suppose  $F'$  is obtained from  $F$  by removing the edge  $\{x, y\}$ . Let us denote by  $N_G(F)$  and  $N_G(F')$  the number of labeled copies of  $F$  and  $F'$  in  $G$ , respectively. Moreover, for every copy  $\widehat{F}'$  of  $F'$  in  $G$  let  $x_{\widehat{F}'}$  and  $y_{\widehat{F}'}$  be the vertices in  $V$  representing  $x$  and  $y$  in that copy. Consequently,

$$N_G(F) = \sum_{\widehat{F}' \subseteq G} \mathbb{1}_E(x_{\widehat{F}'}, y_{\widehat{F}'}),$$

where the sum is taken over all copies  $\widehat{F}'$  of  $F'$  in  $G$ . This way we arrive at

$$N_G(F) = \sum_{\widehat{F}' \subseteq G} (\mathbb{1}_E(x_{\widehat{F}'}, y_{\widehat{F}'}) - d + d) = d \cdot N_G(F') + \sum_{\widehat{F}' \subseteq G} (\mathbb{1}_E(x_{\widehat{F}'}, y_{\widehat{F}'}) - d). \tag{4.4}$$

Owing to the induction assumption we infer that

$$d \cdot N_G(F') = d \cdot \left( d^{|E(F)|-1} \pm \frac{\varepsilon}{2} \right) n^{|V(F)|} \stackrel{d \leq 1}{\cong} \left( d^{|E(F)|} \pm \frac{\varepsilon}{2} \right) n^{|V(F)|}.$$

Therefore, the implication “ $\text{Disc}_d \Rightarrow \text{Count}_d$ ” follows from (4.4) if we show

$$\left| \sum_{\widehat{F}' \subseteq G} (\mathbb{1}_E(x_{\widehat{F}'}, y_{\widehat{F}'}) - d) \right| \leq \frac{\varepsilon}{2} n^{|V(F)|}. \tag{4.5}$$

This inequality follows from the two-set-version of  $\text{Disc}_d$  discussed above. In fact, let  $F''$  be the graph obtained by removing both vertices  $x$  and  $y$  from  $F$ . For a copy  $\widehat{F}''$  of  $F''$  in  $G$  we denote by  $X_{\widehat{F}''} \subseteq V$  the set of vertices that complete  $\widehat{F}''$  to a copy of  $F - y$ , that is, the subgraph of  $F$  obtained by removing the vertex  $y$  only. Similarly, let  $Y_{\widehat{F}''} \subseteq V$  be the set of vertices that complete  $\widehat{F}''$  to a copy of  $F - x$ . Note that every pair of distinct vertices

$\widehat{x} \in X_{\widehat{F}''}$  and  $\widehat{y} \in Y_{\widehat{F}''}$  together with  $\widehat{F}''$  define a copy of  $F'$  in  $G$ , which extends to a copy of  $F$  if the edge  $\{\widehat{x}, \widehat{y}\}$  is present in  $G$ . With this notation we can write

$$\begin{aligned} \left| \sum_{\widehat{F}' \subseteq G} (\mathbb{1}_E(x_{\widehat{F}'}, y_{\widehat{F}'}) - d) \right| &= \left| \sum_{\widehat{F}'' \subseteq G} \sum_{(\widehat{x}, \widehat{y}) \in X_{\widehat{F}''} \times Y_{\widehat{F}''}} (\mathbb{1}_E(\widehat{x}, \widehat{y}) - d) \right| \\ &\leq \sum_{\widehat{F}'' \subseteq G} \left| \sum_{(\widehat{x}, \widehat{y}) \in X_{\widehat{F}''} \times Y_{\widehat{F}''}} (\mathbb{1}_E(\widehat{x}, \widehat{y}) - d) \right| \\ &= \sum_{\widehat{F}'' \subseteq G} \left| e(X_{\widehat{F}''}, Y_{\widehat{F}''}) - d|X_{\widehat{F}''}||Y_{\widehat{F}''}| \right|. \end{aligned}$$

The two-set-version of  $\text{Disc}_d$  lets us bound each of the  $N_G(F'')$  terms of the last sum by  $4\delta n^2$  and since  $F''$  has two vertices fewer than  $F$  we have  $N_G(F'') \leq n^{|V(F)|-2}$ . Consequently, (4.5) follows as long as  $\delta \leq \varepsilon/8$ .  $\square$

We continue with the second implication and deduce  $\text{Dev}_d$  from  $\text{Count}_d$ . In the proof we only use estimates on the number of copies of graphs  $F$  on four vertices.

*Proof of “ $\text{Count}_d \Rightarrow \text{Dev}_d$ ”.* Property  $\text{Dev}_d$  follows from  $\text{Count}_d$  by appealing to all graphs  $F$  with four vertices. Recall that  $\text{Count}_d$  yields estimates for the number of not necessarily induced, labeled copies of a given graph  $F$ . However, with a standard application of the principle of inclusion and exclusion involving the estimates given by  $\text{Count}_d$  for all graphs  $F$  on a given number of vertices we obtain estimates for the number of induced copies. More precisely, this argument shows that  $n$ -vertex graphs  $G = (V, E)$  that satisfy  $\text{Count}_d$  contain

$$\left( d^{|E(F)|} (1-d)^{\binom{|V(F)|}{2} - |E(F)|} \pm o(1) \right) n^{|V(F)|}$$

induced copies of a given graph  $F$ .

For each of the  $2^4$  labeled subgraphs  $F$  of the cycle on four vertices  $C_4$ , we denote by  $N_G^*(F)$  the number of copies of  $F$  that are “induced with respect to  $C_4$  in  $G$ ,” i.e., copies of  $F$  where the corresponding edges in  $E(C_4) \setminus E(F)$  are not present in  $G$  with no restriction on the appearance of the two edges in  $E(K_4) \setminus E(C_4)$  in  $G$ . Adding the four estimates on the number of induced copies of those supergraphs of  $F$  containing some of these two edges in  $E(K_4) \setminus E(C_4)$  yields

$$N_G^*(F) = (d^{e(F)} (1-d)^{4-e(F)} \pm o(1)) n^4. \tag{4.6}$$

These estimates can be used to evaluate the sum in  $\text{Dev}_d$ . In fact, we have

$$\sum_{u_0, u_1 \in V} \sum_{w_0, w_1 \in V} \prod_{i, j \in \{0, 1\}} (\mathbb{1}_E(u_i, w_j) - d) = \sum_{F \subseteq C_4} (1-d)^{e(F)} (-d)^{4-e(F)} \cdot N_G^*(F).$$

Replacing  $N_G^*(F)$  by the estimate from (4.6) and appealing to the binomial theorem yields the desired bound.  $\square$

It is left to discuss the proof “ $\text{Dev}_d \Rightarrow \text{Disc}_d$ ”, which is based on a standard application of the Cauchy-Schwarz inequality.

*Proof of  $\text{Dev}_d \Rightarrow \text{Disc}_d$ .* Given a sufficiently large  $n$ -vertex graph  $G = (V, E)$  satisfying  $\text{Dev}_d(\delta)$ , we shall deduce that  $G$  also satisfies  $\text{Disc}_d(\varepsilon)$ , where  $\varepsilon \rightarrow 0$  as  $\delta \rightarrow 0$ . In fact, we will deduce the two-set-version of  $\text{Disc}_d$ . So let  $U, W \subseteq V$  be arbitrary. We shall show that

$$|e(U, W) - d|U||W|| \leq \varepsilon n^2. \tag{4.7}$$

Recalling that  $e(U) = e(U, U)/2$ , the standard version of  $\text{Disc}_d$  then follows by letting  $U = W$ . For the proof of (4.7) we note that

$$|e(U, W) - d|U||W|| = \left| \sum_{u \in V} \sum_{w \in V} \mathbb{1}_U(u) \mathbb{1}_W(w) (\mathbb{1}_E(u, w) - d) \right|, \tag{4.8}$$

where  $\mathbb{1}_U$  and  $\mathbb{1}_W$  denote the indicator functions of the sets  $U$  and  $W$ . From (4.8) we will derive (4.7) as a consequence of the assumption that  $G$  satisfies  $\text{Dev}_d(\delta)$  by two simple applications of the Cauchy-Schwarz inequality. In fact,

$$\begin{aligned} |e(U, W) - d|U||W||^4 &= \left( \sum_{u \in V} \sum_{w \in V} \mathbb{1}_U(u) \mathbb{1}_W(w) (\mathbb{1}_E(u, w) - d) \right)^4 \\ &= \left( \sum_{u \in V} \mathbb{1}_U(u) \sum_{w \in V} \mathbb{1}_W(w) (\mathbb{1}_E(u, w) - d) \right)^4 \\ &\leq \left( \sum_{u \in V} \mathbb{1}_U^2(u) \sum_{u \in V} \left( \sum_{w \in V} \mathbb{1}_W(w) (\mathbb{1}_E(u, w) - d) \right)^2 \right)^2 \\ &= \left( |U| \cdot \sum_{u \in V} \sum_{w_0, w_1 \in V} \prod_{j=0}^1 \mathbb{1}_W(w_j) (\mathbb{1}_E(u, w_j) - d) \right)^2 \\ &= |U|^2 \cdot \left( \sum_{w_0, w_1 \in V} \mathbb{1}_W(w_0) \mathbb{1}_W(w_1) \sum_{u \in V} \prod_{j=0}^1 (\mathbb{1}_E(u, w_j) - d) \right)^2 \\ &\leq |U|^2 \sum_{w_0, w_1 \in V} \mathbb{1}_W^2(w_0) \mathbb{1}_W^2(w_1) \sum_{w_0, w_1 \in V} \left( \sum_{u \in V} \prod_{j=0}^1 (\mathbb{1}_E(u, w_j) - d) \right)^2 \\ &= |U|^2 |W|^2 \cdot \sum_{w_0, w_1 \in V} \sum_{u_0, u_1 \in V} \prod_{i, j \in \{0, 1\}} (\mathbb{1}_E(u_i, w_j) - d). \end{aligned}$$

From  $|U|, |W| \leq n$  and the assumption that  $G$  satisfies  $\text{Dev}_d(\delta)$  (see (4.2)) we infer

$$|e(U, W) - d|U||W||^4 \leq n^4 \cdot \delta n^4,$$

which yields (4.7) for  $\delta = \varepsilon^4$ . □

This concludes our discussion of the equivalences between  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  and we continue with the discussion of extensions of these properties to the hypergraph case.

**4.3. Quasi-random hypergraphs.** The graph properties  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  have natural counterparts for  $k$ -uniform hypergraphs. In particular, for  $\text{Count}_d$  the generalization is straightforward, i.e., for  $d > 0$  we say a sequence of  $k$ -uniform hypergraphs  $\mathcal{H} = (H_n)_{n \in \mathbb{N}}$  satisfies property

$\text{Count}_d$  if for every  $\varepsilon > 0$  and every  $k$ -uniform hypergraph  $F$  all but finitely many hypergraphs  $H_n = (V_n, E_n)$  of  $\mathcal{H}$  contain  $(d^{|E(F)|} \pm \varepsilon)|V_n|^{|V(F)|}$  copies of  $F$ .

However, the straightforward extension of  $\text{Disc}_d$  is not strong enough to imply  $\text{Count}_d$ . Indeed, for  $d \in [0, 1]$  we say a sequence of  $k$ -uniform hypergraphs  $\mathcal{H} = (H_n)_{n \in \mathbb{N}}$  satisfies property

$\text{Weak-disc}_d$  if for every  $\varepsilon > 0$  all but finitely many hypergraphs  $H_n = (V_n, E_n)$  of  $\mathcal{H}$  have the property that for every  $U \subseteq V_n$  we have

$$\left| e(U) - d \binom{|U|}{k} \right| \leq \varepsilon |V_n|^k,$$

where  $e(U)$  denotes the number of hyperedges of  $H_n$  contained in  $U$ .

It is easy to see that random hypergraphs of density  $d$  satisfy  $\text{Weak-disc}_d$  and  $\text{Count}_d$  with high probability. However, for 3-uniform hypergraphs there are already sequences  $\mathcal{H}$  that satisfy  $\text{Weak-disc}_d$ , but fail to have  $\text{Count}_d$ . We briefly mention two such examples.

Let  $V_n$  be an ordered set of size  $n$  and consider a random graph  $G_n$  on  $V_n$ . For three vertices  $x, y, z \in V$  with  $x < y < z$  we include the hyperedge  $\{x, y, z\}$  in the 3-uniform hypergraph  $H_n$  if in  $G_n$  the vertex  $x$  is connected to exactly one of the vertices  $y$  or  $z$ . It is easy to check that with high probability the sequence  $(H_n)_{n \in \mathbb{N}}$  satisfies  $\text{Weak-disc}_{1/2}$ . On the other hand, for any set of four vertices  $w < x < y < z$  at least two of the three pairs  $\{w, x\}$ ,  $\{w, y\}$ , and  $\{w, z\}$  either form edges in  $G_n$  or not. Hence, at least one of the triples  $\{w, x, y\}$ ,  $\{w, x, z\}$ , and  $\{w, y, z\}$  will be missing in  $H_n$ . Consequently, none of the hypergraphs  $H_n$  contains a copy of  $K_4^{(3)}$  and, therefore,  $(H_n)_{n \in \mathbb{N}}$  fails to satisfy  $\text{Count}_{1/2}$ .

For the second example, we again consider a random graph  $G_n$  on  $n$  vertices. It is a well known fact that with high probability every subset  $U \subseteq V(G_n)$  contains  $(1/8) \binom{|U|}{3} \pm o(n^3)$  (unlabeled) triangles. Consequently, the 3-uniform hypergraphs  $H'_n$  with vertex set  $V(G_n)$  and edge set corresponding to the triangles in the auxiliary random graph  $G_n$  satisfy  $\text{Disc}_d$  for  $d = 1/8$ . Now consider the 3-uniform hypergraph  $F$  consisting of two hyperedges intersecting in 2 vertices. Each copy of  $F$  corresponds to two triangles in the “underlying” random graph that share one edge, i.e., it corresponds to a graph with 4 vertices and 5 edges. Hence the expected number of labeled copies of  $F$  in  $H'_n$  is

$$\left( \left( \frac{1}{2} \right)^5 \pm o(1) \right) n^4 = \left( 2 \left( \frac{1}{8} \right)^2 \pm o(1) \right) n^4$$

and one can show that with high probability this estimate holds for the number of copies of  $F$  in  $H'_n$ . In other words, the hypergraphs  $H'_n$  contain approximately twice as many copies of  $F$  as allowed by  $\text{Count}_d$ . In fact, the constructed hypergraphs  $H'_n$  will contain at least twice as many copies for every 3-uniform hypergraph that contains two hyperedges intersecting in at least 2 vertices.

These two examples can be extended to  $k$ -uniform hypergraphs for any  $k \geq 3$  and they show that uniform edge distribution with respect to the vertex set is not sufficiently strong

to prove counting results for hypergraphs containing edges that intersect in two or more vertices. Property  $\text{Weak-disc}_d$ , however, implies a version of  $\text{Count}_d$  restricted to hypergraphs  $F$  with no such pair of hyperedges, as observed in [41] (see also [15]).

In order to address such examples as  $(H_n)$  and  $(H'_n)$  we require more control over the distribution of the hyperedges. This is achieved by comparing the edge set of our  $k$ -uniform hypergraphs  $H$  against collections of cliques  $K_k^{(k-1)}$  given by arbitrary  $(k-1)$ -uniform hypergraphs  $G$  on the same vertex set. This is made precise in the following definition of  $\text{Disc}_d$  for  $k$ -uniform hypergraphs.

For a  $(k-1)$ -uniform hypergraph  $G$  we denote by  $\mathcal{K}_k(G)$  the set of  $k$ -tuples in  $V(G)$  that span a clique  $K_k^{(k-1)}$  in  $G$ . For  $d \in [0, 1]$  we say a sequence of  $k$ -uniform hypergraphs  $\mathcal{H} = (H_n)_{n \in \mathbb{N}}$  satisfies property

$\text{Disc}_d$  if for every  $\varepsilon > 0$  all but finitely many hypergraphs  $H_n = (V_n, E_n)$  of  $\mathcal{H}$  have the property that for every  $(k-1)$ -uniform hypergraph  $G$  with vertex set  $V_n$  we have

$$||E_n \cap \mathcal{K}_k(G)| - d|\mathcal{K}_k(G)|| \leq \varepsilon |V_n|^k.$$

Note that by viewing subsets  $U \subseteq V_n$  as 1-uniform hypergraphs, for  $k = 2$  this definition of  $\text{Disc}_d$  reduces to the one made for graphs in Definition 4.1. This notion of uniform edge distribution for hypergraphs was first suggested in joint work with Frankl (unpublished) and investigated by Chung [10]. In [42] the implication “ $\text{Disc}_d \Rightarrow \text{Count}_d$ ” for hypergraphs was established. In fact, the proof given for graphs in Section 4.2 extends to the hypergraph case (see, e.g., [50]).

Below we present an extension of  $\text{Dev}_d$  for hypergraphs, which was considered in [11, 12]. For graphs one may view the definition of  $\text{Dev}_d$  as a “weighted version” of  $\text{Count}_d$  for the special graph  $F = C_4$  (the cycle on four vertices). In fact, in (4.2) we consider every potential  $C_4$  and assign to it a weight given by the product of the function  $\mathbb{1}_E(\cdot, \cdot) - d$  evaluated on the four potential edges of that  $C_4$ . As it turns out, viewing  $C_4 = K_{2,2}$  as a complete bipartite graph with vertex classes of size two leads to a useful extension. We denote by  $K_{2,\dots,2}^{(k)}$  the complete  $k$ -partite  $k$ -uniform hypergraph with vertex classes of size 2. In particular,  $K_{2,\dots,2}^{(k)}$  has  $2k$  vertices and  $2^k$  hyperedges. This leads to the following extension of  $\text{Dev}_d$  for hypergraphs. A sequence  $\mathcal{H} = (H_n)_{n \in \mathbb{N}}$  satisfies property

$\text{Dev}_d$  if for every  $\varepsilon > 0$  all but finitely many hypergraphs  $H_n = (V_n, E_n)$  of  $\mathcal{H}$  satisfy

$$\sum_{u_0^1, u_1^1 \in V_n} \cdots \sum_{u_0^k, u_1^k \in V_n} \prod_{i_1, \dots, i_k \in \{0,1\}} \left( \mathbb{1}_{E_n}(u_{i_1}^1, \dots, u_{i_k}^k) - d \right) \leq \varepsilon |V_n|^{2^k},$$

where  $\mathbb{1}_{E_n} : V_n^k \rightarrow \{0, 1\}$  is the indicator function of  $E_n$ ,

The equivalence of  $\text{Dev}_d$  and  $\text{Disc}_d$  appeared in [10] and the equivalence of  $\text{Dev}_d$  and  $\text{Count}_d$  follows from ideas in [11, 12]. Summarizing the discussion above, we arrive at the following extension of Theorem 4.2 for hypergraphs.

**Theorem 4.3.** *For every  $d > 0$  and every integer  $k \geq 2$  the properties  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  for  $k$ -uniform hypergraphs are equivalent.*

We remark that the equivalence of  $\text{Disc}_d$ ,  $\text{Count}_d$ , and  $\text{Dev}_d$  can be established by simply extending the implications in (4.3), which are given in Section 4.2 for the graph case.

The extension of the regularity method to hypergraphs that we discuss in the next section is based on a variant of  $\text{Disc}_d$ . We remark that the regularity method of Gowers [35], on the other hand, is based on a refinement of  $\text{Dev}_d$ . Both approaches yield decompositions of any given  $k$ -uniform hypergraph into blocks with quasi-random properties, which, as discussed in this section, allows one to prove embedding and counting results for hypergraphs of any fixed isomorphism type.

The regularity lemma for hypergraphs deals with a somewhat more complicated notion of quasi-randomness. Just as Szemerédi's regularity lemma yields a partition of the vertex set of a given graph, the discussion above indicates that a regularity lemma for  $k$ -uniform hypergraph should involve a partition of the set of all  $(k-1)$ -tuples. However, in order to “get control” of the partition classes of the  $(k-1)$ -tuples, we impose quasi-randomness on them as well. This quasi-randomness will be provided by a regularity lemma for  $(k-1)$ -uniform hypergraphs, which then leads to a partition of the set of all  $(k-2)$ -tuples. This eventually forces partitions of vertices, pairs, triples,  $\dots$ ,  $(k-1)$ -tuples. The blocks of this family of partitions (see the definition of *complex* in Definition 5.1) consist of a family of  $j$ -uniform hypergraphs for  $j = 2, \dots, k-1$  with two consecutive levels “acting quasi-randomly.”

The regularity lemmas considered in [26, 35, 54] yield that the precision of the quasi-randomness of the  $j$ -tuples (measured by say  $\varepsilon_j$ ) can be chosen as a function of their relative density  $d_j$  (see (5.1)), i.e., the lemma ensures  $d_j \gg \varepsilon_j$  for every  $j$ . On the other hand, the relative density of the  $(j-1)$ -tuples cannot be controlled and the situation when  $\varepsilon_j \gg d_{j-1}$  cannot be avoided. This is similar to the situation in Szemerédi's regularity lemma, where  $\varepsilon \ll 1/t$  cannot be imposed (see the beginning of Section 3.2), and leads to a situation where the implication  $\text{Disc}_d \Rightarrow \text{Count}_d$  cannot be easily extended.

For graphs the “bad” dependency of  $\varepsilon$  and  $t$  could be avoided by slightly changing the graph to which the regularity lemma is applied (see Theorem 3.3). In [50, 51] this approach led to a regularity lemma for hypergraphs which allows a formulation that avoids some of the technical details discussed above. In particular, for the price of perturbing the given hypergraph slightly, the precision of the quasi-randomness for every level can be set to  $\varepsilon \ll d_2, \dots, d_{k-1}$ . We focus on that lemma in the next section.

## 5. Regularity method for hypergraphs

We describe a generalization of the regularity lemma from graphs to hypergraphs. Here we mainly focus on an extension of the regular approximation lemma (see Theorem 3.3). This lemma combined with its corresponding counting lemma yields a proof of the removal lemma (Theorem 2.1) in a straightforward way. We give this proof in Section 5.3. We shall follow the presentation from [50, 51].

**5.1. Regular complexes.** As discussed in Section 4 we consider “nested” families of hypergraphs  $\mathbf{H} = \{H^{(2)}, \dots, H^{(k)}\}$  on the same vertex set  $V$ , where nested means that the edge set of  $H^{(j+1)}$  corresponds to a subset of the cliques on  $j+1$  vertices in  $H^{(j)}$ , i.e.,  $E(H^{(j+1)}) \subseteq \mathcal{K}_{j+1}(H^{(j)})$ , where  $\mathcal{K}_{j+1}(H^{(j)})$  consists of those  $(j+1)$ -tuples of vertices of  $V$  that span a clique in  $H^{(j)}$ . This leads to the definition of a *complex*.

**Definition 5.1.** Let  $\ell \geq k$  be positive integers. A  $(k, \ell)$ -*complex*  $\mathbf{H}$  with vertex set  $V = V_1 \cup \dots \cup V_\ell$  is a collection of  $\ell$ -partite hypergraphs  $\{H^{(j)}\}_{j=2}^k$  with vertex partition

$V_1 \cup \dots \cup V_\ell$  such that

- (a) for every  $j = 2, \dots, k$  the hypergraph  $H^{(j)}$  is  $j$ -uniform and
- (b)  $E(H^{(j+1)}) \subseteq \mathcal{K}_{j+1}(H^{(j)})$  for every  $j = 2, \dots, k - 1$ .

For  $k = 1$  and  $\ell = 2$ , we have a  $(1, 2)$ -complex, which is the same as a pair  $(V_1, V_2)$  of disjoint vertex sets. In the regularity lemma for  $k$ -uniform hypergraphs,  $(k - 1, k)$ -complexes play a role similar to the role played by the pairs in Szemerédi’s regularity lemma. Having this in mind we generalize the notions of *density* and *regular pair*.

For  $j \geq 2$  we define the *density*  $d(H^{(j+1)}|H^{(j)})$  of a  $(j + 1)$ -uniform hypergraph  $H^{(j+1)}$  w.r.t. a  $j$ -uniform hypergraph  $H^{(j)}$  on the same vertex set as the proportion of those cliques in  $H^{(j)}$  whose vertex sets span a hyperedge of  $H^{(j+1)}$ . If  $|\mathcal{K}_{j+1}(H^{(j)})| > 0$ , then we set

$$d(H^{(j+1)}|H^{(j)}) = \frac{|E(H^{(j+1)}) \cap \mathcal{K}_{j+1}(H^{(j)})|}{|\mathcal{K}_{j+1}(H^{(j)})|}, \tag{5.1}$$

and  $d(H^{(j+1)}|H^{(j)}) = 0$  when  $H^{(j)}$  contains no  $(j + 1)$ -clique. Similarly to the graph case we say a  $(j + 1)$ -partite hypergraph  $H^{(j+1)}$  is regular w.r.t. a  $(j + 1)$ -partite hypergraph  $H^{(j)}$  if  $d(H^{(j+1)}|Q^{(j)})$  is approximately the same for all subhypergraphs  $Q^{(j)} \subseteq H^{(j)}$  containing “many” cliques of size  $j + 1$ . More precisely, we say  $H^{(j+1)}$  is  $(\varepsilon, d)$ -regular w.r.t.  $H^{(j)}$  for some  $\varepsilon > 0$  and  $d \geq 0$  if all  $Q^{(j)} \subseteq H^{(j)}$  with

$$|\mathcal{K}_{j+1}(Q^{(j)})| \geq \varepsilon |\mathcal{K}_{j+1}(H^{(j)})|,$$

satisfy

$$d(H^{(j+1)}|Q^{(j)}) = d \pm \varepsilon.$$

We say that a hypergraph  $H^{(j+1)}$  is  $\varepsilon$ -regular w.r.t.  $H^{(j)}$  if it is  $(\varepsilon, d)$ -regular for some  $d \geq 0$ . Moreover, if  $H^{(j+1)}$  and  $H^{(j)}$  are  $\ell$ -partite for some  $\ell > j$  with vertex partition  $V_1 \cup \dots \cup V_\ell$ , then we say that  $H^{(j+1)}$  is  $(\varepsilon, d)$ -regular w.r.t.  $H^{(j)}$  if for every choice of indices  $1 \leq i_1 < \dots < i_{j+1} \leq \ell$  the induced hypergraph  $H^{(j+1)}[V_{i_1} \cup \dots \cup V_{i_{j+1}}]$  is  $(\varepsilon, d)$ -regular w.r.t.  $H^{(j)}[V_{i_1} \cup \dots \cup V_{i_{j+1}}]$ . We also say that such a  $H^{(j+1)}$  is  $\varepsilon$ -regular (respectively  $(\varepsilon, \geq d)$ -regular) w.r.t.  $H^{(j)}$  if all induced hypergraphs  $H^{(j+1)}[V_{i_1} \cup \dots \cup V_{i_{j+1}}]$  are  $(\varepsilon, d_{i_1, \dots, i_{j+1}})$ -regular w.r.t.  $H^{(j)}[V_{i_1} \cup \dots \cup V_{i_{j+1}}]$  for some  $d_{i_1, \dots, i_{j+1}} \geq 0$  (resp. for some  $d_{i_1, \dots, i_{j+1}} \geq d$ ). Finally we define regular complexes.

**Definition 5.2** (regular complex). Let  $\varepsilon > 0$  be given and let  $\mathbf{d} = (d_2, \dots, d_k)$  be a vector of non-negative reals. We say a  $(k, \ell)$ -complex  $\mathbf{H} = \{H^{(j)}\}_{j=2}^k$  with vertex partition  $V_1 \cup \dots \cup V_\ell$  is  $(\varepsilon, \mathbf{d})$ -regular if

- (a)  $(V_h, V_i)$  is  $(\varepsilon, d_2)$ -regular in  $H^{(2)}$  for every  $1 \leq h < i \leq \ell$  and
- (b)  $H^{(j+1)}$  is  $(\varepsilon, d_{j+1})$ -regular w.r.t.  $H^{(j)}$  for every  $j = 2, \dots, k - 1$ .

**5.2. Equitable partitions.** While the regularity lemma for graphs provides a partition of vertices, the regularity lemma for  $k$ -uniform hypergraphs provides a well-structured family of partitions  $\mathcal{P} = \{\mathcal{P}^{(1)}, \dots, \mathcal{P}^{(k-1)}\}$  of vertices, pairs, triples,  $\dots$ , and  $(k - 1)$ -tuples of the vertex set  $V$ . We now describe this structure.

As in Szemerédi’s regularity lemma, we start with a vertex partition

$$\mathcal{P}^{(1)} = \{V_1, \dots, V_{a_1}\}$$

of  $V$ . For every  $j \geq 2$  let  $\text{Cross}_j(\mathcal{P}^{(1)})$  be the family of all crossing  $j$ -tuples  $J$ , i.e., the set of  $j$ -tuples which satisfy  $|J \cap V_i| \leq 1$  for every  $V_i \in \mathcal{P}^{(1)}$ . For every  $j = 2, \dots, k - 1$  the partition  $\mathcal{P}^{(j)}$  will be a partition of  $\text{Cross}_j(\mathcal{P}^{(1)})$  into  $j$ -partite  $j$ -uniform hypergraphs. In particular, for every  $j$ -element set  $J \in \text{Cross}_j(\mathcal{P}^{(1)})$ , there is a unique  $j$ -partite  $j$ -uniform hypergraph  $P^{(j)}(J) \in \mathcal{P}^{(j)}$  that contains  $J$  as a hyperedge. Furthermore, for every  $\ell \geq j$  and every  $\ell$ -element set  $L \in \text{Cross}_\ell(\mathcal{P}^{(1)})$  we may consider  $P^{(j)}(L)$  defined by

$$P^{(j)}(L) = \bigcup \{P^{(j)}(J) \in \mathcal{P}^{(j)} : J \subseteq L \text{ and } |J| = j\}.$$

In other words,  $P^{(j)}(L)$  is the unique union of  $\binom{\ell}{j}$  members of  $\mathcal{P}^{(j)}$  with the property that  $L$  spans a clique on  $P^{(j)}(L)$ , i.e.,  $L \in \mathcal{K}_\ell(P^{(j)}(L))$ . We will also find it convenient to introduce  $P^{(1)}(L)$  to be the partition  $\{V_{i_1}, \dots, V_{i_\ell}\}$  with vertex classes from  $\mathcal{P}^{(1)}$  with  $|L \cap V_{i_j}| = 1$  for every  $j = 1, \dots, \ell$ .

With this notation at hand, we describe the additional structural requirement on  $\mathcal{P}$  in an inductive way. We let  $\mathcal{P}^{(2)}$  be an arbitrary partition of  $\text{Cross}_2(\mathcal{P}^{(1)})$  into bipartite graphs with bipartitions being pairs from  $\mathcal{P}^{(1)}$  and suppose that a partition  $\mathcal{P}^{(j)}$  of  $\text{Cross}_j(\mathcal{P}^{(1)})$  into  $j$ -partite  $j$ -uniform hypergraphs is given. The partition  $\mathcal{P}^{(j)}$  induces a partition  $\mathcal{K}_{j+1}(\mathcal{P}^{(j)})$  of  $\text{Cross}_{j+1}(\mathcal{P}^{(1)})$  defined by

$$\mathcal{K}_{j+1}(\mathcal{P}^{(j)}) := \{\mathcal{K}_{j+1}(P^{(j)}(J)) : J \in \text{Cross}_{j+1}(\mathcal{P}^{(1)})\}.$$

In other words,  $\mathcal{K}_{j+1}(\mathcal{P}^{(j)})$  is a partition of  $\text{Cross}_{j+1}(\mathcal{P}^{(1)})$  with two  $(j + 1)$ -tuples  $J, J' \in \text{Cross}_{j+1}(\mathcal{P}^{(1)})$  belonging to the same class in  $\mathcal{K}_{j+1}(\mathcal{P}^{(j)})$  if  $J' \in \mathcal{K}_{j+1}(P^{(j)}(J))$ , i.e., if  $P^{(j)}(J') = P^{(j)}(J)$ . The structural requirement for the partitions  $\mathcal{P}$  considered here is that  $\mathcal{P}^{(j+1)}$  should refine  $\mathcal{K}_{j+1}(\mathcal{P}^{(j)})$ .

It follows from this requirement for every  $j = 2, \dots, k - 2$  that for every  $k$ -tuple  $K \in \text{Cross}_k(\mathcal{P}^{(1)})$  the family

$$\mathcal{P}(K) = \{P^{(2)}(K), \dots, P^{(k-1)}(K)\} \tag{5.2}$$

is a  $(k - 1, k)$ -complex.

In addition to the structural condition described above we require control over the number of partition classes in  $\mathcal{P}^{(j+1)}$ , and more specifically, over the number of classes contained in  $\mathcal{K}_{j+1}(P^{(j)}(J))$ . This leads to Definition 5.3 below.

**Definition 5.3** (family of partitions). Suppose  $V$  is a set of vertices,  $k \geq 2$  is an integer and  $\mathbf{a} = (a_1, \dots, a_{k-1})$  is a vector of positive integers.

We say  $\mathcal{P} = \mathcal{P}(k - 1, \mathbf{a}) = \{\mathcal{P}^{(1)}, \dots, \mathcal{P}^{(k-1)}\}$  is a family of partitions on  $V$  if it satisfies the following:

- (i)  $\mathcal{P}^{(1)}$  is a partition of  $V$  into  $a_1$  classes and
- (ii) for every  $j = 1, \dots, k - 2$  the partition  $\mathcal{P}^{(j+1)}$  of  $\text{Cross}_{j+1}(\mathcal{P}^{(1)})$  refines  $\mathcal{K}_{j+1}(\mathcal{P}^{(j)})$  and for every  $J \in \text{Cross}_{j+1}(\mathcal{P}^{(1)})$  we have

$$|\{P^{(j+1)} \in \mathcal{P}^{(j+1)} : P^{(j+1)} \subseteq \mathcal{K}_{j+1}(P^{(j)}(J))\}| = a_{j+1}.$$

Moreover, we say  $\mathcal{P} = \mathcal{P}(k - 1, \mathbf{a})$  is  $T_0$ -bounded if  $\max\{a_1, \dots, a_{k-1}\} \leq T_0$ .

Heading towards a generalization of Szemerédi’s regularity lemma, we have just described a notion of  $\varepsilon$ -regularity used in the regular approximation lemma for hypergraphs (see Theorem 5.6 below) and the corresponding structure of the partition. It is convenient to generalize property (ii) of Theorem 3.1, which states that the vertex partition is equitable, i.e., the vertex classes differ in size by at most one. For higher uniformities it will be convenient to express this similarity by requiring the elements of  $\mathcal{P}^{(j)}$  to be of the same density and, moreover, to be  $\varepsilon$ -regular themselves. This leads to the following definition of equitable family of partitions.

**Definition 5.4.** Suppose  $V$  is a set of  $n$  vertices,  $\eta$  and  $\varepsilon$  are positive reals, and  $\mathbf{a} = (a_1, \dots, a_{k-1})$  is a vector of positive integers.

We say a family of partitions  $\mathcal{P} = \mathcal{P}(k-1, \mathbf{a})$  on  $V$  (as defined in Definition 5.3) is  $(\eta, \varepsilon, \mathbf{a})$ -equitable if it satisfies the following:

- (a)  $|\text{Cross}_k(\mathcal{P}^{(1)})| \geq (1 - \eta) \binom{n}{k}$ ,
- (b)  $\mathcal{P}^{(1)} = \{V_i : i \in [a_1]\}$  is equitable, i.e.,  $|V_1| \leq \dots \leq |V_{a_1}| \leq |V_1| + 1$ , and
- (c) for every  $K \in \text{Cross}_k(\mathcal{P}^{(1)})$  the  $(k-1, k)$ -complex  $\mathcal{P}(K) = \{P^{(j)}(K)\}_{j=2}^{k-1}$  is  $(\varepsilon, \mathbf{d})$ -regular for  $\mathbf{d} = (1/a_2, \dots, 1/a_{k-1})$  (see Definition 5.2).

With these definitions at hand we can state the regular approximation lemma from [51].

**5.3. Regular approximation lemma for hypergraphs.** The following theorem is the hypergraph analogue of Theorem 3.3, which was proved in [50] (see also [18] for a different proof).

**Theorem 5.5** (Regular approximation lemma). *Let  $k \geq 2$  be a fixed integer. For all positive constants  $\eta$  and  $\mu$ , and every function  $\varepsilon: \mathbb{N}^{k-1} \rightarrow (0, 1]$  there are integers  $T_0$  and  $n_0$  so that the following holds.*

*For every  $k$ -uniform hypergraph  $H^{(k)} = (V, E)$  with vertex set  $V$  with  $|V| \geq n_0$  there exist a vector  $\mathbf{a} \in \mathbb{N}^{k-1}$  and a  $T_0$ -bounded,  $(\eta, \varepsilon(\mathbf{a}), \mathbf{a})$ -equitable family of partitions  $\mathcal{P} = \mathcal{P}(k-1, \mathbf{a})$  on  $V$  and a  $k$ -uniform hypergraph  $G^{(k)} = (V, E')$  on the same vertex set such that*

- (a) *for every  $K \in \text{Cross}_k(\mathcal{P}^{(1)})$  the hypergraph  $G^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))$  is  $\varepsilon(\mathbf{a})$ -regular w.r.t.  $P^{(k-1)}(K)$  and*
- (b)  $|E \Delta E'| \leq \mu n^k$ .

Similarly to Theorem 3.3, the regular approximation lemma asserts that the edge set of any given hypergraph  $H^{(k)}$  can be altered at  $\mu n^k$  places in such a way that the resulting hypergraph  $G^{(k)}$  admits a regular partition in which one can impose that the  $\varepsilon$  controlling the regularity is as small as one wishes in terms of the size of the partition. We continue with a counting lemma that generalizes Corollary 3.4 and is appropriate for combined applications with the regular approximation lemma.

**Theorem 5.6.** *For all integers  $\ell \geq k \geq 2$  and all constants  $\gamma > 0$  and  $d_k > 0$  there is some  $\nu > 0$  such that for every  $d_0 > 0$  there is  $\varepsilon > 0$  and  $m_0$  so that the following holds.*

*Suppose*

- (a)  $\mathcal{P} = \{P^{(j)}\}_{j=2}^{k-1}$  *is an  $(\varepsilon, (d_2, \dots, d_{k-1}))$ -regular  $(k-1, \ell)$ -complex with  $d_i \geq d_0$  for every  $i = 2, \dots, k-1$  and with vertex classes of size  $m \geq m_0$ ,*

- (b)  $G^{(k)} \subseteq \mathcal{K}_k(P^{(k-1)})$  is  $(\varepsilon, d_k)$ -regular w.r.t.  $P^{(k-1)}$ , and
- (c)  $H^{(k)} \subseteq \mathcal{K}_k(P^{(k-1)})$  is  $\nu$ -close to  $G^{(k)}$ , i.e.,  $|E \Delta E'| \leq \nu |\mathcal{K}_k(P^{(k-1)})|$  for the edges sets  $E$  of  $H^{(k)}$  and  $E'$  of  $G^{(k)}$ .

Then

$$|\mathcal{K}_\ell(H^{(k)})| = (1 \pm \gamma) \prod_{j=2}^k d_j^{\binom{\ell}{j}} \times m^\ell. \tag{5.3}$$

For applications it is convenient to replace (b) by a condition that allows different relative densities of  $G^{(k)}$  for the  $\binom{\ell}{k}$   $k$ -partite subhypergraphs of  $P^{(k-1)}$  and only insists that  $d_k$  is a common lower bound, i.e.,

- (ii')  $G^{(k)} \subseteq \mathcal{K}_k(P^{(k-1)})$  is  $(\varepsilon, \geq d_k)$ -regular w.r.t.  $P^{(k-1)}$ .

In this case, the lower bound of the estimate in (5.3) stays valid:

$$|\mathcal{K}_\ell(H^{(k)})| \geq (1 - \gamma) \prod_{j=2}^k d_j^{\binom{\ell}{j}} \times m^\ell.$$

We remark that Theorem 5.6 can be proved along the lines of the implication of  $\text{Disc}_d \Rightarrow \text{Count}_d$ . More precisely, the proof of that implication can be used to estimate the number of  $K_\ell^{(k)}$  in  $G^{(k)}$  and then one can estimate the difference between the numbers of  $K_\ell^{(k)}$  in  $G^{(k)}$  and in  $H^{(k)}$ . This is trivial in the graph case but requires a more careful argument in the hypergraph case. The details can be found in [50]. We close with a proof of the removal lemma based on Theorems 5.5 and 5.6.

*Proof of Theorem 2.1.* Given  $k$  and  $\delta$  we are supposed to specify  $c > 0$  and  $n_0$  for which Theorem 2.1 holds. To this end we first set

$$\ell = k + 1, \quad \gamma = 1/2, \quad \text{and} \quad d_k = \delta/4. \tag{5.4}$$

For this choice of parameters, Theorem 5.6 yields a certain constant  $\nu > 0$ . Theorem 5.6 also implicitly gives functions  $\varepsilon(\cdot) = \varepsilon_{5.6}(\cdot)$  and  $m_0(\cdot)$  that, for any given  $d_0$ , yield the values of  $\varepsilon(d_0)$  and  $m_0(d_0)$  from that theorem. Without loss of generality we may assume that  $\nu \leq d_k$ . Next we set

$$\mu = \delta\nu/4, \quad \eta = \delta/4, \tag{5.5}$$

and we fix a function  $\varepsilon' = \varepsilon_{5.5}: \mathbb{N}^{k-1} \rightarrow (0, 1]$  defined for  $a_1, \dots, a_{k-1} \in \mathbb{N}$  by

$$\varepsilon'(a_1, \dots, a_{k-1}) = 2^{-1} \varepsilon(\min_{2 \leq i \leq k-1} 1/a_i).$$

For these choices of  $\mu, \eta$ , and  $\varepsilon'(\cdot, \dots, \cdot)$  the regular approximation lemma (Theorem 5.5) yields constants  $T_0$  and  $n_0$ . Finally we set

$$c = 2^{-1} d_k^{k+1} T_0^{-2^{k+1}} \tag{5.6}$$

and let  $n \geq n_0$  be sufficiently large so that  $n \geq \max_{1 \leq t \leq T_0} tm_0(t)$ .

Having fixed the involved parameters, we are given a  $k$ -uniform hypergraph  $H^{(k)} = (V, E)$  that contains at most  $cn^{k+1}$  copies of  $K_{k+1}^{(k)}$ . We apply Theorem 5.5, with constants  $\eta, \mu$  and the function  $\varepsilon'(\cdot, \dots, \cdot)$  chosen above, to  $H^{(k)}$  to obtain a vector  $\mathbf{a} \in \mathbb{N}^{k-1}$  and a  $T_0$ -bounded,  $(\eta, \varepsilon'(\mathbf{a}), \mathbf{a})$ -equitable family of partitions  $\mathcal{P} = \mathcal{P}(k-1, \mathbf{a})$  and a  $k$ -uniform hypergraph  $G^{(k)} = (V, E')$  satisfying properties (a) and (b) of Theorem 5.5.

Next we construct the promised  $K_{k+1}^{(k)}$ -free subhypergraph of  $\widehat{H}^{(k)} \subseteq H^{(k)}$ . For that we remove all hyperedges of  $H^{(k)}$  which are not contained in  $\text{Cross}_k(\mathcal{P}^{(1)})$ ; owing to the equitability of  $\mathcal{P}$ , there are at most  $\eta n^k$  such hyperedges. Moreover, we delete those hyperedges of  $K \in E \cap \text{Cross}_k(\mathcal{P}^{(1)})$  for which

- (A) either the hypergraphs  $G^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))$  and  $H^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))$  differ by more than  $\nu |\mathcal{K}_k(P^{(k-1)}(K))|$  hyperedges, i.e.,

$$\left| (G^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))) \Delta (H^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))) \right| \geq \nu \left| \mathcal{K}_k(P^{(k-1)}(K)) \right|$$

- (B) or the relative density of  $G^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))$  w.r.t.  $P^{(k-1)}(K)$  is less than  $d_k$ .

The number of hyperedges that were deleted because of (A) can be bounded by  $\mu n^k / \nu$ , since property (b) of Theorem 5.5 gives

$$\nu \left| \bigcup \{ \mathcal{K}_k(P^{(k-1)}(K)) : K \text{ was deleted because of (A)} \} \right| \leq \mu n^k.$$

Moreover, note that the relative density of  $H^{(k)} \cap \mathcal{K}_k(P^{(k-1)}(K))$  for a  $k$ -tuple that was deleted because of (B), but which were not deleted because of (A), can be at most  $d_k + \nu$ . Consequently, the number of  $k$ -tuples that were additionally deleted because of (B) is at most  $(d_k + \nu) \binom{n}{k} \leq 2d_k \binom{n}{k} < d_k n^k$ .

Summarizing the above, we have deleted at most

$$\eta n^k + \frac{\mu}{\nu} n^k + d_k n^k \stackrel{(5.4), (5.5)}{<} \delta n^k$$

hyperedges from  $H^{(k)}$  to obtain  $\widehat{H}^{(k)}$ . We claim that  $\widehat{H}^{(k)}$  is  $K_{k+1}^{(k)}$ -free.

Suppose on the contrary that a  $(k + 1)$ -element set  $K' \subseteq V$  spans a copy of  $K_{k+1}^{(k)}$  in  $\widehat{H}^{(k)}$ . This implies that none of the hyperedges of that copy of  $K_{k+1}^{(k)}$  was deleted. Therefore,  $G^{(k)}$  and  $\widehat{H}^{(k)}$  restricted on  $P^{(k-1)}(K')$  satisfy the assumptions of Theorem 5.6 (with (b) replaced by (ii')), whence  $\widehat{H}^{(k)} \subseteq H^{(k)}$  contains at least

$$\frac{1}{2} d_k^{k+1} \prod_{j=2}^{k-1} \left( \frac{1}{a_j} \right)^{\binom{k+1}{j}} \times \left( \frac{n}{a_1} \right)^{k+1} > \frac{1}{2} d_k^{k+1} \frac{n^{k+1}}{T_0^{2k+1}} \stackrel{(5.6)}{=} \delta n^{k+1}$$

copies of  $K_{k+1}^{(k)}$ , which contradicts the assumption on  $H^{(k)}$ . □

**Acknowledgement.** The author was supported by NSF grant DMS 1301698. I thank Mathias Schacht for his selfless and very extensive help in the preparation of this manuscript. Many thanks also to D. Conlon, J. Fox, J. Fuller, H. Hàn, Y. Kohayakawa, J. Nešetřil, and J. Retter for their careful reading and suggestions.

**References**

[1] N. Alon, R. A. Duke, H. Lefmann, V. Rödl, and R. Yuster, *The algorithmic aspects of the regularity lemma*, J. Algorithms **16** (1994), no. 1, 80–109.

- [2] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, *Combinatorica* **20** (2000), no. 4, 451–476.
- [3] N. Alon and A. Shapira, *A characterization of the (natural) graph properties testable with one-sided error*, *SIAM J. Comput.* **37** (2008), no. 6, 1703–1727.
- [4] ———, *Every monotone graph property is testable*, *SIAM J. Comput.* **38** (2008), no. 2, 505–522.
- [5] ———, *A separation theorem in property testing*, *Combinatorica* **28** (2008), no. 3, 261–281.
- [6] T. Austin, *Deducing the density Hales-Jewett theorem from an infinitary removal lemma*, *J. Theoret. Probab.* **24** (2011), no. 3, 615–633.
- [7] T. Austin and T. Tao, *Testability and repair of hereditary hypergraph properties*, *Random Structures Algorithms* **36** (2010), no. 4, 373–463.
- [8] B. Bollobás, P. Erdős, M. Simonovits, and E. Szemerédi, *Extremal graphs without large forbidden subgraphs*, *Ann. Discrete Math.* **3** (1978), 29–41, *Advances in graph theory* (Cambridge Combinatorial Conf., Trinity Coll., Cambridge, 1977).
- [9] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, B. Szegedy, and K. Vesztegombi, *Graph limits and parameter testing*, *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing* (New York), ACM, 2006, pp. 261–270.
- [10] F. R. K. Chung, *Quasi-random classes of hypergraphs*, *Random Structures Algorithms* **1** (1990), no. 4, 363–382.
- [11] F. R. K. Chung and R. L. Graham, *Quasi-random hypergraphs*, *Random Structures Algorithms* **1** (1990), no. 1, 105–124.
- [12] ———, *Quasi-random set systems*, *J. Amer. Math. Soc.* **4** (1991), no. 1, 151–196.
- [13] F. R. K. Chung, R. L. Graham, and R. M. Wilson, *Quasi-random graphs*, *Combinatorica* **9** (1989), no. 4, 345–362.
- [14] D. Conlon and J. Fox, *Graph removal lemmas*, *Surveys in combinatorics 2013*, *London Math. Soc. Lecture Note Ser.*, Cambridge Univ. Press, Cambridge, 2013, pp. 1–50.
- [15] D. Conlon, H. Hàn, Y. Person, and M. Schacht, *Weak quasi-randomness for uniform hypergraphs*, *Random Structures Algorithms* **40** (2012), no. 1, 1–38.
- [16] P. Dodos, V. Kanellopoulos, and K. Tyros, *A simple proof of the density Hales–Jewett theorem*, *Int. Math. Res. Not. IMRN*, to appear.
- [17] R. A. Duke and V. Rödl, *On graphs with small subgraphs of large chromatic number*, *Graphs Combin.* **1** (1985), no. 1, 91–96.
- [18] G. Elek and B. Szegedy, *A measure-theoretic approach to the theory of dense hypergraphs*, *Adv. Math.* **231** (2012), no. 3–4, 1731–1772.
- [19] P. Erdős, *Some remarks on the theory of graphs*, *Bull. Amer. Math. Soc.* **53** (1947), 292–294.
- [20] ———, *Problems and results on graphs and hypergraphs: similarities and differences*, *Mathematics of Ramsey theory* (J. Nešetřil and V. Rödl, eds.), *Algorithms Combin.*, vol. 5, Springer, Berlin, 1990, pp. 12–28.
- [21] P. Erdős, P. Frankl, and V. Rödl, *The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent*, *Graphs Combin.* **2**

- (1986), no. 2, 113–121.
- [22] P. Erdős, J. Nešetřil, and V. Rödl *On Pisier type problems and results (combinatorial applications to number theory)*, Mathematics of Ramsey theory, (J. Nešetřil and V. Rödl, eds.), Algorithms Combin., vol. 5, Springer, Berlin, 1990, pp. 214–231.
- [23] P. Erdős and P. Turán, *On some sequences of integers.*, J. London Math. Soc. **11** (1936), 261–264.
- [24] J. Fox, *A new proof of the graph removal lemma*, Ann. of Math. (2) **174** (2011), no. 1, 561–579.
- [25] J. Fox and L. M. Lovász, *A tight lower bound for Szemerédi’s regularity lemma*, submitted.
- [26] P. Frankl and V. Rödl, *Extremal problems on set systems*, Random Structures Algorithms **20** (2002), no. 2, 131–164.
- [27] Z. Füredi, *Extremal hypergraphs and combinatorial geometry*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994) (Basel), Birkhäuser, 1995, pp. 1343–1352.
- [28] H. Furstenberg, *Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions*, J. Analyse Math. **31** (1977), 204–256.
- [29] H. Furstenberg and Y. Katznelson, *An ergodic Szemerédi theorem for commuting transformations*, J. Analyse Math. **34** (1978), 275–291 (1979).
- [30] ———, *An ergodic Szemerédi theorem for IP-systems and combinatorial theory*, J. Analyse Math. **45** (1985), 117–168.
- [31] ———, *A density version of the Hales-Jewett theorem*, J. Anal. Math. **57** (1991), 64–119.
- [32] O. Goldreich, S. Goldwasser, and D. Ron, *Property testing and its connection to learning and approximation*, J. ACM **45** (1998), no. 4, 653–750.
- [33] W. T. Gowers, *Lower bounds of tower type for Szemerédi’s uniformity lemma*, Geom. Funct. Anal. **7** (1997), no. 2, 322–337.
- [34] ———, *A new proof of Szemerédi’s theorem*, Geom. Funct. Anal. **11** (2001), no. 3, 465–588.
- [35] ———, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Ann. of Math. (2) **166** (2007), no. 3, 897–946.
- [36] R. L. Graham, K. Leeb, and B. L. Rothschild, *Ramsey’s theorem for a class of categories*, Advances in Math. **8** (1972), 417–433.
- [37] B. Green and T. Tao, *New bounds for Szemerédi’s theorem. II. A new bound for  $r_4(N)$* , Analytic number theory, Cambridge Univ. Press, Cambridge, 2009, pp. 180–204.
- [38] ———, *An arithmetic regularity lemma, an associated counting lemma, and applications*, An irregular mind, Bolyai Soc. Math. Stud., vol. 21, János Bolyai Math. Soc., Budapest, 2010, pp. 261–334.
- [39] A. W. Hales and R. I. Jewett, *Regularity and positional games*, Trans. Amer. Math. Soc. **106** (1963), 222–229.
- [40] D. Hilbert, *Ueber die Irreducibilität ganzer rationaler Functionen mit ganzzahligen Coefficienten*, J. Reine Angew. Math. **110** (1892), 104–129.

- [41] Y. Kohayakawa, B. Nagle, V. Rödl, and M. Schacht, *Weak hypergraph regularity and linear hypergraphs*, J. Combin. Theory Ser. B **100** (2010), no. 2, 151–160.
- [42] Y. Kohayakawa, V. Rödl, and J. Skokan, *Hypergraphs, quasi-randomness, and conditions for regularity*, J. Combin. Theory Ser. A **97** (2002), no. 2, 307–352.
- [43] L. Lovász and B. Szegedy, *Graph limits and testing hereditary graph properties*, Tech. Report MSR-TR-2005-110, Microsoft Research, 2005.
- [44] ———, *Szemerédi's lemma for the analyst*, Geom. Funct. Anal. **17** (2007), no. 1, 252–270.
- [45] B. Nagle, V. Rödl, and M. Schacht, *The counting lemma for regular  $k$ -uniform hypergraphs*, Random Structures Algorithms **28** (2006), no. 2, 113–179.
- [46] D. H. J. Polymath, *A new proof of the density Hales-Jewett theorem*, Ann. of Math. (2) **175** (2012), no. 3, 1283–1327.
- [47] R. Rado, *Note on combinatorial analysis*, Proc. London Math. Soc. (2) **48** (1943), 122–160.
- [48] F. P. Ramsey, *On a problem in formal logic*, Proc. London Math. Soc. (2) **30** (1930), 264–286.
- [49] V. Rödl, *Some developments in Ramsey theory*, Proceedings of the International Congress of Mathematicians, Vol. I, II (Kyoto, 1990) (Tokyo), Math. Soc. Japan, 1991, pp. 1455–1466.
- [50] V. Rödl and M. Schacht, *Regular partitions of hypergraphs: counting lemmas*, Combin. Probab. Comput. **16** (2007), no. 6, 887–901.
- [51] ———, *Regular partitions of hypergraphs: regularity lemmas*, Combin. Probab. Comput. **16** (2007), no. 6, 833–885.
- [52] ———, *Generalizations of the removal lemma*, Combinatorica **29** (2009), no. 4, 467–501.
- [53] V. Rödl, M. Schacht, E. Tengan, and N. Tokushige, *Density theorems and extremal hypergraph problems*, Israel J. Math. **152** (2006), 371–380.
- [54] V. Rödl and J. Skokan, *Regularity lemma for  $k$ -uniform hypergraphs*, Random Structures Algorithms **25** (2004), no. 1, 1–42.
- [55] K. F. Roth, *On certain sets of integers*, J. London Math. Soc. **28** (1953), 104–109.
- [56] I. Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II, Colloq. Math. Soc. János Bolyai, vol. 18, North-Holland, Amsterdam, 1978, pp. 939–945.
- [57] T. Sanders, *On Roth's theorem on progressions*, Ann. of Math. (2) **174** (2011), no. 1, 619–636.
- [58] S. Shelah, *Primitive recursive bounds for van der Waerden numbers*, J. Amer. Math. Soc. **1** (1988), no. 3, 683–697.
- [59] J. Solymosi, *A note of a question of Erdős and Graham*, Random Structures Algorithms **25** (2004), no. 2, 263–267.
- [60] E. Szemerédi, *On sets of integers containing no four elements in arithmetic progression*, Acta Math. Acad. Sci. Hungar. **20** (1969), 89–104.
- [61] ———, *On sets of integers containing no  $k$  elements in arithmetic progression*, Acta

- Arith. **27** (1975), 199–245, Collection of articles in memory of Jurij Vladimirovič Linnik.
- [62] ———, *Regular partitions of graphs*, Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), Colloq. Internat. CNRS, vol. 260, CNRS, Paris, 1978, pp. 399–401.
- [63] T. Tao, *A quantitative ergodic theory proof of Szemerédi's theorem*, Electron. J. Combin. **13** (2006), no. 1, Research Paper 99, 49.
- [64] ———, *A variant of the hypergraph removal lemma*, J. Combin. Theory Ser. A **113** (2006), no. 7, 1257–1280.
- [65] A. Thomason, *Pseudorandom graphs*, Random graphs '85 (Poznań, 1985), North-Holland Math. Stud., vol. 144, North-Holland, Amsterdam, 1987, pp. 307–331.
- [66] B. L. van der Waerden, *Beweis einer Baudetschen Vermutung*, Nieuw Archief **15** (1927), 212–216.
- [67] E. Witt, *Ein kombinatorischer Satz der Elementargeometrie*, Math. Nachr. **6** (1952), 261–262.

Department of Mathematics and Computer Science, Emory University, Atlanta, USA and Charles University, Prague, Czech Republic  
E-mail: rodl@mathcs.emory.edu



# Finite dimensional representations of algebraic supergroups

Vera Serganova

**Abstract.** We review recent results and methods in the theory of finite-dimensional representations of classical algebraic supergroups. We discuss connections with Deligne tensor categories, block theory, associated variety, superanalogues of the Borel-Weil-Bott theorem and categorification of tensor representations of infinite-dimensional Lie algebras via representation theory of supergroups.

**Mathematics Subject Classification (2010).** 17B45, 17B10.

**Keywords.** Lie superalgebra, tensor category, flag supervariety, associated variety, superdimension, weight diagram, categorification.

## 1. Introduction

Representation theory of Lie superalgebras was originally motivated by applications in physics and in topology. In recent years duality and categorification unraveled the interplay of representation theory of Lie superalgebras with other branches of mathematics. The goal of this paper is to review recent results and different methods in finite-dimensional representation theory of Lie superalgebras. We completely ignore here a plethora of exciting recent results concerning infinite-dimensional representation theory of Lie superalgebras, (mostly for category  $\mathcal{O}$ ), see [12, 13] and [8].

In this paper  $\mathbb{F}$  is an algebraically closed field of characteristic zero. A *superalgebra*  $A$  is a  $\mathbb{Z}_2$ -graded  $\mathbb{F}$ -algebra, i.e.,  $A = A_0 \oplus A_1$  and  $A_i A_j \subset A_{i+j}$ . If  $x \in A_0$  or  $x \in A_1$ , then  $x$  is called *homogeneous* (even or odd). If  $x \in A_0$  (resp.  $x \in A_1$ ), we write  $\bar{x} = 0$  (resp.  $\bar{x} = 1$ ). If  $x$  and  $y$  are both homogeneous then the product  $xy$  is also homogeneous and  $\overline{xy} = \bar{x} + \bar{y}$ .

A *module* over a superalgebra  $A$  is a  $\mathbb{Z}_2$ -graded  $A$ -module. In particular, for  $A = \mathbb{F}$ , a *vector superspace*  $V$  is an  $\mathbb{F}$ -module, i.e., a  $\mathbb{Z}_2$ -graded vector space  $V = V_0 \oplus V_1$ . Define the dimension of a vector superspace as a pair  $(m|n)$ , where  $m = \dim V_0$ ,  $n = \dim V_1$ . If  $M$  is an  $A$ -module, we denote by  $M^\Pi$  the module obtained from  $M$  by switching parity:  $(M^\Pi)_0 = M_1$ ,  $(M^\Pi)_1 = M_0$ .

Given a classical formula, one superizes it by the following informal sign rule:

- all formulas are written for homogeneous elements only, and then extended to all elements by linearity;
- every term in the formula has an extra “sign” coefficient  $\pm 1$ ;

- if one term is obtained from another by permuting two adjacent letters  $x$  and  $y$ , the extra coefficients differ by the sign  $(-1)^{\bar{x}\bar{y}}$ .<sup>1</sup>

For instance, let us define the *supertrace*. Finite-dimensional trace is a composition of the isomorphism  $\text{End}_{\mathbb{F}}(V) \rightarrow V \otimes V^*$ , defined by<sup>2</sup>

$$(v \otimes l)(u) = v \cdot l(u), \quad \text{for all } v, u \in V, \quad l \in V^*,$$

with the obvious contraction map  $\mathbf{c}: V^* \otimes V \rightarrow \mathbb{F}$ . In the classical case one can ignore the difference between  $V^* \otimes V$  and  $V \otimes V^*$ . In the supercase it becomes important that one must identify  $V \otimes V^*$  with  $V^* \otimes V$  by switching  $v$  and  $l$ . By the sign rule we have

$$\text{str}(v \otimes l) = (-1)^{\bar{v}\bar{l}}l(v).$$

Hence  $\text{str } X$  vanishes for any odd  $X \in \text{End}_{\mathbb{F}}(V)$ , and for even  $X$  we have

$$\text{str } X = \text{tr}_{V_0} X - \text{tr}_{V_1} X.$$

Define the *superdimension*  $\text{sdim } V := \text{str Id}_V = \dim V_0 - \dim V_1$ .

In the language of category theory, vector superspaces are objects of a tensor category with *braiding*  $s: V \otimes W \rightarrow W \otimes V$  given by  $s(v \otimes w) = (-1)^{\bar{v}\bar{w}}w \otimes v$ .

The *supercommutator* of  $a$  and  $b$  is by definition  $ab - (-1)^{\bar{a}\bar{b}}ba$ . A *Lie superalgebra* is a vector superspace  $\mathfrak{g} = \mathfrak{g}_0 \oplus \mathfrak{g}_1$  with an even (i.e., grading preserving) linear map  $[\ , \ ]: \mathfrak{g} \otimes \mathfrak{g} \rightarrow \mathfrak{g}$ , called a *Lie bracket*, satisfying the following conditions:

- (i)  $[a, b] = -(-1)^{\bar{a}\bar{b}}[b, a]$ ,
- (ii)  $[a, [b, c]] = [[a, b], c] + (-1)^{\bar{a}\bar{b}}[b, [a, c]]$ .

Morphisms in the category of Lie superalgebras are parity preserving homomorphisms.

One can define a Lie supergroup in the smooth, analytic or algebraic category. In this paper we will talk about representations of affine algebraic supergroups  $G$ . In the direct approach, one replaces such  $G$  with (super)commutative finitely generated Noetherian Hopf superalgebras  $A = A_0 \oplus A_1$  over  $\mathbb{F}$ ; then  $G$  is associated to  $A$  as the functor  $\text{Hom}(A, \cdot)$  from the category of supercommutative  $\mathbb{F}$ -algebras to the category of groups. A representation of  $G$  is an  $A$ -comodule. One can check that the ideal  $I \subset A$  generated by  $A_1$  is a Hopf ideal and  $A/I$  is a commutative Hopf algebra which is isomorphic to the ring  $\mathbb{F}[G_0]$  of regular functions on some affine algebraic group  $G_0$ . The definition of the Lie algebra  $\mathfrak{g}$  of a supergroup  $G$  is analogous to that in the theory of algebraic groups with modifications given by the sign rule.

In this paper we will use a much easier approach to representation theory going via Harish-Chandra pairs. This approach was suggested by Koszul and Kostant for smooth Lie supergroups, [37, 38], and was recently developed in [58] for complex analytic supergroups and in [42] for algebraic supergroups. Let  $G\text{-mod}$  denote the category of finite-dimensional representations of  $G$  (with odd morphisms allowed). Then  $G\text{-mod}$  is equivalent to the category of finite-dimensional  $(\mathfrak{g}, G_0)$ -modules, i.e.,  $\mathfrak{g}$ -modules for which  $\mathfrak{g}_0$  action can be lifted

<sup>1</sup>It is customary to ignore non-letter symbols when applying the sign rule. This is kosher for symbols which are shortcuts for even operations. For example, if  $m$  and  $\kappa$  are even, then replacing  $a * b$  for  $m(a, b)$  and  $[a, b]$  for  $\kappa(a, b)$  won't affect the sign rule.

<sup>2</sup>Note that here we consider  $\text{End}_{\mathbb{F}}(V)$  acting on the left, and  $\mathbb{F}$  on the right, and two terms of the formula have letters in the same order.

to  $G_0$ . Note that only kernels and cokernels of homogeneous morphisms are defined. If we allow only grading preserving morphisms, then  $G\text{-mod}$  is a rigid symmetric tensor category.

Let us mention finally that most of the basic material on Lie superalgebras and their representations can be found in [44] and [11].

## 2. Classical and strange Lie superalgebras

Let  $V = V_0 \oplus V_1$  be a vector superspace  $\dim V = (m|n)$ . There are four natural Lie superalgebras which are related to  $V$ .

The Lie superalgebra  $\mathfrak{gl}(m, n)$  is by definition  $\text{End}_{\mathbb{F}}(V)$  with the supercommutator  $[X, Y] = XY - (-1)^{\bar{X}\bar{Y}}YX$  as the bracket. It has an ideal  $\mathfrak{sl}(m, n)$  of codimension 1 consisting of (super)traceless linear transformations and a one-dimensional center  $Z$  generated by the identity transformation. If  $m \neq n$ ,  $Z \cap \mathfrak{sl}(m, n) = 0$  and  $\mathfrak{sl}(m, n)$  is a simple Lie superalgebra.

If  $m = n$ , then  $Z \subset \mathfrak{sl}(n, n)$ . In order to obtain a simple Lie superalgebra we should take the quotient  $\mathfrak{psl}(n, n) = \mathfrak{sl}(n, n)/Z$ , which is simple if  $n \geq 2$ .

Assume now that  $V$  is equipped with a non-degenerate even symmetric bilinear form  $b: V \times V \rightarrow \mathbb{F}$ , i.e., for homogeneous  $v, w \in V$  we have

$$b(v, w) = (-1)^{\bar{v}\bar{w}}b(w, v) \quad \text{and} \quad b(v, w) \neq 0 \implies \bar{v} + \bar{w} = 0.$$

Since the restriction of  $b$  to  $V_1$  is skew-symmetric in the usual sense, the non-degeneracy of  $b$  implies that  $\dim V_1 = 2N$  is even. The Lie superalgebra  $\mathfrak{osp}(m, 2N)$  consists of all linear transformations of  $V$  preserving  $b$ :

$$\mathfrak{osp}(m, 2N) = \{X \in \mathfrak{gl}(m, 2N) \mid b(Xv, w) + (-1)^{\bar{X}\bar{v}}b(v, Xw) = 0\}$$

It is simple if  $m + 2N > 2$ . The Lie superalgebra  $\mathfrak{osp}(4, 2)$  has a one-parameter deformation  $D(2, 1; a)$  which is simple if  $a \neq 0, -1$ .

Since  $V$  has a  $\mathbb{Z}_2$ -grading, we can consider some other canonical structures on  $V$  in the case when  $\dim V_0 = \dim V_1 = n$ . For instance, we can equip  $V$  with a non-degenerate odd symmetric form  $b: V \times V \rightarrow \mathbb{F}$ , which defines a non-degenerate pairing  $V_0 \times V_1 \rightarrow \mathbb{F}$ . The linear transformations of  $V$ , which preserve  $b$ , form the Lie subsuperalgebra  $\mathfrak{p}(n)$  of  $\mathfrak{gl}(n, n)$ . To realize  $\mathfrak{p}(n)$  in matrix form choose a basis in  $V$  consisting of a basis in  $V_0$  and the dual basis in  $V_1$ . Then the matrices of elements of  $\mathfrak{p}(n)$  have the form

$$\left( \begin{array}{c|c} A & B \\ \hline C & -A^t \end{array} \right),$$

where  $B$  is symmetric and  $C$  is skew-symmetric. As in the case of  $\mathfrak{gl}(m, n)$ ,  $\mathfrak{p}(n)$  contains the ideal of traceless matrices which is simple for  $n \geq 3$ .

Finally we can fix an invertible odd operator  $\Pi: V \rightarrow V$  such that  $\Pi^2 = -1$  and consider the superalgebra  $\mathfrak{q}(n)$  of all linear transformations of  $V$  commuting with  $\Pi$  in the super sense. It is easy to see that in a suitable basis the matrices of elements from  $\mathfrak{q}(n)$  have the form

$$\left( \begin{array}{c|c} A & B \\ \hline B & A \end{array} \right).$$

The Lie superalgebra  $\mathfrak{q}(n)$  has a one-dimensional center  $Z$  consisting of scalar matrices and the ideal  $\mathfrak{sq}(n)$  of odd codimension 1 given by the condition  $\text{tr } B = 0$ . The Lie superalgebra  $\mathfrak{psq}(n) := \mathfrak{sq}(n)/Z$  is simple for  $n \geq 3$ . Note that  $\mathfrak{gl}(m, n)$  and  $\mathfrak{q}(n)$  are also associative superalgebras.

It was shown by Kac [34] that any finite-dimensional simple Lie superalgebra  $\mathfrak{g}$  with reductive even part  $\mathfrak{g}_0$  and non-trivial odd part  $\mathfrak{g}_1$  is isomorphic to either one of the above superalgebras or to one of two exceptional superalgebra:  $G_3$  of dimension (17|14) or  $F_4$  of dimension (24|16). The Lie superalgebras  $\mathfrak{p}(n)$  and  $\mathfrak{q}(n)$  are sometimes called *strange* superalgebras.

Finite-dimensional simple Lie superalgebras with non-reductive  $\mathfrak{g}_0$  are all obtained as subalgebras of the Lie algebra of vector fields on  $(0|n)$ -dimensional supermanifold; therefore they are called Cartan type superalgebras. We will not discuss them in this paper, although their representation theory is also very interesting, see for instance [33].

The goal of this paper is to discuss finite-dimensional representation theory of Lie superalgebras  $\mathfrak{gl}(m, n)$ ,  $\mathfrak{osp}(m, 2n)$ ,  $\mathfrak{p}(n)$  and  $\mathfrak{q}(n)$ . In fact, we restrict our attention to “integrable” representations, i.e., representations of the corresponding algebraic supergroups  $GL(m, n)$ ,  $OSP(m, 2n)$ ,  $P(n)$  and  $Q(n)$ . Note that  $OSP(m, 2n)$  is not connected if  $m > 0$ . We denote by  $SOSP(m, 2n)$  the connected component of identity. In what follows we denote the Lie superalgebra by  $\mathfrak{g}$  and the corresponding supergroup by  $G$ .

### 3. Superanalogues of Schur–Weyl duality

**3.1. Sergeev–Schur–Weyl duality.** Probably the most classical approach to representation theory is via Schur–Weyl duality. Indeed, any of our supergroups has a natural (or standard) representation  $V$  and it is the most natural thing to study the tensor powers of  $V$  as representations of  $G$ . Define the action of the symmetric group  $S_p$  on  $V^{\otimes p}$  by the formula

$$s_i(v_1 \otimes \cdots \otimes v_p) = (-1)^{\bar{v}_i \bar{v}_{i+1}} v_1 \otimes \cdots \otimes v_{i+1} \otimes v_i \otimes \cdots \otimes v_p. \tag{3.1}$$

This action commutes with the action of  $GL(m, n)$ .

**Theorem 3.1** (A. Sergeev). *Let  $A_p$  and  $B_p$  denote the images of  $\mathbb{F}[S_p]$  and  $U(\mathfrak{gl}(m, n))$  respectively in  $\text{End}_{\mathbb{F}}(V^{\otimes p})$ .*

- (a)  $A_p = \text{End}_{GL(m, n)}(V^{\otimes p})$  and  $B_p = \text{End}_{S_p}(V^{\otimes p})$ .
- (b) The map  $\mathbb{F}[S_p] \rightarrow A_p$  is injective if and only if  $(m + 1)(n + 1) > p$  (the stable case).
- (c) Let  $\Lambda_{m, n}^p$  denote the set of Young diagrams with  $p$  boxes which do not contain the  $(m + 1) \times (n + 1)$ -rectangle. Then  $V^{\otimes p}$  has the following decomposition as a module over  $GL(m, n) \otimes \mathbb{F}[S_p]$

$$V^{\otimes p} = \bigoplus_{\lambda \in \Lambda_{m, n}^p} V_{\lambda} \otimes S(\lambda),$$

where  $S(\lambda)$  denotes the irreducible representation of  $S_p$  associated with  $\lambda$  and  $V_{\lambda}$  is an irreducible representation of  $GL(m, n)$ . Furthermore,  $V_{\lambda}$  and  $V_{\mu}$  are not isomorphic if  $\mu \neq \lambda$ .

As in the classical case,  $V^{\otimes p}$  is a semisimple  $GL(m, n)$ -module, and its irreducible components  $V_\lambda$  are the images of the corresponding Young projectors.

**Remark 3.2.** The Young diagrams in  $\Lambda_{m,n}^p$  are subdiagrams of the  $(m, n)$ -infinite hook (see [3]).

Note that, if  $n = 0$ , we recover the classical Schur–Weyl duality between  $GL(m)$  and  $S_p$ , and, when  $m = 0$ , we obtain a twisted version of the classical Schur–Weyl duality between  $GL(n)$  and  $S_p$ , where the representations of  $GL(n)$  are enumerated by the transposed Young diagrams. Recall that  $\{V_\lambda \mid \lambda \in \Lambda_{m,0}^p, p \in \mathbb{N}\}$  is a complete list of all *polynomial* irreducible representations of  $GL(m)$ . To obtain all irreducible representations one may tensor representations  $V_\lambda$  with tensor powers of 1-dimensional representation  $\Lambda^m V^*$ . Hence one can realize essentially all irreducible representations in tensor powers of  $V$ .

**3.2. The mixed case.** In the supercase, however, one can not make such a shortcut. In order to obtain all irreducible representations of  $GL(m, n)$  we need to consider the mixed tensor powers of  $V$  and  $V^*$ . Here we immediately encounter a new phenomenon: the representation of  $GL(m, n)$  in  $V^{\otimes r} \otimes (V^*)^{\otimes q}$  is *not completely reducible* in general.

As above, the centralizer algebra  $\text{End}_{GL(m,n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$  carries a lot of information about this module. It contains the image of  $\mathbb{F}[S_r \times S_q]$  in  $V^{\otimes r} \otimes (V^*)^{\otimes q}$ , and  $\theta = \mathbf{e} \circ \mathbf{c} \circ \mathbf{s} \in \text{End}_{GL(m,n)}(V \otimes V^*)$ ; here  $\mathbf{c}, \mathbf{e}$  are the obviously defined  $GL(m, n)$ -invariant maps

$$\mathbf{c}: V^* \otimes V \rightarrow \mathbb{F}, \quad \mathbf{e}: \mathbb{F} \rightarrow V \otimes V^*,$$

and  $\mathbf{s}$  is the braiding. By simple calculation,  $\theta^2 = (m-n)\theta$ . Define  $\theta_{r1} \in \text{End}_{GL(m,n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$  as

$$\theta_{r1} = \text{Id}^{\otimes r-1} \otimes \theta \otimes \text{Id}^{\otimes q-1}.$$

The following theorem can be considered as a superanalogue of the first fundamental theorem of invariant theory.

**Theorem 3.3.** *The algebra  $\text{End}_{GL(m,n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$  is generated by  $S_r \times S_q$  and  $\theta_{r1}$ .*

In the case of general linear groups this theorem is a classical result of H. Weyl and it follows from Schur–Weyl duality. For general linear supergroups it follows from Theorem 4.1 below proven in [14].

In fact, one can completely describe generators and relations in  $\text{End}_{GL(m,n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$  if  $r, q \gg m, n$ . The resulting finite-dimensional algebras are called *walled Brauer algebras*, they were first studied in [1]. Before defining them, recall the usual Brauer algebras, which appear if one considers centralizers of orthogonal and symplectic groups (and generalizes to orthosymplectic supergroups).

**3.3. Brauer algebras: orthosymplectic case.** Recall that for orthogonal groups, the analogue of Schur–Weyl duality is given by the Brauer algebra  $B_p(t)$  discovered by R. Brauer on 1937, [5]. It has generators  $s_1, \dots, s_{p-1}$  and  $\tau_1, \dots, \tau_{p-1}$  subject to relations

$$\begin{aligned} s_i^2 &= 1, & \tau_i^2 &= t\tau_i, & s_i\tau_i &= \tau_i s_i = \tau_i, \\ s_i s_j &= s_j s_i, & \tau_i s_j &= s_j \tau_i, & \tau_i \tau_j &= \tau_j \tau_i \quad \text{if } |i-j| \geq 2, \\ s_i s_{i+1} s_i &= s_{i+1} s_i s_{i+1}, & \tau_i \tau_{i+1} \tau_i &= \tau_i, & \tau_{i+1} \tau_i \tau_{i+1} &= \tau_{i+1}, \\ s_i \tau_{i+1} \tau_i &= s_{i+1} \tau_i, & \tau_i s_{i\pm 1} \tau_i &= \tau_i & \tau_{i+1} \tau_i s_{i+1} &= \tau_{i+1} s_i. \end{aligned} \tag{3.2}$$

It is not hard to check that  $B_p(t)$  contains the subalgebra  $\mathbb{F}[S_p]$  generated by  $s_1, \dots, s_{p-1}$ . If  $V$  is a  $(m|2n)$ -dimensional vector space equipped with an even symmetric bilinear form  $b(\cdot, \cdot)$  and  $\{e_k\}, \{f_k\}$  are homogeneous bases in  $V$ , which are dual with respect to the form  $b$ , we can extend the action of  $S_p$  in  $V^{\otimes p}$  given by (3.1) to the action of  $B_p(m - 2n)$  by setting

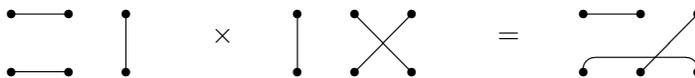
$$\tau_i(v_1 \otimes \dots \otimes v_p) = b(v_i, v_{i+1}) \sum_{k=1}^{m+2n} (-1)^{\bar{e}_k} v_1 \otimes \dots \otimes v_{i-1} \otimes e_k \otimes f_k \otimes v_{i+2} \otimes \dots \otimes v_p. \quad (3.3)$$

One can easily check that the action of  $B_p(t)$  commutes with the action of  $OSP(m, 2n)$ . Note that  $t$  is the superdimension of  $V$ . This action was studied in detail in [2]. The following result was obtained by C. Stroppel and the author independently and is not published yet.

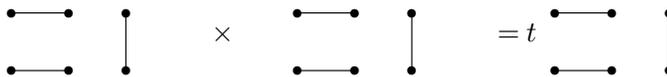
**Theorem 3.4.** *Let  $V$  denote the standard representation of the supergroup  $OSP(m, 2n)$ . Then the homomorphism  $B_p(m - 2n) \rightarrow \text{End}_{OSP(m, 2n)}(V^{\otimes p})$  is surjective.*

Notice that in the case  $m = 0$  we have a surjective homomorphism  $B_p(-2n) \rightarrow \text{End}_{SP(2n)}(V^{\otimes p})$  and if  $n = 0$  a surjective homomorphism  $B_p(m) \rightarrow \text{End}_{O(m)}(V^{\otimes p})$ , as in the classical Brauer duality. This homomorphism is injective if  $p \leq m + 1$  (respectively,  $p \leq 2n + 1$ ).

The Brauer algebra  $B_p(t)$  has a nice diagrammatic presentation. It has a natural basis enumerated by all perfect pairings on the  $2p$ -element set; write its elements as two rows with  $p$  dots each and join the elements in every pair by a string. The resulting picture is called a *Brauer diagram*. The multiplication of the basis elements in  $B_p(t)$  is the composition of diagrams: identify the lower level of the first diagram with the upper level of the second, remove all the cycles and get a new diagram. The product of the corresponding elements in  $B_p(t)$  equals the element corresponding to this new diagram multiplied by  $t^d$ , where  $d$  is the number of removed cycles. In particular,  $\dim B_p(t) = (2p - 1)!! := (2p - 1)(2p - 3) \dots 1$ . For example



and



It is proven in [59] that  $B_p(t)$  is semisimple if  $t \notin \mathbb{Z}$  or  $p < |t|$ , and in [48] a complete criterion for semisimplicity of  $B_p(t)$  is found.

**3.4. Walled Brauer algebras and the  $GL$  case.** Let  $p = r + q$ . Color  $r$  dots in each row black and the other  $q$  white. Call a diagram admissible if any string joins either two points of different colors at the same level or points of the same color at the different levels. It is easy to see that the product of admissible diagrams is admissible. The subalgebra  $WB_{r,q}(t) \subset B_p(t)$  spanned by all admissible diagrams is called the *walled Brauer algebra*. Here are some examples of admissible diagrams:



One can define the action of  $WB_{r,q}(m - n)$  on the mixed tensor power  $V^{\otimes r} \otimes (V^*)^{\otimes q}$  of a  $(m|n)$ -dimensional vector space  $V$  and check that this action commutes with  $GL(m, n)$ . Theorem 3.3 implies that the natural homomorphism  $WB_{r,q} \rightarrow \text{End}_{GL(m,n)}(V^{\otimes p} \otimes (V^*)^{\otimes q})$  is surjective.

**3.5. Duality for strange supergroups.** One can generalize the above duality results to the *strange* supergroups  $P(n)$  and  $Q(n)$ . In the case of  $P(n)$  we obtain an interesting twisted version of the Brauer algebra  $B_p(0)$ . It was described by D. Moon [43]. Let  $C_p$  be the algebra with generators  $s_1, \dots, s_{p-1}$  and  $\epsilon_1, \dots, \epsilon_{p-1}$ , subject to relations

$$\begin{aligned} s_i^2 &= 1, & \epsilon_i^2 &= 0, \quad s_i \epsilon_i = -\epsilon_i s_i = \epsilon_i, \\ s_i s_j &= s_j s_i, & \epsilon_i s_j &= s_j \epsilon_i, \quad \epsilon_i \epsilon_j = \epsilon_j \epsilon_i \text{ if } |i - j| \geq 2, \\ s_i s_{i+1} s_i &= s_{i+1} s_i s_{i+1}, & \epsilon_i \epsilon_{i+1} \epsilon_i &= -\epsilon_i, \quad \epsilon_{i+1} \epsilon_i \epsilon_{i+1} = -\epsilon_{i+1}, \\ s_i \epsilon_{i+1} \epsilon_i &= -s_{i+1} \epsilon_i, & \epsilon_{i+1} \epsilon_i s_{i+1} &= \epsilon_{i+1} s_i, \quad \epsilon_i s_{i \pm 1} \epsilon_i = \epsilon_i. \end{aligned} \tag{3.4}$$

It is shown in [43] that  $C_p$  has the same dimension as  $B_p(t)$ . But, in contrast with  $B_p(t)$ ,  $C_p$  does not have a semisimple deformation. For instance,  $C_2$  is isomorphic to the algebra of upper triangular  $2 \times 2$  matrices, hence it is not commutative; but any semisimple 3-dimensional algebra is commutative. On the other hand,  $B_2(t)$  is commutative and isomorphic to  $\mathbb{F}^3$  if  $t \neq 0$ .

If  $V$  is the standard representation of  $P(n)$ , then we have a homomorphism

$$C_p \rightarrow \text{End}_{P(n)}(V^{\otimes p})$$

defined by the formulas (3.3) with substitution of  $\epsilon$  in the place of  $\tau$  and taking into account that  $\epsilon$  is odd.

**Theorem 3.5** ([43]). *The homomorphism  $C_p \rightarrow \text{End}_{P(n)}(V^{\otimes p})$  is surjective.*

It is important to understand better the structure of the finite-dimensional algebra  $C_p$ . It is not a cellular algebra. However, following [30] one can construct an “almost” cellular basis in  $C_p$ , i.e., a basis satisfying conditions (C1) and (C3) of [30]. (The problem is with the involution  $s_i^* = -s_i$ ,  $\epsilon_i^* = -\epsilon_i$ , which, in this case, permutes the cells.) It is possible to show that, as for the usual Brauer algebra  $B_p(0)$ , primitive idempotents can be enumerated by partitions of length  $p - 2s$  for  $s = 0, \dots, \lfloor \frac{p-1}{2} \rfloor$ .

To understand better the Schur–Weyl duality for  $Q(n)$ , we first formulate the superanalogue of Schur’s Lemma, proven in [34].

**Lemma 3.6.** *Let  $A = A_0 \oplus A_1$  be an associative superalgebra and  $M$  be a simple  $A$ -module. Then  $\text{End}_A(M) = \mathbb{F}$  or  $\text{End}_A(M) = \mathcal{Q}(1)$ , where  $\mathcal{Q}(1)$  is the associative superalgebra whose Lie superalgebra is isomorphic to  $\mathfrak{q}(1)$ .*

In the first case we say that  $M$  is of  $GL$  type and in the second case we say that  $M$  is of  $Q$  type. The irreducible representations of  $GL(m, n)$ ,  $OSP(m, 2n)$  or  $P(n)$  are all of  $GL$  type. Some irreducible representations of  $Q(n)$  are of  $Q$  type, for instance, the standard representation.

The Schur–Weyl duality for  $Q(n)$  in the case of non-mixed tensors was established by Sergeev. Let  $\text{Cliff}_r$  denote the Clifford (super)algebra with generators  $p_i$ ,  $i = 1, \dots, r$ , satisfying the relations

$$p_i^2 = -1, \quad p_i p_j + p_j p_i = 0.$$

Consider the obvious action of  $S_r$  on  $\text{Cliff}_r$ . Let  $H_r$  be the semidirect product of  $\mathbb{F}[S_r]$  and  $\text{Cliff}_r$ .

One can show (see [55]) that  $H_r$  is semisimple. Furthermore, simple  $H_r$ -modules are in bijection with projective representations of  $S_n$ , and can be parametrized by strict partitions<sup>3</sup> of  $r$ . Denote by  $T_\lambda$  the representation of  $H_r$  associated with a strict partition  $\lambda$ . Let  $l(\lambda)$  be the number of non-zero parts in  $\lambda$  and  $\delta(\lambda)$  denote the parity of  $l(\lambda)$ . Then  $T_\lambda$  is of  $GL$  type if  $\delta(\lambda) = 0$  and of  $Q$  type if  $\delta(\lambda) = 1$ . Let  $\Sigma_n^r$  denote the set of strict partitions  $\lambda$  of  $r$  with no more than  $n$  non-zero parts.

Let  $V$  be the standard  $(n|n)$ -dimensional representation of  $Q(n)$ . Extend the action of  $\mathbb{F}[S_r]$  to an action of  $H_r$  on  $V^{\otimes r}$  by setting

$$p_i(v_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes v_n) = (-1)^{\bar{v}_1 + \cdots + \bar{v}_{i-1}} v_1 \otimes \cdots \otimes \Pi(v_i) \otimes \cdots \otimes v_n.$$

Thus, we have defined the homomorphism  $H_r \rightarrow \text{End}_{Q(n)}(V^{\otimes r})$ .

**Theorem 3.7** ([53]).

- (a) *The homomorphism  $H_r \rightarrow \text{End}_{Q(n)}(V^{\otimes r})$  is surjective. It is an isomorphism if and only if  $r \leq \frac{(n+1)(n+2)}{2}$ .*
- (b) *There is a decomposition*

$$V^{\otimes r} = \bigoplus_{\lambda \in \Sigma_n^r, \delta(\lambda)=0} V_\lambda \otimes T_\lambda \oplus \bigoplus_{\lambda \in \Sigma_n^r, \delta(\lambda)=1} V_\lambda \otimes_{Q(1)} T_\lambda, \tag{3.5}$$

here the  $V_\lambda$ -s are mutually non-isomorphic simple  $Q(n)$ -modules,  $V_\lambda$  is of type  $GL$  (resp.  $Q$ ) if  $\delta(\lambda) = 0$  (resp. 1).

Since for  $Q(n)$  we do not have an isomorphism between  $V$  and  $V^*$ , we need to consider mixed tensors  $V^{\otimes r} \otimes (V^*)^{\otimes q}$ . Note that  $\text{End}_{GL(n,n)}(V^{\otimes r} \otimes (V^*)^{\otimes q}) \subset \text{End}_{Q(n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$ , in particular,  $\theta_{r,1} \in \text{End}_{Q(n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$ . It is also clear that the natural action of  $H_r \otimes H_q$  commutes with  $Q(n)$ .

**Theorem 3.8** ([31]). *The operator  $\theta_{r,1}$  and  $H_r \otimes H_q$  generate  $\text{End}_{Q(n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$ .*

Moreover, for all  $n > r + q$  the algebras  $\text{End}_{Q(n)}(V^{\otimes r} \otimes (V^*)^{\otimes q})$  are isomorphic to the same algebra which we denote by  $H_{r,q}$ . This algebra is called the *walled Brauer–Clifford superalgebra* and it also has a beautiful diagrammatic presentation studied in [31].

### 4. Deligne’s categories

In [17, 18] P. Deligne introduced a notion of universal Karoubian  $\mathbb{F}$ -linear tensor category  $\text{Rep } GL(t)$  with one generator, where the parameter  $t \in \mathbb{F}$  is the dimension of the generator. When  $t \notin \mathbb{Z}$  this category is semisimple. For  $t \in \mathbb{Z}$ ,  $t \neq 0$ , the Deligne category has a canonical semisimple quotient equivalent to  $GL(|t|)\text{-mod}$ , [17].

Some other quotients of  $\text{Rep } GL(t)$  are equivalent to the Karoubian subcategories of  $GL(m, n)\text{-mod}$  (with  $m - n = t$ ) generated by the standard representation. Description

---

<sup>3</sup> A partition is strict if all its non-zero parts have different size

of all tensor ideals in  $\underline{\text{Rep}} GL(t)$  is an interesting open problem. Recently J. Comes [15] proved that all non-trivial ideals, which are generated by indecomposable objects, in the Deligne category with integer  $t$  are associated with  $GL(m, n)$ -mod, where  $m - n = t$ .

In this section we explore the interplay between Deligne categories and supergroups and give examples how this interaction enriches both theories.

**4.1. Basic construction.** In [18] Deligne defined a family of symmetric rigid tensor categories  $\underline{\text{Rep}} GL(t)$  and  $\underline{\text{Rep}} O(t)$ . The category  $\underline{\text{Rep}} GL(t)$  is a Karoubian  $\mathbb{F}$ -linear tensor category generated by an object  $X$  and the dual object  $\check{X}$  such that  $c \circ s \circ e = t\text{Id}$  where  $e: \mathbb{F} \rightarrow X \otimes \check{X}$  and  $c: \check{X} \otimes X \rightarrow \mathbb{F}$  are the evaluation and coevaluation morphisms respectively and  $s: X \otimes \check{X} \rightarrow \check{X} \otimes X$  is the braiding,  $\mathbb{F}$  denotes the unit object. The category  $\underline{\text{Rep}} GL(t)$  satisfies the universal property: if  $\mathcal{T}$  is a symmetric rigid  $\mathbb{F}$ -linear tensor category, then the category of  $\mathbb{F}$ -linear tensor functors  $\underline{\text{Rep}} GL(t) \rightarrow \mathcal{T}$  is equivalent to the category of objects in  $\mathcal{T}$  of dimension  $t$ .

The construction starts with building a skeleton category  $\text{Rep}_0 GL(t)$  whose objects are pairs of non-negative integers  $(r, s)$  represented by  $r$  black and  $s$  white nodes. A black node corresponds to a copy of  $X$  and a white node to  $\check{X}$ . The tensor product  $(r, s)$  and  $(r', s')$  is  $(r + r', s + s')$ . The morphism space between two such objects is the formal span of admissible Brauer diagrams with  $r$  black and  $s$  white nodes in the upper row and  $r'$  black and  $s'$  white nodes in the lower row. The composition law on such diagrams is given by the same rule as in Brauer algebras. Then  $\underline{\text{Rep}} GL(t)$  is defined as the Karoubian envelope of the additive envelope of  $\text{Rep}_0 GL(t)$ . Here are the diagrams representing  $c$  and  $e$  respectively:



the first diagram has empty lower row and the second one has empty upper row.

Assume  $t \notin \mathbb{Z}$ . Then category  $\underline{\text{Rep}} GL(t)$  is semisimple and hence abelian. Furthermore, indecomposable=simple subobjects of  $X^{\otimes r} \otimes \check{X}^{\otimes s}$  are in bijection with primitive idempotents of the walled Brauer algebra  $WB_{r,s}(t)$ . They are parametrized by pairs of partitions (equivalently bipartitions). Tensor products of indecomposable objects can be expressed in terms of Littlewood–Richardson coefficients.

Now let  $t$  be an integer. It is still true that indecomposable objects are enumerated by bipartitions, the indecomposable object  $Y_{\lambda,\mu}$  associated with a bipartition  $(\lambda, \mu)$  is defined as the image of a certain primitive idempotent in the walled Brauer algebra  $WB_{r,s}(t)$ , see [14] or [10] for details. However,  $Y_{\lambda,\mu}$  are not simple and tensor product rules change.

**4.2. Basic functors.** It is difficult to describe the ideals of  $\text{Rep} GL(t)$  when  $t \in \mathbb{Z}$ . It is shown in [17] that  $\underline{\text{Rep}} GL(t)$  has a maximal ideal among those in which all endomorphisms have zero trace, it is called the ideal of *negligible* morphisms. The factor category of  $\underline{\text{Rep}} GL(t)$  by this ideal is equivalent to the category  $GL(t, 0)$ -mod (for positive  $t$ ) or  $GL(0, -t)$ -mod (for negative  $t$ ).

On the other hand, for any pair of non-negative integers  $(m, n)$  satisfying  $m - n = t$  one can construct the ideal associated with the category  $GL(m, n)$ -mod. Indeed, by universality there exists a unique (up to isomorphism) tensor functor

$$F_{m,n}: \underline{\text{Rep}} GL(t) \rightarrow GL(m, n)\text{-mod}$$

such that  $F_{m,n}(X) = V$ , and the ideal is its kernel. This functor was studied in [14] by

Comes and Wilson to describe indecomposable summands of the mixed tensor powers of the standard representation of  $GL(m, n)$  and to calculate their characters. For this they deform  $Y_{\lambda, \mu}$  to an object  $Y_{\lambda, \mu}(s) \in \underline{\text{Rep}} GL(s)$  for generic  $s$ . The idempotent  $Y_{\lambda, \mu}(s)$  is not indecomposable but one can compute the multiplicities of indecomposables in it. This computation is based on the deep results of [10] about walled Brauer algebras.

The set of bipartitions  $(\lambda, \mu)$  such that  $F_{m,n}(Y_{\lambda, \mu}) \neq 0$  is described in [14] by a very beautiful combinatorial condition generalizing the hook condition of Remark 3.2. We say  $(\lambda, \mu)$  is an  $(m, n)$ -cross if there exists  $k$  with  $0 \leq k \leq m$  such that  $\lambda_{k+1} + \mu_{m-k+1} \leq n$ . If we draw the Young diagrams  $\lambda$  and  $\mu$  so that their upper left corner coincide and then rotate  $\mu$  on  $180^\circ$ , then  $(\lambda, \mu)$  is an  $(m, n)$ -cross if and only if the latter picture can be covered by an  $m$ -high and  $n$ -wide infinite cross.

**Theorem 4.1** ([14]). *The functor  $F_{m,n}$  is a full tensor functor.  $F_{m,n}(Y_{\lambda, \mu}) \neq 0$  if and only if  $(\lambda, \mu)$  is an  $(m, n)$ -cross.*

This theorem identifies the Karoubian category generated by mixed tensor products in  $GL(m, n)$ -mod with a quotient of  $\underline{\text{Rep}} GL(m - n)$ .

**4.3. Abelinization.** Our next goal is to construct for any integer  $t$  an abelian tensor category  $\overline{\text{Rep}} GL(t)$  that admits a fully faithful tensor functor  $F: \underline{\text{Rep}} GL(t) \rightarrow \overline{\text{Rep}} GL(t)$ . Note there is no canonical way to construct an abelian tensor category from the Karoubian one. One may consider  $\overline{\text{Rep}} GL(t)$  as a good candidate for an ‘‘abelian envelope’’ of  $\underline{\text{Rep}} GL(t)$ .

Let  $x \in \mathfrak{gl}(m, n)_1$  such that  $[x, x] = 0$ . Denote by  $\mathfrak{g}^x = \text{Ker ad}_x$  the centralizer of  $x$  in  $\mathfrak{g} = \mathfrak{gl}(m, n)$ . Then  $[x, \mathfrak{g}] = \text{Im ad}_x$  is the ideal in  $\mathfrak{g}^x$  and  $\mathfrak{g}_x := \mathfrak{g}^x/[x, \mathfrak{g}] \simeq \mathfrak{gl}(m - k, n - k)$ , where  $k$  is the rank of  $x$  in the standard representation  $V$ . Let  $M \in GL(m, n)$ -mod and  $M_x = \text{Ker } x/xM$ . Then  $M_x$  has a natural structure of a  $\mathfrak{g}_x$ -module. Thus,  $M_x \in GL(m - k, n - k)$ -mod. It is a simple exercise to check that  $(M^*)_x \simeq (M_x)^*$  and  $(M \otimes N)_x \simeq M_x \otimes N_x$  for any  $M, N \in GL(m, n)$ -mod. Therefore we have constructed a tensor functor  $E_{m,n}^k: GL(m, n)$ -mod  $\rightarrow GL(m - k, n - k)$ -mod.

**Proposition 4.2.**

- (a) For all  $m, n$  and  $k \leq \min(m, n)$  we have  $E_{m,n}^k \circ F_{m,n} \simeq F_{m-k, n-k}$ .
- (b) Let  $t = m - n$  and  $Y \in \underline{\text{Rep}} GL(t)$ . Then  $F_{m-1, n-1}(Y) = 0$  if and only if  $F_{m,n}(Y)$  is projective in  $GL(m, n)$ -mod.

Let  $\mathcal{F}^p(m, n)$  denote the full subcategory of  $GL(m, n)$ -modules whose objects are subquotients in a direct sum of finitely many copies of  $V^{\otimes r} \otimes (V^*)^{\otimes s}$  with  $r + s \leq p$ . By definition  $\mathcal{F}^p(m, n)$  is an abelian category. It is easy to check that

$$E_{m,n}^1(\mathcal{F}^p(m, n)) \subset \mathcal{F}^p(m - 1, n - 1).$$

It seems plausible, although we don't have a complete proof at the moment, that if  $m+n \gg p$  then  $E_{m,n}^1: \mathcal{F}^p(m, n) \rightarrow \mathcal{F}^p(m - 1, n - 1)$  is an equivalence of abelian categories.

Let  $t \in \mathbb{Z}$ . Consider the family of categories  $\mathcal{F}^p(t + k, k)$  for  $k \geq \min(-t, 0)$  and the family of functors

$$E_{t+k, k}^1: \mathcal{F}^p(t + k, k) \rightarrow \mathcal{F}^p(t + k - 1, k - 1).$$

Let  $\overline{\mathcal{F}}^p(t)$  be the inverse limit of  $\mathcal{F}^p(t + k, k)$  and  $\overline{\text{Rep}} GL(t)$  be the direct limit of  $\overline{\mathcal{F}}^p(t)$ . One can consider now the tensor product bifunctor  $\overline{\mathcal{F}}^p(t) \times \overline{\mathcal{F}}^q(t) \rightarrow \overline{\mathcal{F}}^{p+q}(t)$  defined in

the obvious way and check easily that this bifunctor equips  $\overline{\text{Rep}} GL(t)$  with a monoidal structure. It is also clear that  $\overline{\text{Rep}} GL(t)$  contains a standard object  $\bar{V}$  obtained as the limit of standard objects in  $GL(t+k, k)\text{-mod}$ .

**Theorem 4.3.**

- (a) *The category  $\overline{\text{Rep}} GL(t)$  is a symmetric rigid  $\mathbb{F}$ -linear abelian category. There exists a unique (up to isomorphism) fully faithful tensor functor  $F: \underline{\text{Rep}} GL(t) \rightarrow \overline{\text{Rep}} GL(t)$  such that  $F(X) = \bar{V}$ .*
- (b) *Simple objects of  $\overline{\text{Rep}} GL(t)$  are enumerated by bipartitions  $(\lambda, \mu)$ . Duality switches  $\lambda$  and  $\mu$ .*
- (c) *Let  $L_{\lambda, \mu}$  denote the simple object corresponding to  $(\lambda, \mu)$ ,  $W_{\lambda', \mu'} := F(Y_{\lambda', \mu'})$  and  $[W_{\lambda', \mu'} : L_{\lambda, \mu}]$  denote the multiplicity of  $L_{\lambda, \mu}$  in  $W_{\lambda', \mu'}$ . Then  $[W_{\lambda', \mu'} : L_{\lambda, \mu}] \neq 0$  implies  $|\lambda| - |\lambda'| = |\mu| - |\mu'| \leq 0$  and  $\lambda \leq \lambda', \mu \leq \mu'$  in the dominance order for partitions. Moreover,  $[W_{\lambda, \mu} : L_{\lambda, \mu}] = 1$ .*

It is interesting to calculate the multiplicities  $[W_{(\lambda', \mu')} : L_{\lambda, \mu}]$ . It seems reasonable to conjecture that they are closely related to cellular structure of the walled Brauer algebras and multiplicities calculated in [16].

**4.4. Orthosymplectic case.** The construction of  $\underline{\text{Rep}} O(t)$  is done in a similar way, see [17], by introducing a self-dual object  $X$  and morphisms  $\mathbf{b}: X \otimes X \rightarrow \mathbb{F}$  and  $\check{\mathbf{b}}: \mathbb{F} \rightarrow X \otimes X$  such that

$$\mathbf{b} \circ \mathbf{s} = \mathbf{b}, \mathbf{s} \circ \check{\mathbf{b}} = \check{\mathbf{b}}, \mathbf{b} \circ \check{\mathbf{b}} = t \text{Id},$$

where  $\mathbf{s}: X \otimes X \rightarrow X \otimes X$  is the braiding. The indecomposable objects in  $\underline{\text{Rep}} O(t)$  are enumerated by partitions corresponding to primitive idempotents of the Brauer algebras  $B_r(t)$ . If  $t \notin \mathbb{Z}$ , then  $\underline{\text{Rep}} O(t)$  is semisimple. If  $t \in \mathbb{Z}$ , then the quotient of  $\underline{\text{Rep}} O(t)$  by negligible morphisms is equivalent to the semisimple categories:  $O(t)\text{-mod}$  for  $t > 0$ ,  $SP(-t)\text{-mod}$  for even negative  $t$  and  $OSP(1, 1 - 2t)\text{-mod}$  for odd negative  $t$ . The latter can be considered as the reason why  $OSP(1, 2n)$  is the only algebraic supergroup with semisimple category of finite-dimensional representations which is not a usual group. In a sense  $OSP(1, 2n)$  is a “missing term” in the series of classical reductive groups. For any  $m, n \geq 0$  such that  $t = m - 2n$ , there exists a full tensor functor  $F_{m, 2n}: \underline{\text{Rep}} O(t) \rightarrow OSP(m, 2n)\text{-mod}$ . The question of calculating the kernel of  $F_{m, 2n}$  is still open. We also do not have at the moment classification of indecomposable components of  $OSP(m, 2n)$  in the tensor powers of the standard representation.

As in the case of  $GL(m, n)$ , one can define tensor functors

$$E_{m, 2n}^k: OSP(m, 2n)\text{-mod} \longrightarrow OSP(m - 2k, 2n - 2k)\text{-mod}$$

by taking  $x$  to be an odd selfcommuting element ( $[x, x] = 0$ ) of rank  $2k$  in the standard representation. The obvious analogue of Proposition 4.2 holds. Moreover, one can construct an abelian category  $\overline{\text{Rep}} O(t)$  in the same way as  $\overline{\text{Rep}} GL(t)$ .

**Theorem 4.4.**

- (a) *The abelian category  $\overline{\text{Rep}} O(t)$  is a symmetric rigid  $\mathbb{F}$ -linear abelian category. There exists a unique (up to isomorphism) fully faithful tensor functor  $F: \underline{\text{Rep}} O(t) \rightarrow \overline{\text{Rep}} O(t)$  such that  $F(X) = \bar{V}$ .*

- (b) Simple objects of  $\overline{\text{Rep}} O(t)$  are enumerated by partitions. All simple objects are self-dual.
- (c) Let  $L_\lambda$  denote the simple object corresponding to a partition  $\lambda$ ,  $W_{\lambda'} := F(Y_{\lambda'})$  (where  $Y_{\lambda'}$  is the corresponding indecomposable object in the Deligne's category). Then  $[W_{\lambda'} : L_\lambda] \neq 0$  implies  $|\lambda'| - |\lambda| < 0$  is even and  $\lambda \leq \lambda'$  in the dominance order for partitions. Moreover,  $[W_\lambda : L_\lambda] = 1$ .

**4.5. The case of  $P(n)$ .** For the family  $P(n)$ , we introduce a standard object  $X$  and postulate the existence of odd morphisms

$$\mathbf{b}: X \otimes X \rightarrow \mathbb{F}, \quad \check{\mathbf{b}}: \mathbb{F} \rightarrow X \otimes X$$

satisfying the relations

$$\mathbf{b} \circ \mathbf{s} = \mathbf{b}, \quad \mathbf{s} \circ \check{\mathbf{b}} = -\check{\mathbf{b}}.$$

The sign in the second relation is motivated by representation theory of  $P(n)$ . Indeed, for the standard  $P(n)$ -module  $V$  we have

$$V^* \simeq V^\Pi, \quad \text{and } S^2(V^*) \simeq \Lambda^2(V), \quad \text{Im } \check{\mathbf{b}} \in \Lambda^2(V).$$

Since  $\mathbf{s}^2 = \text{Id}$ , we must have  $\mathbf{b} \circ \check{\mathbf{b}} = 0$ . Note that  $\mathbf{b}$  and  $\check{\mathbf{b}}$  induce also odd morphisms  $B: X \rightarrow \check{X}$  and  $\check{B}: \check{X} \rightarrow X$  defined as the compositions

$$X \xrightarrow{\text{Id} \otimes \mathbf{e}} X \otimes X \otimes \check{X} \xrightarrow{\mathbf{b} \otimes \text{Id}} \check{X}$$

and

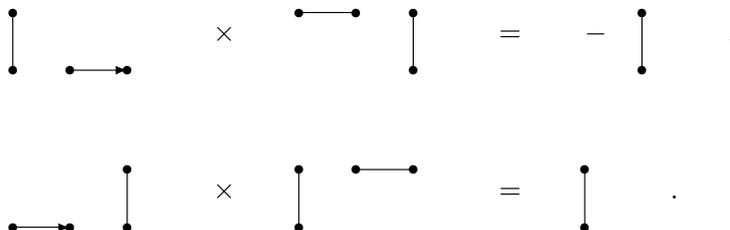
$$\check{X} \xrightarrow{\text{Id} \otimes \check{\mathbf{b}}} \check{X} \otimes X \otimes X \xrightarrow{\mathbf{c} \otimes \text{Id}} X$$

respectively. We also postulate that  $B \circ \check{B} = \check{B} \circ B = \text{Id}$ .

Define  $\text{Rep}_0 P$  to be the category whose objects are non-negative integers, corresponding to  $X^{\otimes r}$ . The  $\mathbb{F}$ -vector space  $\text{Hom}(X^{\otimes r}, X^{\otimes s})$  is the formal span of the set  $D(s, r)$  of all Brauer diagrams with  $r$  nodes in the upper row and  $s$  nodes in the lower row. In particular,  $\text{Hom}(X^{\otimes r}, X^{\otimes s}) \neq 0$  implies  $r + s$  is even. The multiplication of two diagrams  $D_1 \in D(s, r)$  and  $D_2 \in D(r, q)$  is a modification of the Brauer product  $D$  for  $t = 0$  by additional sign. Identifying the lower nodes of  $D_1$  with upper nodes of  $D_2$  we obtain the three-row diagram  $C$ . To determine the sign equip the horizontal strings in the middle row of  $C$  obtained from the bottom row of  $D_1$  with right arrows. Set

$$D_1 \circ D_2 = (-1)^{\sum_\gamma p(\gamma)} D,$$

where  $\gamma$  is a path in  $C$  going from bottom to top or from bottom to bottom from left to right and  $p(\gamma)$  is the number of arrows on the path  $\gamma$  in the “wrong” direction. Examples:



Non-negative integers together with diagrams give us a skeleton category  $\text{Rep}_0 P$ . The category  $\underline{\text{Rep}} P$  is defined as the Karoubian envelope of the additive envelope of  $\text{Rep}_0 P$ .

**Theorem 4.5.** *The category  $\underline{\text{Rep}} P$  is an  $\mathbb{F}$ -linear Karoubian symmetric tensor category. Indecomposable objects of  $\underline{\text{Rep}} P$  are enumerated by partitions up to the change of parity. If  $Y_\lambda$  is an indecomposable object corresponding to a partition  $\lambda$ , then  $Y_\lambda^*$  is isomorphic to  $Y_{\lambda'}$  where  $\lambda'$  is the conjugate of  $\lambda$ . For any  $n \geq 1$  there exists a full tensor functor  $F_n : \underline{\text{Rep}} P \rightarrow P(n)\text{-mod}$  and  $F_n(Y_\lambda)$  is an indecomposable summand in the tensor algebra of the standard representation of  $P(n)$ .*

It is interesting to describe all partitions  $\lambda$  such that  $F_n(Y_\lambda) \neq 0$ . We conjecture that  $\text{sdim } Y_\lambda = 0$  for any non-empty partition  $\lambda$ . Let us also mention that, as in the general linear and orthosymplectic cases, it is possible to construct an abelian tensor category  $\overline{\text{Rep}} P$  and a fully faithful tensor functor  $F : \underline{\text{Rep}} P \rightarrow \overline{\text{Rep}} P$ . Furthermore, the obvious analogue of Theorem 4.4 holds.

Let us finally mention that the formalism of  $Q$ -diagrams developed in [31] seems be just one step away from defining an analogue of the universal tensor category for  $Q(n)$ .

## 5. The category $G\text{-mod}$ , blocks and translation functors

**5.1. General properties of the category  $G\text{-mod}$ .** Now let us assume that  $G$  is an affine algebraic supergroup with reductive  $G_0$ . That implies that  $\mathfrak{g}_0$  is reductive and  $\mathfrak{g}_1$  is a semisimple  $\mathfrak{g}_0$ -module.

For any  $M \in G_0\text{-mod}$  define the induced and coinduced modules by

$$I(M) := U(\mathfrak{g}) \otimes_{U(\mathfrak{g}_0)} M, \quad C(M) := \text{Hom}_{U(\mathfrak{g}_0)}(U(\mathfrak{g}), M).$$

By the supersversion of the Poincaré–Birkhoff–Witt theorem, [34],[44], we have an isomorphism of vector spaces  $U(\mathfrak{g}) \simeq S(\mathfrak{g}_0) \otimes \Lambda(\mathfrak{g}_1)$ , since  $\Lambda(\mathfrak{g}_1)$  in the classical sense coincides with  $S(\mathfrak{g}_1)$  in the “super” sense. Therefore, both  $I(M)$  and  $C(M)$  are finite-dimensional  $\mathfrak{g}$ -modules and they are  $(\mathfrak{g}, G_0)$ -modules. Since  $G_0\text{-mod}$  is semisimple,  $C(M)$  is injective and  $I(M)$  is projective in  $G\text{-mod}$ . Furthermore, we have the following reciprocity, see [51]

$$I(M) = C(M \otimes T),$$

where  $T$  is the one-dimensional representation of  $G_0$  in  $\Lambda^{\text{top}}(\mathfrak{g}_1)$ . It is also shown in [51] that for reductive  $G_0$  the representation  $T$  extends uniquely to a representation of  $G$ . General properties of  $G\text{-mod}$  can be summarized in the following

**Theorem 5.1** ([51]).

- (a) *Every object in  $G\text{-mod}$  has a projective cover and an injective hull.*
- (b) *An object in  $G\text{-mod}$  is injective if and only if it is projective.*
- (c) *If  $L \in G\text{-mod}$  is simple then the projective cover of  $L$  is isomorphic to the injective hull of  $L \otimes T$ .*
- (d) *If  $M \in G\text{-mod}$  has a finite projective resolution, then it is projective.*

Finally, the regular representation of  $G$  is induced from the regular representation of  $G_0$ . Since  $G_0$  is reductive, the structure of its regular representation is well-known. Hence using Frobenius reciprocity we can easily describe the regular representation of  $G$ .

**Theorem 5.2** ([51]). *The regular representation of  $G$  has a decomposition*

$$\mathbb{F}[G] = \bigoplus I_L \otimes \hat{L},$$

where the summation is taken over all non-isomorphic simple  $L \in G\text{-mod}$ ,  $I_L$  is the indecomposable injective hull of  $L$  and  $\hat{L}$  is the trivial  $G$ -module isomorphic to  $L$  as a vector superspace.

**5.2. Highest weight theory.** Here we briefly recall the highest weight theory developed in [34]. Let  $\mathfrak{h}_0 \subset \mathfrak{g}_0$  be a Cartan subalgebra and  $\mathfrak{h}$  denotes the centralizer of  $\mathfrak{h}_0$  in  $\mathfrak{g}$ . Then  $\mathfrak{h} = \mathfrak{h}_0 \oplus \mathfrak{h}_1$  and  $[\mathfrak{h}_0, \mathfrak{h}] = 0$ . The subalgebra  $\mathfrak{h}$  is called a *Cartan subalgebra* of  $\mathfrak{g}$  and by construction all Cartan subalgebras of  $\mathfrak{g}$  are conjugate by the adjoint action of  $G_0$ . Since  $\mathfrak{g}_1$  is a semisimple  $\mathfrak{g}_0$ -module,  $\mathfrak{g}$  has a *root decomposition*

$$\mathfrak{g} = \mathfrak{h} \oplus \bigoplus_{\alpha \in \Delta} \mathfrak{g}_\alpha,$$

where  $\Delta \subset \mathfrak{h}_0^*$  is the set of roots.

Let us look at the root decomposition for the four classical superalgebras discussed in the previous section. If  $\mathfrak{g} = \mathfrak{gl}(m, n)$ ,  $\mathfrak{osp}(m, 2n)$  or  $\mathfrak{p}(n)$ , then  $\mathfrak{h} = \mathfrak{h}_0$  and  $\dim \mathfrak{g}_\alpha = (0|1)$  or  $(1|0)$  for each root  $\alpha \in \Delta$ . Therefore the set of roots  $\Delta$  is naturally equipped with parity  $\Delta = \Delta_0 \cup \Delta_1$  depending on the parity of  $\mathfrak{g}_\alpha$ .

The case  $\mathfrak{g} = \mathfrak{q}(n)$  is different. Here  $\mathfrak{h}$  consists of diagonal matrices in even and odd blocks (in the matrix realization given in the previous section). In this case  $\dim \mathfrak{g}_\alpha = (1|1)$  for every  $\alpha \in \Delta$  and the roots themselves are exactly as in  $\mathfrak{gl}(n)$ .

Next, we define a triangular decomposition by separating roots in positive  $\Delta^+$  and negative  $\Delta^-$  by a generic hyperplane and setting

$$\mathfrak{g} = \mathfrak{n}^- \oplus \mathfrak{h} \oplus \mathfrak{n}^+, \quad \mathfrak{n}^\pm = \bigoplus_{\alpha \in \Delta^\pm} \mathfrak{g}_\alpha.$$

The associated Borel subalgebra  $\mathfrak{b}$  is by definition  $\mathfrak{h} \oplus \mathfrak{n}^+$ . Let us mention that, in contrast with the classical case, not all Borel subalgebras in  $\mathfrak{g}$  are conjugate. There is a way to describe all Borel subalgebras using odd reflections and Weyl groupoid, see [52]. This fact plays an important role in the representation theory of superalgebras.

For any  $\lambda \in \mathfrak{h}_0^*$  there exists a unique simple  $\mathfrak{h}$ -module  $C_\lambda$  such that any  $h \in \mathfrak{h}_0$  acts as scalar multiplication by  $\lambda(h)$ . (If  $\mathfrak{h} = \mathfrak{h}_0$ ,  $C_\lambda$  is one-dimensional.) Define a  $\mathfrak{b}$ -module structure on  $C_\lambda$  be setting  $\mathfrak{n}^+ C_\lambda = 0$  and the Verma module  $M_\lambda := U(\mathfrak{g}) \otimes_{U(\mathfrak{b})} C_\lambda$ . As in the classical highest weight theory,  $M_\lambda$  has a unique simple quotient  $L_\lambda$ . Let  $\Lambda$  be the weight lattice of  $G_0$ . We say that  $\lambda \in \Lambda$  is *dominant* if  $L_\lambda$  is finite-dimensional. We denote by  $\Lambda^+$  the set of all dominant weights.

**Theorem 5.3.** *Let  $G_0$  be connected. Every simple object in  $G\text{-mod}$  is isomorphic to  $L_\lambda$ . Two simple objects  $L_\lambda$  and  $L_\mu$  are isomorphic if and only if  $\mu = \lambda$ .*

The description of the set of dominant weights is not as straightforward as in the classical case. It depends on the choice of a triangular decomposition. For one specific choice of a triangular decomposition it was done in [34]. Using odd reflections one can obtain a general dominance criterion for all triangular decompositions, [52].

**5.3. The center of  $U(\mathfrak{g})$  and blocks.** Assume that  $G$  is connected. Let us denote by  $Z(\mathfrak{g})$  the center of the universal enveloping algebra  $U(\mathfrak{g})$  and let  $\check{Z}(\mathfrak{g}) := \text{Hom}(Z(\mathfrak{g}), \mathbb{F})$  denote the set of central characters. Since all  $M \in G\text{-mod}$  are finite-dimensional,  $M$  is of finite length over  $Z(\mathfrak{g})$ . Therefore

$$G\text{-mod} = \bigoplus_{\chi \in \check{Z}(\mathfrak{g})} G^\chi\text{-mod},$$

where  $G^\chi\text{-mod}$  is the subcategory of modules admitting generalized central character  $\chi$ .

Fix a triangular decomposition and define the Harish-Chandra projection

$$\text{HC}: U(\mathfrak{g}) \rightarrow U(\mathfrak{h})$$

with kernel  $n^-U(\mathfrak{g}) + U(\mathfrak{g})n^+$ . Then the restriction  $\text{HC}: Z(\mathfrak{g}) \rightarrow U(\mathfrak{h})$  is injective and the image lies in the center  $Z(\mathfrak{h})$  of  $U(\mathfrak{h})$ . If  $\mathfrak{h} = \mathfrak{h}_0$  then  $Z(\mathfrak{h}) = U(\mathfrak{h}_0) \simeq S(\mathfrak{h}_0)$ . If  $\mathfrak{g} = \mathfrak{q}(n)$  we still have  $Z(\mathfrak{h}) = U(\mathfrak{h}_0) \simeq S(\mathfrak{h}_0)$ .

Let  $R = \text{HC}(Z(\mathfrak{g}))$ . Recall that, if  $\mathfrak{g}$  is a reductive Lie algebra,  $R$  coincides with the space of the polynomials on  $\mathfrak{h}^*$  invariant under the shifted Weyl group action.

We will briefly review the description of  $R$  for the four classical superalgebras.

If a simple Lie superalgebra  $\mathfrak{g}$  admits an invariant symmetric even form, then  $R$  can be described by two different methods. A. Sergeev, [54], obtained a description of  $R$  by generalizing the Chevalley theorem about invariant polynomials on the adjoint representation. In [32] V. Kac announced another method which uses Verma modules and Schapovalov form. The complete proof using the latter method was published by M. Gorelik in [22]. Let us now formulate the result.

The invariant symmetric form on  $\mathfrak{g}$  induces a non-degenerate symmetric form  $(\cdot, \cdot)$  on  $\mathfrak{h}^*$ . (In this case we always have  $\mathfrak{h} = \mathfrak{h}_0$ .) We call a root  $\alpha$  *isotropic* if  $(\alpha, \alpha) = 0$ . As follows from the structure theory of simple Lie superalgebras any isotropic root is odd. Let  $W$  denote the Weyl group of  $G_0$  and

$$\rho := \frac{1}{2} \sum_{\alpha \in \Delta_0^+} \alpha - \frac{1}{2} \sum_{\alpha \in \Delta_1^+} \alpha.$$

Define the shifted action  $\cdot$  of  $W$  on  $\mathfrak{h}^*$  by  $w \cdot \mu = w(\mu + \rho) - \rho$ . Finally we identify  $S(\mathfrak{h})$  with the ring of polynomial functions  $\mathbb{F}[\mathfrak{h}^*]$ .

**Theorem 5.4.** *Let  $\mathfrak{g} = \mathfrak{gl}(m, n), \mathfrak{osp}(m, 2n), D(2, 1; a), G_3$  or  $F_4$ . A polynomial  $f \in \mathbb{F}[\mathfrak{h}^*]$  belongs to  $R$  if and only if the following two conditions hold.*

- For any  $w \in W$  and  $\mu \in \mathfrak{h}^*$ ,  $f(w \cdot \mu) = f(\mu)$ ;
- If  $\alpha$  is an isotropic root and  $(\mu + \rho, \alpha) = 0$ , then  $f(\mu + t\alpha) = f(\mu)$  for any  $t \in \mathbb{F}$ .

As a corollary,  $Z(\mathfrak{g})$  is not Noetherian if  $\mathfrak{g}$  has at least one isotropic root.

As follows from the definition of HC,  $L(\lambda)$  and  $L(\mu)$  admit the same central character if and only if  $f(\lambda) = f(\mu)$  for any  $f \in R$ .

For any  $\lambda \in \mathfrak{h}^*$  the *degree of atypicality* of  $\lambda$  (notation  $\sharp\lambda$ ) is the maximal number of mutually orthogonal and linearly independent isotropic roots  $\alpha$  such that  $(\lambda + \rho, \alpha) = 0$ . A weight  $\lambda$  is called *typical* if  $\sharp\lambda = 0$ . The *defect* of  $\mathfrak{g}$  is the maximal number of mutually orthogonal and linearly independent isotropic roots. For instance, the defect of  $\mathfrak{gl}(m, n)$  is equal to  $\min(m, n)$  and the defect of  $\mathfrak{osp}(m, 2n)$  equals  $\min(\lfloor \frac{m}{2} \rfloor, n)$ .

By Theorem 5.4, if  $f(\lambda) = f(\mu)$  for any  $f \in R$ , then  $\sharp\lambda = \sharp\mu$ . Hence for any central character  $\chi \in \check{Z}(\mathfrak{g})$  the degree of atypicality  $\sharp\chi$  is well defined.

**Theorem 5.5.** *Assume that  $G = GL(m, n)$  or  $SOSP(m, 2n)$ ,  $\chi \in \check{Z}(\mathfrak{g})$  and  $G^X\text{-mod}$  is not empty. Let  $\chi_0$  denote the central character of the trivial module.*

- (a)  $G^X\text{-mod}$  is a (indecomposable) block in the category  $G\text{-mod}$ .
- (b)  $G^X\text{-mod}$  is semisimple if and only if  $\sharp\chi = 0$ . In this case  $G^X$  has one up to isomorphism simple object.
- (c) Let  $G = GL(m, n)$  and  $\sharp\chi = k$ . Then  $G^X\text{-mod}$  is equivalent to  $GL(k, k)^{X_0}\text{-mod}$ .
- (d) Let  $G = SOSP(2m+1, 2n)$  and  $\sharp\chi = k$ . Then  $G^X\text{-mod}$  is equivalent to  $SOSP(2k+1, 2k)^{X_0}\text{-mod}$ .
- (e) Let  $G = SOSP(2m, 2n)$  and  $\sharp\chi = k$ . Then  $G^X\text{-mod}$  is equivalent to either  $SOSP(2k, 2k)^{X_0}\text{-mod}$  or to  $SOSP(2k+2, 2k)^{X_0}\text{-mod}$ .

The proof of (b) is contained in [35], (c), (d) and (e) is proven in [28]. Finally, (a) follows from combinatorial calculations in [49] and [28].

The situation with center  $Z(\mathfrak{g})$  and blocks for the strange Lie superalgebras is different.

**Theorem 5.6** ([21]). *If  $\mathfrak{g} = \mathfrak{p}(n)$ , then  $Z(\mathfrak{g}) = \mathbb{F}$ .*

As a result the category  $P(n)\text{-mod}$  contains only two blocks if we identify  $M$  with  $M^\Pi$ .

For  $\mathfrak{g} = \mathfrak{q}(n)$  the center  $Z(\mathfrak{g})$  was described by A. Sergeev, see [56]. Let  $x_1, \dots, x_n$  be the basis in  $\mathfrak{h}_0$  consisting of elementary diagonal matrices, and  $\varepsilon_1, \dots, \varepsilon_n$  denote the dual basis in  $\mathfrak{h}_0^*$ . The Weyl group coincides with the group of permutations of the elements in this basis.

**Theorem 5.7** ([56]). *Let  $\mathfrak{g} = \mathfrak{q}(n)$ . Then the image  $R = \text{HC}(Z(\mathfrak{g}))$  in  $\mathbb{F}[x_1, \dots, x_n]$  is generated by  $x_1^m + \dots + x_n^m$  for all odd  $m \in \mathbb{N}$ . Equivalently,  $f$  lies in  $R$  if and only if  $f$  is a symmetric polynomial and for all  $i < j$  and  $t, s \in \mathbb{F}$*

$$\begin{aligned} f(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_{j-1}, -t, x_{j+1}, \dots, x_n) = \\ f(x_1, \dots, x_{i-1}, s, x_{i+1}, \dots, x_{j-1}, -s, x_{j+1}, \dots, x_n). \end{aligned} \quad (5.1)$$

We proceed to classification of blocks  $Q(n)^X\text{-mod}$  in the category  $Q(n)\text{-mod}$ . The weight lattice  $\Lambda = \mathbb{Z}\varepsilon_1 \oplus \dots \oplus \mathbb{Z}\varepsilon_n$ . The set  $\Lambda^+$  of dominant weights is

$$\{\mu = a_1\varepsilon_1 + \dots + a_n\varepsilon_n \mid a_i \in \mathbb{Z}, a_i > a_{i+1} \text{ or } a_i = a_{i+1} = 0 \text{ for all } i = 1, \dots, n-1\}.$$

Define the degree of atypicality  $\sharp\mu$  as the maximal number  $k$  of pairs  $(i_1 < j_1), \dots, (i_k < j_k)$  such that  $x_{i_s} + x_{j_s} = 0$  and all  $i_1, \dots, i_k, j_1, \dots, j_k$  are distinct. Set  $h(\mu)$  be the number of  $i$  such that  $a_i \neq 0$ . Let  $p(\mu)$  be the parity of  $h(\mu)$ . Theorem 5.7 implies that if  $L_\nu$  and  $L_\mu$  admit the same central character, then  $\sharp\nu = \sharp\mu$  and  $p(\nu) = p(\mu)$ . Thus,  $p(\chi)$  and  $\sharp\chi$  are well-defined.

**Theorem 5.8.** *Let  $G = Q(n)$ ,  $\bar{n} = 0, 1$  be the parity of  $n$  and*

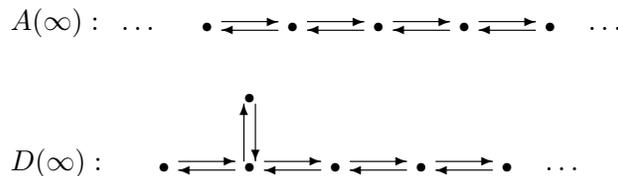
$$k := \begin{cases} 2\sharp\chi & \text{if } p(\chi) = \bar{n} = 0, \\ 2\sharp\chi + 2 & \text{if } p(\chi) = 1, \bar{n} = 0, \\ 2\sharp\chi + 1 & \text{if } p(\chi) = 0, \bar{n} = 1, \\ 2\sharp\chi + 1 & \text{if } p(\chi) = \bar{n} = 1. \end{cases} \tag{5.2}$$

- (a)  $G^X\text{-mod}$  is a block in the category  $G\text{-mod}$ .
- (b) If  $p(\chi) = 0$ , then  $G^X\text{-mod}$  is equivalent to the block in the category  $Q(k)\text{-mod}$  containing the trivial representation.
- (c) If  $p(\chi) = 1$ , then  $G^X\text{-mod}$  is equivalent to the block in the category  $Q(k)\text{-mod}$  containing the standard representation.
- (d) The simple modules of  $G^X\text{-mod}$  are of  $GL$  (resp.  $Q$ ) type if  $p(\chi) = 0$  (resp.  $p(\chi) = 1$ ).

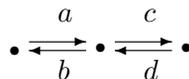
**5.4. Quivers and blocks.** In [25] J. Germoni initiated a detailed study of blocks  $G^X\text{-mod}$  in the case when  $g$  has an even invariant form. He considered the following problems:

- classify tame<sup>4</sup> blocks  $G^X\text{-mod}$ ;
- describe tame blocks in terms of quivers with relations.

Note that the case  $\sharp\chi = 0$  is trivial since  $G^X\text{-mod}$  has only one simple object. It is also not hard to check that  $G^X\text{-mod}$  is wild if  $\sharp\chi \geq 2$ . Germoni conjectured that all blocks of atypicality 1 are tame and verified his conjecture for  $GL(1, 1)$  and  $D(2, 1, a)$ . C. Gruson in [26] described the only atypical block of  $SOSP(3, 2)$  and L. Martirosyan in her thesis described all atypical blocks of  $G_3$  and  $F_4$ . Those results together with Theorem 5.5 imply Germoni’s conjecture for  $GL(m, n)$ ,  $SOSP(m, 2n)$  and all exceptional supergroups. It is interesting that blocks of atypicality 1 are equivalent to one of the following two quivers:



with relations  $ba = cd, ac = 0 = db$  for any subquiver isomorphic to:



**5.5. Translation functors and weight diagrams.** The proof of Theorems 5.5 and 5.8 uses so called *translation functors*. Let  $G = GL(m, n)$ ,  $SOSP(m, 2n)$  or  $Q(n)$  and  $V$  denote the standard module. For any  $M \in G\text{-mod}$  denote by  $M^X$  its projection onto the block  $G^X\text{-mod}$ . Define translation functors  $T_{\chi, \zeta} : G^X\text{-mod} \rightarrow G^\zeta\text{-mod}$  and  $T_{\zeta, \chi}^* : G^\zeta\text{-mod} \rightarrow G^X\text{-mod}$  by setting

$$T_{\chi, \zeta}(M) := (M \otimes V)^\chi, \quad T_{\zeta, \chi}^*(M) := (M \otimes V^*)^\zeta.$$

<sup>4</sup>Here we mean tame wild dichotomy in representation theory



- the simple  $Q(n)$ -modules  $L_\lambda$  and  $L_\mu$  belong to the same block if and only if  $\bar{f}_\lambda = \bar{f}_\mu$ .

**Remark 5.9.** Let  $\sigma$  denote the automorphism of  $\mathfrak{q}(n)$  given by

$$\sigma \begin{pmatrix} A & B \\ B & A \end{pmatrix} = \begin{pmatrix} -A^t & iB^t \\ iB^t & -A^t \end{pmatrix},$$

and  $L_\nu \simeq L_\mu^\sigma$ , then  $f_\nu$  is obtained from  $f_\mu$  by switching  $<$  and  $>$ .

Thus, we have a bijection between diagrams  $f$  without  $\times$ -s, such that  $s(f) \leq n$ , and the blocks  $Q(n)^\chi$ -mod. For any central character  $\chi$  denote by  $\bar{f}_\chi$  the corresponding diagram. Next, we use the following standard result.

**Lemma 5.10.** *An  $\mathfrak{n}^+$ -invariant weight vector in  $L_\mu \otimes V$  (resp.  $L_\mu \otimes V^*$ ) has weight  $\mu + \varepsilon_i$  (resp.  $\mu - \varepsilon_i$ ) for some  $i = 1, \dots, n$ .*

Note the diagram  $f_{\mu+\varepsilon_i}$  is obtained from that of  $f_\mu$  by moving  $>$  to the right or  $<$  to the left. We consider here  $>$  and  $<$  as “halves” of a  $\times$ ; additionally, we consider “an extended line”  $t \geq 0$  with “a pool” of extra symbols at  $t = 0$ , moving to the left from position  $t = 1$  is equivalent to removing and moving  $>$  to the right includes the possibility to add  $>$  at 1. For instance, if  $\mu = \varepsilon_1 - \varepsilon_n$ , then  $f_{\mu+\varepsilon_n}$  is obtained from  $f_\mu$  by removing  $<$  from the  $\times$  at position 1. The second half  $>$  of the  $\times$  remains at 1. Similarly the diagram  $f_{\nu-\varepsilon_i}$  is obtained from that of  $f_\nu$  by moving  $<$  to the right or  $>$  to the left. If  $\nu = \varepsilon_1$ , then  $f_\mu = f_{\nu-\varepsilon_n}$  is obtained from  $f_\nu$  by adding  $<$  at position 1, joined with  $>$  it makes  $\times$  in  $f_\mu$ . After some technical work using Lemma 5.10 one obtains the following.

**Lemma 5.11.** *Let  $\chi, \zeta$  be two central characters such that  $\bar{f}_\zeta$  is obtained from  $\bar{f}_\chi$  by moving  $>$  to the right or  $<$  to the left. Assume also  $h(\chi) = h(\zeta)$ . Then for any  $L_\mu \in Q(n)^\chi$ -mod there exists a unique pair  $L_\nu, L_\nu^\Pi \in Q(n)^\zeta$ -mod such that*

$$T_{\chi,\zeta}(L_\mu) \simeq L_\nu \oplus L_\nu^\Pi \text{ and } T_{\chi,\zeta}^*(L_\nu) \simeq L_\mu \oplus L_\mu^\Pi.$$

The analogous result for  $GL(m, n)$  and  $SOSP(m, n)$  is easier:

$$T_{\chi,\zeta}(L_\mu) \simeq L_\nu \text{ and } T_{\chi,\zeta}^*(L_\nu) \simeq L_\mu.$$

By standard abstract nonsense the pair of functors  $T_{\chi,\zeta}, T_{\chi,\zeta}^*$  establish an equivalence between  $G^\chi$ -mod and  $G^\zeta$ -mod. In the case  $G = Q(n)$  we can't directly use  $T_{\chi,\zeta}$  and  $T_{\chi,\zeta}^*$  since they “double” simple modules.

Let  $M \in G^\chi$ -mod, and  $N := T_{\chi,\zeta}(M)$ . Recall that  $\mathcal{Q}(1)$  is the associative algebra of  $\mathfrak{q}(1)$ . Note that  $N$  has an action of  $\mathcal{Q}(1)$  induced by the  $\mathcal{Q}(1)$ -action on  $V$  generated by  $\Pi$ . This action commutes with the  $U(\mathfrak{g})$ -action. Hence  $M$  is in fact a  $U(\mathfrak{g}) \otimes \mathcal{Q}(1)$ -module. If  $M$  is simple, then Lemma 5.11 implies that  $N \simeq N' \otimes \mathcal{Q}(1)$  for some simple  $\mathfrak{g}$ -module  $N'$ . One can prove by induction on the length of  $M$  that the same holds for any  $M \in G^\chi$ -mod.<sup>5</sup> If  $\text{Hom}^0$  denotes the even part of  $\text{Hom}$ , then  $\text{Hom}_{\mathcal{Q}(1)}^0(\mathcal{Q}(1), N)$  has a well-defined  $\mathfrak{g}$ -module structure and is isomorphic to  $N'$ . We define new functors  $S_{\chi,\zeta}: G^\chi$ -mod  $\rightarrow G^\zeta$ -mod and  $S_{\zeta,\chi}^*: G^\zeta$ -mod  $\rightarrow G^\chi$ -mod by

$$S_{\chi,\zeta}(M) := \text{Hom}_{\mathcal{Q}(1)}^0(\mathcal{Q}(1), T_{\chi,\zeta}(M)) \text{ and } S_{\zeta,\chi}^*(M) := \text{Hom}_{\mathcal{Q}(1)}^0(\mathcal{Q}(1), T_{\zeta,\chi}^*(M)).$$

<sup>5</sup>If  $h(\chi) \neq h(\zeta)$ , this may not be true. Consider for example the block of the trivial module and the block of the standard module.

**Lemma 5.12.** *Let  $\chi, \zeta$  satisfy the conditions of Lemma 5.11. Then  $S_{\chi, \zeta}$  and  $S_{\zeta, \chi}^*$  establish an equivalence between  $G^\chi\text{-mod}$  and  $G^\zeta\text{-mod}$ .*

To finish the proof of Theorem 5.8 we will use translation functors to move all core symbols to the right of all  $\times$ -s.

Let  $t \gg 1$ . By Remark 5.9 we may assume that if  $\bar{f}_\chi$  is non-empty, then its leftmost symbol is  $>$ . Choose a central character  $\chi(t)$  whose core diagram is obtained from  $\bar{f}_\chi$  by moving all symbols to the right of  $t$  if  $p(\chi) = 0$ . If  $p(\chi) = 1$  we move  $>$  to the position 1 and all other symbols to the right of  $t$ . Using Lemma 5.12 one can construct a directed system of functors

$$\Phi^{t,s}: G^{\chi(t)}\text{-mod} \rightarrow G^{\chi(s)}\text{-mod}, \quad t > s \geq 0, \quad \chi(0) = \chi,$$

such that each functor is an equivalence of categories.

Given a weight  $\mu \in P^+$  and the corresponding simple module  $L_\mu$ , call them  $t$ -admissible if  $f_\lambda(s) = \times$  implies  $s \leq t$  (i.e., all crosses are at  $t$  or to the left of  $t$ ). Let  $\mathcal{F}_t(G^{\chi(t)}\text{-mod})$  be the full subcategory of  $G^{\chi(t)}\text{-mod}$  consisting of modules with  $t$ -admissible simple subquotients. Let  $k$  be as in Theorem 5.8,  $p$  be the number of  $>$  and  $r$  be number of  $<$  in  $\bar{f}_{\chi(t)}$  to the right of  $t$ . Then  $k + p + r = n$ . Let  $\kappa$  be the central character of the trivial (resp. standard) module if  $p(\chi) = 0$  (resp.  $p(\chi) = 1$ ). The next step is to define functors

$$\Gamma^t: \mathcal{F}_t(Q(k)^\kappa\text{-mod}) \rightarrow \mathcal{F}_t(G^{\chi(t)}\text{-mod}), \quad R^t: \mathcal{F}_t(G^{\chi(t)}\text{-mod}) \rightarrow \mathcal{F}_t(Q(k)^\kappa\text{-mod}).$$

Let  $\mathfrak{p}$  be the parabolic subalgebra of  $\mathfrak{g}$  defined by

$$\mathfrak{p} := \mathfrak{h} \oplus \bigoplus_{1 \leq i < j \leq n} \mathfrak{g}_{\varepsilon_i - \varepsilon_j} \oplus \bigoplus_{p < i < j \leq n-r} \mathfrak{g}_{\varepsilon_j - \varepsilon_i}.$$

Its Levi subalgebra  $\mathfrak{l} \subset \mathfrak{p}$  is isomorphic to  $\mathfrak{q}(k) \oplus \mathfrak{h}'$ , where  $\mathfrak{h}' \subset \mathfrak{h}$  is the centralizer of  $\mathfrak{q}(k)$  in  $\mathfrak{h}$ . Let

$$\mathfrak{m} := \bigoplus_{i \leq p < j \leq n} \mathfrak{g}_{\varepsilon_i - \varepsilon_j} \oplus \bigoplus_{p < i \leq n-r < j \leq n} \mathfrak{g}_{\varepsilon_i - \varepsilon_j}$$

be the nilpotent radical of  $\mathfrak{p}$ . The part of the weight diagram of  $\chi(t)$  lying to the right of  $t$  induces in a natural way a weight of  $\mathfrak{h}'$  and hence a simple  $\mathfrak{h}'$ -module  $C_t$ . For any  $N \in \mathcal{F}_t(Q(k)^\kappa\text{-mod})$  we define  $\Gamma^t(M)$  to be the maximal finite-dimensional quotient of the parabolically induced module  $U(\mathfrak{g}) \otimes_{U(\mathfrak{p})} (N \otimes C_t)$ . If  $M \in \mathcal{F}_t(G^{\chi(t)}\text{-mod})$  we define  $R^t(M) := M^{\mathfrak{m}}$ .

**Lemma 5.13.** *The functors  $\Gamma^t: \mathcal{F}_t(Q(k)^\kappa\text{-mod}) \rightarrow \mathcal{F}_t(G^{\chi(t)}\text{-mod})$  and  $R^t: \mathcal{F}_t(G^{\chi(t)}\text{-mod}) \rightarrow \mathcal{F}_t(Q(k)^\kappa\text{-mod})$  establish equivalence of categories.*

The proof of Lemma 5.13 requires geometric induction, which we discuss in the next section.

Using combinatorics of weight diagrams it is possible to show that

$$G^\chi\text{-mod} = \bigcup_{t \gg 0} \Phi^{t,0}(\mathcal{F}_t(G^{\chi(t)}\text{-mod})) = \bigcup_{t \gg 0} \Phi^{t,0} \Gamma^t(\mathcal{F}_t(Q(k)^\kappa\text{-mod})). \quad (5.4)$$

This implies Theorem 5.8.

Weight diagrams can be used to determine what happens with simple modules under the equivalence  $G^X\text{-mod} \rightarrow Q(k)^\kappa\text{-mod}$ . For instance, let  $n = 5$  and  $\mu = 3\varepsilon_1 + \varepsilon_2 - 2v_4 - 3\varepsilon_5$ .

Then  $f_\mu = \triangleright \leftarrow \times \circ \circ \circ \text{---}$

We have  $p(\mu) = 0, \sharp\mu = 1, k = 3$ . Move both core symbols to the right using Lemma 5.12.

First, we get  $\triangleright \times \leftarrow \circ \circ \circ \text{---}$

and then  $\times \triangleright \leftarrow \circ \circ \circ \text{---}$

Finally, Lemma 5.13 allows to remove both core symbols and get  $\times \circ \circ \circ \circ \text{---}$   
 This is the diagram of the adjoint representation. Hence, under equivalences of Theorem 5.8,  $L_\mu$  is moved to the adjoint representation of  $\mathfrak{q}(3)$ .

### 6. Geometric methods and categorification

**6.1. The associated variety.** Let  $\mathfrak{g}$  be a finite-dimensional superalgebra and  $X \subset \mathfrak{g}_1$  denote the cone of self-commuting elements, i.e.,

$$X = \{x \in \mathfrak{g}_1 \mid [x, x] = 0\}.$$

For any  $\mathfrak{g}$ -module  $M$  and  $x \in X$  we have  $x^2M = 0$ . We define

$$M_x = \text{Ker } x / \text{Im } x \text{ and } X_M = \{x \in X \mid M_x \neq 0\}.$$

It is clear that  $X_M$  is a Zariski closed conical subvariety of  $X$ . It is called the *associated variety* of  $M$ . Furthermore, if  $M \in G\text{-mod}$ , then  $X_M$  is  $G_0$ -invariant.

In particular, if  $M \simeq \mathfrak{g}$  is the adjoint representation, then  $\mathfrak{g}_x = \mathfrak{g}^x / [x, \mathfrak{g}]$ , where  $\mathfrak{g}^x$  is the centralizer of  $x$  in  $\mathfrak{g}$ . One can easily check that  $[x, \mathfrak{g}]$  is an ideal in  $\mathfrak{g}^x$ . Hence  $\mathfrak{g}_x$  is a Lie superalgebra, by  $G_x$  we denote the corresponding algebraic supergroup. For any  $m \in M, x \in X, g \in \mathfrak{g}$  such that  $xm = 0$  we have  $[x, g]m = xgm \in \text{Im } x$ . Therefore  $M_x$  is equipped with a canonical  $\mathfrak{g}_x$ -module structure. Thus, we have constructed a functor  $F_x$  from the category of  $\mathfrak{g}$ -modules to the category of  $\mathfrak{g}_x$ -modules. The following properties of this functor are straightforward.

- $F_x$  is an additive tensor functor, in particular, we have a canonical isomorphism

$$F_x(M \otimes N) \simeq F_x(M) \otimes F_x(N);$$

- $\text{sdim } F_x(M) = \text{sdim } M$ ;
- If  $M \in G\text{-mod}$ , then  $F_x(M) \in G_x\text{-mod}$ ;
- If  $G_0$  is reductive and  $M \in G\text{-mod}$  is projective, then  $F_x(M) = 0$  for any  $x \in X, x \neq 0$ . Hence  $X_M = \{0\}$ ;
- $X_{M \oplus N} = X_M \cup X_N$  and  $X_{M \otimes N} = X_M \cap X_N$ .

**Remark 6.1.** Note that the functor  $F_x$  for a specific  $x$  was already used for the construction of  $\overline{\text{Rep}} GL(t)$  in Section 4.3.

If we apply  $F_x$  to  $U(\mathfrak{g})$ , considered as the adjoint  $\mathfrak{g}$ -module, we obtain

$$F_x(U(\mathfrak{g})) \simeq U(\mathfrak{g}_x) = U(\mathfrak{g})^x/[x, U(\mathfrak{g})], \tag{6.1}$$

where  $U(\mathfrak{g})^x$  is the set of ad  $x$ -invariants. Therefore we have a homomorphism  $\phi_x : Z(\mathfrak{g}) \rightarrow Z(\mathfrak{g}_x)$  defined as the composition

$$Z(\mathfrak{g}) \rightarrow U(\mathfrak{g})^x \rightarrow U(\mathfrak{g})^x/[x, U(\mathfrak{g})] = U(\mathfrak{g}_x).$$

Let  $\check{\phi}_x : \check{Z}(\mathfrak{g}_x) \rightarrow \check{Z}(\mathfrak{g})$  denote the dual map. For a set  $A \subset \check{Z}(\mathfrak{g}_x)$  let  $I(A)$  denote the annihilator of  $A$  in  $Z(\mathfrak{g}_x)$ . Then (6.1) implies that if a  $\mathfrak{g}$ -module  $M$  admits central character  $\chi \in \check{Z}(\mathfrak{g})$  then  $I(\check{\phi}_x^{-1}(\chi))M_x = 0$ . In particular, if  $\chi \notin \text{Im } \check{\phi}_x$ , then  $M_x = 0$ . Therefore it seems important to study the fibers and the image of  $\check{\phi}_x$ . It happens that for the classical and exceptional superalgebras the fibers are discrete. Let us explain details for the classical supergroups  $GL(m, n)$ ,  $SOSP(m, 2n)$  and  $Q(n)$ . (The case of  $P(n)$  is not included since the center of the universal enveloping algebra is trivial.)

First, we summarize the results about the geometry of the self-commuting cone. The study of the self-commuting cone was initiated in [27]. It was motivated by applications to cohomology theory of Lie superalgebras.

**Theorem 6.2** ([19, 27]). *Let  $G = GL(m, n)$ ,  $SOSP(m, 2n)$  or  $Q(n)$ ,  $X \subset \mathfrak{g}_1$  be the self-commuting cone and  $d$  denote the defect of  $\mathfrak{g}$ .*

- (a)  $X$  has finitely many  $G_0$  orbits.
- (b) There is a stratification

$$X = \bigsqcup_{k=0}^d X_k, \quad \bar{X}_k = \bigsqcup_{i=0}^k X_i,$$

such that each  $X_k$  is a union of  $G_0$ -orbits of the same dimension.

- (c) If  $x, y \in X_k$ , then  $\mathfrak{g}_x \simeq \mathfrak{g}_y$ . Moreover, if  $\mathfrak{g} = \mathfrak{gl}(m, n)$ , then  $\mathfrak{g}_x = \mathfrak{gl}(m - k, n - k)$ , if  $\mathfrak{g} = \mathfrak{osp}(m, 2n)$ , then  $\mathfrak{g}_x = \mathfrak{osp}(m - 2k, n - 2k)$ , and if  $\mathfrak{g} = \mathfrak{q}(n)$ , then  $\mathfrak{g}_x = \mathfrak{q}(n - 2k)$ .

In what follows  $d$  always denotes the defect of  $\mathfrak{g}$ .

**Theorem 6.3** ([19]). *Let  $G = GL(m, n)$ ,  $SOSP(m, 2n)$  or  $Q(n)$ ,  $x \in X_k \subset \mathfrak{g}_1$ .*

- (a) If  $\check{\phi}_x(\zeta) = \chi$ , then  $\sharp\chi = \sharp\zeta + k$ .
- (b) The fiber  $\check{\phi}_x^{-1}(\chi)$  is not empty if and only if  $\sharp\chi \geq k$ .
- (c) Let  $G = SOSP(2M, 2n)$  with  $M > n$  and  $k = d = n$ . Then  $\mathfrak{g}_x \simeq \mathfrak{so}(2M - 2n)$  and either  $\check{\phi}_x^{-1}(\chi)$  consists of one  $\varphi$ -invariant point or  $\check{\phi}_x^{-1}(\chi) = \{\zeta, \varphi(\zeta)\}$ , where  $\varphi$  is the involution of  $\mathfrak{g}_x$  induced by the symmetry of the Dynkin diagram.
- (d) In all other cases, if  $\sharp\chi \geq k$ , then the fiber  $\check{\phi}_x^{-1}(\chi)$  consists of one point.

The map  $\check{\phi}_x$  has a very simple description in terms of weight diagrams. For instance, if  $G = Q(n)$ ,  $\phi_x(\zeta) = \chi$  if and only if  $\zeta$  and  $\chi$  have the same core diagram.

Let us mention some further results and applications of the associated variety and  $F_x$ .

**Theorem 6.4.** *Let  $G = GL(m, n)$  or  $SOSP(m, 2n)$  and  $L \in G^x\text{-mod}$  be a simple module. Then  $X_L = \bar{X}_k$ , where  $k = \sharp\chi$ .*

At the moment we do not have a proof of the analogous result for  $G = Q(n)$  although we suspect that it is true.

In [36] V. Kac and M. Wakimoto defined defect and atypicality degree and made several conjectures relating those numbers with characters and dimensions of irreducible modules. One of them was proven in [50].

**Theorem 6.5.** *Let  $G = GL(m, n)$  or  $SOSP(m, 2n)$  and  $L \in G^x\text{-mod}$  be a simple module. Then  $\text{sdim } L \neq 0$  if and only if  $\sharp\chi = d$ .*

The calculation in [40] of superdimensions of simple modules for  $G_3$  and  $F_4$  confirm the Kac–Wakimoto conjecture (Theorem 6.5) for these superalgebras. It is an open problem to find all irreducible representations of  $Q(n)$  and  $P(n)$  of non-zero superdimension.

Let  $L \in G^x\text{-mod}$  and  $k := \sharp\chi < d$ . It is interesting to study  $F_x(L)$  for  $x \in X_k$ , since then, by Theorem 6.3,  $F_x(L)$  lies in a typical block of  $G_x\text{-mod}$ . Hence  $F_x(L)$  equals a direct sum of several copies of the same simple  $G_x$ -module  $L_\chi$ . Note that the number  $\text{sdim Hom}_{G_x}(L_\chi, F_x(L))$  coincides with modified dimension of  $L$ , introduced in [24] and [23]. In [50] and [39] this connection was used to prove a generalized Kac–Wakimoto conjecture for  $GL(m, n)$  and  $SOSP(m, 2n)$ .

One can also relate analytic properties of the character of a finite-dimensional module, see [19], and its complexity, see [4], with the dimension of its associated variety.

**6.2. Geometric induction and Borel–Weil–Bott theory.** Let  $G$  be an algebraic supergroup and  $K$  be a closed subgroup. Then one can construct a homogeneous supervariety  $G/K$ , [41]. Then any  $M \in K\text{-mod}$  induces a  $G$ -equivariant vector bundle  $\mathcal{M}$  on  $G/K$ . The space of sections of this bundle  $\Gamma(G/K, \mathcal{M})$  has a natural structure of a  $G$ -module and we call the functor  $\Gamma$  the geometric induction functor. Note that one can define  $\Gamma$  using the language of Hopf superalgebras if one wants to avoid a rather technical question of existence of  $G/K$ . As in the case of usual groups  $\Gamma$  is the right adjoint to the restriction functor:  $G\text{-mod} \rightarrow K\text{-mod}$ . Note that  $\Gamma$  is left exact but not exact in general. The right derived functor  $R^i(\Gamma)$  coincides with the cohomology group  $H^i(G/K, \mathcal{M})$ .

In what follows we assume that  $G_0$  is reductive. We are interested in the case when  $K = P$  is a parabolic subgroup of  $G$ , i.e., a subgroup containing some Borel subgroup  $B \subset G$ . Without loss of generality we may assume that the Lie superalgebra  $\mathfrak{p}$  contains a fixed Cartan subalgebra  $\mathfrak{h}$ . Then  $\mathfrak{p}$  can be described as the non-negative part  $\bigoplus_{i \geq 0} \mathfrak{g}^i$  of the  $\mathbb{Z}$ -

grading given by eigenvalues of  $\text{ad } h$  for some  $h \in \mathfrak{h}_0$ . The zero part  $\mathfrak{g}^0$  would automatically have a reductive even part. We call the corresponding subgroup of  $P$  the Levi subgroup, and denote it by  $P_{red}$ . The subalgebra  $\mathfrak{m} = \bigoplus_{i > 0} \mathfrak{g}^i$  is the nilpotent radical of  $\mathfrak{p}$ . Since  $G_0/P_0$  is a

projective variety, if  $M$  is finite-dimensional, then  $H^i(G/P, \mathcal{M})$  is finite-dimensional for all  $i$ . If  $G$  is a reductive algebraic group and  $M$  is a simple  $P$ -module the cohomology groups  $H^i(G/P, \mathcal{M})$  are described by the famous Borel–Weil–Bott theorem.

The superanalogue of Borel–Weil–Bott theory was initiated by I. Penkov. He proved a superanalogue of the Borel–Weil–Bott theorem for the flag supermanifold  $G/B$  for classical and exceptional supergroups in the typical case. To formulate it we need to introduce a couple of notations. Recall that  $\Lambda$  is the weight lattice. We consider the shifted Weyl group

action  $w \cdot \lambda = w(\lambda + \rho) - \rho$ . Let  $\Lambda^{++}$  denote the positive Weyl chamber. It is important that  $\Lambda^{++}$  does not coincide with the set of dominant weights  $\Lambda^+$  but we have  $\Lambda^{++} \subset \Lambda^+$ .

In order to avoid an uncomfortable twist in calculations we define

$$\Gamma_i(G/P, M) := H^i(G/P, \mathcal{M}^*)^*.$$

**Theorem 6.6** ([46]). *Let  $\lambda \in \Lambda$  and  $C_\lambda$  denote the irreducible representation of  $P = B$  with character  $\lambda$  (it is one-dimensional if  $\mathfrak{h}_0 = \mathfrak{h}$ ). Assume also that  $\sharp\lambda = 0$ .*

*If  $\lambda + \rho$  is not regular, then  $\Gamma_i(G/B, C_\lambda) = 0$  for all  $i \geq 0$ .*

*If  $\lambda + \rho$  is regular, then there exists a unique  $w \in W$  such that  $\mu = w \cdot \lambda \in \Lambda^{++}$ . Then  $\Gamma_i(G/B, C_\lambda) = 0$  if  $i \neq l(w)$  and  $\Gamma_{l(w)}(G/B, C_\lambda) \simeq L_\mu$ .*

The above result can be easily generalized for an arbitrary parabolic subgroup as far as the central character remains typical. If  $\lambda$  is atypical both irreducibility of the cohomology groups and vanishing of all but one cohomology do not hold anymore.

On the other hand, I. Penkov has shown that for any  $\lambda$  the character of the Euler characteristic

$$\text{ch } \mathcal{E}_\lambda(P) = \sum (-1)^i \text{ch } \Gamma_i(G/P, C_\lambda)$$

is given by a superanalogue of the Weyl formula, see for instance [29].

To avoid introducing additional notations we write this formula only in the case  $P = B$ :

$$\text{ch } \mathcal{E}_\lambda(B) = D \sum_{w \in W} \epsilon(w) e^{w(\lambda + \rho)},$$

where

$$\rho := \frac{1}{2} \sum_{\alpha \in \Delta^+} \text{sdim } \mathfrak{g}_\alpha \alpha,$$

(for instance  $\rho = 0$  for  $G = Q(n)$ ) and

$$D = \frac{\prod_{\alpha \in \Delta^+} (e^{\alpha/2} + e^{-\alpha/2})^{\dim(\mathfrak{g}_\alpha)_1}}{\prod_{\alpha \in \Delta^+} (e^{\alpha/2} - e^{-\alpha/2})^{\dim(\mathfrak{g}_\alpha)_0}} \dim C_\lambda.$$

In the case when  $\lambda$  is a typical dominant weight, Theorem 6.6 implies  $\text{ch } L_\lambda = \text{ch } \mathcal{E}_\lambda(B)$ . That gives a geometric proof of Kac typical character formula, established in [35].

Calculating  $\text{ch } L_\lambda$  for atypical  $\lambda$  turned out to be a difficult problem which was open for a while. It was solved for  $GL(m, n)$  in [49], for  $Q(n)$  in [47] and for  $SOSP(m, 2n)$  in [28] using Borel–Weil–Bott theory for supergroups. Below we explain the main idea.

Consider the Grothendieck group  $\mathcal{K}_G$  of the category  $G\text{-mod}$ . For any  $\lambda \in \Lambda^+$  there is the unique maximal parabolic subgroup  $Q_\lambda \subset G$  such that the irreducible  $B$ -module  $C_\lambda$  extends to a  $Q_\lambda$ -module. In particular,  $Q_\lambda = B$  if  $\lambda \in \Lambda^{++}$ . Let

$$\mathcal{E}_\lambda := \sum_{i \geq 0} (-1)^i [\Gamma_i(G/Q_\lambda, C_\lambda)]$$

be the class of the Euler characteristic in  $\mathcal{K}_G$ . Introduce a partial order on  $\Lambda^+$  by setting  $\mu \leq \lambda$  if  $\lambda - \mu$  is a sum of positive roots. It is possible to show that

$$\mathcal{E}_\lambda = \sum_{\mu \leq \lambda} d_{\lambda, \mu} [L_\mu],$$

for some integers  $d_{\lambda,\mu}$  such that  $d_{\lambda,\lambda} = 1$ . Since the characters of  $\mathcal{E}_\lambda$  are computable, one can reduce computing  $\text{ch } L_\lambda$  to computing  $d_{\lambda,\mu}$ . The latter can be calculated by rather involved combinatorial algorithm.

Note that the problem of computing  $\Gamma_i(G/B, C_\lambda)$  for atypical  $\lambda$  is still open. Let us finish this subsection with two general results of interest. First, one can express the multiplicities  $[\Gamma^i(G/P, M : L_\mu)]$  using Lie algebra cohomology as in the classical Kostant theorem.

**Theorem 6.7** ([29]). *If  $P_\lambda$  denotes the projective cover of  $L_\lambda$  and  $M$  is a simple  $P$ -module, then*

$$[\Gamma_i(G/P, M) : L_\lambda] = \dim \text{Hom}_{P_{red}}(M, H^i(\mathfrak{m}, P_\lambda)).$$

Second, a certain weak analogue of the BGG reciprocity holds.

**Theorem 6.8** ([29]). *The Euler characteristics characters  $\text{ch } \mathcal{E}_\lambda(B)$  for all  $\lambda \in \Lambda^{++}$  are linearly independent in the ring of characters of  $G\text{-mod}$ . For any indecomposable projective module  $P_\mu$  there is a unique decomposition*

$$\text{ch } P_\mu = \sum_{\lambda \in \Lambda^{++}} a_{\mu,\lambda} \text{ch } \mathcal{E}_\lambda(B),$$

and

$$\text{ch } \mathcal{E}_\lambda(B) = \sum_{\mu \in \Lambda^+} a_{\mu,\lambda} \text{ch } L_\mu.$$

Finally, we should state here for the sake of next section the following

**Lemma 6.9** ([28]). *If  $M$  is a simple  $P$ -module, then all the spaces  $\Gamma_i(G/P, M)$  when  $i$  varies belong to the same block of  $G\text{-mod}$ .*

**6.3. Categorification.** In [6] J. Brundan suggested a new remarkable approach to the problem of computing irreducible characters for  $GL(m, n)$  and obtained an easier algorithm for calculating multiplicities  $d_{\lambda,\mu}$ . The combinatorial proof of equivalence of two algorithms, [49] and [6], can be found in [45]. Later Brundan applied the same method to reprove the results of [47] for  $Q(n)$ . To some extent the same method was applied to  $SOSP(m, 2n)$  in [29] for computing multiplicities  $a_{\mu,\lambda}$  (in notations of Theorem 6.8).

The main idea of Brundan’s approach is to categorify some representation  $F$  of an infinite-dimensional Lie algebra  $\mathcal{L}$  (both depending on  $G$ ). One would like to identify the integral form  $\mathbf{F}$  of  $F$  with the Grothendieck group  $\mathcal{K}_G$  in such a way that translation functors categorify the Chevalley generators of  $\mathcal{L}$  and the projective and simple modules categorify canonical and dual to canonical bases in  $\mathbf{F}$ . We now proceed with details in the case  $G = Q(n)$ , following [7].

In this case  $\mathcal{L}$  is the infinite-dimensional Lie superalgebra  $\mathfrak{o}(\infty)$  which can be defined as the direct limit  $\lim_{\rightarrow} \mathfrak{o}(2N + 1)$ . This Lie algebra has the infinite Dynkin diagram  $B_\infty$

$$\circ \leftarrow \circ - \circ - \circ - \dots,$$

with Chevalley generators  $\{E_i, F_i \mid i \in \mathbb{Z}_{>0}\}$  and the usual Chevalley–Serre relations. We fix the Cartan subalgebra and the Chevalley generators. Let  $\mathbb{V}$  denote the standard representation of  $\mathcal{L}$ . We can choose a basis  $\{v_i \mid i \in \mathbb{Z}\}$  in  $\mathbb{V}$  such that the action of the Chevalley

generators is given by the formulas

$$\begin{aligned}
 \text{for } i > 1 \quad & E_i(v_i) = v_{i+1}, \quad E_i(v_{-i-1}) = v_{-i}, \quad E_i(v_j) = 0 \text{ if } j \neq i, -i - 1, \\
 & F_i(v_{i+1}) = v_i, \quad F_i(v_{-i}) = v_{-i-1}, \quad F_i(v_j) = 0 \text{ if } j \neq i + 1, -i, \\
 & E_1(v_0) = 2v_1, \quad E_1(v_{-1}) = v_0, \quad E_1(v_j) = 0 \text{ if } j \neq 0, -1, \\
 & F_1(v_0) = 2v_{-1}, \quad F_1(v_1) = v_0, \quad F_1(v_j) = 0 \text{ if } j \neq 0, 1.
 \end{aligned} \tag{6.2}$$

Next we define  $R$  to be the quotient of the tensor algebra  $T(\mathbb{V})$  by the relations

$$\begin{aligned}
 v_i^2 &= 0 \text{ if } i \neq 0, \\
 v_i v_j + v_j v_i &= 0 \text{ if } i > j, j + i \neq 0, \\
 v_0^2 &= (-1)^i v_i v_{-i} + v_{-i} v_i.
 \end{aligned} \tag{6.3}$$

If  $V_N$  denotes the standard representation of  $\mathfrak{o}(2N + 1)$ , and  $R(N)$  is the Koszul dual of  $S(V_N)/(r)$ , where  $r \in S^2(V_N)$  is the quadratic form preserved by  $\mathfrak{o}(2N + 1)$ , then  $R = \varinjlim R(N)$ . Hence  $R$  has a natural structure of  $\mathfrak{L}$ -module. Consider the grading  $R = \bigoplus_{n \geq 0} R^n$

induced by the obvious grading of  $T(\mathbb{V})$ . It is clear that  $R^n$  is  $\mathfrak{L}$ -invariant.

The homogeneous component  $R^n$  has a natural monomial basis  $v_{a_1} \dots v_{a_n}$ , where  $(a_1, \dots, a_n)$  satisfy the condition

$$a_i > a_{i+1} \quad \text{if } a_i \neq 0, \quad a_i \geq a_{i+1} \quad \text{if } a_i = 0.$$

Note that these are precisely the conditions of dominance of  $\lambda = \sum_{i=1}^n a_i \varepsilon_i$  for  $G = Q(n)$ . So

for each  $\lambda \in \Lambda^+$  we may set  $v_\lambda := v_{a_1} \dots v_{a_n}$ . One can check that  $\mathbf{F} = \bigoplus \mathbb{Z} v_\lambda$  is invariant under the action of the Chevalley  $\mathbb{Z}$ -form of  $\mathfrak{L}$ .

Consider the weight decomposition of  $R^n$  with respect to the Cartan subalgebra of  $\mathfrak{L}$ . Let  $\text{wt } v_\lambda$  denote the weight of  $v_\lambda$ , then  $\text{wt } v_i = -\text{wt } v_{-i}$  and  $\text{wt } v_0 = 0$ . Set  $\delta_i = \text{wt } v_i$ , then

$$\text{wt } v_\lambda = \sum b_j \delta_j,$$

where  $b_j = 0, \pm 1$ . The easiest way to calculate  $b_j$  is by means of the weight diagram  $f_\lambda$ :

$$b_j = \begin{cases} 1 & \text{if } f_\lambda(i) = > \\ -1 & \text{if } f_\lambda(i) = < \\ 0 & \text{if } f_\lambda(i) = \circ, \times \end{cases} \tag{6.4}$$

Define the  $\mathbb{Z}$ -linear map  $\gamma: \mathcal{K}_G \rightarrow \mathbf{F}$  by setting  $\gamma(\mathcal{E}_\lambda) := v_\lambda$ . By (6.4)  $\text{wt } \gamma(\mathcal{E}_\lambda) = \text{wt } \gamma(\mathcal{E}_\mu)$  if and only if  $\mathcal{E}_\lambda$  and  $\mathcal{E}_\mu$  belong to the Grothendieck group  $\mathcal{K}_G^\chi$  of same block  $G^\chi\text{-mod}$  (that makes sense by Lemma 6.9). Thus, we have a bijection

$$\text{wt: } \{ \text{integral central characters of } \mathfrak{g} \} \leftrightarrow \{ \text{weights of } \mathbf{F} \}.$$

Define endofunctors  $T_i$  and  $T_i^*$  of  $G\text{-mod}$  by setting

$$T_i = \bigoplus_{\text{wt}(\zeta) - \text{wt}(\chi) = \delta_{i+1} - \delta_i} T_{\chi, \zeta}, \quad T_i^* = \bigoplus_{\text{wt}(\zeta) - \text{wt}(\chi) = \delta_i - \delta_{i+1}} T_{\chi, \zeta}^*,$$

here we assume  $\delta_0 = 0$ .

Since  $T_i, T_i^*$  are exact, they induce  $\mathbb{Z}$ -linear operators on  $\mathcal{K}_G$ . Denote those operators by  $t_i$  and  $t_i^*$  respectively. Computing the action of  $t_i$  and  $t_i^*$  on  $\mathcal{E}_\lambda$  is not difficult since one knows the characters of  $\mathcal{E}_\lambda$ . The following can be checked by direct calculation in the basis  $\{\mathcal{E}_\lambda\}$ :

- if  $i > 0$ , then  $\gamma \circ t_i = 2E_i \circ \gamma$ ,  $\gamma \circ t_i^* = 2F_i \circ \gamma$ ;
- if  $M \in G^\times\text{-mod}$  and  $p(\chi) = 1$ , then  $\gamma \circ t_0([M]) = 2E_0 \circ \gamma([M])$ ,  $\gamma \circ t_0^*([M]) = 2F_0 \circ \gamma([M])$ ;
- if  $M \in G^\times\text{-mod}$  and  $p(\chi) = 0$ , then  $\gamma \circ t_0([M]) = E_0 \circ \gamma([M])$ ,  $\gamma \circ t_0^*([M]) = F_0 \circ \gamma([M])$ .

Thus, translation functors act in  $\mathcal{K}_G$  in the same way as Chevalley generators in  $\mathbf{F}$ . Next step is to consider the quantized universal enveloping  $U_q(\mathfrak{g})$  in order to construct the canonical Lusztig basis in  $\mathbf{F}$ . We omit the details here, since a lot of computations is involved, and just state the result. It is proven in [7] that there exist a canonical topological basis  $\{u_\lambda \mid \lambda \in \Lambda^+\}$  in the completion of  $\mathbf{F}$  and the dual basis  $\{l_\lambda \mid \lambda \in \Lambda^+\}$ .

**Theorem 6.10** ([7]). *For any  $\lambda \in \Lambda^+$  the following relation holds:*

$$\gamma([L_\lambda]) = l_\lambda, \quad \gamma(P_\lambda) = u_\lambda.$$

By theorem 6.10 we have

$$v_\lambda = \sum_{\mu \leq \lambda} d_{\lambda, \mu} l_\mu.$$

Since the crystal structure in  $\mathbf{F}$  is relatively simple, one can find a combinatorial algorithm for calculating a  $d_{\lambda, \mu}$ , see [7]. In [57] this algorithm is formulated in terms of weight diagrams.

Let us finish with the remark that the above method can be applied to the problem of finding Kazhdan–Lusztig polynomials for the category  $\mathcal{O}$ , if, instead of  $R$ , one works with the tensor algebra  $T(\mathbb{V})$ . The analogue of Theorem 6.10 in this more complicated situation was conjectured in [7]. This conjecture is still open for  $Q(n)$ . For  $GL(m, n)$  there are now two proofs: in [8] and in [12].

## References

- [1] G. Benkart, M. Chakrabarti, T. Halverson, and R. Leduc, C. Lee, J. Stroome *Tensor product representations of general linear groups and their connections with Brauer algebras*, J. of Algebra, **166** (1994), 529–567.
- [2] G. Benkart, C. Lee Shader, and A. Ram, *Tensor product representations for orthosymplectic Lie superalgebras*, J. of Pure and Applied Algebra, **130** (1998), No. 1, 1–48.
- [3] A. Berele and A. Regev, *Hook Young diagrams with application to combinatorics and representations of Lie superalgebras*, Adv. in Math. **64** (1987), 118–175.
- [4] B. Boe, J. Kujawa, and D. Nakano, *Complexity for modules over the classical Lie superalgebra  $gl(m|n)$* , Compos. Math. **148** (2012), no. 5, 1561–1592.
- [5] R. Brauer, *On algebras which are connected with simple Lie groups*, Ann. of Math. **38** (1937), 857–872.

- [6] J. Brundan, *Kazhdan-Lusztig polynomials and character formulae for the Lie superalgebra  $gl(m|n)$* , J. Amer. Math. Soc. **16** (2003), no. 1, 185–231.
- [7] J. Brundan, *Kazhdan-Lusztig polynomials and character formulae for the Lie superalgebra  $q(n)$* , Adv. Math. **182** (2004), no. 1, 28–77.
- [8] J. Brundan, I. Losev, and B. Webster, *Tensor product categorifications and the super Kazhdan-Lusztig conjecture* arXiv:1310.0349v2
- [9] J. Brundan and C. Stroppel, *Highest weight categories arising from Khovanov’s diagram algebra IV: the general linear supergroup*, J. Eur. Math. Soc. (JEMS) **14** (2012), no. 2, 373–419.
- [10] J. Brundan and C. Stroppel, *Gradings on walled Brauer algebras and Khovanov’s arc algebra*. Adv. Math. **231** (2012), no. 2, 709–773.
- [11] S.J. Cheng and W. Wang, *Dualities and representations of Lie superalgebras*, Graduate Studies in Mathematics, 144. American Mathematical Society, Providence, RI, 2012.
- [12] S.J. Cheng, N. Lam, and W. Wang, *Brundan–Kazhdan–Lusztig conjecture for general linear Lie superalgebra* arXiv:1203.0092v3.
- [13] S.J. Cheng, N. Lam, and W. Wang, *Super duality and irreducible characters of orthosymplectic Lie superalgebras*, Invent. Math. **183** (2011), no. 1, 189–224.
- [14] J. Comes and B. Wilson, *Deligne’s category  $\text{Rep } GL(\delta)$  and representations of general linear supergroups*. Represent. Theory **16** (2012), 568–609.
- [15] J. Comes, *Ideals in Deligne’s tensor category  $\text{Rep } GL(\delta)$* , arXiv:1201.5669.
- [16] A. Cox and M. De Visscher, *Diagrammatic Kazhdan–Lusztig theorem for the (walled) Brauer algebra*, J. Algebra **340** (2011), 151–181.
- [17] P. Deligne. *La catégorie des représentations du groupe symétrique  $S_t$ , lorsque  $t$  n’est pas un entier naturel*. In Algebraic Groups and homogeneous spaces, 209–273, TIFR, Mumbai, 2007.
- [18] P. Deligne, *La série exceptionnelle de groupes de Lie*, C. R. Acad. Sci. Paris Ser. I. Math. **323** (1996), 577–582.
- [19] M. Duflo and V. Serganova, *On associated variety for Lie superalgebras*, arXiv:math/0507198
- [20] D. Grantcharov, J.-H. Jung, S.-J. Kang, M. Kashiwara, and M. Kim, *Quantum queer superalgebra and crystal bases*, Proc. Japan Acad. Ser. A Math. Sci. **86** (2010) 177–182.
- [21] M. Gorelik, *The center of a simple  $P$ -type Lie superalgebra*, J. Algebra **246** (2001), no. 1, 414–428.
- [22] M. Gorelik, *The Kac construction of the centre of  $U(\mathfrak{g})$  for Lie superalgebras*, J. Non-linear Math. Phys. **11** (2004), no. 3, 325–349.
- [23] N. Geer, J. Kujawa, and B. Patureau-Mirand, *Generalized trace and modified dimension functions on ribbon categories*, Selecta Math. (N.S.) **17** (2011), no. 2, 453–504.
- [24] N. Geer and B. Patureau-Mirand, *An invariant supertrace for the category of representations of Lie superalgebras*, Pacific J. Math. **238** (2008), no.2, 331–348.
- [25] J. Germoni, *Représentations indécomposables des algèbres de Lie spéciales linéaires*, Thèse de l’université de Strasbourg, 1997

- [26] C. Gruson, *Cohomologie des modules de dimension finie sur la super algèbre  $\mathfrak{osp}(3|2)$* , J. Algebra **259** (2003), 581–598.
- [27] C. Gruson, *Sur l’ideal du cone autocommutant des super algèbres de Lie basiques classiques et étranges*, Ann. Inst. Fourier (Grenoble) **50** (2000), no. 3, 807–831.
- [28] C. Gruson and V. Serganova, *Cohomology of generalized supergrassmannians and character formulae for basic classical Lie superalgebras*, Proc. Lond. Math. Soc. (3) **101** (2010), no. 3, 852–892.
- [29] C. Gruson and V. Serganova, *Bernstein-Gelfand-Gelfand reciprocity and indecomposable projective modules for classical algebraic supergroups*, Mosc. Math. J. **13** (2013), no. 2, 281–313.
- [30] J. Graham and G. Lehrer, *Cellular algebras*, Inv. Math. **123** (1996), 1–34.
- [31] J. H. Jung and S.-J. Kang, *Mixed Schur-Weyl-Sergeev duality for queer Lie superalgebras*, J. Algebra **399** (2014), 516–545.
- [32] V. G. Kac, *Laplace operators of infinite-dimensional Lie algebras and theta functions*, Proc. Nat. Acad. Sci. U.S.A. **81** (1984), no. 2, Phys. Sci., 645–647.
- [33] V. G. Kac, *Classification of supersymmetries*. Proceedings of the International Congress of Mathematicians, Vol. I (Beijing, 2002), 319–344, Higher Ed. Press, Beijing, 2002.
- [34] V. G. Kac, *Lie superalgebras*, Adv. Math. **26** (1977), 8–96.
- [35] V. G. Kac, *Representations of classical Lie superalgebras*. Differential geometrical methods in mathematical physics, II (Proc. Conf., Univ. Bonn, Bonn, 1977), 597–626, Lecture Notes in Math., 676, Springer, Berlin, 1978.
- [36] V. G. Kac and M. Wakimoto, *Integrable highest weight modules over affine superalgebras and number theory*. Lie theory and geometry, 415–456, Progr. Math., **123**, Birkhäuser Boston, Boston, MA, 1994.
- [37] B. Kostant, *Graded manifolds, graded Lie theory, and prequantization*. Lecture Notes in Mathematics 570. Berlin, Springer-Verlag, 1977, 177–306.
- [38] J.L. Koszul, *Graded manifolds and graded Lie algebras*, International Meeting on Geometry and Physics (Bologna), Pitagora, 1982, 71–84.
- [39] J. Kujawa, *The generalized Kac-Wakimoto conjecture and support varieties for the Lie superalgebra  $\mathfrak{osp}(m|2n)$* . Recent developments in Lie algebras, groups and representation theory, 201–215, Proc. Sympos. Pure Math., 86, Amer. Math. Soc., Providence, RI, 2012.
- [40] L. Martirosyan, *Representations of the exceptional Lie superalgebras  $G_3$  and  $F_4$* , PhD thesis, Berkeley, 2013.
- [41] A. Masuoka and A. Zubkov, *Quotient sheaves of algebraic supergroups are superschemes*, J. Algebra **348** (2011), 135–170.
- [42] A. Masuoka, *Harish-Chandra pairs for algebraic affine supergroup schemes over an arbitrary field*. Transform. Groups **17** (2012), no. 4, 1085–1121.
- [43] D. Moon, *Tensor product representations of the Lie superalgebra  $\mathfrak{p}(n)$  and their centralizers*, Comm. Algebra **31** (2003), no. 5, 2095–2140.
- [44] I. Musson, *Lie superalgebras and enveloping algebras*, Graduate Studies in Mathematics, 131. American Mathematical Society, Providence, RI, 2012.

- [45] I. Musson and V. Serganova, *Combinatorics of character formulas for the Lie superalgebra  $gl(m, n)$* , Transform. Groups **16** (2011), no. 2, 555–578.
- [46] I. Penkov, *Borel-Weil-Bott theory for classical Lie supergroups*, (Russian) Translated in J. Soviet Math. **51** (1990), no. 1, 2108–2140.
- [47] I. Penkov and V. Serganova, *Characters of irreducible  $G$ -modules and cohomology of  $G/P$  for the Lie supergroup  $G = Q(N)$* . Algebraic geometry, 7. J. Math. Sci. (New York) **84** (1997), no. 5, 1382–1412.
- [48] H. Rui, *A criterion on the semisimple Brauer algebra*, J. of Combinatorial theory, Ser. A **111** (2005), 78–88.
- [49] V. Serganova, *Kazhdan-Lusztig polynomials and character formula for the Lie superalgebra  $gl(m|n)$* , Selecta Math. (N.S.) **2** (1996), no. 4, 607–651.
- [50] V. Serganova, *On the superdimension of an irreducible representation of a basic classical Lie superalgebra*, Supersymmetry in mathematics and physics, 253–273, Lecture Notes in Math. **2027**, Springer, Heidelberg, 2011.
- [51] V. Serganova, *Quasireductive supergroups*, New developments in Lie theory and its applications, 141–159, Contemp. Math. **544**, Amer. Math. Soc., Providence, RI, 2011.
- [52] V. Serganova, *Kac-Moody superalgebras and integrability*, Developments and trends in infinite-dimensional Lie theory, 169–218, Progr. Math. **288**, Birkhäuser Boston, Inc., Boston, MA, 2011.
- [53] A. Sergeev, *The tensor algebra of the tautological representation as a module over the Lie superalgebras  $gl(n, m)$  and  $Q(n)$* , Mat. Sb. **123** (1984) 422–430 (in Russian).
- [54] A. Sergeev, *The invariant polynomials on simple Lie superalgebras*, Represent. Theory **3** (1999), 250–280.
- [55] A. Sergeev, *The Howe duality and the projective representations of symmetric groups*, Represent. Theory **3** (1999), 416–434 (electronic).
- [56] A. Sergeev, *The centre of enveloping algebra for Lie superalgebra  $Q(n, C)$* , Lett. Math. Phys. **7** (1983), no. 3, 177–179.
- [57] Y. Su and R. Zhang, *Characters and character formulae for queer Lie superalgebras*, arXiv:1305.2906
- [58] E. Vishnyakova, *On complex Lie supergroups and split homogeneous supermanifolds*, Transform. Groups **16** (2011), no. 1, 265–285.
- [59] H. Wenzl, *On the structure of Brauer’s centralizer algebras*, Ann. of Math, **128** (1988), 173–193.

Vera Serganova, UC Berkeley

E-mail: serganov@math.berkeley.edu

# **Special Lectures**



# Connecting the McKay correspondence and Schur-Weyl duality

Georgia Benkart

**Abstract.** The McKay correspondence and Schur-Weyl duality have inspired a vast amount of research in mathematics and physics. The McKay correspondence establishes a bijection between the finite subgroups of the special unitary 2-by-2 matrices and the simply laced affine Dynkin diagrams from Lie theory. It has led to the discovery of many other remarkable A-D-E phenomena. Schur-Weyl duality reveals hidden connections between the representation theories of two algebras that centralize one another in their actions on the same space. We merge these two notions and explain how this gives new insights and results. Our approach uses the combinatorics of walks on graphs, the Jones basic construction, and partition algebras.

**Mathematics Subject Classification (2010).** Primary 14E16, 05E10, 20C05.

**Keywords.** McKay correspondence, Schur-Weyl duality.

## 1. Introduction

Walks on graphs have found widespread applications in modeling networks, biological and random processes, information flow, and many other phenomena. Typically, the walker (a person, particle, or impulse) transitions from one node to another along an edge which may have an assigned probability. The adjacency matrix of the graph and the Bratteli diagram facilitate answering questions such as: How many different walks of  $k$  steps are there from point  $a$  to point  $b$  on the graph? What is the probability that a particle moves from  $a$  to  $b$  in  $k$  steps? When the graph arises from representations of groups, a much richer structure is available to answer such questions.

Let  $G$  be a group and  $V$  be a finite-dimensional  $G$ -module over the complex field  $\mathbb{C}$ . The *representation graph*  $\mathcal{R}_V(G)$  of  $G$  associated to  $V$  has nodes  $\lambda$  corresponding to the irreducible  $G$ -modules  $\{G^\lambda \mid \lambda \in \Lambda(G)\}$  over  $\mathbb{C}$ . For  $\lambda, \mu \in \Lambda(G)$ , there are  $a_{\mu,\lambda}$  edges from  $\mu$  to  $\lambda$  in  $\mathcal{R}_V(G)$  if

$$G^\mu \otimes V = \bigoplus_{\lambda \in \Lambda(G)} a_{\mu,\lambda} G^\lambda.$$

Thus, the number of edges  $a_{\mu,\lambda}$  from  $\mu$  to  $\lambda$  in  $\mathcal{R}_V(G)$  is the multiplicity of  $G^\lambda$  as a summand of  $G^\mu \otimes V$ .

In the particular case that  $G$  is a finite group, the representation graph and character table of  $G$  are closely related. Assume  $\chi_V$  is the character of  $V$  and  $\chi_\lambda$  is the character of  $G^\lambda$  for  $\lambda \in \Lambda(G)$ , and let  $d = \dim V = \chi_V(1)$ . Steinberg [24] has shown that when the action of

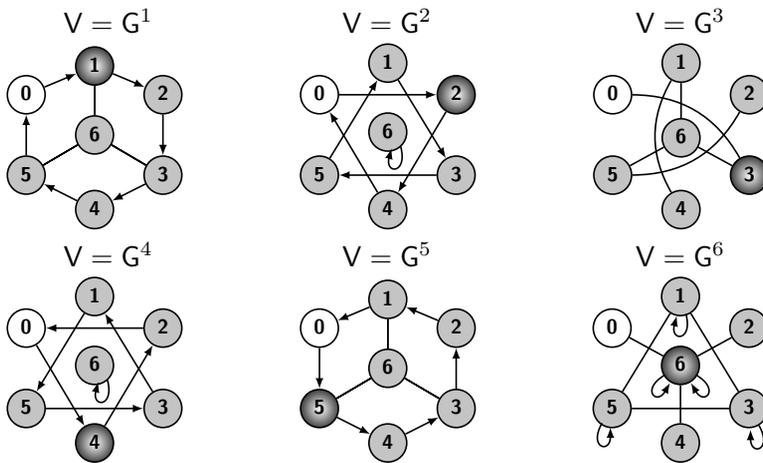
---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

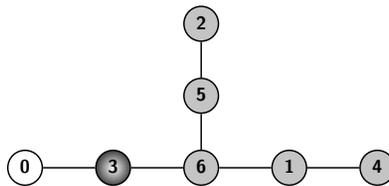
$G$  on  $V$  is faithful, the eigenvalues of the matrix  $(d \delta_{\mu,\lambda} - a_{\mu,\lambda})$  are  $d - \chi_V(g)$  as  $g$  ranges over conjugacy class representatives of  $G$ , and the eigenvector corresponding to  $d - \chi_V(g)$  is  $(\chi_\lambda(g))$ . These vectors form the columns of the character table of  $G$ . The vector  $(d^\lambda)$  whose entries are the dimensions  $d^\lambda = \dim G^\lambda = \chi_\lambda(1)$  corresponds to the eigenvalue 0.

Let  $G^0$  be the one-dimensional trivial  $G$ -module on which every element of the group acts as the identity transformation, and let  $m_k^\lambda$  be the number of walks of  $k$  steps from 0 to  $\lambda$  on the graph  $\mathcal{R}_V(G)$ . Since each step on the graph is accomplished by tensoring with  $V$ ,  $m_k^\lambda$  is exactly the multiplicity of the irreducible  $G$ -module  $G^\lambda$  in  $G^0 \otimes V^{\otimes k} \cong V^{\otimes k}$ .

**Example 1.1.** The group  $G = \text{SL}_2(\mathbb{F}_3)$  of  $2 \times 2$  matrices of determinant 1 over the field  $\mathbb{F}_3$  of 3 elements has six non-trivial irreducible modules, denoted here  $G^j, j = 1, \dots, 6$ . The representation graph  $\mathcal{R}_V(G)$  for each  $V = G^j$  is pictured in the figure below, where the white node corresponds to the trivial module  $G^0$ , and the black node indicates  $V$ . Edges without directional arrows are two-way streets.



In particular, we can read off from the last graph that  $G^6 \otimes G^6 = G^0 \oplus G^2 \oplus G^4 \oplus 2 \cdot G^6$ . Graphs that appear similar except for reversal of the directional arrows correspond to dual modules, and there are two such dual module pairs  $\{G^1, G^5\}$  and  $\{G^2, G^4\}$ . The module labeled  $G^3$  has an especially intriguing representation graph, as unraveling it gives



which is the affine Dynkin diagram of type  $\hat{E}_6$ .

**Example 1.2.** The special unitary group  $\text{SU}_2 = \left\{ \begin{pmatrix} x & y \\ -\bar{y} & \bar{x} \end{pmatrix} \mid x, y \in \mathbb{C}, x\bar{x} + y\bar{y} = 1 \right\}$ , where “ $\bar{\phantom{x}}$ ” denotes complex conjugate, has infinitely many finite-dimensional irreducible modules,  $V(r), r = 0, 1, \dots$ , indexed by the nonnegative integers, and  $\dim V(r) = r + 1$ . The natural module for  $\text{SU}_2$  is the space  $V = V(1) = \mathbb{C}^2 = \left\{ \begin{pmatrix} \cdot \\ \cdot \end{pmatrix} \right\}$  of  $2 \times 1$  column vectors, upon which  $\text{SU}_2$  acts by matrix multiplication. The well-known Clebsch-Gordan formula,

$$V(r) \otimes V = V(r - 1) \oplus V(r + 1) \quad (V(-1) = 0) \tag{1.1}$$

gives the rule for tensoring with  $V$ , and the representation graph  $\mathcal{R}_V(\mathrm{SU}_2)$  is

$$\textcircled{0} \text{---} \textcircled{1} \text{---} \textcircled{2} \text{---} \textcircled{3} \text{---} \textcircled{4} \text{---} \dots \tag{1.2}$$

Walks on this graph are equivalent to walks on the set of nonnegative integers.

In 1980, J. McKay [21] made the striking discovery that there is a natural one-to-one correspondence between the isomorphism classes of finite subgroups of the special unitary group  $\mathrm{SU}_2$  and the simply laced affine Dynkin diagrams. This discovery has led to an immense literature that connects the McKay correspondence to a wide array of topics in mathematics and physics such as Kleinian singularities, Coxeter transformations, Hilbert schemes, the cohomology of Calabi-Yau manifolds, and mirror symmetry, to name just a few. Almost a century earlier, F. Klein had determined that a finite subgroup of  $\mathrm{SU}_2$  must be isomorphic to one of the following: (a) a cyclic group  $\mathbf{C}_n$  of order  $n$ , (b) a binary dihedral group  $\mathbf{D}_n$  of order  $4n$ , or (c) one of the 3 exceptional groups: the binary tetrahedral group  $\mathbf{T}$  of order 24 (which is isomorphic to the group  $\mathrm{SL}_2(\mathbb{F}_3)$  in Example 1.1), the binary octahedral group  $\mathbf{O}$  of order 48, or the binary icosahedral group  $\mathbf{I}$  of order 120. Binary here refers to the fact that the center is  $\{\pm I\}$ , where  $I$  is the  $2 \times 2$  identity matrix, and the group modulo its center is the dihedral group or the rotational symmetry group of the tetrahedron, octahedron, or icosahedron in the exceptional cases.

Let  $G$  be a finite subgroup of  $\mathrm{SU}_2$  and  $V = \mathbb{C}^2$ . McKay’s observation was that the representation graph  $\mathcal{R}_V(G)$  for  $G = \mathbf{C}_n, \mathbf{D}_n, \mathbf{T}, \mathbf{O}, \mathbf{I}$  is exactly the affine Dynkin diagram  $\hat{A}_{n-1}, \hat{D}_{n+2}, \hat{E}_6, \hat{E}_7, \hat{E}_8$ , respectively, with the vertex 0 being the affine node (see Section A.1 of the Appendix for the diagrams). Moreover, the following hold: (i) the “marks” that appear above the nodes on the affine Dynkin diagram are the dimensions of the irreducible  $G$ -modules; (ii) the sum of those dimensions is the Coxeter number of the corresponding finite Dynkin diagram obtained by removing the affine node; (iii) the Cartan matrix of the Dynkin diagram is  $C = 2I - A$ , where  $A = (a_{\mu,\lambda})$  is the adjacency matrix of  $\mathcal{R}_V(G)$  and  $I$  is the identity matrix of the appropriate size; (iv) the marks are the coordinates of the Perron-Frobenius eigenvector of  $A$ ; and (v) the eigenvectors of  $C$  form the character table of  $G$ . (Part (v) inspired Steinberg’s result mentioned earlier.)

A walk on the graph  $\mathcal{R}_V(G)$  corresponds to tensoring by  $V$ , so it is natural to consider the centralizer algebra  $Z_k(G) = \mathrm{End}_G(V^{\otimes k})$  of transformations that commute with the action of  $G$  on  $V^{\otimes k}$ . The algebra  $Z_k(G)$  encodes essential information about the structure of  $V^{\otimes k}$  as a  $G$ -module. The projection maps from  $V^{\otimes k}$  onto its irreducible  $G$ -summands are idempotents in  $Z_k(G)$ , and the multiplicity of  $G^\lambda$  in  $V^{\otimes k}$  (hence, the number of walks of  $k$  steps from 0 to  $\lambda$  on  $\mathcal{R}_V(G)$ ) is the dimension of the irreducible  $Z_k(G)$ -module corresponding to  $\lambda$ . The structure and representation theory of the algebras  $Z_k(G)$  control, and are controlled by, the combinatorics of the representation graph  $\mathcal{R}_V(G)$  via Schur-Weyl duality.

Schur-Weyl duality has been one of the most prolific concepts in representation theory, uncovering hidden connections between the representations of seemingly unrelated algebraic objects. This paper combines the McKay correspondence and Schur-Weyl duality. Among the results presented here is a new way of relating the McKay correspondence for the groups  $\mathbf{T}$  and  $\mathbf{O}$  to partition algebras and partitions (see Sections 4.1 and 4.2). The work is based on investigations with J. Barnes and T. Halverson [1], and for the exceptional cases with T. Halverson [2]. To them I extend my sincere thanks.

## 2. Centralizers and Schur-Weyl duality

The *centralizer algebra* of the action of a group  $G$  on the  $k$ -fold tensor power of a finite-dimensional  $G$ -module  $V$  over  $\mathbb{C}$  is the semisimple associative algebra

$$Z_k(G) = \text{End}_G(V^{\otimes k}) = \{X \in \text{End}(V^{\otimes k}) \mid X(gw) = gX(w) \ \forall g \in G, w \in V^{\otimes k}\}. \quad (2.1)$$

Let  $\Lambda_k(G)$  denote the subset of  $\lambda \in \Lambda(G)$  such that  $G^\lambda$  occurs in  $V^{\otimes k}$  with multiplicity  $m_k^\lambda \geq 1$ . The irreducible  $Z_k(G)$ -modules  $Z_k^\lambda$  are in bijection with the elements  $\lambda$  of  $\Lambda_k(G)$ . *Schur-Weyl duality* relates the decomposition of  $V^{\otimes k}$  as a  $G$ -module to the decomposition of  $V^{\otimes k}$  as a  $Z_k(G)$ -module establishing deep connections between the representation theories of  $G$  and  $Z_k(G)$ :

- $V^{\otimes k} \cong \bigoplus_{\lambda \in \Lambda_k(G)} m_k^\lambda G^\lambda$     and     $V^{\otimes k} \cong \bigoplus_{\lambda \in \Lambda_k(G)} d^\lambda Z_k^\lambda$ ;
- $\dim Z_k^\lambda = m_k^\lambda =$  number of walks of  $k$  steps from  $0$  to  $\lambda$  on  $\mathcal{R}_V(G)$ ;
- $\dim G^\lambda = d^\lambda$ ;
- $\dim Z_k(G) = \sum_{\lambda \in \Lambda_k(G)} (\dim Z_k^\lambda)^2 = \sum_{\lambda \in \Lambda_k(G)} (m_k^\lambda)^2$   
 $=$  number of walks of  $2k$  steps from  $0$  to  $0$  on  $\mathcal{R}_V(G)$   
 $= \dim Z_{2k}^0$ ;
- as a  $(G, Z_k(G))$ -bimodule,  $V^{\otimes k}$  has a multiplicity-free decomposition,

$$V^{\otimes k} \cong \bigoplus_{\lambda \in \Lambda_k(G)} (G^\lambda \otimes Z_k^\lambda).$$

By applying idempotents in the algebras  $Z_k(G)$  to project onto the irreducible  $G$ -summands, often one is able to build the entire family of finite-dimensional irreducible  $G$ -modules from a single well-chosen module  $V$  and its tensor powers. Indeed, Schur’s groundbreaking 1901 doctoral thesis constructed the finite-dimensional irreducible polynomial representations for the general linear group  $GL_n(\mathbb{C})$  from tensor powers of its natural module  $V = \mathbb{C}^n$  in exactly this way. The algebra  $Z_k(GL_n(\mathbb{C}))$  is a homomorphic image of the group algebra  $\mathbb{C}\mathbf{S}_k$  of the symmetric group  $\mathbf{S}_k$  for  $k \geq 1$ , which acts by permuting the factors of  $V^{\otimes k}$ . Idempotents in  $\mathbb{C}\mathbf{S}_k$  provided the necessary projection maps. Schur-Weyl duality has been applied in many different settings in the intervening years.

**2.1. Bratteli diagrams.** The *Bratteli diagram*  $\mathcal{B}_V(G)$  associated to the group  $G$  and the module  $V$  gives an effective tool for determining information about walks on the representation graph  $\mathcal{R}_V(G)$ , the tensor powers  $V^{\otimes k}$ , and the family of centralizer algebras  $Z_k(G)$  and their irreducible modules. The diagram  $\mathcal{B}_V(G)$  is the infinite graph with vertices labeled by the elements of  $\Lambda_k(G)$  on level  $k$ . A walk of  $k$  steps on the representation graph  $\mathcal{R}_V(G)$  from  $0$  to  $\lambda$  is a sequence  $(0, \lambda^1, \lambda^2, \dots, \lambda^k = \lambda)$  starting at  $\lambda^0 = 0$ , such that  $\lambda^j \in \Lambda(G)$  for each  $1 \leq j \leq k$ , and  $\lambda^{j-1}$  is connected to  $\lambda^j$  by an edge in  $\mathcal{R}_V(G)$ . Such a walk is equivalent to a unique path of length  $k$  on the Bratteli diagram  $\mathcal{B}_V(G)$  from  $0$  at the top to  $\lambda \in \Lambda_k(G)$  on level  $k$ .

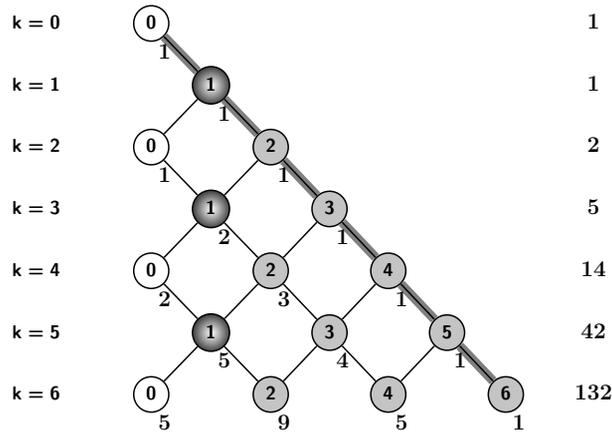


Figure 2.1. Levels  $k = 0, 1, \dots, 6$  of the Bratteli diagram for  $SU_2$

For the group  $G = SU_2$  and the  $G$ -module  $V = \mathbb{C}^2$ , the top levels of the Bratteli diagram are displayed in Figure 2.1 above. For the finite subgroups  $G$  of  $SU_2$ , the top levels of the diagrams  $\mathcal{B}_V(G)$  can be found in Section A.2 of the Appendix.

The subscript on vertex  $\lambda \in \Lambda_k(G)$  in  $\mathcal{B}_V(G)$  indicates the number  $m_k^\lambda$  of paths (walks) from 0 on the top to  $\lambda$  at level  $k$ . It can be easily computed by summing, in a Pascal triangle fashion, the subscripts of the vertices at level  $k - 1$  that are connected to  $\lambda$ . This is dimension of the irreducible  $Z_k(G)$ -module  $Z_k^\lambda$ , which is also the multiplicity of  $G^\lambda$  in  $V^{\otimes k}$ . The sum of the squares of those dimensions at level  $k$  is the number on the right, which is the dimension of the centralizer algebra  $Z_k(G)$ . In the particular case  $G = SU_2$ , those dimensions are the familiar Catalan numbers  $C_k$ . Hence,

$$\dim Z_k(SU_2) = C_k = \frac{1}{k+1} \binom{2k}{k},$$

which is the number of walks on  $\{0, 1, 2, \dots\}$  of  $2k$  steps that begin and end at 0 – one of the hundreds of objects that the Catalan number  $C_k$  enumerates. (See [23, Exercise 6.19].)

**2.2.  $SU_2$  and Temperley-Lieb algebras.** Let  $V = \mathbb{C}^2$ , and let  $\sigma$  be the linear transformation on  $V^{\otimes 2}$  interchanging the tensor factors,  $\sigma(v \otimes w) = w \otimes v$ . The symmetric group  $S_k$  acts as place permutations on  $V^{\otimes k}$ , with the simple transposition generator  $s_i = (i \ i + 1)$ ,  $1 \leq i < k$ , of  $S_k$  acting via the transformation

$$s_i = \underbrace{\mathbf{1} \otimes \dots \otimes \mathbf{1}}_{i-1 \text{ factors}} \otimes \sigma \otimes \underbrace{\mathbf{1} \otimes \dots \otimes \mathbf{1}}_{k-i-1 \text{ factors}}, \tag{2.2}$$

where  $\mathbf{1}$  is the identity map of  $V$ . Under this action,  $S_k$  commutes with  $SU_2$ . Thus, there is a representation  $\Phi_k : \mathbb{C}S_k \rightarrow \text{End}_{SU_2}(V^{\otimes k}) = Z_k(SU_2)$ ,  $\Phi_k(s_i) = s_i$ , of the group algebra  $\mathbb{C}S_k$ . However, this map is injective only for  $k = 1, 2$ .

The transformation  $e = 1 - \sigma \in \mathbb{C}S_2$  acts on  $V^{\otimes 2}$  by  $e(v \otimes w) = v \otimes w - w \otimes v$ , and  $\frac{1}{2}e$  is the idempotent projecting  $V^{\otimes 2}$  onto the antisymmetric tensors in  $V^{\otimes 2}$ , which form a copy of the trivial  $SU_2$ -module  $V(0)$ . The elements

$$e_i = \underbrace{\mathbf{1} \otimes \dots \otimes \mathbf{1}}_{i-1 \text{ factors}} \otimes e \otimes \underbrace{\mathbf{1} \otimes \dots \otimes \mathbf{1}}_{k-i-1 \text{ factors}}, \quad 1 \leq i \leq k-1 \tag{2.3}$$

generate  $Z_k(\text{SU}_2)$ . The map  $\Phi_k : \mathbb{C}\text{S}_k \rightarrow Z_k(\text{SU}_2)$  is surjective, and its image can be identified with the Temperley-Lieb algebra  $\text{TL}_k(2)$ .

Let  $x$  be an element of  $\mathbb{C}$  or a parameter. The *Temperley-Lieb algebra*  $\text{TL}_k(x)$  is the unital associative algebra with generators  $e_1, \dots, e_{k-1}$  and relations

$$\begin{aligned} \text{(TL1)} \quad & e_i^2 = xe_i, & 1 \leq i \leq k-1, \\ \text{(TL2)} \quad & e_i e_{i\pm 1} e_i = e_i, & 1 \leq i \leq k-1, \\ \text{(TL3)} \quad & e_i e_j = e_j e_i, & |i-j| > 1. \end{aligned} \tag{2.4}$$

Since  $Z_k(\text{SU}_2) = \text{End}_{\text{SU}_2}(\mathbb{V}^{\otimes k}) \cong \text{TL}_k(2)$  ( $x = 2$  here), we identify the generator  $e_i$  in  $\text{TL}_k(2)$  with the map in (2.3) and use the same notation for both. The set

$$\Lambda_k(\text{SU}_2) = \begin{cases} \{0, 2, \dots, k\} & \text{if } k \text{ is even,} \\ \{1, 3, \dots, k\} & \text{if } k \text{ is odd,} \end{cases}$$

indexes the irreducible  $\text{TL}_k(2)$ -modules. The number of walks of  $k$  steps from 0 to  $k - 2\ell$  on  $\mathcal{R}_V(\text{SU}_2)$  equals the number of walks of  $k$  steps from 0 to  $k - 2\ell$  on the nonnegative integers and is known to be the difference of adjacent binomial coefficients,

$$\left\langle \begin{matrix} k \\ \ell \end{matrix} \right\rangle := \binom{k}{\ell} - \binom{k}{\ell-1}$$

(see [9, Sec. 2.8] or [31]). For each  $k - 2\ell \in \Lambda_k(\text{SU}_2)$ , where  $\ell = 0, 1, \dots, \lfloor k/2 \rfloor$ , let  $\text{TL}_k^{(k-2\ell)} = Z_k^{(k-2\ell)}$  be the irreducible  $\text{TL}_k(2)$  module labeled by  $k - 2\ell$ . Then  $\text{TL}_k^{(k-2\ell)}$  has dimension  $\left\langle \begin{matrix} k \\ \ell \end{matrix} \right\rangle$ , and these modules are constructed explicitly in [31]. Moreover, by Schur-Weyl duality,

$$\begin{aligned} \mathbb{V}^{\otimes k} &\cong \bigoplus_{k-2\ell \in \Lambda_k(\text{SU}_2)} \left\langle \begin{matrix} k \\ \ell \end{matrix} \right\rangle \mathbb{V}(k-2\ell), && \text{as an } \text{SU}_2\text{-module,} \\ &\cong \bigoplus_{k-2\ell \in \Lambda_k(\text{SU}_2)} (k-2\ell+1) \text{TL}_k^{(k-2\ell)}, && \text{as a } \text{TL}_k(2)\text{-module,} \\ &\cong \bigoplus_{k-2\ell \in \Lambda_k(\text{SU}_2)} \left( \mathbb{V}(k-2\ell) \otimes \text{TL}_k^{(k-2\ell)} \right), && \text{as an } (\text{SU}_2, \text{TL}_k(2))\text{-bimodule.} \end{aligned}$$

The Temperley-Lieb algebras  $\text{TL}_k(x)$  have appeared in numerous contexts in mathematics and physics. They first arose in statistical mechanics [26], where the  $e_i$  occur as transfer matrices corresponding to addition of a single interaction between spins on a lattice. They have played a critical role in the work of Jones [12] and Wenzl [27] on subfactors of von Neumann algebras, where the  $e_i$  are the orthogonal projections arising in a tower of algebras (compare Sec. 2.3 below and [9, Chap. 2]). Since the Temperley-Lieb algebras are also quotients of the Hecke algebras of type A, they are related to integrable models, braid groups, quantum groups, categorification, and quantum computing. Their connections with braid groups have led to important invariants of knots and links such as the Jones and HOMFLYPT polynomials (see [14, 28], and the references cited therein).

**2.3. Jones basic construction.** Let  $G$  be a subgroup of  $SU_2$  such that  $G \neq \{I\}$  or  $\{\pm I\}$ . Any transformation that commutes with  $SU_2$  on  $V^{\otimes k}$  also commutes with  $G$ , so there is a reverse inclusion of centralizers  $TL_k(2) = \text{End}_{SU_2}(V^{\otimes k}) \subseteq \text{End}_G(V^{\otimes k}) = Z_k(G)$ . We identify the subalgebra of  $\text{End}_G(V^{\otimes k})$  generated by  $\mathbf{1}$  and the  $e_i$  in (2.3) with  $TL_k(2)$  and apply a construction due to Jones [12, Sec. 3] to locate additional generators for the centralizer algebra  $Z_k = Z_k(G) = \text{End}_G(V^{\otimes k})$  for each  $k$ . The construction uses the natural embedding of  $Z_k$  into  $Z_{k+1}$  given by  $a \mapsto a \otimes \mathbf{1}$ , which holds for any  $k \geq 1$ .

**Proposition 2.1.** (Cf. [1, Props. 1.26 and 1.27]) *Assume  $Z_k = Z_k(G)$  for all  $k \geq 0$ .*

- (a) *For  $a \in \text{End}(V^{\otimes(k+1)})$ , there is a unique  $b \in \text{End}(V^{\otimes k})$  so  $ae_k = (b \otimes \mathbf{1})e_k$ . Hence,  $Z_{k+1}e_k = Z_k e_k$ , where  $Z_k$  is identified with  $Z_k \otimes \mathbf{1}$ . The map  $Z_k \rightarrow Z_k e_k \subseteq Z_{k+1}$  given by  $a \mapsto ae_k$  is injective.*
- (b)  *$Z_k e_k Z_k = Z_{k+1} e_k Z_{k+1}$  is an ideal of  $Z_{k+1}$ .*

The Jones basic construction for  $Z_k \subseteq Z_{k+1}$  is based on the ideal  $Z_k e_k Z_k$  of  $Z_{k+1}$  and the fact that  $\Lambda_{k-1}(G) \subseteq \Lambda_{k+1}(G)$ , and it involves two key ideas:

- As a  $G$ -module  $V^{\otimes(k+1)} = V_{\text{old}}^{\otimes(k+1)} \oplus V_{\text{new}}^{\otimes(k+1)}$ , where

$$V_{\text{old}}^{\otimes(k+1)} = \bigoplus_{\lambda \in \Lambda_{k-1}(G)} m_{k+1}^\lambda G^\lambda \quad \text{and} \quad V_{\text{new}}^{\otimes(k+1)} = \bigoplus_{\lambda \in \Lambda_{k+1}(G) \setminus \Lambda_{k-1}(G)} m_{k+1}^\lambda G^\lambda. \tag{2.5}$$

Using the fact that  $\frac{1}{2}e_k$  corresponds to the projection onto the trivial  $G$ -module in the last two tensor slots of  $V^{\otimes(k+1)}$ , Wenzl ([29, Prop. 4.10], [30, Prop. 2.2]) proves that  $Z_k e_k Z_k \cong \text{End}_G(V_{\text{old}}^{\otimes(k+1)})$ , and therefore

$$\begin{aligned} Z_{k+1} &= \text{End}_G(V^{\otimes(k+1)}) \cong \text{End}_G(V_{\text{old}}^{\otimes(k+1)}) \oplus \text{End}_G(V_{\text{new}}^{\otimes(k+1)}) \\ &\cong Z_k e_k Z_k \oplus \text{End}_G(V_{\text{new}}^{\otimes(k+1)}). \end{aligned} \tag{2.6}$$

- There is an algebra isomorphism  $Z_{k-1} \cong e_k Z_k e_k$  sending  $a \in Z_{k-1}$  to  $e_k a e_k = 2ae_k = 2e_k a \in Z_{k+1}$ . Viewing  $Z_k e_k$  as a module for both  $Z_k e_k Z_k$  and  $Z_{k-1} \cong e_k Z_k e_k$  by multiplication on the left and right, respectively, we have that these actions commute and centralize one another:

$$Z_k e_k Z_k \cong \text{End}_{Z_{k-1}}(Z_k e_k) \quad \text{and} \quad Z_{k-1} \cong \text{End}_{Z_k e_k Z_k}(Z_k e_k).$$

Schur-Weyl duality implies that the simple summands of the semisimple algebras  $Z_k e_k Z_k$  and  $Z_{k-1}$  (and also their irreducible modules) can be indexed by the same set  $\Lambda_{k-1}(G)$ .

By restriction, the irreducible  $Z_k$ -module  $Z_k^\lambda$  becomes a  $Z_{k-1}$ -module,

$$\text{Res}_{Z_{k-1}}^{Z_k}(Z_k^\lambda) = \bigoplus_{\mu \in \Lambda_{k-1}(G)} \Theta_{\lambda, \mu} Z_{k-1}^\mu.$$

The multiplicity  $\Theta_{\lambda, \mu}$  of  $Z_{k-1}^\mu$  in  $Z_k^\lambda$  is 1 or 0 for all groups considered here. The inclusion matrix for  $Z_{k-1} \subseteq Z_k$  is  $\Theta = (\Theta_{\lambda, \mu})$ , and for  $\text{End}_{Z_k}(Z_k e_k) \subseteq \text{End}_{Z_{k-1}}(Z_k e_k)$  is the transpose  $\Theta^t$ , which allows us to conclude the following:

- For the tower  $\mathbb{C} = Z_0 \subset Z_1 \subset \dots \subset Z_k \subset Z_{k+1} \subset \dots$  of centralizer algebras, the edges in the Bratteli diagram between levels  $k$  and  $k + 1$  corresponding to  $Z_k e_k Z_k \subseteq Z_{k+1}$  are the reflection over level  $k$  of the edges between  $k - 1$  and  $k$  corresponding to  $Z_{k-1} \subseteq Z_k$ .
- The edges which are NOT obtained by reflection give a copy of the representation graph  $\mathcal{R}_V(G)$  embedded in the Bratteli diagram  $\mathcal{B}_V(G)$ .

**Remark 2.2.** For  $SU_2$ , the shaded edges in Figure 2.1 above indicate the embedding of the representation graph  $\mathcal{R}_V(SU_2)$  in  $\mathcal{B}_V(SU_2)$ . For the finite subgroups  $G$  of  $SU_2$ , the top levels of the Bratteli diagrams are displayed in Section A.2 of the Appendix, and the shaded edges give the representation graph  $\mathcal{R}_V(G)$  (the corresponding affine Dynkin diagram) embedded in  $\mathcal{B}_V(G)$ . All other edges in  $\mathcal{B}_V(G)$  are obtained by the Jones basic construction and reflection. A similar phenomenon occurs in the theory of subfactors, where the principal graph embeds in the Bratteli diagram (see for example [14]).

**2.4. Jones-Wenzl projection maps for  $SU_2$ .** The Jones-Wenzl idempotents in  $TL_k(2)$  are defined recursively by setting  $f_1 = 1$  and letting

$$f_n = f_{n-1} - \frac{n-1}{n} f_{n-1} e_{n-1} f_{n-1}, \quad 1 < n \leq k. \tag{2.7}$$

These idempotents satisfy the following properties (see [5, 7, 12, 27]):

- (JW1)  $f_n^2 = f_n, \quad 1 \leq n \leq k,$
- (JW2)  $e_i f_n = f_n e_i = 0, \quad 1 \leq i < n \leq k,$
- (JW3)  $e_i f_n = f_n e_i, \quad 1 \leq n < i \leq k - 1,$
- (JW4)  $e_n f_n e_n = \frac{n+1}{n} f_{n-1} e_n, \quad 1 \leq n \leq k - 1,$
- (JW5)  $1 - f_n \in \langle e_1, \dots, e_{n-1} \rangle,$
- (JW6)  $f_m f_n = f_n f_m \quad 1 \leq m, n \leq k,$

where  $\langle e_1, \dots, e_{n-1} \rangle$  stands for the subalgebra of  $TL_k(2)$  generated by  $e_1, \dots, e_{n-1}$ . Expressions for the  $f_n$  as linear combinations of words in  $e_1, \dots, e_{k-1}$  can be found in [7].

The irreducible  $SU_2$ -module  $V(k)$  appears in  $V^{\otimes k}$  with multiplicity 1, and it does not occur as a summand of  $V^{\otimes \ell}$  for any  $\ell < k$  (i.e. node  $k$  is reached for the first time after  $k$  steps on  $\mathcal{R}_V(SU_2)$ ). It is well known [8, Sec. 11.1] that the totally symmetric tensors  $S(V^{\otimes k})$  of  $V^{\otimes k}$  satisfy  $S(V^{\otimes k}) \cong V(k)$  as an  $SU_2$ -module, and that  $f_k(V^{\otimes k}) = S(V^{\otimes k})$  ([7, Prop. 1.3, Cor. 1.4]). In particular, if  $\{v_{-1}, v_1\}$  is a basis for  $V$ , then  $S(V^{\otimes 2}) = \text{span}_{\mathbb{C}}\{v_{-1} \otimes v_{-1}, v_{-1} \otimes v_1 + v_1 \otimes v_{-1}, v_1 \otimes v_1\} \cong V(2)$ , and  $f_2 = 1 - \frac{1}{2}e_1$  projects  $V^{\otimes 2}$  onto that space.

**2.5. Projection maps for  $G$ .** Assume  $G$  is a subgroup of  $SU_2$ ,  $Z_0 = \mathbb{C}$ , and  $Z_k = Z_k(G)$  for all  $k \geq 1$ . Let  $d^\lambda = \dim G^\lambda$  for  $\lambda \in \Lambda(G)$ . The next proposition describes the interaction between idempotents for different tensor powers and gives a recursive procedure for constructing idempotents. This result uses the tower of embeddings  $\dots \subset Z_k \subset Z_{k+1} \subset Z_{k+2} \subset \dots$  and the containment  $TL_k(2) \subseteq Z_k$ .

**Proposition 2.3.** [1, Lem.1.40, Prop. 1.42].

- (a) For  $\lambda \in \Lambda_k(G)$ , assume  $G^\lambda$  is a summand of  $V_{\text{new}}^{\otimes k}$ , and  $G^\lambda \otimes V = \bigoplus_i G^{\mu_i}$  in  $V^{\otimes(k+1)}$ . Let  $f_\lambda$  and  $f_{\mu_i}$  be the corresponding projection maps. Then the following hold for all  $\mu_i$  such that  $G^{\mu_i} \subseteq V_{\text{new}}^{\otimes(k+1)}$ , i.e. for all  $\mu_i \in \Lambda_{k+1}(G) \setminus \Lambda_{k-1}(G)$ :

- (i)  $f_\lambda f_{\mu_i} = f_{\mu_i} = f_{\mu_i} f_\lambda$ ;
  - (ii)  $f_{\mu_i}$  commutes with  $e_j$  for  $j > k + 1$ ;
  - (iii)  $e_{k+1} f_{\mu_i} e_{k+1} = \frac{d^{\mu_i}}{d^\lambda} f_\lambda e_{k+1}$ .
- (b) Assume  $\mu = \mu_i \in \Lambda_{k+1}(G) \setminus \Lambda_{k-1}(G)$  for some  $i$ , and  $G^\mu \otimes V = G^\lambda \oplus G^\nu$ , where  $\nu \in \Lambda_{k+2}(G) \setminus \Lambda_k(G)$ . Let

$$f_\nu = f_\mu - \frac{d^\lambda}{d^\mu} f_\mu e_{k+1} f_\mu. \tag{2.8}$$

Then

- (i)  $f_\nu$  is an idempotent in  $Z_{k+2}$  and  $f_\mu f_\nu = f_\nu = f_\nu f_\mu$ ;
- (ii)  $f_\nu$  commutes with  $e_j$  for  $j > k + 2$ ;
- (iii)  $e_{k+2} f_\nu e_{k+2} = \frac{d^\nu}{d^\mu} f_\mu e_{k+2}$ ;
- (iv)  $f_\nu$  projects  $V^{\otimes(k+2)}$  onto  $G^\nu$ .

**2.6. Projections related to branch nodes and generators.** Let  $G$  be one of the finite subgroups  $C_n, D_n, T, O, I$  of  $SU_2$ . (In [1], the infinite cyclic and dihedral subgroups  $C_\infty$  and  $D_\infty$  of  $SU_2$  are also discussed, but we will not consider them here.) A *branch node* in the representation graph  $\mathcal{R}_V(G)$  is any vertex of degree greater than 2. Let  $br(G)$  denote the branch node in  $\mathcal{R}_V(G)$ , and in the case of  $D_n (n > 2)$ , which has 2 branch nodes, set  $br(D_n) = 1$ . We consider the affine node of the cyclic graph  $\mathcal{R}_V(C_n)$  to be the branch node, so that  $br(C_n) = 0$ . We say that the *diameter* of  $\mathcal{R}_V(G)$ , denoted by  $di(G)$ , is the maximum distance between 0 and any vertex  $\lambda \in \Lambda(G)$  in  $\mathcal{R}_V(G)$ . For  $G = C_n$ , we set  $di(G) = \tilde{n}$ , where  $\tilde{n}$  is as in (2.9).

$G$	$SU_2$	$C_n$	$D_n$	$T$	$O$	$I$	where $\tilde{n} = \begin{cases} \frac{1}{2}n & \text{(if } n \text{ is even),} \\ n & \text{(if } n \text{ is odd).} \end{cases}$	(2.9)
$di(G)$	$\infty$	$\tilde{n}$	$n$	$4$	$6$	$7$		
$br(G)$	$-$	$0$	$1$	$2$	$3$	$5$		

When  $k \leq b := br(G)$ , then  $V_{new}^{\otimes k} = G^{(k)} = V(k)$ , the irreducible  $SU_2$ -module. In this case, the projection of  $V^{\otimes k}$  onto  $G^{(k)}$  is given by  $f_{(k)} := f_k$ , where  $f_k$  is the Jones-Wenzl idempotent. The irreducible  $SU_2$ -module  $V(b + 1)$  is reducible as a  $G$ -module. Let  $V_{new}^{\otimes(b+1)} = V(b + 1) = \bigoplus_j G^{\mu_j}$  be its decomposition into irreducible  $G$ -modules (each summand corresponds to a node connected to the branch node), and let  $f_{b+1} = \sum_j f_{\mu_j}$  be the corresponding decomposition of the Jones-Wenzl idempotent  $f_{b+1}$  into minimal orthogonal idempotents that commute with  $G$  and project  $V_{new}^{\otimes(b+1)}$  onto the irreducible  $G$ -summands  $G^{\mu_j}$ .

For finite subgroups  $G$ , the idempotents  $f_{\mu_j}$  can be constructed as in [8, (2.32)] using the corresponding irreducible characters  $\chi_{\mu_j}$  of the group,

$$f_{\mu_j} = \frac{\dim G^{\mu_j}}{|G|} \sum_{g \in G} \overline{\chi_{\mu_j}(g)} g^{\otimes(b+1)}, \tag{2.10}$$

where  $g^{\otimes(b+1)}$  is the matrix of  $g$  on  $V^{\otimes(b+1)}$  and “ $\bar{\phantom{x}}$ ” denotes complex conjugate.

Proposition 2.3 enables us to construct the other projection maps. We illustrate this for the binary octahedral group  $O$ . Our numbering of the nodes in  $\mathcal{R}_V(O)$  (the Dynkin diagram  $\hat{E}_7$ ) is that given in Section A.1 of the Appendix.

**Example 2.4.** Assume  $G = \mathbf{O}$ , and let  $b = \text{br}(\mathbf{O}) = 3$ . Set  $f_{(k)} = f_k$  for  $1 \leq k \leq 3$ , where  $f_k$  is the Jones-Wenzl idempotent given in (2.7). For  $\mu = (b + 1) = (4)$ , let  $\mu' = (4')$ , and let  $f_\mu$  (resp.  $f_{\mu'}$ ) denote the projection map of  $V^{\otimes 4}$  onto  $\mathbf{O}^\mu = \mathbf{O}^{(4)}$  (resp. onto  $\mathbf{O}^{\mu'} = \mathbf{O}^{(4')}$ ), which can be constructed using (2.10) if an explicit expression is needed. Then the Jones-Wenzl idempotent  $f_4$  decomposes as  $f_4 = f_{(4)} + f_{(4')}$ , and correspondingly  $f_4(V^{\otimes 4}) = V(4)$  decomposes as an  $\mathbf{O}$ -module  $V(4) = \mathbf{O}^{(4)} \oplus \mathbf{O}^{(4')}$ . The idempotents projecting onto the other irreducible  $\mathbf{O}$ -modules can be computed from these idempotents by applying Proposition 2.3 (b) with the following choices:  $\lambda = (3), \mu = (4), \nu = (5)$  produces  $f_{(5)} = f_{(4)} - \frac{4}{3}f_{(4)}e_4f_{(4)}$ , which projects  $V^{\otimes 5}$  onto  $\mathbf{O}^{(5)}$ , and then  $\lambda = (4), \mu = (5), \nu = (6)$  gives  $f_{(6)} = f_{(5)} - \frac{3}{2}f_{(5)}e_5f_{(5)}$ , which projects  $V^{\otimes 6}$  onto  $\mathbf{O}^{(6)}$ .

The idempotents constructed above enable us to identify generators for  $Z_k$ . The only case not covered by the next theorem ( $G = \mathbf{C}_n, k \geq \tilde{n} - 1$  in (d)) requires some additional notation and is omitted here, but it can be found in [1].

**Theorem 2.5.** [1, Thm. 1.45] *Let  $G, \text{br}(G)$ , and  $\text{di}(G)$  be as in (2.9), and let  $Z_k = Z_k(G)$ . Then  $Z_1 = \mathbb{C}1 \cong Z_0$ . Moreover,*

- (a) *if  $1 \leq k \leq \text{br}(G)$ , then  $Z_k = \text{TL}_k(2)$ ;*
- (b) *if  $k \geq \text{di}(G)$ , then  $Z_{k+1} = Z_k e_k Z_k$ ;*
- (c) *if  $k = \text{br}(G)$  and  $G \neq \mathbf{D}_2$ , then  $Z_{k+1} = \langle Z_k, e_k, f_\mu \rangle$ , where  $\mu$  is either of the two elements in  $\Lambda_{b+1}(G) \setminus \Lambda_{b-1}(G)$  for  $b = \text{br}(G)$ , and  $f_\mu$  is the projection of  $V_{\text{new}}^{\otimes (b+1)}$  onto  $G^\mu$ . If  $G = \mathbf{D}_2$ , then  $Z_2 = \langle Z_1, e_1, f_{\mu_1}, f_{\mu_2} \rangle$ , where  $\mu_1, \mu_2 \in \{(0'), (2), (2')\}$  and  $\mu_1 \neq \mu_2$ .*
- (d) *if  $\text{br}(G) < k$ , (with  $k < n - 1$  when  $G = \mathbf{D}_n$ , and  $k < \tilde{n} - 1$  when  $G = \mathbf{C}_n$ ), then  $Z_{k+1} = \langle Z_k, e_k \rangle$*
- (e) *if  $G = \mathbf{D}_n$  and  $n > 2$ , then  $Z_n = \langle Z_{n-1}, e_{n-1}, f_\nu \rangle$ , where  $\nu \in \{(n), (n')\}$  and  $f_\nu$  is the projection of  $V_{\text{new}}^{\otimes n}$  onto  $G^\nu$ , and  $Z_{k+1} = \langle Z_k, e_k \rangle$  for all  $k \geq n$ .*

In [1] we give a set of generators and relations for the algebras  $Z_k(G)$ , but leave open the problem of finding a presentation for these algebras.

### 3. Cyclic and Binary Dihedral Subgroups of $\text{SU}_2$

**3.1. The centralizer algebra  $Z_k(\mathbf{C}_n)$ .** Let

$$g = \begin{pmatrix} \zeta^{-1} & 0 \\ 0 & \zeta \end{pmatrix} \quad \text{and} \quad h = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \tag{3.1}$$

where  $i = \sqrt{-1}$  and  $\zeta \in \mathbb{C}$ . Set  $\zeta = \zeta_n$ , a primitive  $n$ th root of unity, and let  $\mathbf{C}_n$  be the cyclic subgroup of  $\text{SU}_2$  generated by  $g$ . The irreducible modules for  $\mathbf{C}_n$  are all one-dimensional and are given by  $\mathbf{C}_n^{(\ell)} = \mathbb{C}v_\ell$  for  $\ell = 0, 1, \dots, n - 1$ , where  $gv_\ell = \zeta^\ell v_\ell$ . Thus,  $\Lambda(\mathbf{C}_n) = \{0, 1, \dots, n - 1\}$ , and  $\mathbf{C}_n^{(j)} \cong \mathbf{C}_n^{(\ell)}$  whenever  $j \equiv \ell \pmod n$ . The natural  $\mathbf{C}_n$ -module  $V$  of  $2 \times 1$  column vectors, which  $\mathbf{C}_n$  acts on by matrix multiplication, can be identified with the module  $\mathbf{C}_n^{(-1)} \oplus \mathbf{C}_n^{(1)} = \mathbf{C}_n^{(n-1)} \oplus \mathbf{C}_n^{(1)}$ . Since  $\mathbf{C}_n^{(\ell)} \otimes \mathbf{C}_n^{(m)} \cong \mathbf{C}_n^{(\ell+m)}$ , a modular Clebsch-Gordan formula holds:

$$\mathbf{C}_n^{(\ell)} \otimes V = \mathbf{C}_n^{(\ell-1)} \oplus \mathbf{C}_n^{(\ell+1)} \quad (\text{superscripts mod } n). \tag{3.2}$$

We assume that  $\{v_{-1} = (1, 0)^t, v_1 = (0, 1)^t\}$  is the standard basis for  $V$ . Let  $r = (r_1, \dots, r_k) \in \{-1, 1\}^k$ , and set  $|r| = |\{r_i \mid r_i = -1\}|$ . Corresponding to  $r$  is the vector  $v_r = v_{r_1} \otimes \dots \otimes v_{r_k} \in V^{\otimes k}$ , and  $gv_r = \zeta^{k-2|r|}v_r$ . For two such  $k$ -tuples  $r$  and  $s$ ,

$$k - 2|r| \equiv k - 2|s| \pmod n \iff |r| \equiv |s| \pmod{\tilde{n}}, \tag{3.3}$$

where  $\tilde{n} = n$  if  $n$  is odd and  $\tilde{n} = \frac{1}{2}n$  if  $n$  is even, as in (2.9).

The Bratteli diagram for tensoring with  $V$  in the cyclic case corresponds to Pascal’s triangle on a cylinder of diameter  $\tilde{n}$ , and this behavior is reflected in the next result. (Compare  $\mathcal{B}_V(\mathbf{C}_5)$  in Figure A.1 and  $\mathcal{B}_V(\mathbf{C}_{10})$  in Figure A.2 in the Appendix.)

**Theorem 3.1.** [1, Thms. 2.7 and 2.16] *For  $k \geq 1, n \geq 3$ , and  $Z_k(\mathbf{C}_n) = \text{End}_{\mathbf{C}_n}(V^{\otimes k})$ ,*

(a)  $\{E_{r,s} \mid r, s \in \{-1, 1\}^k, |r| \equiv |s| \pmod{\tilde{n}}\}$  is a basis for  $Z_k(\mathbf{C}_n)$ , where  $E_{r,s}$  is the linear transformation defined by  $E_{r,s}v_t = \delta_{s,t}v_r$ . In particular, if  $n = 2\tilde{n}$  and  $\tilde{n}$  is odd, then  $Z_k(\mathbf{C}_n) \cong Z_k(\mathbf{C}_{\tilde{n}})$ .

(b) The dimension of  $Z_k(\mathbf{C}_n)$  is the coefficient of  $z^k$  in  $(1 + z)^{2k}|_{z^{\tilde{n}}=1}$ ; hence, it is given by

$$\dim Z_k(\mathbf{C}_n) = \sum_{\substack{0 \leq a, b \leq k \\ a \equiv b \pmod{\tilde{n}}}} \binom{k}{a} \binom{k}{b}.$$

(c) The irreducible  $Z_k(\mathbf{C}_n)$ -modules are in bijection with the elements of the set  $\Lambda_k(\mathbf{C}_n) = \{\ell \in \{0, 1, \dots, n - 1\} \mid \ell \equiv k - 2a_\ell \pmod n \text{ for some } 0 \leq a_\ell \leq k\}$  ( $a_\ell$  is always taken to be the minimal value in  $\{0, 1, \dots, k\}$  with that property).

(d) For  $\ell \in \Lambda_k(\mathbf{C}_n)$ , the irreducible  $Z_k(\mathbf{C}_n)$ -module  $Z_k^{(\ell)}$  is the  $\zeta^\ell$ -eigenspace of  $g$ ,

$$Z_k^{(\ell)} = \text{span}_{\mathbf{C}}\{v_r \in V^{\otimes k} \mid k - 2|r| \equiv \ell \pmod n\} = \text{span}_{\mathbf{C}}\{v_r \in V^{\otimes k} \mid |r| \equiv a_\ell \pmod{\tilde{n}}\},$$

and  $\dim Z_k^{(\ell)} = \sum_{\substack{0 \leq b \leq k \\ b \equiv a_\ell \pmod{\tilde{n}}}} \binom{k}{b}$ , which is the coefficient of  $z^{a_\ell}$  in  $(1 + z)^k|_{z^{\tilde{n}}=1}$ .

**Example 3.2.** Assume  $k = 5$  and  $n = 6$ , so that  $\tilde{n} = 3$ . Then  $\Lambda_5(\mathbf{C}_6) = \{1, 3, 5\}$ , and for  $\ell = 1, 3, 5$ , we have  $a_\ell = 2, 1, 0$  respectively. Since

$$(1 + z)^5|_{z^3=1} = 1 + 5z + 10z^2 + 10 + 5z + z^2,$$

it follows that  $\dim Z_5^{(1)} = 11, \dim Z_5^{(3)} = 10$ , and  $\dim Z_5^{(5)} = 11$ . Now  $k = 5 \equiv 2 \pmod 3$ , and the coefficient of  $z^2$  in

$$(1 + z)^{10}|_{z^3=1} = 1 + 10z + 45z^2 + 120 + 210z + 252z^2 + 210 + 120z + 45z^2 + 10 + z$$

is  $45 + 252 + 45 = 342$ , so  $\dim Z_5(\mathbf{C}_6) = 342 = 11^2 + 10^2 + 11^2$ .

**3.2. The centralizer algebra  $Z_k(\mathbf{D}_n)$ .** Next we discuss the centralizer algebra  $Z_k(\mathbf{D}_n) = \text{End}_{\mathbf{D}_n}(V^{\otimes k})$  for  $V = \mathbb{C}^2$ . The binary dihedral group  $\mathbf{D}_n$  is generated by the elements  $g, h$  in (3.1), where now  $\zeta = \zeta_{2n}$ , a primitive  $2n$ th root of 1. The element  $g$  generates a cyclic subgroup  $\mathbf{C}_{2n}$  of  $\mathbf{D}_n$  of order  $2n$ , which implies that  $\text{End}_{\mathbf{D}_n}(V^{\otimes k}) = Z_k(\mathbf{D}_n) \subseteq Z_k(\mathbf{C}_{2n}) = \text{End}_{\mathbf{C}_{2n}}(V^{\otimes k})$ . We will exploit that in our considerations.

To describe a basis for the algebra  $Z_k(\mathbf{D}_n)$ , we impose the following order on  $k$ -tuples in  $\{-1, 1\}^k$ . Say  $r \succeq s$  if  $|r| \leq |s|$ , and if  $|r| = |s|$ , then  $r$  is greater than or equal to  $s$  in the lexicographic order coming from the relation  $1 > -1$ . It follows from Theorem 3.1 (a) that a basis for  $Z_k(\mathbf{C}_{2n})$  is given by  $\mathcal{B}^k(\mathbf{C}_{2n}) = \{E_{r,s} \mid r, s \in \{-1, 1\}^k, |r| \equiv |s| \pmod n\}$ , where  $|r|, |s| \in \{0, 1, \dots, k\}$ . Now

$$|r| \equiv |s| \pmod n \iff |-r| = k - |r| \equiv k - |s| = |-s| \pmod n,$$

so  $E_{r,s} \in \mathcal{B}^k(\mathbf{C}_{2n})$  if and only if  $E_{-r,-s} \in \mathcal{B}^k(\mathbf{C}_{2n})$ .

**Theorem 3.3** ([1, Thm. 3.13]). *Let  $Z_k(\mathbf{D}_n) = \text{End}_{\mathbf{D}_n}(\mathbb{V}^{\otimes k})$  for  $\mathbb{V} = \mathbb{C}^2$  and  $n \geq 2$ . Then*

(a)  $Z_k(\mathbf{D}_n) = \{X \in Z_k(\mathbf{C}_{2n}) \mid hX = Xh\}$ , where the matrix  $h$  is given in (3.1).

(b) A basis for  $Z_k(\mathbf{D}_n)$  is the set

$$\mathcal{B}^k(\mathbf{D}_n) = \{E_{r,s} + E_{-r,-s} \mid r, s \in \{-1, 1\}^k, r \succ -r, |r| \equiv |s| \pmod n\}. \tag{3.4}$$

(c) The dimension of  $Z_k(\mathbf{D}_n)$  is given by

$$\begin{aligned} \dim Z_k(\mathbf{D}_n) &= \frac{1}{2} \dim Z_k(\mathbf{C}_{2n}) = \frac{1}{2} \sum_{\substack{0 \leq a, b \leq k \\ a \equiv b \pmod n}} \binom{k}{a} \binom{k}{b} \\ &= \frac{1}{2} (\text{coefficient of } z^k \text{ in } (1+z)^{2k} \Big|_{z^n=1}). \end{aligned} \tag{3.5}$$

Results on the irreducible modules for  $Z_k(\mathbf{D}_n)$  can be found in [1, Thm. 3.29]. In [1, Remark 2.13 and Sec. 3.6], we give realizations of  $Z_k(\mathbf{C}_n)$  and  $Z_k(\mathbf{D}_n)$  as diagram algebras.

### 4. Centralizer algebras in the exceptional cases

In this section, we establish connections between the exceptional centralizer algebras and partition algebras.

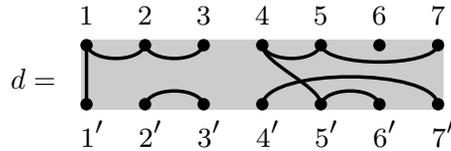
**4.1. Partition algebras.** The *partition algebras*  $P_k(x)$ ,  $x \in \mathbb{C}$ , were introduced by Martin [17] as generalized Temperley-Lieb algebras to study the Potts lattice model of interacting spins in statistical mechanics. As shown by Jones [13], there is a Schur-Weyl duality relation between the partition algebra  $P_k(n)$  and the symmetric group  $S_n$  as centralizer algebras of each other on the  $k$ -fold tensor power  $M_n^{\otimes k}$  of the  $n$ -dimensional permutation module  $M_n$  of  $S_n$ , where  $S_n$  acts diagonally on  $M_n^{\otimes k}$ . More specifically, there is a surjective algebra homomorphism

$$P_k(n) \rightarrow Z_k(S_n) := \text{End}_{S_n}(M_n^{\otimes k}) = \{X \in \text{End}(M_n^{\otimes k}) \mid X\sigma = \sigma X \ \forall \sigma \in S_n\},$$

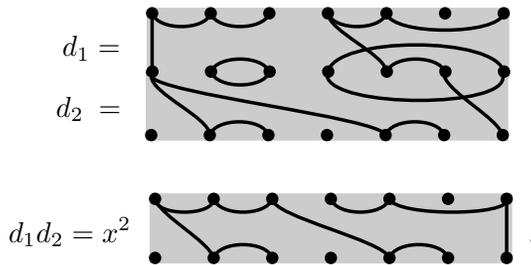
which is an isomorphism when  $n \geq 2k$ . Schur-Weyl duality enables results on the partition algebras  $P_k(n)$  and the symmetric groups  $S_n$  to be transported in either direction. On one hand, the papers [10, 18, 19] have exploited the representation theory of the symmetric groups to establish results about partition algebras and their representations; while on the

other, the recent work of Bowman, DeVisscher, and Orellana [4] has applied the representation theory of the partition algebras  $P_k(n)$  to derive information about the long-standing problem of determining the Kronecker coefficients that arise when the tensor product of two irreducible  $S_n$ -modules is decomposed into irreducible  $S_n$ -summands.

The partition algebra  $P_k(x)$  has a basis over  $\mathbb{C}$  indexed by set partitions of the set  $\{1, 2, \dots, k, 1', 2', \dots, k'\}$  into nonempty blocks. An example of such a set partition for  $k = 7$  is  $\{\{1, 2, 3, 1'\}, \{4, 5, 7, 5', 6'\}, \{6\}, \{2', 3'\}, \{4', 7'\}\}$ , which has 5 blocks. The set partitions can be represented as diagrams having two rows of  $k$  nodes each, with the top nodes indexed by  $1, 2, \dots, k$  and the bottom nodes indexed by  $1', 2', \dots, k'$  from left to right. Nodes are connected by an edge if they lie in the same block. To the set partition above, we associate the diagram



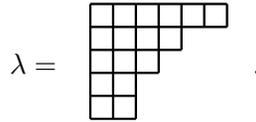
The way the edges are drawn is immaterial; what matters is that the nodes in each block of the set partition are connected, and there are no edges between nodes belonging to different blocks. Multiplication of two diagrams  $d_1, d_2$  is accomplished by placing  $d_1$  above  $d_2$ ; identifying the nodes in the bottom row of  $d_1$  with those in the top row of  $d_2$ ; concatenating the edges; deleting all connected components that lie entirely in the middle row of the joined diagrams; and multiplying the concatenated diagram by a factor of  $x$  for each deleted middle row component. For example,



The *Stirling number*  $\left\{ \begin{matrix} k \\ j \end{matrix} \right\}$  of the 2nd kind counts the number of ways to partition a set of  $k$  elements into  $j$  nonempty blocks. In particular,  $\left\{ \begin{matrix} k \\ 0 \end{matrix} \right\} = 0$  for all  $k \geq 1$ , and  $\left\{ \begin{matrix} k \\ j \end{matrix} \right\} = 0$  if  $j > k$ . By convention,  $\left\{ \begin{matrix} 0 \\ 0 \end{matrix} \right\} = 1$ . The sum  $\sum_{j=0}^k \left\{ \begin{matrix} k \\ j \end{matrix} \right\} = B(k)$ , where  $B(k)$  is the  $k$ th *Bell number*. Identifying  $P_0(x)$  with  $\mathbb{C}$ , we have

$$\dim P_k(x) = B(2k) \quad \text{for all } k \geq 0. \tag{4.1}$$

The irreducible modules  $S_n^\lambda$  for  $S_n$  are indexed by partitions  $\lambda$  of  $n$ , written  $\lambda \vdash n$ . Thus,  $\lambda$  is a sequence  $(\lambda_1, \dots, \lambda_\ell)$  of weakly decreasing nonnegative integers such that the sum  $|\lambda| := \sum_{i=1}^\ell \lambda_i = n$ . We identify  $\lambda$  with its Young diagram, so that for  $\lambda = (6, 4, 3, 2, 2) \vdash 17$ ,



The conjugate partition obtained by interchanging the rows and columns of  $\lambda$  is  $\lambda^t = (5, 5, 3, 2, 1, 1)$  for this example.

The permutation module for  $S_n$  decomposes into irreducible summands according to  $M_n = S_n^{(n)} \oplus S_n^{(n-1,1)}$ , where the module  $S_n^{(n)}$  indexed by the one-part partition  $(n)$  is the trivial  $S_n$ -module, and  $S_n^{(n-1,1)}$  is the  $(n - 1)$ -dimensional reflection module. Tensoring with  $M_n$  obeys the following rule,

$$S_n^\mu \otimes M_n = \bigoplus_{\lambda=(\mu-\square)+\square} a_{\mu,\lambda} S_n^\lambda, \tag{4.2}$$

which says, “First remove a box from  $\mu$  to get a partition  $\nu$  of  $n - 1$ , and then add a box to  $\nu$  to get a partition  $\lambda$  of  $n$  in all possible ways.” It is derived from the restriction and induction formulas

$$\text{Res}_{S_{n-1}}^{S_n}(S_n^\mu) = \bigoplus_{\nu=\mu-\square} S_{n-1}^\nu \quad \text{and} \quad \text{Ind}_{S_{n-1}}^{S_n}(S_{n-1}^\nu) = \bigoplus_{\lambda=\nu+\square} S_n^\lambda.$$

This two-step process inspired the introduction of the intermediate centralizer algebras  $Z_{k+\frac{1}{2}}(S_n) := \text{End}_{S_{n-1}}(M_n^{\otimes k})$ , which play an important role in understanding the structure and representation theory of partition algebras (see for example, [10, 20]). The diagrams in  $P_{k+1}(x)$  corresponding to set partitions which have  $k + 1$  and  $(k + 1)'$  in the same block form a subalgebra  $P_{k+\frac{1}{2}}(x)$  of  $P_{k+1}(x)$ , and there is a surjective algebra homomorphism  $P_{k+\frac{1}{2}}(n) \rightarrow Z_{k+\frac{1}{2}}(S_n)$ , which is an isomorphism if  $n \geq 2k$ .

Irreducible modules for  $P_k(x)$  and  $P_{k+\frac{1}{2}}(x)$  are labeled by partitions  $\varrho$  of  $r$ , where  $r$  is an integer satisfying  $0 \leq r \leq k$ . Since the irreducible modules for  $S_n$  are indexed by the partitions of  $n$ , Schur-Weyl duality implies that the irreducible modules for  $Z_k(S_n)$  are also indexed by partitions  $\lambda$  of  $n$ , and those for  $Z_{k+\frac{1}{2}}(S_n)$  by partitions  $\nu$  of  $n - 1$ . The partition  $\varrho$  that results from removing the first part of  $\lambda$  or  $\nu$  must satisfy  $0 \leq |\varrho| \leq k$ .

**4.2. The octahedral and tetrahedral cases.** Viewing the exceptional binary polyhedral groups  $G = \mathbf{T}, \mathbf{O}, \mathbf{I}$  as subgroups of  $SU_2$ , we have that the center of  $G$  is  $\mathcal{Z}(G) = \{\pm \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\}$ , and  $G/\mathcal{Z}(G) \cong \mathbf{A}_4, \mathbf{S}_4, \mathbf{A}_5$ , respectively, where  $\mathbf{A}_n \subset S_n$  is the alternating subgroup of  $S_n$  of even permutations. As the center  $\mathcal{Z}(G)$  acts trivially on even tensor powers  $V^{\otimes 2\ell}$  of  $V = \mathbb{C}^2$ , there is an induced representation  $G/\mathcal{Z}(G) \rightarrow \text{End}(V^{\otimes 2\ell})$ . In particular,  $V^{\otimes 2}$  as a module for  $\mathbf{T}/\mathcal{Z}(\mathbf{T}) \cong \mathbf{A}_4$  is the permutation module  $M_4 \cong S_4^{(4)} \oplus S_4^{(3,1)}$  regarded as an  $\mathbf{A}_4$ -module by restriction. For  $\mathbf{O}/\mathcal{Z}(\mathbf{O}) \cong \mathbf{S}_4$ ,  $V^{\otimes 2} \cong S_4^{(4)} \oplus S_4^{(2,1,1)}$ ; and for  $\mathbf{I}/\mathcal{Z}(\mathbf{I})$ ,  $V^{\otimes 2}$  is the sum of the trivial  $\mathbf{A}_5$ -module and a 3-dimensional irreducible  $\mathbf{A}_5$ -module.

The induction and restriction rules for the pairs  $(\mathbf{T}, \mathbf{C}_6)$  and  $(\mathbf{A}_4, \mathbf{A}_3)$  are the same, as they also are for the pairs  $(\mathbf{O}, \mathbf{D}_3)$  and  $(\mathbf{S}_4, \mathbf{S}_3)$ , where  $\mathbf{C}_6$  is a cyclic subgroup of  $\mathbf{T}$  of order 6 and  $\mathbf{D}_3$  is a binary dihedral subgroup of  $\mathbf{O}$  of order 12. The representation graphs for the  $S_4$ -modules  $V^{\otimes 2}$  and  $M_4$  are exactly the same except for interchanging the labels on the  $S_4$ -irreducible modules  $S_4^{(3,1)}$  and  $S_4^{(2,1,1)}$ . These results imply the following, where we assume the 0th tensor power is the trivial module for the group.

**Theorem 4.1** ([2]). For  $\ell \in \frac{1}{2}\mathbb{Z}_{>0}$ ,

- (a)  $Z_{2\ell}(\mathbf{O}) = \text{End}_{\mathbf{O}}(V^{\otimes 2\ell}) \cong Z_{\ell}(\mathbf{S}_4) = \begin{cases} \text{End}_{\mathbf{S}_4}(M_4^{\otimes \ell}) & \text{if } \ell \in \mathbb{Z}_{>0} \\ \text{End}_{\mathbf{S}_3}(M_4^{\otimes(\ell-\frac{1}{2})}) & \text{if } \ell \in \frac{1}{2}\mathbb{Z}_{>0} \setminus \mathbb{Z}_{>0}. \end{cases}$
- (b)  $Z_{2\ell}(\mathbf{T}) = \text{End}_{\mathbf{T}}(V^{\otimes 2\ell}) \cong Z_{\ell}(\mathbf{A}_4) = \begin{cases} \text{End}_{\mathbf{A}_4}(M_4^{\otimes \ell}) & \text{if } \ell \in \mathbb{Z}_{>0} \\ \text{End}_{\mathbf{A}_3}(M_4^{\otimes(\ell-\frac{1}{2})}) & \text{if } \ell \in \frac{1}{2}\mathbb{Z}_{>0} \setminus \mathbb{Z}_{>0}. \end{cases}$
- (c)  $\dim Z_{2\ell}(\mathbf{O}) = \dim Z_{\ell}(\mathbf{S}_4) = \sum_{r=1}^4 \left\{ \begin{matrix} 2\ell \\ r \end{matrix} \right\} = \frac{4^{2\ell} + 6 \cdot 4^{\ell} + 8}{24}.$
- (d)  $\dim Z_{2\ell}(\mathbf{T}) = \dim Z_{\ell}(\mathbf{A}_4) = \sum_{r=1}^4 \left\{ \begin{matrix} 2\ell \\ r \end{matrix} \right\} + \left\{ \begin{matrix} 2\ell \\ 3 \end{matrix} \right\} + \left\{ \begin{matrix} 2\ell \\ 4 \end{matrix} \right\} = \frac{4^{2\ell} + 8}{12}.$

Part (c) is a consequence of the fact that  $Z_{\ell}(\mathbf{S}_4)$  is a homomorphic image of the partition algebra  $P_{\ell}(4)$ , and the dimension is a sum of Stirling numbers of the 2nd kind. Part (d) follows from the fact that the algebra  $Z_{\ell}(\mathbf{A}_n)$  is a certain relative of the partition algebra  $P_{\ell}(n)$  studied by Bloss [3], who computed the dimension of  $Z_{\ell}(\mathbf{A}_n)$  for  $\ell \in \mathbb{Z}_{\geq 0}$ . The dimension expressions in (c) and (d) were also given in [1, Sec. 4.3] and can be obtained by inductive arguments using the Bratteli diagrams of  $\mathbf{O}$  and  $\mathbf{T}$ .

The irreducible modules for  $P_{\ell}(4)$  are indexed by partitions of 4 (resp. 3) when  $\ell \in \mathbb{Z}_{\geq 0}$  (resp.  $\ell \in \frac{1}{2}\mathbb{Z}_{>0} \setminus \mathbb{Z}_{>0}$ ). It follows from Theorem 4.1 (a) that the Bratteli diagram  $\mathcal{B}_V(\mathbf{O})$  is the same as the Bratteli diagram  $\mathcal{B}_{M_4}(\mathbf{S}_4)$  pictured in Figure 4.1 below, where level  $\ell \in \frac{1}{2}\mathbb{Z}_{>0}$  corresponds to  $Z_{2\ell}(\mathbf{O})$  in  $\mathcal{B}_V(\mathbf{O})$  and to  $Z_{\ell}(\mathbf{S}_4)$  in  $\mathcal{B}_{M_4}(\mathbf{S}_4)$ . (Compare  $\mathcal{B}_V(\mathbf{O})$  in Figure A.4 of the Appendix.)

When restricted to  $\mathbf{A}_n$ , the irreducible modules for the symmetric group  $\mathbf{S}_n$  that are indexed by conjugate partitions are isomorphic, and the irreducible  $\mathbf{S}_n$ -modules indexed by partitions that are self-conjugate split into a direct sum of two irreducible  $\mathbf{A}_n$ -modules having equal dimensions, which we distinguish with  $\pm$ . Thus, for  $\mathbf{A}_4$  there are 4 irreducible modules indexed by  $(4), (3, 1), (2, 2)^+, (2, 2)^-$ , and for  $\mathbf{A}_3$  there are 3, which are labeled by  $(3), (2, 1)^+, (2, 1)^-$ . Theorem 4.1 (b) implies that the Bratteli diagrams  $\mathcal{B}_V(\mathbf{T})$  and  $\mathcal{B}_{M_4}(\mathbf{A}_4)$  are the same (see Figure 4.2 below).

*Remarkably, the marks on the Dynkin diagrams for  $\mathbf{O}$  and  $\mathbf{T}$  in Figures 4.1 and 4.2 (compare Section A.1) turn out to be exactly the dimensions of the irreducible modules for the symmetric group  $\mathbf{S}_4$  (resp. the alternating group  $\mathbf{A}_4$ ) indexed by the corresponding partitions of 4 and are twice the dimensions of the irreducible modules for  $\mathbf{S}_3$  (resp.  $\mathbf{A}_3$ ) indexed by partitions of 3. A walk on the Dynkin diagram or on the Bratteli diagram for  $\mathbf{O}$  or  $\mathbf{T}$  just amounts to removing a box or adjoining a box to a partition on alternate steps.*

**4.3. The icosahedral case.** A different approach is required for the binary icosahedral group  $\mathbf{I}$ , as  $V^{\otimes 2}$  has dimension 4, is the direct sum of irreducible  $\mathbf{A}_5$ -modules of dimension 1 and 3, and is *not* the permutation module for  $\mathbf{I}/\mathcal{Z}(\mathbf{I}) \cong \mathbf{A}_5$ . The details are too lengthy to include here, and some are still a work in progress, but we briefly mention a few facts about this case. Set  $L_1 = 1$  and  $L_2 = 3$  (note the coincidence with the dimensions above), and let  $L_{n+1} = L_n + L_{n-1}$  for  $n \geq 2$ . Then  $L_n$  is *n*th Lucas number given by  $L_n = \varphi^n + (-\varphi)^{-n}$ , where  $\varphi = \frac{1}{2}(1 + \sqrt{5})$ , the golden ratio. We have the following:

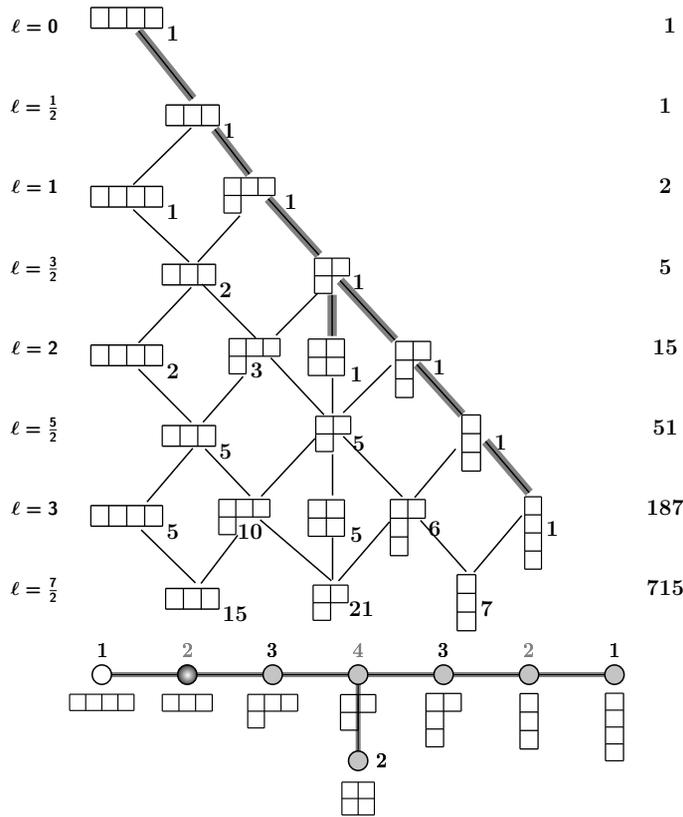


Figure 4.1. Bratteli and Dynkin diagrams for  $O$

**Proposition 4.2** (See [1]). *For  $k \geq 1$ , let  $Z_k(\mathbf{I}) = \text{End}_{\mathbf{I}}(V^{\otimes k})$ , where  $V = \mathbb{C}^2$ . Then,*

$$\dim Z_k(\mathbf{I}) = \frac{4^k + 12L_{2k} + 20}{60}.$$

Expressions for the dimensions of the irreducible modules for  $Z_k(\mathbf{I})$  also involve Lucas numbers and can be found in [1, Sec. 4.3]. Even tensor powers  $V^{\otimes 2\ell}$  can be viewed as  $A_5$ -modules, and the representation theory of  $A_5$  plays an essential role in studying them.

### 5. McKay correspondence for non-simply laced diagrams

McKay’s correspondence has been extended to affine Dynkin diagrams with multiple edges in several different ways [11, 22]. Slodowy’s correspondence [22], which was motivated by the study of singularities, starts with a finite subgroup  $G$  of  $SU_2$  and an automorphism  $\sigma$  of  $G$  stabilizing the defining representation  $V = \mathbb{C}^2$ . (The automorphism  $\sigma$  can be identified with a graph automorphism of the affine Dynkin diagram associated to  $G$  that fixes the node corresponding to  $V$ .) Set  $\tilde{G} = \langle \sigma \rangle \rtimes G$ , and let  $\{\tilde{G}^\alpha\}$  (resp.  $\{G^\lambda\}$ ) denote the set of irreducible modules for  $\tilde{G}$  (resp. for  $G$ ). Each irreducible  $\tilde{G}$ -module can be regarded as a module for  $G$  by restriction, and each irreducible  $G$ -module can be induced to a module for  $\tilde{G}$ . Since  $V$  is

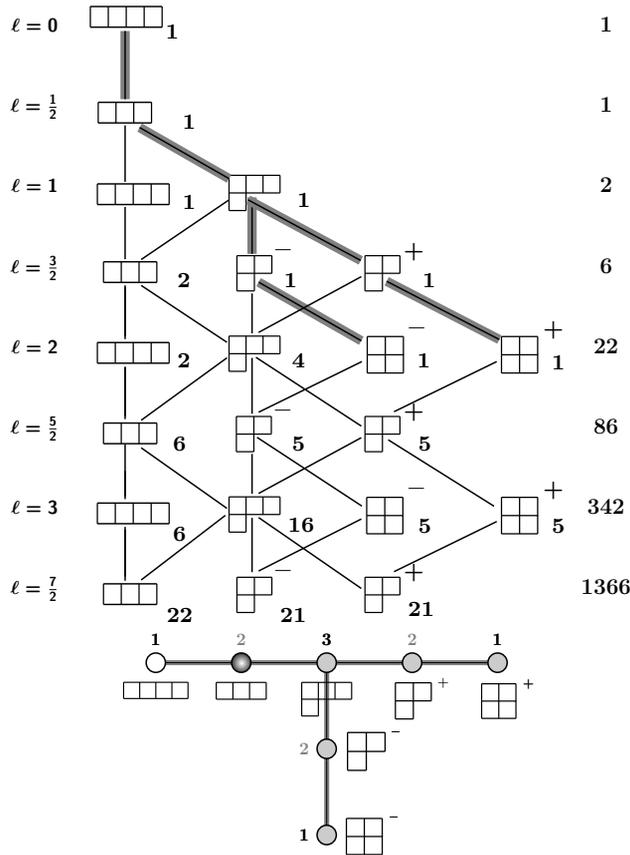


Figure 4.2. Bratteli and Dynkin diagrams for  $\mathbf{T}$

a module for both  $G$  and  $\tilde{G}$ , we can consider the following tensor products:

$$\text{Res}_{\tilde{G}}(\tilde{G}^\beta) \otimes V = \sum_{\alpha} b_{\beta,\alpha} \text{Res}_{\tilde{G}}(\tilde{G}^\alpha) \quad \text{and} \quad \text{Ind}_{\tilde{G}}(G^\mu) \otimes V = \sum_{\lambda} b_{\mu,\lambda}^\vee \text{Ind}_{\tilde{G}}(G^\lambda). \quad (5.1)$$

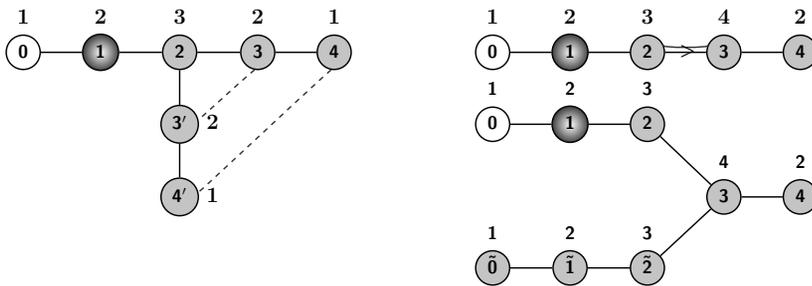
Then for  $B = (b_{\beta,\alpha})$  (resp.  $B^\vee = (b_{\mu,\lambda}^\vee)$ ), the matrix  $C = 2I - B$  (resp.  $C^\vee = 2I - B^\vee$ ) is the Cartan matrix of the affine Dynkin diagram obtained from folding the diagram corresponding to  $G$  (resp. the dual Dynkin diagram with the arrow reversed). In the table below we have displayed this information and have included in the last column the alternate notation commonly used for the diagrams, which comes from considering the associated twisted affine Lie algebras (see [15, pp. 54-55]). Further details can be found in [22] and [25].

$(\tilde{G}, \text{affine diagram})$	$(G, \text{affine diagram})$	affine diagrams for $(C, C^\vee)$
$(\mathbf{O}, \hat{E}_7)$	$(\mathbf{T}, \hat{E}_6)$	$(\hat{F}_4, \hat{F}_4^\vee = \hat{E}_6^{(2)})$
$(\mathbf{T}, \hat{E}_6)$	$(\mathbf{D}_2, \hat{D}_4)$	$(\hat{G}_2, \hat{G}_2^\vee = \hat{D}_4^{(3)})$
$(\mathbf{D}_{2(n-1)}, \hat{D}_{2n})$	$(\mathbf{D}_{n-1}, \hat{D}_{n+1})$	$(\hat{B}_n, \hat{B}_n^\vee = \hat{A}_{2n-1}^{(2)})$
$(\mathbf{D}_{2n}, \hat{A}_{2n-1})$	$(\mathbf{D}_n, \hat{D}_{n+2})$	$(\hat{C}_n^\vee = \hat{D}_{n+1}^{(2)}, \hat{C}_n)$

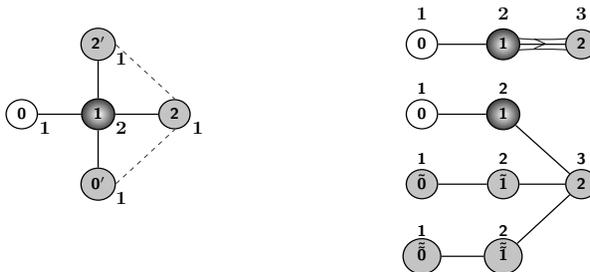
In the two examples discussed below, which correspond to the first two lines of the table,

the foldings are indicated by dashed lines. The  $\tilde{G}$ -modules corresponding to the nodes to the left of the multiple bond in the folded diagram are not irreducible. For example, in the  $(\hat{E}_6, \hat{F}_4)$ -case there are 8 irreducible  $\tilde{G}$ -modules:  $\tilde{G}^{(j)}, \tilde{G}^{(\bar{j})}, 0 \leq j \leq 2, \tilde{G}^{(3)}, \tilde{G}^{(4)}$ , with  $\dim \tilde{G}^{(j)} = \dim \tilde{G}^{(\bar{j})} = j + 1$  for  $j = 0, 1, 2$ ,  $\dim \tilde{G}^{(3)} = 4$ , and  $\dim \tilde{G}^{(4)} = 2$ . The representation graph  $\mathcal{R}_V(\tilde{G})$  is pictured beneath the Dynkin diagram  $\hat{F}_4$ , and the Bratteli diagram  $\mathcal{B}_V(\tilde{G})$  is displayed in Figure A.5 of Section A.2 of the Appendix. Because  $\tilde{G}$  is isomorphic to the binary octahedral group  $\mathbf{O}$ , the Bratteli diagrams  $\mathcal{B}_V(\tilde{G})$  and  $\mathcal{B}_V(\mathbf{O})$  are essentially the same up to relabeling of the indices for the irreducible modules. The only edges in  $\mathcal{B}_V(\tilde{G})$  that are not obtained from the Jones basic construction via induction and restriction are exactly the affine Dynkin diagram of type  $\hat{E}_7$ .

$(\hat{E}_6, \hat{F}_4)$ :



$(\hat{D}_4, \hat{G}_2)$ :



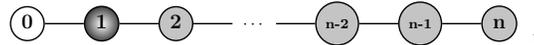
Similarly, in the  $(\hat{D}_4, \hat{G}_2)$ -case, there are 7 irreducible  $\tilde{G}$ -modules:  $\tilde{G}^{(2)}$  and the ones corresponding to the nodes to the left of the arrow on the  $\hat{G}_2$  diagram,  $\tilde{G}^{(j)}, \tilde{G}^{(\bar{j})}, \tilde{G}^{(\bar{\bar{j})}), 0 \leq j \leq 1$ , with  $\dim \tilde{G}^{(2)} = 3$  and  $\dim \tilde{G}^{(j)} = \dim \tilde{G}^{(\bar{j})} = \dim \tilde{G}^{(\bar{\bar{j})}) = j + 1$  for  $j = 0, 1$ . The representation graph  $\mathcal{R}_V(\tilde{G})$  for  $\tilde{G}$  is shown beneath the Dynkin diagram  $\hat{G}_2$ . The group  $\tilde{G}$  is the semidirect product of the cyclic group of order 3 generated by the graph automorphism  $\sigma$  with the binary dihedral group  $\mathbf{D}_2$  (which is the quaternion group of order 8), and  $\tilde{G}$  is isomorphic to the binary tetrahedral group  $\mathbf{T}$ . The Bratteli diagrams  $\mathcal{B}_V(\tilde{G})$  and  $\mathcal{B}_V(\mathbf{T})$  are basically the same up to relabeling of the indices for the irreducible modules (compare Figures A.3 and A.5 of the Appendix), and the only edges in the Bratteli diagram  $\mathcal{B}_V(\tilde{G})$  not coming from the Jones basic construction form a copy of the affine Dynkin diagram of type  $\hat{E}_6$ .

**Concluding Remarks.** The results discussed in this paper can be applied in a wide variety of different settings, where they give essential combinatorial and representation-theoretic information. For example, the Grothendieck ring of the tensor category of finite-dimensional

representations of the quantum group  $U_q(\mathfrak{sl}_2)$  at a root of unity  $q = e^{\pi i/(n+2)}$  modulo representations of quantum dimension 0 is the commutative, associative Verlinde algebra  $\mathcal{V}_n$  having basis  $\chi_j, j = 0, 1, \dots, n$ , which satisfies a truncated version of the Clebsch-Gordan formula from (1.1),

$$\chi_1 \otimes \chi_j = \chi_{j-1} + \chi_{j+1} \quad (\chi_{-1} = 0 = \chi_{n+1}). \tag{5.2}$$

The matrix  $\mathcal{X}$  of multiplication by  $\chi_1$  in  $\mathcal{V}_n$  is the adjacency matrix of the finite Dynkin diagram of type  $A_{n+1}$ ,



and it satisfies  $p_{n+1}(\mathcal{X}) = 0$ , where the polynomial  $p_{n+1}(x)$  is defined by the recursion

$$p_0(x) = 1, \quad p_1(x) = x, \quad \text{and} \quad p_{n+1}(x) = xp_n(x) - p_{n-1}(x) \quad \text{for } n \geq 1. \tag{5.3}$$

(Compare [6], which explores further connections with tensor categories and conformal field theory.) Thus,  $p_n(x) = U_n(x/2)$ , where  $U_n(x)$  is the ubiquitous Chebyshev polynomial of the 2nd kind.

Using (5.2), we associate a Bratteli diagram to powers  $\chi_1^{\otimes k}$  of the generator  $\chi_1$  of  $\mathcal{V}_n$ . Walks of  $2k$  steps from 0 at level 0 to 0 at level  $2k$  on this Bratteli diagram are in bijection with lattice paths (Dyck paths) in the first quadrant with up steps  $(1, 1)$  and down steps  $(1, -1)$ , starting at  $(0, 0)$ , ending at  $(2k, 0)$ , and having bounded height  $n$  (illustrated in Figure 5.1 below for  $n = 3$ ). The first coordinate of a point in the lattice path records the step number; the second corresponds to  $\chi_j$ . The dashed path in the Bratteli diagram below corresponds to the lattice path on the right. Especially noteworthy in this particular example is the appearance of Fibonacci numbers giving the coefficients of the  $\chi_j$  in  $\chi_1^{\otimes k}$ .

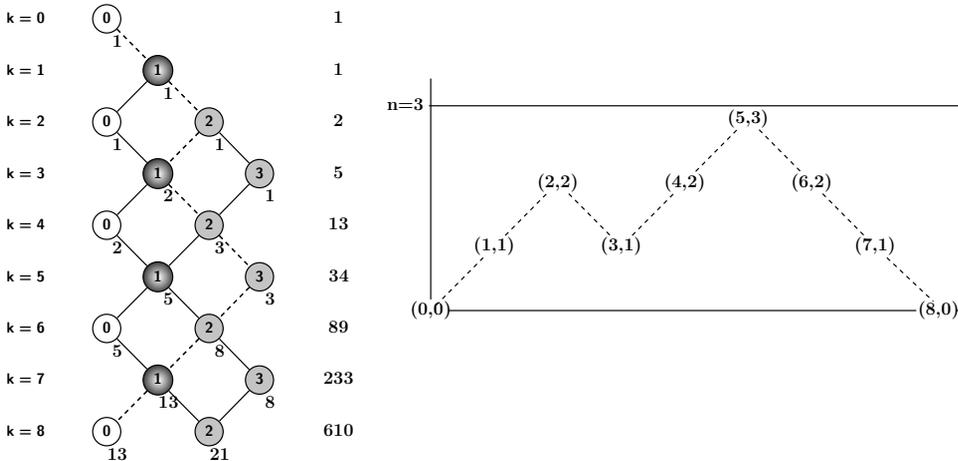


Figure 5.1.  $\chi_1^{\otimes k}$  in the Verlinde algebra  $\mathcal{V}_3$  and a Dyck path

Setting  $q_n(x) = x^{n/2}p_n(x^{-1/2})$ , we have from [16] that the generating function for the number  $d(k, n)$  of Dyck paths of bounded height  $n$  with  $2k$  steps is

$$\sum_{k=0}^{\infty} d(k, n)x^k = \frac{q_n(x)}{q_{n+1}(x)}.$$

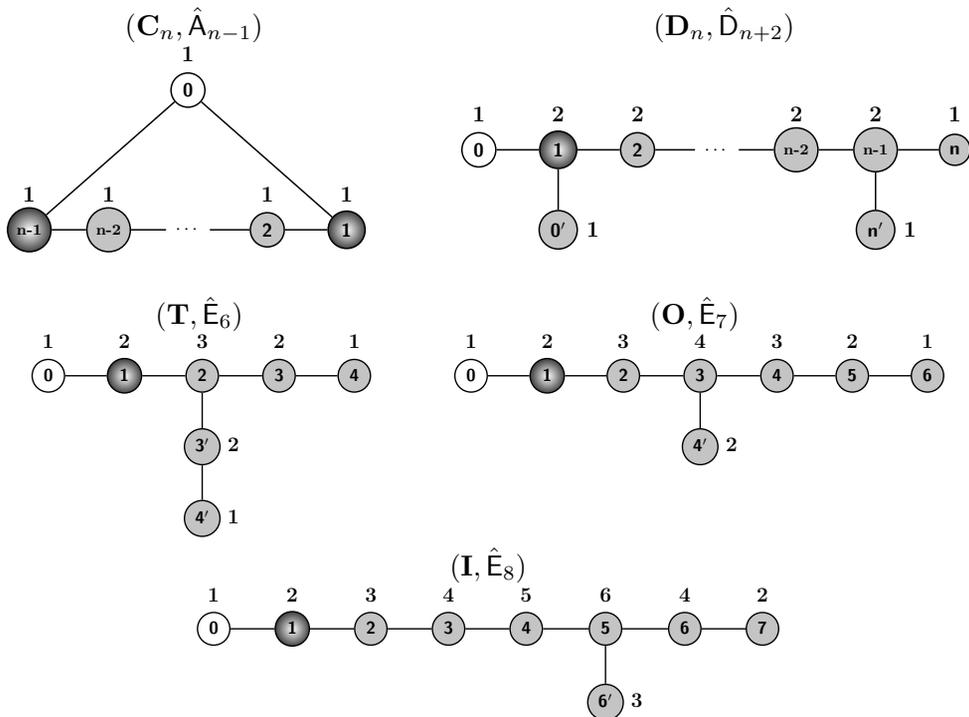
Therefore, the coefficient of  $\chi_0$  in the product  $\chi_1^{\otimes 2k}$  is given by  $d(k, n)$ . In the example,

$$\sum_{k=0}^{\infty} d(k, 3)x^k = \frac{1 - 2x}{1 - 3x + x^2} = 1 + x + 2x^2 + 5x^3 + 13x^4 + 34x^5 + 89x^6 + \dots,$$

so that  $d(4, 3) = 13$  when  $n = 3$ . From the Bratteli diagram,  $\chi_1^{\otimes 8} = 13\chi_0 + 21\chi_2$ .

**A. Appendix**

**A.1. Simply laced affine Dynkin diagrams.** The representation graph  $\mathcal{R}_V(G)$  for a finite subgroup  $G$  of  $SU_2$  is the corresponding affine Dynkin diagram of type  $\hat{A}_{n-1}, \hat{D}_{n+2}, \hat{E}_6, \hat{E}_7, \hat{E}_8$ . In the figures below, the label inside the node is the index of the irreducible  $G$ -module, and the label above the node is its dimension, which is the mark on the Dynkin diagram. The trivial module is indicated in white and the module  $V = \mathbb{C}^2$  in black. In the cyclic case  $V = \mathbb{C}_n^{(n-1)} \oplus \mathbb{C}_n^{(1)}$  and in all other cases  $V = G^{(1)}$ .



**A.2. Bratteli diagrams.** The first few rows of the Bratteli diagrams  $\mathcal{B}_V(G)$  for finite subgroups  $G$  of  $SU_2$  are displayed here. The nodes at level  $k$  label the irreducible  $G$ -modules that appear in  $V^{\otimes k}$ . The number beneath each node at level  $k$  is the multiplicity of the corresponding  $G$ -module in  $V^{\otimes k}$ , and it is also the dimension of the irreducible  $Z_k(G)$ -module having the same label as the node. The right-hand column contains the sum of the squares of these dimensions and equals  $\dim Z_k(G)$ . An edge between level  $k$  and level  $k + 1$  is shaded if it cannot be obtained as the reflection over level  $k$  of an edge between level  $k - 1$  and  $k$ .

The shaded edges give an embedding of the representation graph  $\mathcal{R}_V(G)$  (the affine Dynkin diagram) into the Bratteli diagram  $\mathcal{B}_V(G)$ .

The cyclic groups  $C_5$  and  $C_{10}$  have isomorphic Bratteli diagrams; each corresponds to Pascal's triangle on a cylinder of "diameter" 5:

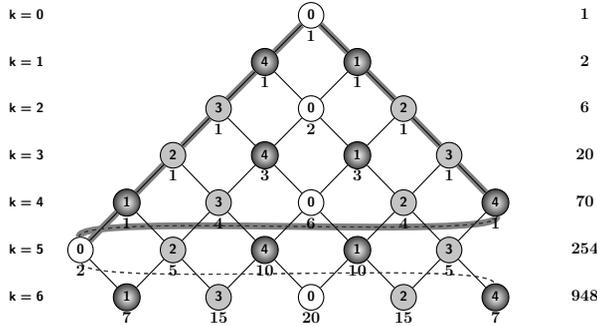


Figure A.1. Bratteli diagram for the cyclic group  $C_5$

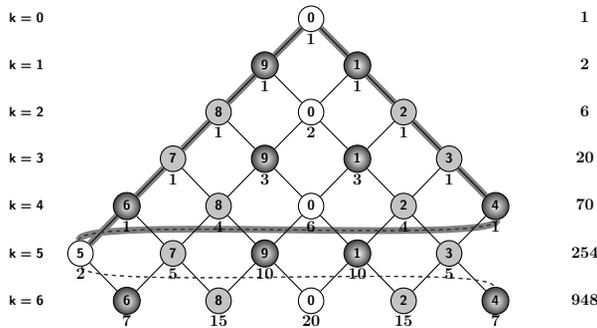


Figure A.2. Bratteli diagram for the cyclic group  $C_{10}$

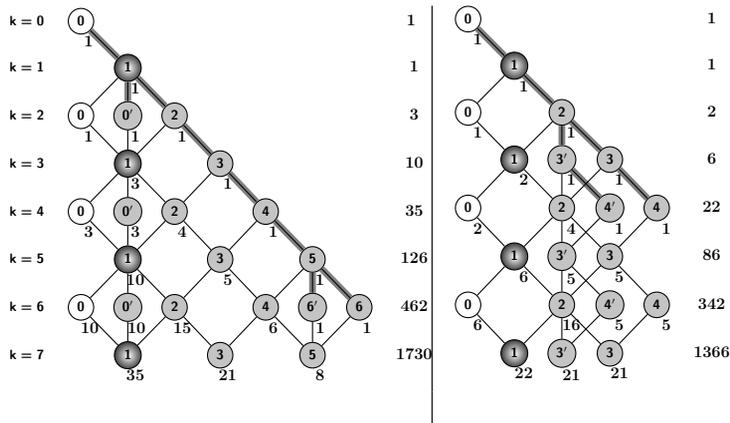


Figure A.3. Bratteli diagrams for the binary dihedral group  $D_6$  and binary tetrahedral group  $T$

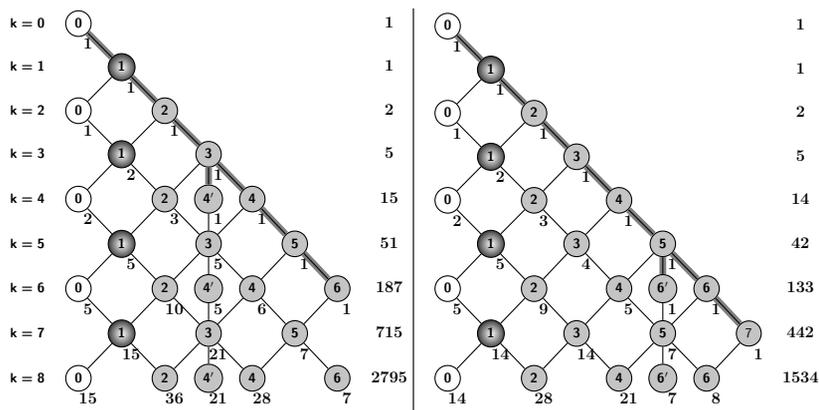


Figure A.4. Bratteli diagrams for the binary octahedral group  $O$  and binary icosahedral group  $I$

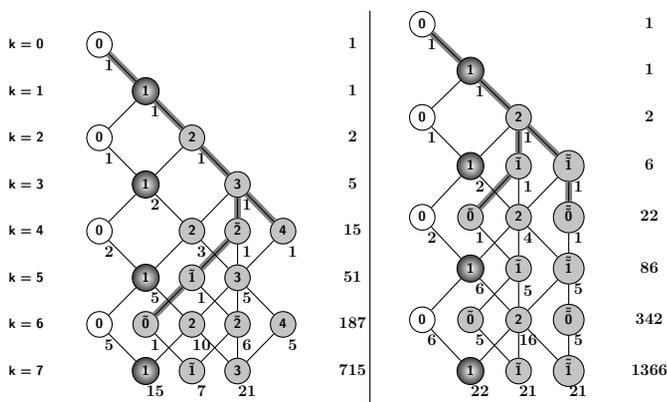


Figure A.5. Bratteli diagrams associated to  $\hat{F}_4$  and  $\hat{G}_2$

### References

- [1] Barnes, J.M., Benkart, G., and Halverson, T., *McKay centralizer algebras*, submitted; arXiv:1213.5254.
- [2] Benkart, G. and Halverson, T., *Exceptional McKay centralizer algebras*, to appear.
- [3] Bloss, M., *The partition algebra as a centralizer algebra of the alternating group*, Comm. Algebra **33** (2005), no. 7, 2219–2229.
- [4] Bowman, C., DeVisscher, M., and Orellana, R., *The partition algebra and the Kronecker coefficients*, Trans. Amer. Math. Soc. to appear; arXiv:1210.5579.
- [5] Cautis, S. and Jackson, D.M., *The matrix of chromatic joins and the Temperley-Lieb algebra*, J. Combin. Theory Ser. B **89** (2003), no. 1, 109–155.

- [6] Etingof, P. and Khovanov, M., *Representations of tensor categories and Dynkin diagrams*, Internat. Math. Res. Notices **5** (1995), 235–247.
- [7] Frenkel, I. and Khovanov, M., *Canonical bases in tensor products and graphical calculus for  $U_q(\mathfrak{sl}_2)$* , Duke Math. J. **87** (1997), no. 3, 409–480.
- [8] Fulton, W. and Harris, J., *Representation Theory, A First Course*, Graduate Texts in Mathematics **129**, Springer-Verlag, New York, 1991.
- [9] Goodman, F.M., de la Harpe, P., and Jones, V.F.R., *Coxeter Graphs and Towers of Algebras*, Springer, New York, 1989.
- [10] Halverson, T. and Ram, A., *Partition algebras*, European J. Combin. **26** (2005), no. 6, 869–921.
- [11] Happel, D., Preiser, U., and Ringel, C.M., *Binary polyhedral groups and Euclidean diagrams*, Manuscripta Math. **31** (1980), no. 1-3, 317–329.
- [12] Jones, V.F.R., *Index for subfactors*, Invent. Math. **72** (1983), 1–25.
- [13] ———, *The Potts model and the symmetric group*, in: Subfactors: Proceedings of the Taniguchi Symposium on Operator Algebras (Kyuzeso, 1993), World Scientific Publishing, River Edge, N.J., 1994, pp. 259–267.
- [14] ———, *On the origin and development of subfactors and quantum topology*, Bull. Amer. Math. Soc. **46** (2009) no. 2, 309–326.
- [15] Kac, V.G., *Infinite-dimensional Lie Algebras*, 3rd ed., Cambridge Univ. Press, Cambridge, 1990.
- [16] Krattenthaler, C., *Permutations with restricted patterns and Dyck paths. Special issue in honor of Dominique Foata’s 65th birthday* (Philadelphia, PA, 2000), Adv. in Appl. Math. **27** (2001), no. 2-3, 510–530.
- [17] Martin, P., *Representations of graph Temperley-Lieb algebras*, Publ. Res. Inst. Math. Sci. **26** (1990) no. 3, 485–503.
- [18] ———, *Temperley-Lieb algebras for non-planar statistical mechanics – the partition algebra construction*, J. Knot Theory Ramifications **3** (1994), 51–82.
- [19] ———, *The structure of the partition algebra*, J. Algebra **183** (1996), 319–358.
- [20] Martin, P. and Rollet, G., *The Potts model representation and a Robinson-Schensted correspondence for the partition algebra*, Compositio Math. **112** (1998), 237–254.
- [21] McKay, J., *Graphs, singularities, and finite groups*, The Santa Cruz Conference on Finite Groups (Univ. California, Santa Cruz, Calif., 1979), pp. 183–186, Proc. Sympos. Pure Math. **37**, Amer. Math. Soc., Providence, R.I., 1980.
- [22] Slodowy, P., *Simple Singularities and Simple Algebraic Groups*, Lecture Notes in Mathematics **815**, Springer, Berlin, 1980.
- [23] Stanley, R.P., *Enumerative Combinatorics* Vol. 2, Cambridge Studies in Adv. Math. **62**, Cambridge, 1999.
- [24] Steinberg, R., *Finite subgroups of  $SU_2$ , Dynkin diagrams and affine Coxeter elements*, Pacific J. Math. **118** (1985), no. 2, 587–598.
- [25] Stekolshchik, R., *Notes on Coxeter Transformations and the McKay Correspondence*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2008.
- [26] Temperley, H.N.V. and Lieb, E.H., *Relations between the “percolation” and the*

- “colouring” problem and other graph-theoretical problems associated with regular planar lattices: some exact results for the “percolation” problem*, Proc. Roy. Soc. London Ser. A **322** (1971), 251–280.
- [27] Wenzl, H., *On sequences of projections*, C.R. Math. Rep. Acad. Sci. Canada **9** (1987), no. 1, 5–9.
- [28] ———, *Braids and invariants of 3-manifolds*, Invent. Math. **114** (1993), no. 2, 235–275.
- [29] ———, *On tensor categories of Lie type  $E_N$ ,  $N \neq 9$* , Adv. Math. **177** (2003), no. 1, 66–104.
- [30] ———, *On centralizer algebras for spin representations*, Comm. Math. Phys. **314** (2012), no. 1, 243–263.
- [31] B.W. Westbury, *The representation theory of the Temperley-Lieb algebras*, Math. Z. **219** (1995), 539–565.

Department of Mathematics, University of Wisconsin-Madison, 480 Lincoln Dr., Madison, Wisconsin 53706, USA

E-mail: benkart@math.wisc.edu

# Rational points on elliptic and hyperelliptic curves

Manjul Bhargava

**Abstract.** A *hyperelliptic curve*  $C$  over  $\mathbb{Q}$  is the graph of an equation of the form  $y^2 = f(x)$ , where  $f$  is a polynomial having coefficients in the rational numbers  $\mathbb{Q}$  and distinct roots in  $\mathbb{C}$ . The special case where the degree of  $f$  is 3 is called an *elliptic curve*  $E$  over  $\mathbb{Q}$  which, as we will discuss, has many special properties not shared by general hyperelliptic curves  $C$ . A solution  $(x, y)$  to  $C : y^2 = f(x)$ , with  $x$  and  $y$  rational numbers, is called a *rational point* on  $C$ .

Given a random elliptic or hyperelliptic curve  $C : y^2 = f(x)$  over  $\mathbb{Q}$  with  $f(x)$  of a given degree  $n$ , how many rational points do we expect on the curve  $C$ ? Equivalently, how often do we expect a random polynomial  $f(x)$  of degree  $n$  to take a square value over the rational numbers? In this article, we give an overview of a number of recent conjectures and theorems giving some answers and partial answers to this question.

**Mathematics Subject Classification (2010).** Primary 11G05, 11G30; Secondary 11R45, 14H25, 20G30.

**Keywords.** Elliptic curve, rank, hyperelliptic curve, rational points, Hasse principle, Birch–Swinnerton-Dyer Conjecture.

## 1. Introduction

An ancient question in mathematics, and in number theory in particular, is that of understanding the square values taken by a given mathematical expression over the integers. Perhaps the oldest example is that of constructing square values  $c^2$  of the expression  $a^2 + b^2$ . Solutions to  $c^2 = a^2 + b^2$  already seem to occur as far back as 2500 B.C. in the huge right-angled integer-sided stone structures of the megalithic monuments of Northern Europe and Egypt. The first known explicitly recorded integer solutions to the equation  $c^2 = a^2 + b^2$  are found on the Babylonian tablet known as “Plimpton 322,” which dates back to 1800 B.C.; some of the largest solutions on this clay tablet are made up of five-digit numbers for  $a, b, c$  (e.g.,  $18541^2 = 12709^2 + 13500^2$ )! It seems evident that the writer of this tablet had some method of producing these solutions  $(a, b, c)$ . The first known recorded formula for constructing all integer solutions  $(a, b, c)$  to  $c^2 = a^2 + b^2$  was given by Euclid c. 300 B.C.

Another ancient and famous example of such a “square equation” is that in which, for a fixed positive integer  $n$ , we wish to find all the integer solutions  $(x, y)$  to  $x^2 = ny^2 \pm 1$ . This equation has also arisen in many contexts for well over a millennium. In c. 628, Brahmagupta showed that if  $x^2 = ny^2 \pm 1$  has at least one solution, then it has infinitely many solutions, and provided an explicit method for generating these solutions. In c. 1150, Bhaskara provided a method for producing an integer solution to  $x^2 = ny^2 \pm 1$  for any fixed nonsquare

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

integer  $n > 0$ ; that this method indeed always terminates with a solution was proved by Lagrange in 1768. Thus, for any nonsquare integer  $n > 0$ , the equation  $x^2 = ny^2 \pm 1$  always possesses infinitely many solutions.

A further important example of a square equation that originates “only” about 470 years ago is  $c^2 = a^4 + b^4$ . In c. 1640, Fermat proved that this equation does not have any nonzero integer solutions, i.e., the sum of two nonzero fourth powers is never a square. In particular, this implies that when  $n = 4$ , the famous equation  $c^n = a^n + b^n$  of Fermat does not have a solution. The latter equation has of course played a very influential role in number theory, both due to the techniques that were used to solve it as well as in the context of the many ensuing developments surrounding “Fermat’s Last Theorem”.

There have been many other examples of square equations over the years, but the three described above are typical examples that are very well-known and have had a major influence, not just in number theory, but in other related areas as well. For example, the equation  $x^2 = ny^2 \pm 1$  and its solutions played an important role in the negative solution of Hilbert’s 10th problem, due to Davis, Matiyasevich, Putnam, and Robinson. In addition, the methods that were originally used to solve these equations still play a very important role today. For example, Fermat used his method of “infinite descent” to prove his result, a method that has influenced number theory very heavily ever since.

A more modern but still very influential example of a square equation originates in a problem that Ramanujan posed in the Journal of the Indian Mathematical Society in 1913. He asked: What are the values of  $n$  such that  $2^n - 7$  is a square number? More precisely, he asked if all such values of  $n$  are given by  $n = 3, 4, 5, 7, \text{ and } 15$ , yielding  $2^n - 7 = 1^2, 3^2, 5^2, 11^2, \text{ and } 181^2$ , respectively. This problem was eventually solved in the positive by Nagell in 1948 using methods of diophantine analysis. Equations such as  $m^2 = 2^n - 7$  are called *exponential equations*, because one of the variables, namely  $n$ , is in the exponent; this can make analysis of the integer solutions to the equation quite different and often much more difficult. This problem posed back in 1913, and its solution by Nagell, continues to generate much research in the area. Exponential equations of the form  $m^2 = a^n - b$ , where  $a, b$  are fixed and  $m, n$  are the variables, are called equations of Ramanujan–Nagell-type. It is now known [65] that such equations always have at most two solutions  $(m, n)$ , *unless*  $a = 2$  and  $b = 7$ ; so Ramanujan clearly chose the constants  $a, b$  in a special and very atypical way! Typically, such an equation should have at most two solutions, and indeed, usually should have no solutions whatsoever.

In this article, rather than exponential equations, we wish to consider the simplest type of square equation, namely, where we wish to determine when an integer polynomial in one variable takes a square value.

## 2. Hyperelliptic curves

More precisely, we wish to consider the equation

$$y^2 = f(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n \quad (2.1)$$

where  $f$  is a polynomial in one variable with integer coefficients  $a_0, a_1, \dots, a_n$  having distinct roots in  $\mathbb{C}$ .<sup>1</sup>

The graph of such an equation (2.1) is called a *hyperelliptic curve*. It is called a “curve” because its real solutions trace out a curve in  $\mathbb{R}^2$ ; the word “hyperelliptic” refers to the fact that this curve is symmetric about the  $x$ -axis: if  $(x, y)$  is a solution to  $y^2 = f(x)$ , then so is  $(x, -y)$ . Solutions to  $y^2 = f(x)$  are thus called *points* on the corresponding hyperelliptic curve  $C$ .

We are interested in finding all integer solutions  $(x, y)$  to  $y^2 = f(x)$ . More generally, we may look for all rational solutions  $(x, y)$  to  $y^2 = f(x)$ ; such solutions  $(x, y)$  are then called *rational points* on the corresponding hyperelliptic curve  $C$ , as such a point  $(x, y)$  on  $C$  has the property that both  $x$  and  $y$  are rational numbers.

### 3. The possible number of rational points on a general algebraic curve

The first question that one may ask is: how many rational points can such a hyperelliptic curve  $C : y^2 = f(x)$  have, where  $f$  is a polynomial of degree  $n$ ?

More generally, one may ask: how many rational points can a general algebraic curve have? That is, if we take a general polynomial  $h(x, y)$  in two variables with rational coefficients, and graph its zeroes in  $\mathbb{R}^2$ , then we may ask how many rational points can lie on this curve. The case of hyperelliptic curves is that where  $h(x, y) = y^2 - f(x)$ .

This question, about the possible number of rational solutions to  $h(x, y) = 0$ , has a beautiful answer in terms of the topology of the graph of the equation—but we must graph the solutions to the equation  $h(x, y) = 0$  in  $\mathbb{C}^2$  rather than  $\mathbb{R}^2$  to get the full picture! This is not normally done in school because in the world we live, our paper is two-dimensional over  $\mathbb{R}$  and not four-dimensional. But  $\mathbb{C}$  instead of  $\mathbb{R}$  is the natural place to look for solutions to polynomial equations. This is already evident when studying one-variable polynomial equations: the Fundamental Theorem of Algebra states that any degree  $n$  real polynomial  $f$  has  $n$  roots over  $\mathbb{C}$ , but the theorem certainly does not hold if we only count the roots over  $\mathbb{R}$ .

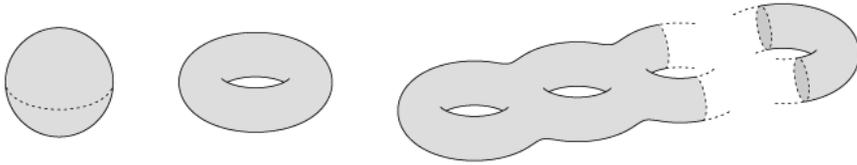
If one thinks about the graph of  $h(x, y) = 0$  in  $\mathbb{C}^2$ , then the answer to the question of how many rational solutions  $(x, y)$  that the equation  $h(x, y) = 0$  can have is very much tied to the *topology* of its graph in  $\mathbb{C}^2$ ! This is one of the really beautiful connections between number theory and other branches of mathematics, particularly geometry and topology.

To state this connection with topology, we must first classify what this graph of  $h(x, y) = 0$  can look like in  $\mathbb{C}^2$ . At first, it might seem difficult to imagine this graph in  $\mathbb{C}^2$ , since  $\mathbb{C}^2$  has four real dimensions; however, it is easy to see that the graph itself will be a two-dimensional real *surface*, so in the end it is not so difficult to visualize. (Algebraic curves  $C : h(x, y) = 0$ , when graphed over  $\mathbb{C}$ , are also then called *Riemann surfaces*.)

When we graph an equation of the form  $h(x, y) = 0$  in  $\mathbb{C}^2$  to obtain a curve  $C$ , we obtain a (Riemann) surface. This leads to the notion of the “genus” of a curve. The basic theorem here is that  $C$  will topologically be a compact surface with  $g$  “donut holes”, perhaps

<sup>1</sup>There is no loss of generality in the assumption of distinct roots. Indeed, if  $f$  had a repeated root over  $\mathbb{C}$ , then it would have a square factor  $g^2$ , where  $g$  and  $h = f/g^2$  are polynomials with integer coefficients with  $h$  squarefree. The problem of finding solutions  $(x, y)$  to the equation  $y^2 = f(x)$  over any field (or indeed any integral domain) would then be equivalent to finding solutions  $(x, y)$  to the equation  $y^2 = h(x)$ , where now  $h(x)$  has distinct roots over  $\mathbb{C}$ .

with finitely many points removed.<sup>2</sup> The number  $g$  occurring here is called the *genus* of the algebraic curve  $C$ . Here are what the graphs in  $\mathbb{C}^2$  of algebraic curves of genus 0, 1, and  $\geq 3$  would look like (more or less):



The significance of the genus is the following general theorem that relates the genus of a curve to the structure of its rational points:

**Theorem 3.1.** *The set of rational points on an algebraic curve of genus  $g$  is:*

- (a) *either empty or infinite if  $g = 0$ ;*
- (b) *finite or infinite if  $g = 1$ ; and*
- (c) *finite if  $g > 1$  (Mordell’s Conjecture 1922, Faltings’ Theorem 1983).*

This theorem is truly remarkable in that it relates the genus (a topological invariant of the locus of solutions over  $\mathbb{C}$ !) to the structure of rational points. Part (a) follows from the Hasse–Minkowski Theorem, while part (b) is trivial. Part (c) is the deepest; it was conjectured by Mordell in 1922, and proved by Faltings in 1983.

It is interesting to note that Theorem 3.1 is effective in the case of  $g = 0$  (again, using the Hasse–Minkowski Theorem), but there is no known algorithm to determine how many rational points a curve has once  $g \geq 1$ . In particular, in the case  $g = 1$ , it is not even known in general how to determine whether a curve has finitely or infinitely many rational points (although the Birch and Swinnerton-Dyer Conjecture does give a conjectural method)!

#### 4. The possible number of rational points on a hyperelliptic curve

It is not hard to visualize the graph in  $\mathbb{C}^2$  and thus compute the genus of any hyperelliptic curve:

**Theorem 4.1.** *Let  $C : y^2 = f(x)$  define a hyperelliptic curve over  $\mathbb{C}$ , where  $f$  is a polynomial in one variable of degree  $n$  having no repeated roots in  $\mathbb{C}$ . Then the genus  $g$  of the curve  $C$  is given by*

$$g = \left\lfloor \frac{n-1}{2} \right\rfloor.$$

In other words, the genus  $g$  and degree  $n$  are related by:  $n = 2g + 1$  or  $n = 2g + 2$ . Thus, if  $f$  has degree 1 or 2, then the hyperelliptic curve  $C : y^2 = f(x)$  has genus 0; and if  $f$  has degree 3 or 4, then the genus of  $C$  is 1; and if  $f$  has degree 5 or 6, then the genus of  $C$  is 2, etc.<sup>3</sup>

<sup>2</sup>These finitely many points can be filled in by adding in the solutions “at infinity” in the projective plane; see [42, §§4–5] for a beautiful treatment.

<sup>3</sup>It is also worth noting here that, in order to make  $C$  a compact Riemann surface, one adds in one point at infinity when the degree  $n$  of  $f$  is odd, and two points at infinity when the degree  $n$  of  $f$  is even (see, e.g., [69, §II, Exercise 2.14(a)–(b)]).

We then immediately obtain from Theorem 3.1:

**Corollary 4.2.** *Suppose  $f$  is a polynomial in one variable of degree  $n$  having rational coefficients and no repeated roots over  $\mathbb{C}$ . Then the set of rational points on the hyperelliptic curve  $y^2 = f(x)$  is:*

- (a) *either empty or infinite if  $n = 1$  or  $2$ ;*
- (b) *finite or infinite if  $n = 3$  or  $4$ ; and*
- (c) *finite if  $n \geq 5$ .*

That is, once the degree of the polynomial  $f$  is at least 5, then  $y^2 = f(x)$  can have at most finitely many solutions in rational numbers. Meanwhile, if the degree of  $f$  is either 3 or 4, then the number of rational solutions can be finite or infinite, and there is no known algorithm to determine whether  $y^2 = f(x)$  has finitely or infinitely many solutions. If  $f$  has degree 3, then the curve  $C : y^2 = f(x)$  is called an *elliptic curve*.

In this article, we wish to study the question: how many rational points do we expect a *typical* hyperelliptic curve  $C : y^2 = f(x)$  of degree  $n$  to have? We begin in the next section by studying the case  $n = 3$  (and thus  $g = 1$ )—i.e., the important case of elliptic curves. This case is quite different and plays a special role in the theory for a number of reasons, as we now explain.

## 5. The case of elliptic curves

An *elliptic curve* over  $\mathbb{Q}$  is a genus one curve that has an equation of the form

$$E : y^2 = a_0x^3 + a_1x^2 + a_2x + a_3 \quad (5.1)$$

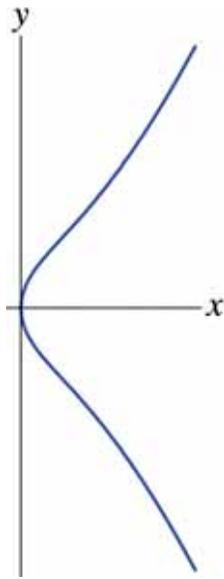
where  $a_0, a_1, a_2, a_3$  are rational numbers. By replacing  $y$  by  $y/a_0$  and  $x$  by  $x/a_0 - a_1/(3a_0)$  in (5.1), we may assume that  $a_0 = 1$  and  $a_1 = 0$ . Thus we may write any elliptic curve over  $\mathbb{Q}$  in the form

$$E_{A,B} : y^2 = x^3 + Ax + B. \quad (5.2)$$

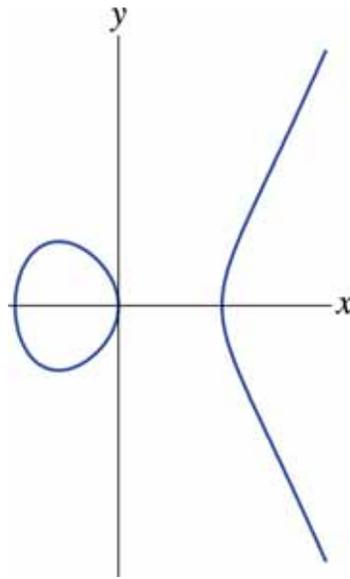
Moreover, by scaling  $y$  by  $c^3$  and  $x$  by  $c^2$  for appropriate  $c \in \mathbb{Q}$ , we may assume that  $A$  and  $B$  are integers with the property that, for every prime  $p$ , if  $p^4$  is a factor of  $A$  then  $p^6$  is not a factor of  $B$ . Every elliptic curve  $E$  over  $\mathbb{Q}$  can be expressed uniquely as such an  $E_{A,B}$ .

The case of elliptic curves is very special for a number of reasons. First, it is the first case where we do not know in general how to find all rational points. Second, it is the first case where we do not know in general how to determine whether there are finitely many or infinitely many rational points. Finally—and perhaps most importantly—it is the only case where the set of rational points on the curve always has an extra additional group structure, as we will explain shortly.

The graph of an elliptic curve in  $\mathbb{R}^2$  tends to look like one of the following two pictures:



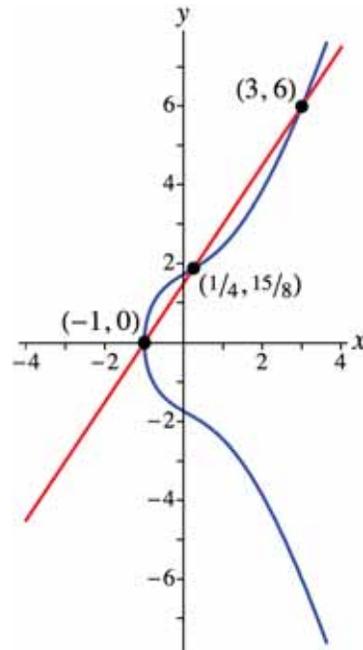
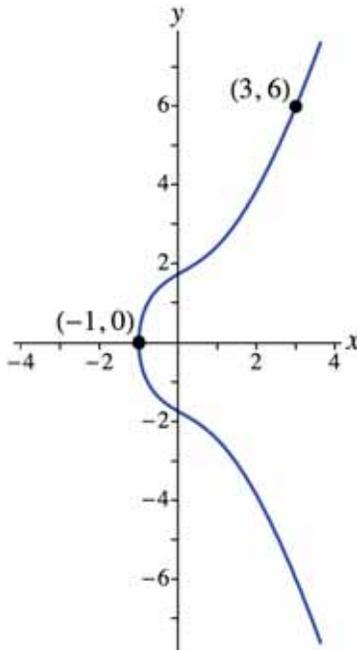
$$y^2 = x^3 + x$$



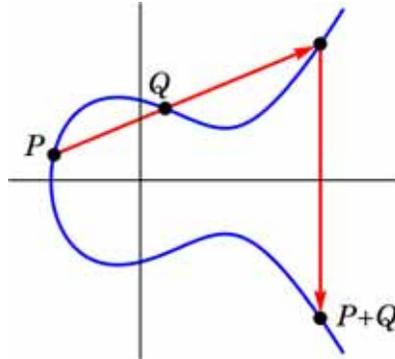
$$y^2 = x^3 - x$$

depending on whether the cubic polynomial in  $x$  has 1 real root or 3 real roots, respectively.

To define a group law on the set of rational points on an elliptic curve, we first note that if one has two rational points on a plane elliptic curve  $E$  over  $\mathbb{Q}$ , then the line connecting those two rational points always intersects  $E$  in a third rational point. (This is easy to prove: solving for the coordinates of this third point involves finding the root of a cubic polynomial over  $\mathbb{Q}$ , two of whose roots are already known to be rational; hence the third root must also be rational!) For example, knowing the two rational points  $(-1, 0)$  and  $(3, 6)$  on  $E : y^2 = x^3 + 2x + 3$  leads to a new rational point  $(1/4, 15/8)$  on  $E$  as seen below:

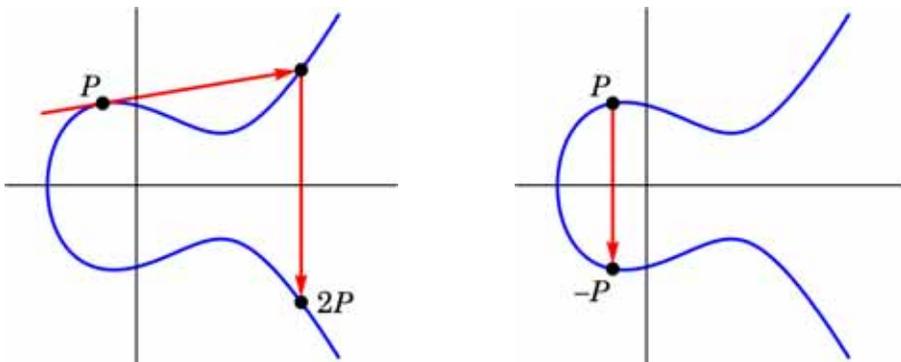


Given rational points  $P$  and  $Q$  on an elliptic curve  $E$ , we may define a rational point  $P + Q$  on  $E$  by taking the third point of intersection of  $E$  with the line connecting  $P$  and  $Q$ , and then reflecting this point across the  $x$ -axis:



One checks that—together with the point at infinity as the identity (which is the point one reaches when one follows the curve all the way up, or equivalently all the way down)—this law of addition endows the set of rational points on  $E$  with the structure of an abelian group.

How does one add  $P + P$ ? The line connecting  $P$  and  $P$  is the tangent to the curve at  $P$ ; hence to add  $P$  and  $P$ , one constructs the tangent to the curve at  $P$ , finds the other point of intersection with the curve, and then reflects it across the  $x$ -axis. Vertical lines in the plane seem to intersect only two (finite) points on the curve  $E$ ; the third point of intersection is the point at infinity. Therefore, reflecting a point  $P$  across the  $x$ -axis yields  $-P$ . (Note that the reflection of the point at infinity across the  $x$ -axis is still the point at infinity!) These two special cases of the addition law, where  $Q$  is equal to  $P$  or equal to the point at infinity, respectively, are illustrated below:



With these simple rules for addition, it is immediately clear that addition is commutative. It is not nearly as obvious that it is associative, but this can also be checked by working out the algebra using coordinates. Hence, with the point at infinity as the identity, this law of addition endows the set of rational points on  $E$  with the structure of an abelian group. (The modern viewpoint of algebraic geometry makes this fact clear—the group we are getting here is called the “divisor class group” of the elliptic curve  $E$  over  $\mathbb{Q}$ , i.e., the group of degree 0 divisors on  $E$ , modulo linear equivalence.)

## 6. Mordell's Theorem and the rank of an elliptic curve

For an elliptic curve  $E$ , the group of rational points on  $E$  is denoted by  $E(\mathbb{Q})$ . The basic theorem about the structure of this abelian group  $E(\mathbb{Q})$  of rational points on  $E$  is due to Mordell:

**Theorem 6.1** (Mordell). *The group  $E(\mathbb{Q})$  of rational points on  $E$  is finitely generated.*

Since  $E(\mathbb{Q})$  is a finitely generated abelian group, by the Fundamental Theorem of Finitely Generated Abelian Groups, we have

$$E(\mathbb{Q}) \cong \mathbb{Z}^r \oplus T$$

for some  $r \geq 0$  and  $T$  a finite abelian group. It is a theorem of Mazur that the group  $T$  is bounded in size by 16. Thus  $r$  measures how “big” the group  $E(\mathbb{Q})$  is. In particular, when  $r = 0$ , the group  $E(\mathbb{Q})$  is finite, and otherwise it is infinite. The quantity  $r$  is called the *rank* of  $E$ .

The rank of  $E$  in essence measures the number of points needed to generate all rational points on the curve, as the group  $T$  is usually trivial. By Mordell's Theorem, this number  $r$  is always finite.

The behavior of this fundamental invariant  $r$ , the rank, remains mysterious. For example, it is not known what ranks can occur for elliptic curves over  $\mathbb{Q}$ , or even whether ranks can take arbitrarily large values! The current record for the largest rank known is associated to a certain elliptic curve over  $\mathbb{Q}$  of rank  $\geq 28$ , found by Elkies in 2006.

Another question that naturally arises: Given an elliptic curve  $E$  over  $\mathbb{Q}$ , is there an algorithm that provably determines the rank of  $E$ ? This is an unsolved problem, although the Birch and Swinnerton-Dyer Conjecture does give a conjectural (positive!) answer to this question.

The next natural questions that arise are related to the *typical* behavior of the rank. Namely, what is the *expected* size of the rank? Do most curves have small rank (e.g., 0 or 1)? These are the two basic questions concerning the typical behavior of rational points on elliptic curves that we wish to address in this article.

## 7. The behavior of rank on average

Recall that any elliptic curve  $E$  over  $\mathbb{Q}$  can be expressed by a cubic equation

$$E : y^2 = x^3 + Ax + B, \tag{7.1}$$

where  $A$  and  $B$  are integers. We may define the *height* of  $E$  by the size of the coefficients of its defining equation (7.1). Since it is natural to compare the size of  $|A^3|$  to  $|B^2|$  (e.g., the discriminant of the cubic polynomial  $x^3 + Ax + B$  is  $-4A^3 - 27B^2$ ), the (*naive*) height  $H(E)$  of  $E$  is usually defined by  $H(E) := \max\{|4A^3|, 27B^2\}$ . The constants 4 and 27 are not of much importance here; some authors replace both of these constants with 1, which would not affect any of the discussion that follows.

There are various other notions of height for an elliptic curve, such as the Faltings height, the discriminant, and the conductor. Any of these notions of height could be used instead for the questions that follow (but are not expected to change any of the answers):

**Question 7.1.** When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, what is the average size of the rank?

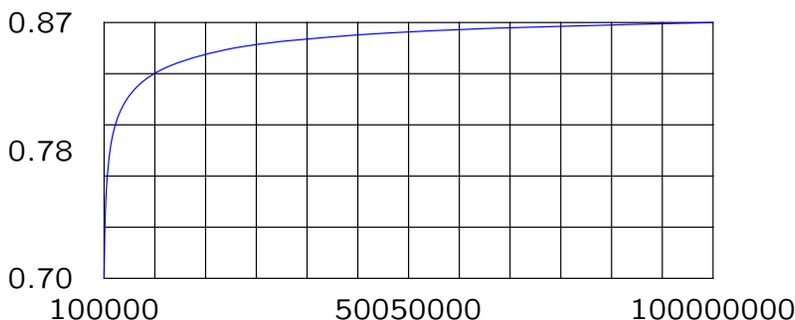
**Question 7.2.** When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, do most elliptic curves  $E$  have small rank, e.g., 0 or 1?

A conjectural answer to these questions was first given for certain families of elliptic curves by Goldfeld [43], and in a more general context by Katz and Sarnak [50]. They predicted:

**Conjecture 7.3** (Goldfeld, Katz–Sarnak). *The average rank of elliptic curves is  $1/2$ . (More precisely, one expects 50% of curves to have rank 0, and 50% to have rank 1.)*<sup>4</sup>

However, as far as proofs, previously this average has not even been known to be finite (let alone  $1/2$ )! Computations do not currently give much support to the conjecture either. It was observed by Brumer and McGuinness [27] in their 1990 computations that rank 2 curves seem to occur surprisingly often, and with *increasing* frequency! These computations were extended recently by Bektimirov, Stein, and Watkins; below is the resulting graph obtained from their computations, which plots the average rank of all elliptic curves of conductor less than  $N$ , where  $N$  ranges between  $10^5$  and  $10^8$ :

**Average rank of elliptic curves ordered by conductor**



Note that the average rank already starts at a value greater than  $1/2$  (indeed, at approximately .7), and then only increases from there; in particular, it is not clear from the graph whether it is increasing to some bounded constant greater than .86, or to infinity! In any case, it certainly does not seem to be approaching  $1/2$ . Moreover, the computations find that the proportion of rank 2 curves (conjectured to be 0% of all curves) is steadily *increasing* in this range.

When I first saw this graph, I immediately went and showed it to my colleague Peter Sarnak, and asked him how he could possibly believe his conjecture given this data! After a cursory glance at the picture, he thought for a second, and then told me that we must not have computed far enough and that the graph is clearly going to turn around and approach  $1/2$ ! For a beautiful survey explaining some of the reasons behind this conjecture, and the tension with the existing data, see [1].

Computations have not corroborated the conjecture to date. However, the first theoretical result towards the boundedness of average rank was given by Brumer [26]. In 1992, Brumer

<sup>4</sup>Thus the conjecture predicts that 100% of elliptic curves over  $\mathbb{Q}$  have rank 0 or 1; this is not to say that *no* elliptic curves have rank  $\geq 2$ , but only that such curves ought to be rare and have zero density.

showed that the Generalized Riemann Hypothesis (GRH) and the Birch and Swinnerton-Dyer Conjecture (BSD) together imply that the average rank is bounded (in fact, bounded by 2.3). In 2004, Heath-Brown [46] (still assuming GRH + BSD) improved this to average rank  $\leq 2.0$ . In 2006, Young [81] further improved this (again assuming GRH + BSD) to  $\leq \frac{25}{14} \approx 1.79$ . This latter achievement was significant in that it implied (assuming GRH + BSD) that a positive proportion of elliptic curves have rank 0 or 1.

Unconditionally, however, this problem on the boundedness of the average rank remained open. The fact that standard conjectures like GRH and BSD implied the boundedness of the average rank—but that the data did not support it—was a great motivation to study this problem more closely, and from a different point of view, and it finally led to an unconditional proof of the boundedness of average rank.

## 8. The average rank of elliptic curves is bounded

In recent work, an unconditional proof of the boundedness of the average rank of elliptic curves was obtained. In joint work with Arul Shankar [15], we showed that the (limsup of the) average rank of elliptic curves is bounded, and in fact bounded by 1.5. In a series of subsequent papers [16–18], the average rank bound has been improved to less than 1:

**Theorem 8.1** (joint work with Arul Shankar). *When elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, the average rank is bounded; in fact, the average rank is less than .885.*

Since the *average* rank is less than 1, it follows that a positive proportion (in fact, at least 11.5%) of elliptic curves must have rank 0. Since it is easy to see (e.g., by the Hilbert irreducibility theorem) that 100% of elliptic curves  $E$  have torsion subgroup  $T \subset E(\mathbb{Q})$  trivial, this shows for the first time the result that a positive proportion of elliptic curves have no rational points (other than the point at infinity).

Through a closer analysis, we in fact prove:

**Theorem 8.2** (joint work with Arul Shankar). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a proportion of at least 20.62% have rank 0.*

Thus at least 20.62% of elliptic curves have no rational points (except for the point at infinity)!

Recall that Conjecture 7.3 implies that 100% of elliptic curves have rank 0 or 1. In this direction, we obtain the following result.

**Theorem 8.3** (joint work with Arul Shankar). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a proportion of at least 83.75% have rank 0 or 1.*

Thus the vast majority of elliptic curves have small rank.

Note that the above results are primarily about, or are consequences of, *upper bounds* for the average rank. What about lower bounds? Previously the best known lower bound on the average rank was 0. We already have seen that a positive proportion of elliptic curves have only finitely many rational points (in fact, just the one rational point at infinity). The question naturally arises: do a positive proportion of elliptic curves have infinitely many rational points?

## 9. The average rank of elliptic curves is positive

In recent joint work with Christopher Skinner [19], we give an unconditional proof that a positive proportion of elliptic curves over  $\mathbb{Q}$  have rank 1.

**Theorem 9.1** (joint work with Christopher Skinner). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a positive proportion of them have rank 1.*

It follows that a positive proportion of elliptic curves over  $\mathbb{Q}$  have infinitely many rational points. In particular, the (liminf of the) average rank of elliptic curves is strictly positive:

**Corollary 9.2.** *When elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, the average rank is strictly positive.*

In forthcoming joint work with Christopher Skinner and Wei Zhang [20], we have made the above results more quantitative.

**Theorem 9.3** (joint work with Christopher Skinner and Wei Zhang). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a proportion of at least 20.68% have rank 1. Thus the average rank of elliptic curves is at least .2068.*

Thus at least 20.68% of elliptic curves over  $\mathbb{Q}$  have infinitely many rational points.

## 10. Most elliptic curves satisfy the Birch and Swinnerton-Dyer Conjecture

We may ask the analogous questions about what is called the “analytic rank” of an elliptic curve. Given an elliptic curve  $E$  over  $\mathbb{Q}$ , the *analytic rank* of  $E$  is defined as the order of vanishing at  $s = 1$  of an analytic function on  $\mathbb{C}$  associated to  $E$  called the *L-function*  $L(E, s)$  of  $E$ . The *L-function*  $L(E, s)$  of  $E$  is defined by a convergent product in the half-plane  $\operatorname{Re}(s) > 3/2$ ; the product is taken over all primes  $p$ , where the factor corresponding to a prime  $p$  is determined by the number of points on the curve  $E$  over the finite field  $\mathbb{Z}/p\mathbb{Z}$ . It was proven only in the 1990’s that  $L(E, s)$  could be analytically continued to the whole complex plane (in particular, to  $s = 1$ !); this follows from the works of Wiles, Taylor–Wiles, Diamond, Fujiwara, Conrad–Diamond–Taylor, and Breuil–Conrad–Diamond–Taylor.

The *Birch and Swinnerton-Dyer Conjecture* predicts that the analytic rank of the elliptic curve  $E$ —i.e. the order of vanishing at  $s = 1$  of the *L-function*  $L(E, s)$  of  $E$ —is equal to the rank of  $E$ . In other words, a certain analytic invariant of  $E$ —the analytic rank—is equal to the fundamental algebraic invariant of  $E$ —the rank.

The work of Gross–Zagier and Kolyvagin showed that the Birch and Swinnerton-Dyer conjecture holds for all curves having analytic rank  $\leq 1$  (which ought to be the case for 100% of elliptic curves). Nevertheless, it has not been known previously that this conjecture, or the condition of having analytic rank  $\leq 1$ , holds for more than 0% of elliptic curves.

Combining our results in §8 with the work of Skinner and Urban [71] on the Iwasawa Main Conjecture for  $\operatorname{GL}(2)$ , in [16] we were able to deduce:

**Theorem 10.1** (joint work with Arul Shankar). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a positive proportion of them have analytic rank 0; that is, a positive proportion of elliptic curves have nonvanishing *L-function* at  $s = 1$ .*

Meanwhile, the proof of Theorem 9.1 in fact entailed proving the following analogue of Theorem 10.1 for the case of analytic rank 1:

**Theorem 10.2** (joint work with Christopher Skinner). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a positive proportion of them have analytic rank 1; that is, a positive proportion of elliptic curves have vanishing  $L$ -function at  $s = 1$ .*

Thus a positive proportion of elliptic curves have analytic rank 0, and a positive proportion have analytic rank 1.

In particular, Theorem 10.1 (or Theorem 10.2) therefore implies:

**Corollary 10.3.** *A positive proportion of elliptic curves satisfy the Birch and Swinnerton-Dyer conjecture.*

In forthcoming joint work with Christopher Skinner and Wei Zhang [20], we have made the above results more quantitative:

**Theorem 10.4** (joint work with Christopher Skinner and Wei Zhang). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a proportion of at least 16.50% have both rank and analytic rank equal to 0; and a proportion of at least 20.68% have both rank and analytic rank equal to 1.*

Theorem 10.4 implies that at least 37.18% of elliptic curves satisfy the Birch and Swinnerton-Dyer Conjecture. The methods of [20], however, allow us to prove better lower bounds on the proportion of curves that have analytic rank 0 or 1 than for the sum of individual lower bounds on the proportions of curves having analytic rank 0 and those having analytic rank 1. We prove in [20]:

**Theorem 10.5** (joint work with Christopher Skinner and Wei Zhang). *When all elliptic curves  $E$  over  $\mathbb{Q}$  are ordered by height, a proportion of at least 66.48% satisfy the Birch and Swinnerton–Dyer Conjecture.*

Thus Theorem 10.5 implies that *most* elliptic curves  $E$  over  $\mathbb{Q}$  satisfy the Birch and Swinnerton-Dyer Conjecture.

## 11. Most hyperelliptic curves of higher genus and odd degree have only one rational point

What about rational points on hyperelliptic curves of higher genus  $g$ ? We consider first the case of odd degree  $n$ , where  $n = 2g + 1$ , which is the case that most naturally extends that of elliptic curves (which occurs when  $g = 1$ ). A hyperelliptic curve of odd degree  $n = 2g + 1$  can be expressed as

$$C : y^2 = a_0x^{2g+1} + a_1x^{2g} + a_2x^{2g-1} + \cdots + a_{2g}x + a_{2g+1} \quad (11.1)$$

where  $a_0, a_1, \dots, a_{2g+1}$  are rational numbers. As in the elliptic curve case, by replacing  $y$  by  $y/a_0^g$  and  $x$  by  $x/a_0 - a_1/((2g+1)a_0)$  in (11.1), we may assume that  $a_0 = 1$  and  $a_1 = 0$ . Thus we may write any hyperelliptic curve over  $\mathbb{Q}$  of odd degree  $n = 2g + 1$  in the form

$$C : y^2 = f(x) = x^{2g+1} + a_2x^{2g-1} + \cdots + a_{2g}x + a_{2g+1}. \quad (11.2)$$

Again, by scaling  $y$  by  $c^{2g+1}$  and  $x$  by  $c^2$  for appropriate  $c \in \mathbb{Q}$ , we may assume that  $a_2, \dots, a_{2g+1}$  are integers with the property that, for all primes  $p$ , it is not the case that  $a_i$  is divisible by  $p^{2i}$  for all  $i$ . Every hyperelliptic curve  $C$  over  $\mathbb{Q}$  of odd degree  $n = 2g + 1$  can be expressed uniquely in this manner.

With this normalization, we may again define the *height* of the hyperelliptic curve  $C$  by the size of the coefficients of its defining equation (11.2). In this case, it is natural to compare the sizes of  $|a_i^{1/i}|$  over all  $i$ , and so we define the height of the hyperelliptic curve (11.2) by  $H(C) := \max\{|a_i|^{2g(2g+1)/i}\}_{i=2, \dots, 2g+1}$ . We include the expression  $2g(2g + 1)$  in the definition so that the degree of the height function  $H$  is the same as that of the discriminant  $\Delta$  of the polynomial  $f(x)$  in (11.2).

The case of degree  $n = 3$  (i.e.,  $g = 1$ ) corresponds to the case of elliptic curves, and the above normalization (11.2) and definition of height are equivalent to those given for elliptic curves in the first paragraphs of §5 and §7, respectively. Like elliptic curves, hyperelliptic curves over  $\mathbb{Q}$  of odd degree have exactly one rational point at infinity, again obtained by following the curve all the way up (equivalently, all the way down).

What is the analogue, for higher genus curves, of the group structure on the set of rational points on an elliptic curve? Although the rational points on a hyperelliptic curve  $C$  of degree  $n$  do *not* in general have any natural group structure once  $n > 3$ , the “Jacobian”  $J(C)$  of  $C$  does. The *Jacobian* of an algebraic curve  $C$  over  $\mathbb{Q}$  of genus  $g$  is a dimension  $g$  algebraic variety (a higher dimensional analogue of an algebraic curve) over  $\mathbb{Q}$  whose rational points also naturally possess an abelian group structure. (More precisely, the Jacobian of  $C$  may be viewed as the group of degree 0 divisors on  $C$ , modulo linear equivalence; that this construction yields an algebraic variety over  $\mathbb{Q}$  was demonstrated by Weil.) The Jacobian of an elliptic curve is the elliptic curve itself—with group structure given as in §5—but for a general algebraic curve the dimension of the Jacobian variety is given by the genus  $g$ . A geometric construction of the Jacobian of a hyperelliptic curve over  $\mathbb{C}$ —as well as a geometric interpretation of the group law on the Jacobian (generalizing the construction in §5 for elliptic curves)—was given in a beautiful paper of Donagi [37] (see also [63], [35]).

The Mordell–Weil theorem, a generalization of Mordell’s Theorem 6.1, states that the group of rational points on the Jacobian of an algebraic curve over  $\mathbb{Q}$  (or indeed, any *abelian variety* over  $\mathbb{Q}$ ) is finitely generated. Thus, we may define the rank of the Jacobian  $J = J(C)$  of any algebraic curve  $C$  over  $\mathbb{Q}$  just as we defined the rank of an elliptic curve over  $\mathbb{Q}$ , namely, as the rank of the abelian group of rational points on  $J$ .

We may then ask the analogues of Questions 7.1 and 7.2 for Jacobians of odd degree hyperelliptic curves. When hyperelliptic curves over  $\mathbb{Q}$  of odd degree  $n$  are ordered by height, what is the average rank of their Jacobians? Do most of these Jacobians have rank at most 0 or 1?

In joint work with Benedict Gross [10], we proved the following generalization of the work in [15]:

**Theorem 11.1** (joint work with Benedict Gross). *When hyperelliptic curves over  $\mathbb{Q}$  of any fixed odd degree  $n \geq 3$  are ordered by height, the average rank of their Jacobians is at most  $3/2$ .*

Our method is uniform for all odd degrees  $n$  (including  $n = 3$ ); note also that our bound on the average rank is independent of the genus  $g$  (i.e., the dimension of the Jacobian!). This strongly suggests that the average rank of the Jacobians of odd degree hyperelliptic curves over  $\mathbb{Q}$  should be independent of the genus/dimension. Given Theorem 11.1, one may again

naturally conjecture that the average rank should always be  $1/2$ , with 100% having rank 0 or 1. In this direction, it follows from Theorem 11.1 that:

**Corollary 11.2** (joint work with Benedict Gross). *When hyperelliptic curves over  $\mathbb{Q}$  of any fixed odd degree  $n \geq 3$  are ordered by height, at least 50% of their Jacobians have rank 0 or 1.*

How are ranks of the Jacobians of hyperelliptic curves over  $\mathbb{Q}$  related to the rational points on these hyperelliptic curves? The method of Chabauty [29], as refined by Coleman [30], yields a finite and effective bound on the number of rational points on a curve over  $\mathbb{Q}$  whenever its genus is greater than the rank of its Jacobian. Since the average rank of the Jacobians of odd degree hyperelliptic curves of genus  $g$  is at most  $3/2$  (independent of the genus  $g$ ), this immediately implies, for the first time, that the number of rational points on most hyperelliptic curves of genus  $g > 1$  must be bounded by some absolute constant independent of  $g$ . Moreover, the proportion  $> 50\%$  implied by “most” in the latter sentence only improves as the genus  $g$  tends to infinity (since  $g$  is much larger than the average rank as  $g$  tends to infinity).

This Chabauty–Coleman-style analysis, in combination with the methods of the proof of Theorem 11.1, has recently been extended further by Bjorn Poonen and Michael Stoll [60], in order to prove the following theorem:

**Theorem 11.3** (Bjorn Poonen and Michael Stoll). *When hyperelliptic curves over  $\mathbb{Q}$  of odd degree  $n \geq 7$  are ordered by height, a positive proportion have only one rational point (namely, the rational point at infinity); moreover, the proportion of hyperelliptic curves of odd degree  $n$  having only one rational point approaches 100% as  $n$  tends to infinity.*

Thus “most” hyperelliptic curves of sufficiently large odd degree have only one rational point. Furthermore, a positive proportion of all hyperelliptic curves of any fixed odd degree  $n \geq 7$  have just the one rational point at infinity; note that the analogue of this result for  $n = 3$  was given by Theorem 8.2, while the case  $n = 5$  remains open.

## 12. Most hyperelliptic curves of higher genus and even degree have no rational points

We next turn to the case of hyperelliptic curves

$$C : y^2 = f(x) = a_0x^{2g+2} + a_1x^{2g+1} + \cdots + a_{2g+2} \quad (12.1)$$

over  $\mathbb{Q}$  having even degree  $n = 2g + 2$ . This is in a certain sense the “general case”, since the rational points on any hyperelliptic curve  $C : y^2 = f(x)$  of odd degree  $n - 1 = 2g + 1$  are in one-to-one correspondence with the rational points on the associated hyperelliptic curve  $C' : y^2 = g(x) := x^n f(1/x)$  of even degree  $n = 2g + 2$ . Hence the problem of studying rational points on odd degree hyperelliptic curves is a special case of the corresponding problem for even degree hyperelliptic curves.<sup>5</sup>

<sup>5</sup>In more geometric language, by a *hyperelliptic curve over  $\mathbb{Q}$*  we mean here a smooth, geometrically irreducible, complete curve  $C$  over  $\mathbb{Q}$  equipped with a fixed map of degree 2 to  $\mathbb{P}^1$  defined over  $\mathbb{Q}$ . Since any such map of degree 2 always has an even number of branch points, any hyperelliptic curve over  $\mathbb{Q}$  can be expressed in the form (12.1). Note that an odd degree hyperelliptic curve always has a branch point at infinity; hence an odd degree hyperelliptic curve is a hyperelliptic curve that possesses a marked rational Weierstrass point, namely, the rational branch point at infinity. (The two points at infinity on an even degree hyperelliptic curve are generically not rational.)

By a suitable change of variable, we may assume that all the  $a_i$  are integers in (12.1). Let us order all hyperelliptic curves  $C$  in (12.1) by the height  $H(C) := \max\{|a_i|\}$ . Our question is then: if hyperelliptic curves (12.1) over  $\mathbb{Q}$  are ordered by height, how many rational points do we expect these curves to have? In [6], we prove the following:

**Theorem 12.1.** *Fix  $g \geq 2$ . When ordered by height, most (i.e., more than 50% of) hyperelliptic curves (12.1) over  $\mathbb{Q}$  have no rational points. Moreover, the proportion of hyperelliptic curves having no rational points approaches 100% as  $g \rightarrow \infty$ .*

Thus Theorem 12.1 states that, as  $g \rightarrow \infty$ , a density approaching 100% of general hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  possess no rational points.

More precisely, let  $\rho_g$  denote the lower density of hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$ , when ordered by height, that have no rational points. Then  $\rho_g \rightarrow 1$  as  $g \rightarrow \infty$ . In fact, we prove that  $\rho_g = 1 - o(2^{-g})$ , so the convergence to 1 is quite rapid. Theorem 12.1 may thus be viewed as a “strong asymptotic form of Faltings’ Theorem” for hyperelliptic curves over  $\mathbb{Q}$ .

### 13. Hasse principle

The *Hasse local-global principle* for quadratic forms states that a quadratic form  $f(x_0, \dots, x_n)$  over  $\mathbb{Q}$  has a nontrivial rational zero if and only if it has a nontrivial zero over the real numbers  $\mathbb{R}$  and over the  $p$ -adic numbers  $\mathbb{Q}_p$  for all primes  $p$ . In other words, the quadratic form  $f(x_0, \dots, x_n) = 0$  has a nontrivial solution “globally” if and only if it does so “locally”. This principle is indeed what lies behind much of the rich arithmetic theory of quadratic forms. The question then naturally arises: to what extent does the Hasse principle hold for cubic (or higher degree) polynomials?

If the number of variables is large compared to the degree (e.g., cubic forms in fourteen variables—see [47]), then the Hasse principle holds. This follows from the Hardy–Littlewood–Ramanujan circle method. However, hyperelliptic curves involve only two variables, so it is not clear from such methods whether we would expect the Hasse principle to hold in general for such curves/equations.

By work of Poonen and Stoll [59] (see also [9]), a positive proportion of hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  are *locally soluble*, i.e., have points locally over  $\mathbb{R}$  and over  $\mathbb{Q}_p$  for every  $p$ . The methods of Theorem 12.1—which states that most hyperelliptic curves over  $\mathbb{Q}$  have *no* rational points—then allow us in [6] to show:

**Theorem 13.1.** *Most (i.e., more than 50% of) locally soluble hyperelliptic curves (12.1) of genus  $g > 1$  fail the Hasse principle. Moreover, the density of locally soluble hyperelliptic curves of genus  $g$  failing the Hasse principle approaches 100% as  $g \rightarrow \infty$ .*

By extending the techniques used to prove the results discussed in §8 and §9, we can also now show that a positive proportion of plane cubics (i.e., genus one curves in the plane) fail the Hasse principle; in addition, a positive proportion nontrivially satisfy the Hasse principle. See [7] for details.

## 14. Method of proof

Although the methods behind the proofs of all of these theorems vary, they do have one important common feature. Namely, in each case, we:

- 1) Identify a suitable representation  $V$  of an algebraic group  $G$  defined over  $\mathbb{Z}$ .
- 2) Find an explicit mapping from rational points on our curves  $C$  to orbits  $O$  of  $G(\mathbb{Z})$  on  $V(\mathbb{Z})$ , i.e., a mapping from rational points on  $C$  to lattice points  $v \in V(\mathbb{R}) \cong \mathbb{R}^n$  up to the action of  $G(\mathbb{Z})$ .

This mapping should also have the property that the coefficients of the equation defining the curve  $C$  exactly match up with the values on  $O$  (or  $v$ ) of a fixed set of generators for the polynomial invariants for the action of  $G$  on  $V$ .

- 3) Show that lattice points in  $V(\mathbb{R})$  in the image of this mapping—up to  $G(\mathbb{Z})$ -equivalence—are relatively *rare*, by using geometry-of-numbers arguments to asymptotically count the number of such lattice points having bounded invariants in a fundamental domain for the action of  $G(\mathbb{Z})$  on  $V(\mathbb{R})$ .

In general, all three of these steps can be quite nontrivial. In many cases, constructions from algebraic geometry may be used to identify a suitable representation  $(G, V)$  in Step 1). To carry out Step 2), one must carry out these geometric constructions first over a non-algebraically-closed field, and then—even more importantly—over  $\mathbb{Z}$ , so that one has the desired interpretation of the lattice points in  $V(\mathbb{R})$ .

To carry out the third step, we first count lattice points having bounded invariants in a fundamental region for the action of  $G(\mathbb{Z})$  on  $V(\mathbb{R})$ . The difficulty in carrying out this count is that the region in which we are counting lattice points is not bounded but has multiple cusps going off to infinity. Thus the situation is similar to that described in [3, 4] (see [5, §6] for a discussion); in the current cases, these difficulties are further magnified by the sheer sizes and numbers of the cusps involved. To count the number of points in the cusps, and to show that there are not too many lattice points of interest in these cusps, we use an averaging method introduced in [4] and developed further in the papers [15–18] and [6]. The basic idea is to note that all fundamental domains for  $G(\mathbb{Z})$  on  $V(\mathbb{R})$  contain the same number of lattice points having given invariants; thus we may average over a bounded continuum of fundamental domains (this is a “smoothing” technique) to simplify the problem. To make this idea work requires a suitable decomposition of the cuspidal regions of the fundamental domains; see [4, 6, 18] for examples of lattice point counts in such regions and a discussion of some general methods for carrying out such counts.

Once this count of lattice points having bounded invariants is complete, one must then sieve to obtain the (smaller) count of just those lattice points lying in the image of the map in Step 2). A general method to carry out such a sieve, which we refer to as the “geometric sieve”, is discussed in [8]; it is an adaptation and strengthening of a sieve introduced by Ekedahl [39] (see also Poonen [56, 57]).

**Example 14.1.** The representation  $(G, V)$  that we use for genus 2 curves  $C : y^2 = f(x) = a_0x^6 + a_1x^5 + \cdots + a_6$  is the action of  $G = \mathrm{SL}_6$  on the space  $V$  of pairs  $(A, B)$  of symmetric  $6 \times 6$  matrices. Indeed, a set of generators for the polynomial invariants of this action of  $G$  on  $(A, B) \in V$  consists of the coefficients of the degree 6 polynomial  $\mathrm{Det}(Ax - B)$  in  $x$ ; we would like these seven coefficients to be equal to the seven coefficients of  $f$  via the mapping in 2).

Suppose that we have a rational point on our curve  $C$ . By a suitable change of variable, there is no loss of generality in assuming that this rational point is located at  $x = 0$ , so that  $a_6 = c^2$  with  $c \in \mathbb{Z}$ . We then associate to this rational point  $(0, c)$  on our hyperelliptic curve the pair of symmetric matrices:

$$(A, B) = \left( \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & a_0 & a_1 & a_2 \\ 0 & 0 & 1 & a_1 & a_2 & a_3 \\ 0 & 1 & 0 & a_2 & a_3 & a_4 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & c \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & a_1 & a_2 & 0 \\ 0 & 1 & 0 & a_2 & a_3 & 0 \\ c & 0 & 0 & 0 & 0 & -a_5 \end{bmatrix} \right). \tag{14.1}$$

Note that the construction of this pair of matrices really involves  $c$ —i.e., it involves taking a square root of  $a_6$ . One easily verifies that this pair of symmetric  $6 \times 6$  matrices  $(A, B)$  has the remarkable property that

$$\text{Det}(Ax - B) = a_0x^6 + a_1x^5 + \dots + a_6.$$

On the other hand, we show by geometry-of-numbers arguments and a sieve [6, §§4–6] that many polynomials  $f(x)$  with integer coefficients cannot be expressed as  $\text{Det}(Ax - B)$  for any integer  $6 \times 6$  symmetric matrices  $A$  and  $B$ ! Furthermore, even those  $f$  that can be so expressed are often not expressible via a pair of integer symmetric matrices whose orbit contains an element  $(A, B)$  of the shape (14.1). This is what then allows us to deduce that most hyperelliptic curves over  $\mathbb{Q}$  do not possess any rational points, as stated in Theorems 12.1 and 13.1.

The construction of the map given by (14.1) involves first associating to a rational point on  $C$  an ideal class of order 2 in a certain special ring of rank 6 determined by  $f$ , and then associating to the latter data a pair of symmetric matrices; see [6, §2] for details.

We note that, in most cases, Steps 1)–3) as outlined above in fact yield a great deal of information beyond rational points. We now discuss the proofs of each of the results in Sections 8 through 13 in some more detail, and also describe this additional information obtained.

**14.1. Elliptic curves.** In the case of elliptic curves, we use the following representations: a)  $\text{SL}_2(\mathbb{Z})$  acting on the space  $\text{Sym}^4\mathbb{Z}^2$  of binary quartic forms; b)  $\text{SL}_3(\mathbb{Z})$  acting on the space  $\text{Sym}^3\mathbb{Z}^3$  of ternary cubic forms; c)  $\text{SL}_2(\mathbb{Z}) \times \text{SL}_4(\mathbb{Z})$  acting on the space  $\mathbb{Z}^2 \otimes \text{Sym}_2\mathbb{Z}^4$  of pairs of quaternary quadratic forms; and d)  $\text{SL}_5(\mathbb{Z}) \times \text{SL}_5(\mathbb{Z})$  acting on the space  $\mathbb{Z}^5 \otimes \wedge^2\mathbb{Z}^5$  of quintuples of  $5 \times 5$  skew-symmetric matrices. For each of these representations, the ring of invariants over  $\mathbb{C}$  is freely generated by two invariants, which we may call  $A$  and  $B$ , respectively. In each case, we wish to construct a mapping from rational points on  $E = E_{A,B} : y^2 = x^3 + Ax + B$  to orbits, for which the  $A$  and  $B$  of the elliptic curve match the invariants  $A$  and  $B$  associated to the elements of the orbit!

In the case of the representation a) above, this mapping from rational points on elliptic curves to orbits of  $\text{SL}_2(\mathbb{Z})$  on  $\text{Sym}^4\mathbb{Z}^2$  was first studied by Birch and Swinnerton-Dyer in [22]. They not only found a map from rational points of  $E = E_{A,B}$  to integral binary quartic forms, but they showed that the domain of this map naturally extends to elements of the 2-Selmer group  $\text{Sel}_2(E)$  of  $E$ .

For any integer  $m \geq 1$ , the  $m$ -Selmer group  $\text{Sel}_m(E)$  of an elliptic curve  $E$  over  $\mathbb{Q}$  is a finite abelian group of exponent  $m$  that fits into an exact sequence

$$0 \rightarrow E(\mathbb{Q})/mE(\mathbb{Q}) \rightarrow \text{Sel}_m(E) \rightarrow \text{III}_E[m] \rightarrow 0; \quad (14.2)$$

here  $\text{III}_E[m]$  denotes the  $m$ -torsion subgroup of a certain mysterious group called the *Tate–Shafarevich group*  $\text{III}_E$  of  $E$ . Thus the order of the  $m$ -Selmer group of an elliptic curve  $E$  over  $\mathbb{Q}$  gives an upper bound for  $m^r$ , where  $r$  is the rank of  $E$ .

Birch and Swinnerton-Dyer constructed a mapping as in Step 2)—not just from  $E(\mathbb{Q})$ —but from  $\text{Sel}_2(E)$  to  $\text{SL}_2(\mathbb{Z})$ -orbits of integral binary quartic forms. This connection between 2-Selmer group elements and integral binary quartic forms was first introduced and used in the original elliptic curve computations of Birch and Swinnerton-Dyer in order to efficiently compute ranks of elliptic curves, which led them to their celebrated conjecture. Indeed, this interpretation of binary quartic forms in terms of elements of 2-Selmer groups is still one of the fastest ways of computing and enumerating ranks of elliptic curves in practice, as in, e.g., Cremona’s influential `mwr` program.

The representations b), c), and d) above behave analogously, but with the 2-Selmer group replaced by the  $m$ -Selmer group for  $m = 3, 4$ , and  $5$  respectively. The relevant mappings in 2) were constructed by Cremona, Fisher, and Stoll [33] for  $m = 3, 4$  and by Fisher [40] for  $m = 5$  (building on earlier work of Cassels [28]). These results of Birch–Swinnerton-Dyer, Cremona–Fisher–Stoll, and Fisher had previously been used primarily, and very effectively, for the purposes of efficient arithmetic computation with elliptic curves.

By carrying out the above steps (particularly Step 3)) on these spaces, we showed in [15–18]:

**Theorem 14.2** (joint work with Arul Shankar). *Let  $m \leq 5$  be a positive integer. When all elliptic curves  $E$  are ordered by height, the average size of the  $m$ -Selmer group  $\text{Sel}_m(E)$  is  $\sigma(m)$ , the sum of the divisors of  $m$ .*

Theorem 14.2, and the sieve methods behind it, naturally lead us to conjecture that the average size of the  $m$ -Selmer group should be  $\sigma(m)$  for all positive integers  $m$  (cf. [17, Conjecture 4], [14, §1.3]).

It is the case  $m = 5$  of this theorem that ends up being the most useful as far as average rank bounds are concerned. First, we note that the case  $m = 5$  of Theorem 14.2 immediately yields an upper bound of 1.05 on the average rank of elliptic curves. If  $r$  denotes the rank of an elliptic curve  $E$ , then the size of the 5-Selmer group of  $E$  is an upper bound for  $5^r$ . Since  $20r - 15 \leq 5^r$  for any nonnegative integer  $r$ , we conclude by Theorem 14.2 that (the limsup of) the average rank  $\bar{r}$  of elliptic curves, when ordered by height, must satisfy  $20\bar{r} - 15 \leq 6$ , i.e.,  $\bar{r} \leq 21/20$ .

To improve this bound further, we observe that the bound of 1.05 can be attained only if 95% of elliptic curves have rank 1 and 5% have rank 2. However, it is widely expected that 50% of elliptic curves should have even rank and 50% should have odd rank. Indeed, the  $L$ -function of an elliptic curve  $E$  over  $\mathbb{Q}$  satisfies a functional equation, and the sign  $\pm 1$  of this functional equation is called the *root number* of  $E$ . The parity conjecture (implied by the Birch and Swinnerton–Dyer conjecture) states that the rank of an elliptic curve is even if and only if its root number is  $+1$ ; furthermore, one expects that the root numbers of elliptic curves should be equidistributed. Thus one expects that 50% of elliptic curves should have even rank, and 50% odd rank.

The parity conjecture has not been proven, but there is a remarkable result of Dokchitser and Dokchitser [36] which states that the parity of the  $p$ -Selmer rank of an elliptic curve is determined by the root number. This result suffices for our purposes because Theorem 14.2 yields bounds on not just the rank but the 5-Selmer group sizes of elliptic curves.

Thus a suitable result towards the equidistribution of root numbers of elliptic curves would imply a better upper bound on the average rank. In [18] (building on the works [16, 45, 64, 68]), we established such a result. Specifically, we proved that a majority of elliptic curves in fact do have equidistributed root number. More precisely, we showed that there exists a family  $F$  of elliptic curves  $E_{A,B}$ , having density greater than 55.01% among all elliptic curves when ordered by height and defined by congruence conditions on  $A$  and  $B$ , such that the root number of elliptic curves in  $F$  is equidistributed.

This result is then sufficient to prove the improved upper bound on the average rank of elliptic curves contained in Theorem 8.1. Similar arguments also then yield Theorems 8.2 and 8.3, namely, that at least 20.62% of elliptic curves over  $\mathbb{Q}$  have rank 0, and at least 83.75% have rank 0 or 1. We may further combine these methods with the work of Skinner and Urban [71] on the Iwasawa Main Conjecture for  $GL_2$  which implies, in particular, that if  $E$  is an elliptic curve over  $\mathbb{Q}$  satisfying certain mild congruence properties, and  $E$  has 5-Selmer rank 0, then  $E$  has analytic rank 0. This then allows us to deduce Theorem 10.1, namely, that a positive proportion of elliptic curves over  $\mathbb{Q}$  also have analytic rank 0.

Obtaining positive *lower* bounds on the average rank of elliptic curves requires additional ingredients. First, the parity argument we discussed above immediately implies also that a sizable proportion of elliptic curves  $E$  over  $\mathbb{Q}$  have 5-Selmer rank 1; since most such elliptic curves also have trivial 5-torsion in  $E(\mathbb{Q})$ , such elliptic curves have root number  $-1$  by the theorem of Dokchitser–Dokchitser, and thus should have rank 1 by the parity conjecture. To prove that such elliptic curves do in fact have rank 1, we require a theorem that assures us that an elliptic curve has rank 1 when it has trivial 5-torsion and 5-Selmer rank 1. A theorem of this nature, applicable to a sufficiently general class of elliptic curves, is proven in joint work with Christopher Skinner [19], by making use of a classical construction of Heegner together with a number of very recent advances in the subject: the Gross–Zagier formula, in the general form proved by Yuan, Zhang, and Zhang [82]; a  $p$ -adic variant of the Gross–Zagier formula due to Bertolini, Darmon, and Prasanna [2] and Brooks [24]; work on the Iwasawa Main Conjecture for  $GL_2$  by Wan [75], building on the earlier work of Skinner–Urban [71]; and a converse to a theorem of Gross, Zagier, and Kolyvagin due to Skinner [70]. Together with the methods of Theorem 14.2 and the parity arguments as described above, this then allows us in [19] to prove that a positive proportion of elliptic curves over  $\mathbb{Q}$  have rank 1 and thus infinitely many rational points (Theorem 9.1).

Through the additional use of the works of Zhang [83] and Skinner–Zhang [72] on indivisibility of Heegner points, in joint work with Skinner and Zhang [20] we are able to nontrivially quantify the results on analytic rank contained in Theorems 10.1 and 10.2. This leads to a proof that at least 16.50% of elliptic curves over  $\mathbb{Q}$  have both rank and analytic rank equal to 0; at least 20.68% of elliptic curves over  $\mathbb{Q}$  have rank and analytic rank equal to 1; and at least 66.48% of elliptic curves over  $\mathbb{Q}$  have analytic rank  $\leq 1$ , and thus at least 66.48% of elliptic curves over  $\mathbb{Q}$  satisfy the Birch and Swinnerton-Dyer Conjecture (Theorems 10.4 and 10.5).

**14.2. Odd degree hyperelliptic curves.** We now turn to the case of hyperelliptic curves of higher genus. Recall that, in the case of elliptic curves, the main initial idea was to use the

classical parametrization of 2-Selmer elements of elliptic curves by certain binary quartic forms ([22, Lemma 2]) to transform the problem into one of counting the integral orbits of the group  $SL_2$  on the space of binary quartic forms.

What is the analogue of the space of binary quartic forms for odd degree hyperelliptic curves of higher genus? A first guess might be binary  $(2g + 2)$ -ic forms, but this does not work. Indeed, such forms basically parametrize even degree hyperelliptic curves, rather than homogeneous spaces for the Jacobians of odd degree hyperelliptic curves. (It is a coincidence in the case of genus one that hyperelliptic curves of degree 4 give torsors for the Jacobians of curves of degree 3; in particular, it is a coincidence in the case of genus one that the Jacobian of a genus one curve is again a curve!)

To understand what is the generalization of the space of binary quartic forms to the case of higher genus, we observe that  $SL_2$  (or rather  $PSL_2$ ) may also be viewed as  $SO_3$ . This is because when  $SL_2$  acts on the space of binary quadratic forms  $ax^2 + bxy + cy^2$ , it fixes the discriminant  $A_0 = b^2 - 4ac$  (a ternary quadratic form!).

Consider the action of  $SO_3$  on the space of all ternary quadratic forms (i.e., on the symmetric square of its three-dimensional standard representation). This representation is six-dimensional. However, it is not irreducible, since the quadratic form  $A_0$  is fixed! The complementary 5-dimensional representation is irreducible, and indeed this is the representation of binary quartic forms (when the group is viewed as  $PSL_2$  rather than  $SO_3$ )!

We consider now the split quadratic form

$$A = \begin{bmatrix} & & & & & 1 \\ & & & & 1 & \\ & & & & & \\ & & & \ddots & & \\ & & 1 & & & \\ 1 & & & & & \end{bmatrix}$$

in  $2g + 1$  variables. Let  $SO_{2g+1}$  denote the orthogonal group of this quadratic form, and consider the action of this group on the symmetric square of its standard representation  $W$ . The quadratic form  $A$  is again fixed, and the complementary representation  $V$  (of dimension  $g(2g + 3)$ ) is irreducible. The elements  $B$  of  $V$  (the space of quadratic forms in  $2g + 1$  variables, modulo translation by  $A$ ) yield the desired generalization of binary quartic forms.

The action of  $SO_{2g+1}$  on  $V$  has  $2g$  independent invariants, given by the coefficients of the polynomial  $f(x)$  defined by

$$f(x) = \text{Disc}(Ax - B) = (-1)^g \det(Ax - B) = x^{2g+1} + c_2x^{2g-1} + \dots + c_{2g+1}.$$

We say that  $B$  is *nondegenerate* if  $\text{Disc}(f) \neq 0$ .

On the geometric side, we may associate to any nondegenerate element  $B$  in  $V(\mathbb{Q})$  a pencil of quadrics in projective space  $\mathbb{P}(W \oplus \mathbb{Q}) = \mathbb{P}^{2n+1}$ : two quadrics generating this pencil are

$$A' = \begin{bmatrix} A & \\ & 0 \end{bmatrix} \quad \text{and} \quad B' = \begin{bmatrix} B & \\ & 1 \end{bmatrix}.$$

The discriminant  $\text{Disc}(A'x - B'y)$  of this pencil is a homogeneous polynomial  $h(x, y)$  of degree  $2g + 2$  satisfying  $h(1, 0) = 0$  and  $h(x, 1) = f(x)$ . The Fano variety  $F_B$  of maximal common linear isotropic subspaces of this pencil of quadrics is a smooth variety of dimension  $g$  over  $\mathbb{Q}$ , and turns out to naturally form a principal homogeneous space for the

Jacobian  $J$  of the curve  $C : y^2 = f(x)$ ! Over an algebraically closed field, this was proven in work of Reid [63], Desale and Ramanan [35], and Donagi [37]. Over a non-algebraically closed field, it turns out that this homogeneous space always has order dividing 2. A general treatment of pencils of quadrics over arithmetic fields was given recently in the Ph.D. thesis of Wang [77].

Recall that that the 2-Selmer group  $\text{Sel}_2(J)$  of  $J$  is an elementary abelian 2-group that fits into an exact sequence

$$0 \rightarrow J(\mathbb{Q})/2J(\mathbb{Q}) \rightarrow \text{Sel}_2(J) \rightarrow \text{III}_J[2] \rightarrow 0, \tag{14.3}$$

where  $\text{III}_J[2]$  denotes the 2-torsion subgroup of the mysterious Tate-Shafarevich group  $\text{III}_J$  of  $J$ . In joint work with Benedict Gross [10], we prove that there is an injective map from the 2-Selmer group of  $J$  to the set of orbits of  $\text{SO}_{2g+1}(\mathbb{Z})$  on  $V(\mathbb{Z})$  having characteristic polynomial  $f(x)$ .

This therefore yields Steps 1) and 2) above for general odd degree hyperelliptic curves; in particular, the case  $g = 1$  recovers the correspondence of Birch and Swinnerton-Dyer between 2-Selmer elements of elliptic curves and binary quartic forms.

Step 3) is a bit more difficult to carry out in this case, because we must handle the cusps of not just one representation, but an infinite sequence of representations indexed by  $g$ ! A uniform argument for all  $g$  is developed in [10]. Together with a sieve, we obtain the following theorem:

**Theorem 14.3** (joint work with Benedict Gross). *Let  $g \geq 1$  be any integer. When all odd degree hyperelliptic curves  $C$  over  $\mathbb{Q}$  of genus  $g$  are ordered by height, the average size of the 2-Selmer group  $\text{Sel}_2(J)$  of their Jacobians  $J$  is at most 3, independent of  $g$ .*

Since  $|\text{Sel}_2(J)|$  is an upper bound for  $2^r$ , where  $r$  denotes the rank of  $J$ , and since  $2r \leq 2^r$ , we immediately obtain Theorem 11.1; namely, the average rank of the Jacobians of odd degree hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  is bounded above by  $3/2$ , independent of  $g$ .

Recall that the method of Chabauty [29], as refined by Coleman [30], gives a finite and effective bound on the number of rational points on a curve over  $\mathbb{Q}$  whenever its genus is greater than the rank of its Jacobian. Theorem 11.1 thus immediately implies that the density of odd degree hyperelliptic curves of genus  $g$  satisfying Chabauty’s condition tends to 1 as  $g$  tends to infinity! It follows, in particular, that one can effectively bound the number of rational points on most odd degree hyperelliptic curves.

Using a local equidistribution result for the upper bound in Theorem 14.3 (see [10, Theorem 12.4]), together with the method of Chabauty and Coleman (including a refinement due to McCallum [52]) and some additional local  $p$ -adic arguments, Poonen and Stoll [60] show that in fact Chabauty’s method yields only one point (namely, the point at infinity) on a positive proportion of odd degree hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$ , provided that  $g \geq 3$ ; moreover, the density of such curves having only one rational point tends to 1 as  $g$  tends to infinity! This is the content of Theorem 11.3. See [60] for details.

**14.3. General hyperelliptic curves.** Finally, how to treat general (i.e., even degree) hyperelliptic curves over  $\mathbb{Q}$ ? As explained in Footnote 5, such curves do not have any marked rational point, while odd degree hyperelliptic curves have a marked rational point at infinity. To get a handle on general hyperelliptic curves, we observe that, in the parametrization of 2-Selmer elements of odd degree hyperelliptic curves by pairs  $(A, B)$  of quadratic forms,

the marked point on the corresponding curve  $C : y^2 = \text{Disc}(Ax - B)$  comes about from the fact that  $A$  is a fixed split quadratic form of discriminant 1.

To “unmark” the point, we also allow  $A$  to vary! This leads to the representation of  $G(\mathbb{Z}) = \text{SL}_n(\mathbb{Z})$  on the space  $V(\mathbb{Z}) = \mathbb{Z}^2 \otimes \text{Sym}^2 \mathbb{Z}^n$  of pairs of quadratic forms in  $n$  variables, where  $n = 2g + 2$ . This is precisely the representation used in [6] to get a handle on rational points on general even degree hyperelliptic curves. Associated to the  $G(\mathbb{Z})$ -orbit of an element  $(A, B) \in V(\mathbb{Z})$  is again the curve  $C : y^2 = \text{Disc}(Ax - B)$ , which will now be a general, even degree hyperelliptic curve over  $\mathbb{Q}$ , as desired.

The algebra required to fully carry out Step 2) is perhaps the most subtle of all the cases discussed in this section. It requires using the ring of rank  $n$  associated to integral binary  $n$ -ic forms as in the work of Birch and Merriman [27], Nakagawa [54] (see also Wood [79]), and the association of pairs of integral  $n$ -ary quadratic forms to ideal classes of order 2 in such rings as in [5], the work of Morales [53], and especially the general work of Wood [80]. The end result, however, is very simple, and leads to the explicit mapping described in Example 14.1.

Unlike previous cases, the general orbits of the representation of  $G(\mathbb{Z})$  on  $V(\mathbb{Z})$  in this case are not used to parametrize elements of the 2-Selmer group of these hyperelliptic curves  $C$ , but rather the *fake 2-Selmer set* (see the work of Bruin and Stoll [25] for the precise definition); the fake 2-Selmer set is empty when there are no rational points. This is a new feature, which also must be dealt with in Step 3). Indeed, in previous cases, there was always at least one orbit associated to every curve (namely, the identity element of the associated Selmer group); in the case of even hyperelliptic curves, we use geometry-of-numbers arguments to prove that most even hyperelliptic curves over  $\mathbb{Q}$  have empty fake 2-Selmer sets, by showing in particular that the cusps of the fundamental domains contain a negligible number of points. This yields Theorem 12.1. Upon showing that most hyperelliptic curves over  $\mathbb{Q}$  have points over  $\mathbb{Q}_\nu$  for all places  $\nu$  of  $\mathbb{Q}$  as in [9, 59], this also shows that most locally soluble hyperelliptic curves over  $\mathbb{Q}$  fail the Hasse principle (Theorem 13.1). Moreover, we show that these failures of the Hasse principle are explained by what is called a *Brauer–Manin obstruction* to the existence of rational points. See [6] for details.

**14.4. Related works.** We end by describing how the methods in Steps 1)–3) also apply in other related works. In joint work with Benedict Gross and Xiaoheng Wang [11], we carry out a thorough study of the relationship between the arithmetic constructions of [6] and the geometric constructions of [35, 37, 63, 77]. This allows us to obtain not just a count of the average size of the fake 2-Selmer set of  $C$ , but in fact the fake 2-Selmer set of  $J^1 = \text{Pic}_{C/\mathbb{Q}}^1$  (see [11, §1]). Using the counting results of [6], together with a result of Dokchitser and Dokchitser [11, Appendix A] on parities of 2-Selmer ranks of the Jacobians of hyperelliptic curves  $C$  over  $\mathbb{Q}$ , allows us to prove that a positive proportion of general hyperelliptic curves of any genus  $g \geq 1$  have *no rational points over any odd degree extension of  $\mathbb{Q}$* . See [11] for details.

There have also been a number of recent results on curves having various types of marked rational points. In joint work with Wei Ho [13, 23], we determine the average sizes of 2- and 3-Selmer groups in various families of elliptic curves with marked rational points, such as curves with one or two general marked rational points, or curves with a marked rational 2-torsion or 3-torsion point. In particular, we prove that the average rank is bounded in all of these families of elliptic curves.

For hyperelliptic curves of higher genus, we have already mentioned the works [10, 60]

on odd degree hyperelliptic curves, i.e., hyperelliptic curves with a marked rational Weierstrass point. The case of hyperelliptic curves with a marked *non-Weierstrass point* has been studied thoroughly in recent work of Shankar and Wang [68]. Other families of hyperelliptic curves over  $\mathbb{Q}$  with various other types of marked rational points are also currently being pursued by a number of authors.

Finally, there has been much recent work on families of curves that are not necessarily hyperelliptic! Works of Ho [48], Gruson–Sam–Weyman [44], and especially Thorne [73] all identify representations corresponding to various families of non-hyperelliptic curves. Analogues of Steps 2) and 3) as described above are currently being worked out, and should eventually lead to results of the type described in this article also for various families of non-hyperelliptic curves.

## 15. Summary

- A hyperelliptic curve over  $\mathbb{Q}$  of degree  $n$  can be expressed by an equation of the form

$$y^2 = f(x)$$

where  $f(x)$  is a squarefree polynomial of degree  $n$  with integer coefficients. The genus  $g$  of such a curve is given by  $g = \lfloor \frac{n-1}{2} \rfloor$ , i.e.,  $n = 2g + 1$  or  $n = 2g + 2$ . The case  $n = 3$  corresponds to the case of *elliptic curves*.

- (joint work with Arul Shankar [18]) The average rank of elliptic curves over  $\mathbb{Q}$  is less than 1 (in fact, less than .885). At least 20.62% of elliptic curves over  $\mathbb{Q}$  have rank 0 and have no rational points (except that at infinity). At least 83.75% of elliptic curves over  $\mathbb{Q}$  have rank 0 or 1.
- (joint work with Christopher Skinner [19]) The average rank of elliptic curves over  $\mathbb{Q}$  is greater than 0. A positive proportion of elliptic curves over  $\mathbb{Q}$  have rank 1 and thus have infinitely many rational points.
- ([7]) Similar results hold also for degree  $n = 4$ ; namely, a positive proportion of hyperelliptic curves over  $\mathbb{Q}$  of degree 4 have no rational points, and a positive proportion have infinitely many rational points.
- (joint work with Christopher Skinner, and Wei Zhang [20]) At least 16.50% of elliptic curves over  $\mathbb{Q}$  have both rank and analytic rank equal to 0, and at least 20.68% have rank and analytic rank equal to 1. At least 66.48% of elliptic curves over  $\mathbb{Q}$  satisfy the Birch and Swinnerton-Dyer Conjecture.
- (joint work with Benedict Gross [10]) The average rank of the Jacobians of odd degree hyperelliptic curves over  $\mathbb{Q}$  of any fixed genus  $g \geq 1$  is bounded above by  $3/2$ , independent of the genus.
- (Bjorn Poonen and Michael Stoll [60]) For each  $g \geq 3$ , a positive proportion of odd degree hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  have only the one rational point at infinity. As  $g \rightarrow \infty$ , a density approaching 100% of odd degree hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  have only the one rational point at infinity.
- ([6]) For  $g > 1$ , most (i.e.,  $> 50\%$  of) general hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  have *no rational points*. Moreover, most such hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$

fail the Hasse principle. As  $g \rightarrow \infty$ , a density approaching 100% of hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  have no rational points. Furthermore, as  $g \rightarrow \infty$ , a density approaching 100% of locally soluble hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  fail the Hasse principle. These failures of the Hasse principle can be explained by a Brauer–Manin obstruction.

**Acknowledgements.** The author was partially funded by NSF grant DMS-1001828 and a Simons Investigator Grant. I thank Benedict Gross, Wei Ho, Bjorn Poonen, Arul Shankar, Christopher Skinner, Michael Stoll, Xiaoheng Wang, and Wei Zhang for their help.

## References

- [1] Bektimirov, B., Mazur, B., Stein, W., and Watkins, M., *Average ranks of elliptic curves: tension between data and conjecture*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 2, 233–254 (electronic).
- [2] Bertolini, M., Darmon, H., and Prasanna, K., *Generalized Heegner cycles and  $p$ -adic Rankin  $L$ -series*, Duke Math. J. **162** (2013), no. 6, 1033–1148.
- [3] Bhargava, M., *The density of discriminants of quartic rings and fields*, Ann. of Math. **162** (2005), 1031–1063.
- [4] ———, *The density of discriminants of quintic rings and fields*, Ann. of Math. **172** (2010), no. 3, 1559–1591.
- [5] ———, *Higher composition laws and applications*, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006.
- [6] ———, *Most hyperelliptic curves over  $\mathbb{Q}$  have no rational points*, <http://arxiv.org/abs/1308.0395>.
- [7] ———, *A positive proportion of plane cubics fail the Hasse principle*, <http://arxiv.org/abs/1402.1131>.
- [8] ———, *The geometric sieve and the density of squarefree values of invariant polynomials*, <http://arxiv.org/abs/1402.0031>.
- [9] Bhargava, M., Cremona, J., and Fisher, T., *The density of hyperelliptic curves over  $\mathbb{Q}$  of genus  $g$  that have points everywhere locally*, Preprint.
- [10] Bhargava, M. and Gross, B., *The average size of the 2-Selmer group of the Jacobians of hyperelliptic curves with a rational Weierstrass point*, <http://arxiv.org/abs/1208.1007>, in: Automorphic Representations and  $L$ -functions, TIFR Studies in Math. **22** (2013), 23–91.
- [11] Bhargava, M., Gross, B., and Wang, X., *Pencils of quadrics and the arithmetic of hyperelliptic curves*, <http://arxiv.org/abs/1310.7692>.
- [12] Bhargava, M. and Ho, W., *Coregular spaces and genus one curves*, <http://arxiv.org/abs/1306.4424>.
- [13] ———, *Average sizes of Selmer groups in families of elliptic curves*, Preprint.
- [14] Bhargava, M., Kane, D., Lenstra, H., Poonen, B., and Rains, E., *Modeling the distribution of ranks, Selmer groups, and Shafarevich–Tate groups of elliptic curves*, <http://arxiv.org/abs/1304.3971>.

- [15] Bhargava, M. and Shankar, A., *Binary quartic forms having bounded invariants, and the boundedness of the average rank of elliptic curves*, <http://arxiv.org/abs/1006.1002>, Ann. of Math., to appear.
- [16] ———, *Ternary cubic forms having bounded invariants and the existence of a positive proportion of elliptic curves having rank 0*, <http://arxiv.org/abs/1007.0052>, Ann. of Math., to appear.
- [17] ———, *The average number of elements in the 4-Selmer groups of elliptic curves is 7*, <http://arxiv.org/abs/1312.7333>.
- [18] ———, *The average size of the 5-Selmer group of elliptic curves is 6, and the average rank is less than 1*, <http://arxiv.org/abs/1312.7859>.
- [19] Bhargava, M. and Skinner, C., *A positive proportion of elliptic curves over  $\mathbb{Q}$  have rank one*, <http://arxiv.org/abs/1401.0233>.
- [20] Bhargava, M., Skinner, C., and Zhang, W., *A majority of elliptic curves over  $\mathbb{Q}$  satisfy the Birch and Swinnerton-Dyer Conjecture*, Preprint.
- [21] Birch, B. J. and Merriman, J. R., *Finiteness theorems for binary forms*, Proc. London Math. Soc. **s3-24** (1972), 385–394.
- [22] Birch, B. J. and Swinnerton-Dyer, H. P. F., *Notes on elliptic curves. I*, J. Reine Angew. Math. **212** (1963), 7–25.
- [23] Borel, A. and Harish-Chandra, *Arithmetic subgroups of algebraic groups*, Ann. of Math. **75** (1962), 485–535.
- [24] Brooks, E. H., *Generalized Heegner cycles, Shimura curves, and special values of  $p$ -adic  $L$ -functions*, Ph.D. thesis, University of Michigan, Ann Arbor, 2013.
- [25] Bruin, N., Stoll, M., *Two-cover descent on hyperelliptic curves*, Math. Comp. **78** (2009), no. 268, 2347–2370.
- [26] Brumer, A., *The average rank of elliptic curves I*, Invent. Math. **109** (1992), no. 3, 445–472.
- [27] Brumer, A. and McGuinness, O., *The behavior of the Mordell–Weil group of elliptic curves*, Bull. A.M.S. **23** (1990), no. 2, 375–382.
- [28] Cassels, J. W. S., *Arithmetic on curves of genus 1, IV, Proof of the Hauptvermutung*, J. Reine Angew. Math. **211** (1962), 95–112.
- [29] Chabauty, C., *Sur les points rationnels des courbes algébriques de genre supérieur à l'unité*, C. R. Acad. Sci. Paris **212** (1941), 882–885.
- [30] Coleman, R. F., *Effective Chabauty*, Duke Math. J. **52** (1985), no. 3, 765–770.
- [31] Colliot-Thélène, J.-L., and Poonen, B., *Algebraic families of nonzero elements of Shafarevich-Tate groups*, J. Amer. Math. Soc. **13** (2000), no. 1, 83–99.
- [32] Colliot-Thélène, J.-L. and Sansuc, J.-J., *La descente sur les variétés rationnelles II*, Duke Math. J. **54** (1987), 375–492.
- [33] Cremona, J., Fisher, T., and Stoll, M., *Minimisation and reduction of 2-, 3- and 4-coverings of elliptic curves*, Algebra & Number Theory **4** (2010), no. 6, 763–820.
- [34] de Jong, A. J., *Counting elliptic surfaces over finite fields*, Mosc. Math. J. **2** (2002), no. 2, 281–311.
- [35] Desale, U. V. and Ramanan, S., *Classification of vector bundles of rank 2 on hyperel-*

- liptic curves*, *Invent. Math.* **38** (1976), 161–185.
- [36] Dokchitser, T. and Dokchitser, V., *On the Birch–Swinnerton-Dyer quotients modulo squares*, *Ann. of Math.* **172** (2010), no. 1, 567–596.
- [37] Donagi, R., *Group law on the intersection of two quadrics*, *Annali della Scuola Normale Superiore di Pisa* **7** (1980), 217–239.
- [38] Dong Quan, N. N., *Algebraic families of hyperelliptic curves violating the Hasse principle*, Preprint.
- [39] Ekedahl, T., *An infinite version of the Chinese remainder theorem*, *Comment. Math. Univ. St. Paul.* **40** (1991), 53–59.
- [40] Fisher, T., *The invariants of a genus one curve*, *Proc. Lond. Math. Soc.* (3) **97** (2008), 753–782.
- [41] Fouvry, É., *Sur le comportement en moyenne du rang des courbes  $y^2 = x^3 + k$* , *Séminaire de Théorie des Nombres, Paris, 1990–91*, *Progr. Math.* **108**, Birkhäuser Boston, Boston, MA, 1993.
- [42] Fulton, W., *Algebraic Curves*, W. A. Benjamin, New York–Amsterdam, 1969.
- [43] Goldfeld, D., *Conjectures on elliptic curves over quadratic fields*, *Number theory, Carbondale 1979* (*Proc. Southern Illinois Conf., Southern Illinois Univ., Carbondale, Ill., 1979*), pp. 108–118, *Lecture Notes in Math.*, **751**, Springer, Berlin, 1979.
- [44] Gruson, L., Sam, S., and Weyman, J., *Moduli of abelian varieties, Vinberg theta-groups, and free resolutions*, <http://arxiv.org/abs/1203.2575>.
- [45] Halberstadt, E., *Signes locaux des courbes elliptiques en 2 et 3*, *C. R. Acad. Sci. Paris Sér. I Math.* **326** (1998), no. 9, 1047–1052.
- [46] Heath-Brown, D. R., *The average analytic rank of elliptic curves*, *Duke Math. J.* **122** (2004), no. 3, 591–623.
- [47] ———, *Cubic forms in 14 variables*, *Invent. Math.* **170**, 199–230.
- [48] Ho, W., *Orbit Parametrizations of Curves*, Ph.D. thesis, Princeton University, 2009.
- [49] Idoneal, *Are most cubic plane curves over the rationals elliptic?*, Jan. 10, 2010, <http://mathoverflow.net/questions/11349/>.
- [50] Katz, N. M. and Sarnak, P., *Random matrices, Frobenius eigenvalues, and monodromy*. *American Mathematical Society Colloquium Publications* **45**, American Mathematical Society, Providence, RI, 1999.
- [51] Littelmann, P., *Koreguläre und äquidimensionale Darstellungen*, *J. Algebra* **123** (1989), 193–222.
- [52] McCallum, W. G., *On the method of Coleman and Chabauty*, *Math. Ann.* **299** (1994), no. 3, 565–596.
- [53] Morales, J., *On some invariants for systems of quadratic forms over the integers*, *J. Reine Angew. Math.* **426** (1992), 107–116.
- [54] Nakagawa, J., *Binary forms and orders of algebraic number fields*, *Invent. Math.* **97** (1989), 219–235.
- [55] Panyushev, D., *On invariant theory of  $\theta$ -groups*, *J. Algebra* **283** (2005), 655–670.
- [56] Poonen, B., *Squarefree values of multivariable polynomials*, *Duke Math. J.* **118** (2003), no. 2, 353–373.

- [57] Poonen, B., *Bertini theorems over finite fields*, Ann. of Math. **160** (2004), no. 3, 1099–1127.
- [58] Poonen, B. and Rains, E., *Random maximal isotropic subspaces and Selmer groups*, J. Amer. Math. Soc. **25** (2012), 245–269.
- [59] B. Poonen and M. Stoll, *A local-global principle for densities*, Topics in number theory (University Park, PA, 1997), 241–244, Math. Appl. **467**, Kluwer Acad. Publ., Dordrecht, 1999.
- [60] Poonen, B. and Stoll, M., *Most odd degree hyperelliptic curves have only one rational point*, <http://arxiv.org/abs/1302.0061>.
- [61] Poonen, B. and Voloch, J. F., *Random diophantine equations*, in Arithmetic of higher-dimensional algebraic varieties, Progress in Math. (2004), Birkhäuser, 175–184.
- [62] Popov, V. L. and Vinberg, E. B., *Invariant Theory*, in Algebraic Geometry IV, Encyclopaedia of Mathematical Sciences **55**, Springer-Verlag, 1994.
- [63] Reid, M., *The complete intersection of two or more quadrics*, Ph.D. Thesis, Trinity College, Cambridge (1972).
- [64] Rohrlich, D. E., *Variation of the root number in families of elliptic curves*, Compositio Math. **87** (1993), no. 2, 119–151.
- [65] Saradha, N. and Srinivasan, A., *Generalized Lebesgue–Ramanujan–Nagell equations*, in Diophantine Equations (2008), Narosa, 207–223.
- [66] Schaefer, E. F., *2-descent on the Jacobians of hyperelliptic curves*, J. Number Theory **51** (1995), 219–232.
- [67] Selmer, E. S., *The Diophantine equation  $ax^3 + by^3 + cz^3 = 0$* , Acta Math. **85** (1957), 203–362.
- [68] Shankar, A. and Wang, X., *Average size of the 2-Selmer group for monic even hyperelliptic curves*, <http://arxiv.org/abs/1307.3531>.
- [69] Silverman, J. H., *The arithmetic of elliptic curves*, GTM **106**, Springer-Verlag, 1986.
- [70] Skinner, C., *A converse to a theorem of Gross, Zagier, and Kolyvagin*, Preprint.
- [71] Skinner, C., Urban, E., *The Iwasawa main conjectures for  $GL_2$* , Invent. Math **195** (2014), no. 1, 1–277.
- [72] Skinner, C. and Zhang, W., *Indivisibility of Heegner points in multiplicative cases*, Preprint.
- [73] Thorne, J., *The arithmetic of simple singularities*, Ph.D. Thesis, Harvard University, 2012.
- [74] ———, *Vinberg’s representations and arithmetic invariant theory*, Alg. & Num. Th. **7** (2013), No. 9, 2331–2368.
- [75] Wan, X., *The Iwasawa Main Conjecture for some Rankin products*, Preprint.
- [76] ———, *Pencils of quadrics and Jacobians of hyperelliptic curves*, Ph.D. thesis, Harvard University, 2013.
- [77] ———, *Maximal linear spaces contained in the base loci of pencils of quadrics*, <http://arxiv.org/abs/1302.2385>.
- [78] Wong, S., *On the density of elliptic curves*, Compositio Math. **127** (2001), no. 1, 23–54.

- [79] Wood, M., *Rings and ideals parametrized by binary  $n$ -ic forms*, J. London Math. Soc. (2) **83** (2011), 208–231.
- [80] ———, *Parametrization of ideal classes in rings associated to binary forms*, J. reine angew. Math. **689** (2014), 169–199.
- [81] Young, M. P., *Low-lying zeros of families of elliptic curves*, J. Amer. Math. Soc. **19** (2006), no. 1, 205–250.
- [82] Yuan, X., Zhang, S., and Zhang, W., *The Gross–Zagier Formula on Shimura Curves*, Annals of Math. Studies **184**, 2013.
- [83] Zhang, W., *Selmer groups and the indivisibility of Heegner points*, Preprint.

Department of Mathematics, Princeton University, Princeton, NJ 08544, USA

E-mail: bhargava@math.princeton.edu

# Singular stochastic PDEs

Martin Hairer

**Abstract.** We present a series of recent results on the well-posedness of very singular parabolic stochastic partial differential equations. These equations are such that the question of what it even means to be a solution is highly non-trivial. This problem can be addressed within the framework of the recently developed theory of “regularity structures”, which allows to describe candidate solutions locally by a “jet”, but where the usual Taylor polynomials are replaced by a sequence of custom-built objects. In order to illustrate the theory, we focus on the particular example of the Kardar-Parisi-Zhang equation, a popular model for interface propagation.

**Mathematics Subject Classification (2010).** 60H15, 81S20, 82C28.

**Keywords.** Regularity structures, renormalisation, stochastic PDEs.

## 1. Introduction

In this article, we report on a recently developed theory [23] allowing to give a robust meaning to a large class of stochastic partial differential equations (SPDEs) that have traditionally been considered to be ill-posed. The general structure of these equations is

$$\mathcal{L}u = F(u) + G(u)\xi, \quad (1.1)$$

where the dominant linear operator  $\mathcal{L}$  is of parabolic (or possibly elliptic) type,  $F$  and  $G$  are local nonlinearities depending on  $u$  and its derivatives of sufficiently low order, and  $\xi$  is some driving noise. Problems arise when  $\xi$  (and therefore also  $u$ ) is so singular that some of the terms appearing in  $F$  and / or the product between  $G$  and  $\xi$  are ill-posed. For simplicity, we will consider all of our equations in a *bounded* spatial region with periodic boundary conditions.

One relatively simple example of an ill-posed equation of the type (1.1) is that of a system of equations with a nonlinearity of Burgers type driven by space-time white noise:

$$\partial_t u = \partial_x^2 u + F(u) \partial_x u + \xi. \quad (1.2)$$

(See Section 2.2 below for a definition of the space-time white noise  $\xi$ .) Here,  $u(x, t) \in \mathbf{R}^n$  and  $F$  is a smooth matrix-valued function, so that one can in general not rewrite the nonlinearity as a total derivative. In this example, which was originally studied in [20] but then further analysed in the series of articles [24, 25, 29], solutions at any fixed instant of time have exactly the same regularity (in space) as Brownian motion. As a consequence,  $\partial_x u$  is expected to “look like” white noise. It is of course very well-known from the study of

ordinary stochastic differential equations (SDEs) that in this case the product  $F(u) \partial_x u$  is “unstable”: one can get different answers depending on the type of limiting procedure used to define it. This is the reason why one has different solution theories for SDEs: one obtains different answers, depending on whether they are interpreted in the Itô or in the Stratonovich sense [30, 43, 44].

Another example is given by the KPZ equation [32] which can formally be written as

$$\partial_t h = \partial_x^2 h + (\partial_x h)^2 - C + \xi, \quad (1.3)$$

and is a very popular model of one-dimensional interface propagation. As in the case of (1.2), one expects solutions to this equation to “look like” Brownian motion (in space) for any fixed instant of time. Now the situation is much worse however: the nonlinearity looks like the square of white noise, which really shouldn’t make any sense! In this particular case however, one can use a “trick”, the Cole-Hopf transform, to reduce the problem to an equation that has an interpretation within the framework of classical SPDE theory [4]. Furthermore, this “Cole-Hopf solution” was shown in [4] to be the physically relevant solution since it describes the mesoscopic fluctuations of a certain microscopic interface growth model, see also [17]. On the other hand, the problem of interpreting these solutions directly at the level of (1.3) and to show their stability under suitable approximations had been open for a long time, before being addressed in [21].

Both examples mentioned so far have only one space dimension. This particular feature (together with some additional structure in the case of the KPZ equation, see Remark 5.17 below) allowed to treat them by borrowing estimates and techniques from the theory of (controlled) rough paths [15, 18, 34]. This approach breaks down in higher spatial dimensions. More recently, a general theory of “regularity structures” was developed in [23], which unifies many previous approaches and allows in particular to treat higher dimensional problems.

Two nice examples of equations that can be treated with this new approach are given by

$$\partial_t \Phi = \Delta \Phi + C\Phi - \Phi^3 + \xi, \quad (1.4a)$$

$$\partial_t \Psi = -\Delta(\Delta \Psi + C\Psi - \Psi^3) + \operatorname{div} \xi, \quad (1.4b)$$

in space dimension  $d = 3$ . These equations can be interpreted as the natural “Glauber” and “Kawasaki” dynamics associated to Euclidean  $\Phi^4$  field theory in the context of stochastic quantisation [40]. It is also expected to describe the dynamical mesoscale fluctuations for phase coexistence models that are “almost mean-field”, see [5]. These equations cease to have function-valued solutions in dimension  $d \geq 2$ , so that the classical interpretation of the cubic nonlinearity loses its meaning there. In two dimensions, a solution theory for these equations was developed in [1], which was later improved in [10–12], see Section 3.1 below. The case  $d = 3$  (which is the physically relevant one in the interpretation as dynamical fluctuations for phase coexistence models) had remained open and was eventually addressed in [23].

A final example of the kind of equations that can be addressed by the theory exposed in these notes (but this list is of course not exhaustive) is a continuous analogue to the classical parabolic Anderson model [8]:

$$\partial_t u = \Delta u + u\eta + Cu, \quad (1.5)$$

in dimensions  $d \in \{2, 3\}$ . In this equation,  $\eta$  denotes a noise term that is white in space, but constant in time. This time, the problem is that in dimension  $d \geq 2$ , the product  $u\eta$  ceases to make sense classically, as a consequence of the lack of regularity of  $u$ .

The following “meta-theorem” (formulated in a somewhat vague sense, precise formulations differ slightly from problem to problem and can be found in the abovementioned articles) shows in which sense one can give meaning to all of these equations.

**Theorem 1.1.** *Consider the sequence of classical solutions to any of the equations (1.2)–(1.5) with  $\xi$  (resp.  $\eta$ ) replaced by a smooth regularised noise  $\xi_\varepsilon$  and  $C = C_\varepsilon$  depending on  $\varepsilon$ . Then, there exists a choice  $C_\varepsilon \rightarrow \infty$  such that this sequence of solutions converges to a limit in probability, locally in time. Furthermore, this limit is universal, i.e. does not depend on the details of the regularisation  $\xi_\varepsilon$ .*

Besides these convergence results, the important fact here is that the limit is *independent* of the precise details of the regularisation mechanism. In addition, the theory of regularity structures also yields rates of convergence, as well as an intrinsic description of these limits. It also provides automatically a very detailed local description of these limits.

The aim of this article is to give an overview of the ingredients involved in the proof of a result like Theorem 1.1. We structure this as follows. In Section 2, we recall a number of properties and definitions of Hölder spaces of positive (and negative!) order that will be useful for our argument. In Section 3, we then explain how, using only standard tools, it is possible to provide a robust solution theory for not-so-singular SPDEs, like for example (1.4) in dimension  $d = 2$ . Section 4 is devoted to a short overview of the main definitions and concepts of the abstract theory of regularity structures which is a completely general way of formalising the properties of objects that behave “like Taylor polynomials”. Section 5 then finally shows how one can apply this general theory to the specific context of the type of parabolic SPDEs considered above, how renormalisation procedures can be built into the theory, and how this affects the equations.

Throughout the whole article, our argumentation will remain mostly at the heuristic level, but we will make the statements and definitions as precise as possible.

**1.1. An alternative approach.** A different approach to building solution theories for singular PDEs was developed simultaneously to the one presented here by Gubinelli & Al in [19]. That approach is based on the properties of Bony’s paraproduct [2, 3, 7], in particular on the paraproduct formula. One advantage is that in the paraproduct-based approach one generally deals with globally defined objects rather than the “jets” used in the theory of regularity structures. This comes at the expense of achieving a less clean break between the analytical and the algebraic aspects of a given problem and obtaining less detailed information about the solutions. Furthermore, its scope is not as wide as that of the theory of regularity structures, see also Remark 5.17 below for more details.

## 2. Some properties of Hölder spaces

We recall in this section a few standard results from harmonic analysis that are very useful to have in mind. Note first that the linear part of all of the equations described in the introduction is invariant under some space-time scaling. In the case of the heat equation, this is the parabolic scaling. In other words, if  $u$  is a solution to the heat equation, then  $\tilde{u}(t, x) = u(\lambda^{-2}t, \lambda^{-1}x)$  is also a solution to the heat equation.

This suggests that we should look for solutions in function / distribution spaces respecting this scaling. Given a smooth compactly supported test function  $\varphi$  and a space-time

coordinate  $z = (t, x)$ , we henceforth write  $\varphi_z^\lambda(s, y) = \lambda^{-d-2}\varphi(\lambda^{-2}(s - t), \lambda^{-1}(y - x))$ , where  $d$  denotes the spatial dimension and the factor  $\lambda^{-d-2}$  is chosen so that the integral of  $\varphi_z^\lambda$  is the same as that of  $\varphi$ . In the case of the stochastic Cahn-Hilliard equation (1.4b), we would naturally use instead a temporal scaling of  $\lambda^{-4}$  and the prefactor would then be  $\lambda^{-d-4}$ .

With these notations at hand, we define spaces of distributions  $\mathcal{C}^\alpha$  for  $\alpha < 0$  in the following way. Denoting by  $\mathcal{B}_\alpha$  the set of smooth test functions  $\varphi: \mathbf{R}^{d+1} \rightarrow \mathbf{R}$  that are supported in the centred ball of radius 1 and such that their derivatives of order up to  $1 + |\alpha|$  are uniformly bounded by 1, we set

**Definition 2.1.** Let  $\eta$  be a distribution on  $d + 1$ -dimensional space-time and let  $\alpha < 0$ . We say that  $\eta \in \mathcal{C}^\alpha$  if the bound  $|\eta(\varphi_z^\lambda)| \lesssim \lambda^\alpha$  holds uniformly over all  $\lambda \in (0, 1]$ , all  $\varphi \in \mathcal{B}_\alpha$ , and locally uniformly over  $z \in \mathbf{R}^{d+1}$ .

For  $\alpha \geq 0$ , we say that a function  $f: \mathbf{R}^{d+1} \rightarrow \mathbf{R}$  belongs to  $\mathcal{C}^\alpha$  if, for every  $z \in \mathbf{R}^{d+1}$  there exists a polynomial  $P_z$  of (parabolic) degree at most  $\alpha$  and such that the bound

$$|f(z') - P_z(z')| \lesssim |z - z'|^\alpha,$$

holds locally uniformly over  $z$  and uniformly over all  $z'$  with  $|z' - z| \leq 1$ . Here, we say that a polynomial  $P$  in  $z = (t, x)$  is of parabolic degree  $n$  if each monomial is of the form  $z^k$  with  $|k| = 2|k_0| + \sum_{i \neq 0} k_i \leq n$ . In other words, the degree of the time variable ‘‘counts double’’. For  $z = (t, x)$ , we furthermore write  $|z| = |t|^{1/2} + |x|$ . (When treating (1.4b), powers of  $t$  count four times and one writes  $|z| = |t|^{1/4} + |x|$ .)

We now collect a few important properties of the spaces  $\mathcal{C}^\alpha$ .

**2.1. Analytical properties.** First, given a function and a distribution (or two distributions) it is natural to ask under what regularity assumptions one can give an unambiguous meaning to their product. It is well-known, at least in the Euclidean case but the extension to the parabolic case is straightforward, that the following result yields a sharp criterion for when, in the absence of any other structural knowledge, one can multiply a function and distribution of prescribed regularity [2, Thm 2.52].

**Theorem 2.2.** *Let  $\alpha, \beta \neq 0$ . Then, the map  $(f, g) \mapsto f \cdot g$  defined on all pairs of continuous functions extends to a continuous bilinear map from  $\mathcal{C}^\alpha \times \mathcal{C}^\beta$  to the space of all distributions if and only if  $\alpha + \beta > 0$ . Furthermore, if  $\alpha + \beta > 0$ , the image of the multiplication operator is  $\mathcal{C}^{\alpha \wedge \beta}$ .*

Another important property of these spaces is given by how they transform under convolution with singular kernels. Let  $K: \mathbf{R}^{d+1} \rightarrow \mathbf{R}$  be a function that is smooth away from the origin and supported in the centred ball of radius 1. One should think of  $K$  as being a truncation of the heat kernel  $\mathcal{G}$  in the sense that  $\mathcal{G} = K + R$  where  $R$  is a smooth space-time function. We then say that  $K$  is of order  $\beta$  (in the case of a truncation of the heat kernel one has  $\beta = 2$ ) if one can write  $K = \sum_{n \geq 0} K_n$  for kernels  $K_n$  which are supported in the centred ball of radius  $2^{-n}$  and such that

$$\sup_z |D^k K_n(z)| \lesssim 2^{((d+2)+|k|-\beta)n}, \tag{2.1}$$

for any fixed multiindex  $k$ , uniformly in  $n$ . Multiplying the heat kernel with a suitable partition of the identity, it is straightforward to verify that this bound is indeed satisfied.

With these notations at hand, one has the following very general Schauder estimate, see for example [41, 42] for special cases.

**Theorem 2.3.** *Let  $\beta > 0$ , let  $K$  be a kernel of order  $\beta$ , and let  $\alpha \in \mathbf{R}$  be such that  $\alpha + \beta \notin \mathbf{N}$ . Then, the convolution operator  $\eta \mapsto K \star \eta$  is continuous from  $\mathcal{C}^\alpha$  into  $\mathcal{C}^{\alpha+\beta}$ .*

**Remark 2.4.** The condition  $\alpha + \beta \notin \mathbf{N}$  seems somewhat artificial. It can actually be dispensed with by slightly changing the definition of  $\mathcal{C}^\alpha$ .

**2.2. Probabilistic properties.** Let now  $\eta$  be a random distribution, which we define in general as a continuous linear map  $\varphi \mapsto \eta(\varphi)$  from the space of compactly supported smooth test functions into the space of square integrable random variables on some fixed probability space  $(\Omega, \mathbf{P})$ . We say that it satisfies *equivalence of moments* if, for every  $p \geq 1$  there exists a constant  $C_p$  such that the bound

$$\mathbf{E}|\eta(\varphi)|^{2p} \leq C_p (\mathbf{E}|\eta(\varphi)|^2)^p,$$

holds for uniformly over all test functions  $\varphi$ . This is of course the case if the random variables  $\eta(\varphi)$  are Gaussian, but it also holds if they take values in an inhomogeneous Wiener chaos of fixed order [39].

Given a stationary random distribution  $\eta$  and a (deterministic) distribution  $C$ , we say that  $\eta$  has covariance  $C$  if  $\mathbf{E}\eta(\varphi)\eta(\psi) = \langle C \star \varphi, \psi \rangle$ , where  $\langle \cdot, \cdot \rangle$  denotes the  $L^2$ -scalar product. With this notation at hand, space-time white noise  $\xi$  is the Gaussian random distribution on  $\mathbf{R}^{d+1}$  with covariance given by the delta distribution. In other words,  $\xi(\varphi)$  is centred Gaussian for every  $\varphi$  and  $\mathbf{E}\xi(\varphi)\xi(\psi) = \langle \varphi, \psi \rangle_{L^2}$ .

Similarly to the case of stochastic processes, a random distribution  $\tilde{\eta}$  is said to be a *version* of  $\eta$  if, for every fixed test function  $\varphi$ ,  $\tilde{\eta}(\varphi) = \eta(\varphi)$  almost surely. One then has the following Kolmogorov criterion, a proof of which can be found for example in [23].

**Theorem 2.5.** *Let  $\eta$  be a stationary random distribution satisfying equivalence of moments and such that, for some  $\alpha < 0$ , the bound*

$$\mathbf{E}|\eta(\varphi_z^\lambda)|^2 \lesssim \lambda^{2\alpha},$$

*holds uniformly over  $\lambda \in (0, 1]$  and  $\varphi \in \mathcal{B}_\alpha$ . Then, for any  $\kappa > 0$ , there exists a  $\mathcal{C}^{\alpha-\kappa}$ -valued random variable  $\tilde{\eta}$  which is a version of  $\eta$ .*

From now on, we will make the usual abuse of terminology and not distinguish between different versions of a random distribution.

**Remark 2.6.** It follows immediately from the scaling properties of the  $L^2$  norm that one can realise space-time white noise as a random variable in  $\mathcal{C}^{-\frac{d}{2}-1-\kappa}$  for every  $\kappa > 0$ . This is sharp in the sense that it can *not* be realised as a random variable in  $\mathcal{C}^{-\frac{d}{2}-1}$ . This is akin to the fact that Brownian motion has sample paths belonging to  $\mathcal{C}^\alpha$  for every  $\alpha < \frac{1}{2}$ , but *not* for  $\alpha = \frac{1}{2}$ .

Let now  $K$  be a kernel of order  $\beta$  as before, let  $\xi$  be space-time white noise, and set  $\eta = K \star \xi$ . It then follows from either Theorem 2.5 directly, or from Theorem 2.3 combined with Remark 2.6, that  $\eta$  belongs almost surely to  $\mathcal{C}^\alpha$  for every  $\alpha < \beta - \frac{d}{2} - 1$ . We now turn to the question of how to define powers of  $\eta$ . If  $\beta \leq \frac{d}{2} + 1$ ,  $\eta$  is not a random function, so that its powers are in general undefined.

Recall that if  $\xi$  is space-time white noise and  $L^2(\xi)$  denotes the space of square-integrable random variables that are measurable with respect to the  $\sigma$ -algebra generated by  $\xi$ , then  $L^2(\xi)$  can be decomposed into a direct sum  $L^2(\xi) = \bigoplus_{m \geq 0} \mathcal{H}^m(\xi)$  so that  $\mathcal{H}^0$  contains constants,  $\mathcal{H}^1$  contains random variables of the form  $\xi(\varphi)$  with  $\varphi \in L^2$ , and  $\mathcal{H}^m$  contains suitable generalised Hermite polynomials of order  $m$  in the elements of  $\mathcal{H}^1$ , see [37, 39] for details. Elements of  $\mathcal{H}^m$  have a representation by square-integrable kernels of  $m$  variables, and this representation is unique if we impose that the kernel is symmetric under permutation of its arguments. In other words, one has a surjection  $I^{(m)} : L^2(\mathbf{R}^{d+1})^{\otimes m} \rightarrow \mathcal{H}^m$  and  $I^{(m)}(L) = I^{(m)}(L')$  if and only if the symmetrisations of  $L$  and  $L'$  coincide.

In the particular case where  $K$  is non-singular,  $\eta$  is a random function and its  $n$ th power  $\eta^n$  can be represented as

$$\eta^n(\varphi) = \sum_{2m < n} P_{m,n} C^m I^{(n-2m)}(K_\varphi^{(n-2m)}), \tag{2.2}$$

where

$$K_\varphi^{(r)}(z_1, \dots, z_r) := \int K(z - z_1) \cdots K(z - z_r) \varphi(z) dz,$$

for some combinatorial factors  $P_{m,n}$ . Here we have set  $C = \int K^2(z) dz$ . A simple calculation then shows that

**Proposition 2.7.** *If  $K$  is compactly supported, then  $K_\varphi^{(n)}$  is square integrable if the function  $(K \star \hat{K})^n$ , where  $\hat{K}(z) = K(-z)$ , is integrable.*

We now define the  $n$ th Wick power  $\eta^{\diamond n}$  of  $\eta$  as the random distribution given by only keeping the dominant term in (2.2):

$$\eta^{\diamond n}(\varphi) = I^{(n)}(K_\varphi^{(n)}).$$

By Proposition 2.8, this makes sense as soon as  $K \star \hat{K} \in L^n(\mathbf{R}^{d+1})$ . One then has the following result, a version of which can be found for example in [14].

**Proposition 2.8.** *Let  $K$  be a compactly supported kernel of order  $\beta \in (\frac{d+2}{2}(1 - \frac{1}{n}), \frac{d+2}{2})$  and let  $\eta = K \star \xi$  as above. Then,  $\eta^{\diamond n}$  is well-defined and belongs almost surely to  $C^\alpha$  for every  $\alpha < (2\beta - d - 2) \frac{n}{2}$ .*

*Proof.* A simple calculation shows that

$$|(K \star \hat{K})(z)|^n \lesssim |z|^{(2\beta - d - 2)n},$$

so that  $\|K_\varphi^{(n)}\|_{L^2}^2 \lesssim \lambda^{(2\beta - d - 2)n}$ . The claim then follows from Theorem 2.5, noting that random variables belonging to a Wiener-Itô chaos of finite order satisfy the equivalence of moments.  $\square$

It is important to note that this result is stable: replacing  $K$  by a smoothed kernel  $K_\varepsilon$  and letting  $\varepsilon \rightarrow 0$  yields convergence in probability of  $\eta_\varepsilon^{\diamond n}$  to  $\eta^{\diamond n}$  in  $C^\alpha$  (with  $\alpha$  as in the statement of the proposition) for most “reasonable” choices of  $K_\varepsilon$ . Furthermore, for fixed  $\varepsilon > 0$ , one has an explicit formula relating  $\eta_\varepsilon^{\diamond n}$  to  $\eta_\varepsilon$ :

$$\eta_\varepsilon^{\diamond n}(z) = H_n(\eta_\varepsilon(z), C_\varepsilon), \tag{2.3}$$

where the rescaled Hermite polynomials  $H_n(\cdot, C)$  are related to the standard Hermite polynomials by  $H_n(u, C) = C^{n/2} H_n(C^{-1/2}u)$  and we have set  $C_\varepsilon = \int K_\varepsilon^2(z) dz$ .

### 3. General methodology

The general methodology for providing a robust meaning to equations of the type presented in the introduction is as follows. We remark that the main reason why these equations seem to be ill-posed is that there is no canonical way of multiplying arbitrary distributions. The distributions appearing in our setting are however not arbitrary. For instance, one would expect solutions to semilinear equations of this type to locally “look like” the solutions to the corresponding linear problems. This is because, unlike hyperbolic or dispersive equations, parabolic (or elliptic) equations do not transport singularities. This gives hope that if one could somehow make sense of the nonlinearity, when applied to the solution to the linearised equation (which is a Gaussian process and therefore amenable to explicit calculations), then one could maybe give meaning to the equations themselves.

**3.1. The Da Prato-Debussche trick.** In some situations, one can apply this idea directly, and this was originally exploited in the series of articles [10–12]. Let us focus on the example of the dynamical  $\Phi^4$  model in dimension 2, which is formally given by

$$\partial_t \Phi = \Delta \Phi + C \Phi - \Phi^3 + \xi ,$$

where  $\xi$  is (spatially periodic) space-time white noise in space dimension 2.

Let now  $\xi_\varepsilon$  denote a smoothed version of  $\xi$  given for example by  $\xi_\varepsilon = \rho_\varepsilon \star \xi$ , where  $\rho_\varepsilon(t, x) = \varepsilon^{-4} \rho(\varepsilon^{-2}t, \varepsilon^{-1}x)$ , for some smooth compactly supported space-time mollifier  $\rho$ . In this case, denoting again by  $K$  a cut-off version of the heat kernel and noting that  $K$  is of order 2 (and therefore also of every order less than 2), it is immediate that  $\eta = K \star \xi$  satisfies the assumptions of Proposition 2.8 for every integer  $n$ .

In view of (2.3), this suggests that it might be possible to show that the solutions to

$$\begin{aligned} \partial_t \Phi_\varepsilon &= \Delta \Phi_\varepsilon + 3C_\varepsilon \Phi_\varepsilon - \Phi_\varepsilon^3 + \xi_\varepsilon \\ &= \Delta \Phi_\varepsilon - H_3(\Phi_\varepsilon, C_\varepsilon) + \xi_\varepsilon , \end{aligned} \tag{3.1}$$

with  $C_\varepsilon = \int K_\varepsilon^2(z) dz$  as above, where  $K_\varepsilon = \rho_\varepsilon \star K$ , converge to a distributional limit as  $\varepsilon \rightarrow 0$ . This is indeed the case, and the argument goes as follows. Writing  $\eta_\varepsilon = K_\varepsilon \star \xi$  and  $v_\varepsilon = \Phi_\varepsilon - \eta_\varepsilon$  with  $\Phi_\varepsilon$  the solution to (3.1), we deduce that  $v_\varepsilon$  solves the equation

$$\partial_t v_\varepsilon = \Delta v_\varepsilon - H_3(\eta_\varepsilon + v_\varepsilon, C_\varepsilon) + R_\varepsilon ,$$

for some smooth function  $R_\varepsilon$  that converges to a smooth limit  $R$  as  $\varepsilon \rightarrow 0$ . We then use elementary properties of Hermite polynomials to rewrite this as

$$\begin{aligned} \partial_t v_\varepsilon &= \Delta v_\varepsilon - (H_3(\eta_\varepsilon, C_\varepsilon) + 3v_\varepsilon H_2(\eta_\varepsilon, C_\varepsilon) + 3v_\varepsilon^2 \eta_\varepsilon + v_\varepsilon^3) + R_\varepsilon \\ &= \Delta v_\varepsilon - (\eta_\varepsilon^{\circ 3} + 3v_\varepsilon \eta_\varepsilon^{\circ 2} + 3v_\varepsilon^2 \eta_\varepsilon + v_\varepsilon^3) + R_\varepsilon . \end{aligned}$$

By Proposition 2.8 (and the remarks that follow), we see that  $\eta_\varepsilon^{\circ n}$  converges in probability to a limit  $\eta^{\circ n}$  in every space  $\mathcal{C}^\alpha$  for  $\alpha < 0$ . We can then *define* a random distribution  $\Phi$  by  $\Phi = \eta + v$ , where  $v$  is the solution to

$$\partial_t v = \Delta v - (\eta^{\circ 3} + 3v \eta^{\circ 2} + 3v^2 \eta + v^3) + R . \tag{3.2}$$

As a consequence of Theorem 2.3 (combined with additional estimates showing that the  $\mathcal{C}^\gamma$ -norm of  $K \star (f \mathbf{1}_{t>0})$  is small over short times provided that  $f \in \mathcal{C}^\alpha$  for  $\alpha \in (-2, 0)$ )

and  $\gamma < \alpha + \beta$ ), it is relatively easy to show that (3.2) has local solutions, and that these solutions are robust with respect to approximations of  $\eta^{\diamond n}$  in  $\mathcal{C}^\alpha$  for  $\alpha$  sufficiently close to 0. In particular, this shows that one has  $\Phi_\varepsilon \rightarrow \Phi$  in probability, at least locally in time for short times.

**Remark 3.1.** The dynamical  $\Phi^4$  model in dimension 2 was previously constructed in [1] (see also the earlier work [31] where a related but different process was constructed), but that construction relied heavily on *a priori* knowledge about its invariant measure and it was not clear how robust the construction was with respect to perturbations.

**3.2. Breakdown of the argument and a strategy to rescue it.** While the argument outlined above works very well for a number of equations, it unfortunately breaks down for the equations mentioned in the introduction. Indeed, consider again (1.4a), but this time in space dimension  $d = 3$ . In this case, one has  $\eta \in \mathcal{C}^{-\frac{1}{2}-\kappa}$  for every  $\kappa > 0$  and, by Proposition 2.8, one can still make sense of  $\eta^{\diamond n}$  for  $n < 5$ . One could therefore hope to define again a solution  $\Phi$  by setting  $\Phi = \eta + v$  with  $v$  the solution to (3.2). Unfortunately, this is doomed to failure: since  $\eta^{\diamond 3} \in \mathcal{C}^{-\frac{3}{2}-\kappa}$  (but no better), one can at best hope to have  $v \in \mathcal{C}^{\frac{1}{2}-\kappa}$ . As a consequence, both products  $v \cdot \eta^{\diamond 2}$  and  $v^2 \cdot \eta$  fall outside of the scope of Theorem 2.2 and we cannot make sense of (3.2).

One might hope at this stage that the Da Prato-Debussche trick could be iterated to improve things: identify the “worst” term in the right hand side of (3.2), make sense of it “by hand”, and try to obtain a well-posed equation for the remainder. While this strategy can indeed be fruitful and allows us to deal with slightly more singular problems, it turns out to fail in this situation. Indeed, no matter how many times we iterate this trick, the right hand side of the equation for the remainder  $v$  will *always* contain a term proportional to  $v \cdot \eta^{\diamond 2}$ . As a consequence, one can *never* hope to obtain a remainder of regularity better than  $\mathcal{C}^{1-\kappa}$  which, since  $\eta^{\diamond 2} \in \mathcal{C}^{-1-\kappa}$ , shows that it is not possible to obtain a well-posed equation by this method. See also Remark 5.17 below for a more systematic explanation of when this trick fails.

In some cases, one does not even know how to get started: consider the class of “classical” one-dimensional stochastic PDEs given by

$$\partial_t u = \partial_x^2 u + f(u) + g(u)\xi, \tag{3.3}$$

where  $\xi$  denotes space-time white noise,  $f$  and  $g$  are fixed smooth functions from  $\mathbf{R}$  to  $\mathbf{R}$ , and the spatial variable  $x$  takes values on the circle. Then, we know in principle how to use Itô calculus to make sense of (3.3) by rewriting it as an integral equation and interpreting the integral against  $\xi$  as an Itô integral, see [13]. However, this notion of solution is not very robust under approximations since space-time regularisations of the driving noise  $\xi$  typically destroy the probabilistic structure required for Itô integration. This is in contrast to the solution theory sketched in Section 3.1 which was very stable under approximations of the driving noise, even though it required suitable adjustments to the equation itself. Unfortunately, the argument of Section 3.1 (try to find some function / distribution  $\eta$  so that  $v = u - \eta$  has better regularity properties and then obtain a well-posed equation for  $v$ ) appears to break down completely.

The main idea now is that even though we may not be able to find a global object  $\eta$  so that  $u - \eta$  has better regularity, it might be possible to find a *local* object that does the trick at any one point. More precisely, setting  $\eta = K \star \xi$  as above (this time  $\eta$  is a Hölder continuous

function in  $\mathcal{C}^{\frac{1}{2}-\kappa}$  for every  $\kappa > 0$  by Theorems 2.3 and 2.5), one would expect solutions to (3.3) to be well approximated by

$$u(z') \approx u(z) + g(u(z))(\eta(z') - \eta(z)) . \quad (3.4)$$

The intuition is that since  $K$  is regular everywhere except at the origin, convolution with  $K$  is “almost” a local operator, modulo more regular parts. Since, near any fixed point  $z$ , we would expect  $g(u)\xi$  to “look like”  $g(u(z))\xi$  this suggests that near that point  $z$ , the function  $K \star (g(u)\xi)$  should “look like”  $g(u(z))\eta$ , which is what (3.4) formalises.

Note that this looks very much like a first-order Taylor expansion, but with  $\eta(z') - \eta(z)$  playing the role of the linear part  $z' - z$ . If we assume that (3.4) yields a good approximation to  $u$ , then one would also expect that

$$g(u(z')) \approx g(u(z)) + g'(u(z))g(u(z))(\eta(z') - \eta(z)) ,$$

so that  $g(u)$  has again a “first-order Taylor expansion” of the same type as the one for  $u$ . One could then hope that if we know somehow how to multiply  $\eta$  with  $\xi$ , this knowledge could be leveraged to define the product between  $g(u)$  and  $\xi$  in a robust way. It turns out that this is *not* quite enough for the situation considered here. However, this general strategy turns out to be very fruitful, provided that we also control higher-order local expansions of  $u$ , and this is precisely what the theory of regularity structures formalises [23, 26]. In particular, besides being applicable to (3.3), it also applies to all of the equations mentioned in the introduction.

## 4. Regularity structures

We now describe a very general framework in which one can formulate “Taylor expansions” of the type (3.4). We would like to formalise the following features of Taylor expansions. First, the coefficients of a Taylor expansion (i.e. the value and derivatives of a given function in the classical case or the coefficients  $u(z)$  and  $g(u(z))$  in the case (3.4)) correspond to terms of different degree / homogeneity and should therefore naturally be thought of as elements in some graded vector space. Second, an expansion around a given point can be reexpanded around a different point at the expense of changing coefficients, like so:

$$\begin{aligned} a \cdot 1 + b \cdot x + c \cdot x^2 &= (a + bh + ch^2) \cdot 1 + (b + 2ch) \cdot (x - h) + c \cdot (x - h)^2 , \\ u \cdot 1 + g(u) \cdot (\eta(z') - \eta(z)) &= (u + g(u)(\eta(z'') - \eta(z))) \cdot 1 + g(u) \cdot (\eta(z') - \eta(z'')) . \end{aligned}$$

Lastly, we see from these expressions that if we order coefficients by increasing homogeneity, then the linear transformation performing the reexpansion has an upper triangular structure with the identity on the diagonal.

**4.1. Basic definitions.** The properties just discussed are reflected in the following algebraic structure.

**Definition 4.1.** A regularity structure  $\mathcal{T} = (A, T, G)$  consists of the following elements:

1. A discrete index set  $A \subset \mathbf{R}$  such that  $0 \in A$  and  $A$  is bounded from below.
2. A *model space*  $T = \bigoplus_{\alpha \in A} T_\alpha$ , with each  $T_\alpha$  a Banach space; elements in  $T_\alpha$  are said to have *homogeneity*  $\alpha$ . Furthermore  $T_0$  is one-dimensional and has a distinguished basis vector  $\mathbf{1}$ . Given  $\tau \in T$ , we write  $\|\tau\|_\alpha$  for the norm of its component in  $T_\alpha$ .

3. A structure group  $G$  of (continuous) linear operators acting on  $T$  such that, for every  $\Gamma \in G$ , every  $\alpha \in A$ , and every  $\tau_\alpha \in T_\alpha$ , one has

$$\Gamma\tau_\alpha - \tau_\alpha \in T_{<\alpha} := \bigoplus_{\beta < \alpha} T_\beta . \tag{4.1}$$

Furthermore,  $\Gamma\mathbf{1} = \mathbf{1}$  for every  $\Gamma \in G$ .

The prime example of a regularity structure one should keep in mind is the one associated to Taylor polynomials on space-time  $\mathbf{R}^{d+1}$ . In this case, the space  $T$  is given by all polynomials in  $d + 1$  indeterminates  $X_0, \dots, X_d$ , with  $X_0$  representing the “time” coordinate. It comes with a canonical basis given by all monomials of the type  $X^k = X_0^{k_0} \dots X_d^{k_d}$  with  $k$  an arbitrary multiindex. The basis vector  $\mathbf{1}$  is the one corresponding to the zero multiindex. The space  $T$  has a natural grading by postulating that the homogeneity of  $X^k$  is  $|k| = 2k_0 + \sum_{i \neq 0} k_i$  and a natural norm by postulating that  $\|X^k\| = 1$ . In the case of the polynomial regularity structure, the structure group  $G$  is simply given by  $\mathbf{R}^{d+1}$ , endowed with addition, and acting on monomials by

$$\hat{\Gamma}_h X^k = (X - h)^k = (X_0 - h_0)^{k_0} \dots (X_d - h_d)^{k_d} . \tag{4.2}$$

It is immediate that all axioms of a regularity structure are satisfied in this case.

In the case of polynomials, there is a natural “realisation” of the structure  $\mathcal{T}$  at each space-time point  $z$ , which is obtained by turning an abstract polynomial into the corresponding concrete polynomial (viewed now as a real-valued function on  $\mathbf{R}^{d+1}$ ) based at  $z$ . In other words, we naturally have a family of linear maps  $\Pi_z : T \rightarrow \mathcal{C}^\infty(\mathbf{R}^d)$  given by

$$(\Pi_z X^k)(z') = (z'_0 - z_0)^{k_0} \dots (z'_d - z_d)^{k_d} . \tag{4.3}$$

It is immediate that the group  $G$  transforms these maps into each other in the sense that  $\Pi_z \hat{\Gamma}_h = \Pi_{z+h}$ . It is furthermore an immediate consequence of the scaling properties of monomials that the maps  $\Pi_z$  and the representation  $h \mapsto \hat{\Gamma}_h$  of  $\mathbf{R}^{d+1}$  are “compatible” with our grading for the model space  $T$ . More precisely, one has

$$\langle \varphi_z^\lambda, \Pi_z X^k \rangle = \lambda^{|k|} \langle \varphi, \Pi_0 X^k \rangle , \quad \|\hat{\Gamma}_h X^k\|_\ell = C_{k,\ell} |h|^{|k|-\ell} ,$$

for some constants  $C_{k,\ell}$  and every  $\ell \leq |k|$ . Here,  $\langle \cdot, \cdot \rangle$  denotes again the usual  $L^2$ -scalar product.

These observations suggest the following definition of a “model” for  $\mathcal{T}$ , where we impose properties similar to the ones we just found for the polynomial model. A model always requires the specification of an ambient space, together with a possibly inhomogeneous scaling. For definiteness, we will fix our ambient space to be  $\mathbf{R}^{d+1}$  endowed with the parabolic scaling as above. We also denote by  $\mathcal{S}'$  the space of all distributions (the letter  $\mathcal{D}$  is reserved for a different usage below). We also denote by  $L(E, F)$  the set of all continuous linear maps between the topological vector spaces  $E$  and  $F$ .

**Definition 4.2.** Given a regularity structure  $\mathcal{T}$ , a model for  $\mathcal{T}$  consists of maps

$$\mathbf{R}^{d+1} \ni z \mapsto \Pi_z \in L(T, \mathcal{S}') , \quad \mathbf{R}^{d+1} \times \mathbf{R}^{d+1} \ni (z, z') \mapsto \Gamma_{zz'} \in G ,$$

satisfying the algebraic compatibility conditions

$$\Pi_z \Gamma_{zz'} = \Pi_{z'} , \quad \Gamma_{zz'} \circ \Gamma_{z'z''} = \Gamma_{zz''} , \tag{4.4}$$

as well as the analytical bounds

$$|\langle \Pi_z \tau, \varphi_z^\lambda \rangle| \lesssim \lambda^\alpha \|\tau\|, \quad \|\Gamma_{zz'} \tau\|_\beta \lesssim |z - z'|^{\alpha - \beta} \|\tau\|. \quad (4.5)$$

Here, the bounds are imposed uniformly over all  $\tau \in T_\alpha$ , all  $\beta < \alpha \in A$ , and all test functions  $\varphi \in \mathcal{B}_r$  with  $r = \inf A$ , and locally uniformly in  $z$  and  $z'$ .

**Remark 4.3.** These definitions suggest a natural topology for the space  $\mathcal{M}$  of all models for a given regularity structure, generated by the following family of pseudo-metrics indexed by compact sets  $K$ :

$$\sup_{z \in K} \left( \sup_{\varphi, \lambda, \alpha, \tau} \lambda^{-\alpha} |\langle \Pi_z \tau - \bar{\Pi}_z \tau, \varphi_z^\lambda \rangle| + \sup_{|z - z'| \leq 1} \sup_{\alpha, \beta, \tau} |z - z'|^{\beta - \alpha} \|\Gamma_{zz'} \tau - \bar{\Gamma}_{zz'} \tau\|_\beta \right). \quad (4.6)$$

Here the inner suprema run over the same sets as before, but with  $\|\tau\| = 1$ .

**4.2. Hölder classes.** It is clear from the above discussion that if  $\mathcal{S}$  is the polynomial structure,  $\Pi$  is defined as in (4.3), and  $\Gamma_{zz'} = \hat{\Gamma}_{z' - z}$  with  $\hat{\Gamma}_h$  as in (4.2), then  $(\Pi, \Gamma)$  is a model for  $\mathcal{S}$  in the sense of Definition 4.2. Given an arbitrary regularity structure  $\mathcal{S}$  and an arbitrary model  $(\Pi, \Gamma)$ , it is now natural to define the corresponding ‘‘Hölder spaces’’ as spaces of distributions that can locally (near any space-time point  $z$ ) be approximated by  $\Pi_z \tau$  for some  $\tau \in T$ . This would be the analogue to the statement that a smooth function is one that can locally be approximated by a polynomial.

There is however one major difference with the case of smooth functions. It is of course the case that if  $f$  is smooth, then the coefficients of the Taylor expansion of  $f$  at any point are uniquely determined by the behaviour of  $f$  in the vicinity of that point. This is in general *not* the case anymore in the context of the framework we just described. To appreciate this fact, consider the following example. Fix  $\alpha \in (0, 1)$  and  $m \in \mathbf{N}$ , and take for  $\mathcal{S}$  the regularity structure where  $A = \{0, \alpha\}$ ,  $T_0 \cong \mathbf{R}$  with basis vector  $\mathbf{1}$ ,  $T_\alpha \cong \mathbf{R}^m$  with basis vectors  $(e_i)_{i \leq m}$ , and structure group  $G \cong \mathbf{R}^m$  acting on  $T$  via  $\hat{\Gamma}_h e_i = e_i - h_i \mathbf{1}$ . Let then  $W$  be an  $\mathbf{R}^m$ -valued  $\alpha$ -Hölder continuous function defined on the ambient space and set

$$\Pi_z \mathbf{1} = 1, \quad (\Pi_z e_i)(z') = W_i(z') - W_i(z), \quad \Gamma_{zz'} = \hat{\Gamma}_{W(z) - W(z')}.$$

Again, it is straightforward to verify that this does indeed define a model for  $\mathcal{S}$ . In fact, setting  $m = 1$  and  $W = \eta$ , this is precisely the structure one would use to formalise the expansion (3.4).

Let now  $F: \mathbf{R}^m \rightarrow \mathbf{R}$  be a smooth function and consider the function  $f$  on the ambient space given by  $f(z) = F(W(z))$ . For any  $z$ , we furthermore set

$$T \ni \hat{f}(z) = F(W(z)) \mathbf{1} + \sum_{i=1}^m (\partial_i F)(W(z)) e_i.$$

It then follows immediately from the usual Taylor expansion of  $F$  and the definition of the model  $(\Pi, \Gamma)$  that one has the bound

$$|f(z') - (\Pi_z \hat{f}(z))(z')| \lesssim |z - z'|^{2\alpha}, \quad (4.7)$$

so that in this context and with respect to this specific model, the function  $f$  behaves as if it were of class  $\mathcal{C}^{2\alpha}$  with ‘‘Taylor series’’ given by  $\hat{f}$ . In the case where the underlying

space is one-dimensional, this is precisely the insight exploited in the theory of rough paths [16, 35, 36] in order to develop a pathwise approach to stochastic calculus. More specifically, the perspective given here (i.e. controlling functions via analogues to Taylor expansion) is that of the theory of controlled rough paths developed in [18].

It is now very natural to ask whether, just like in the case of smooth functions, a bound of the type (4.7) is sufficient to uniquely specify  $\hat{f}(z)$  for every point  $z$ . Unfortunately, the answer to this question is that “it depends”. The reason is that while (4.5) imposes an upper bound on the behaviour of  $\Pi_z$  in the vicinity of  $z$ , it does *not* impose any corresponding lower bound. For example,  $W \equiv 0$  is an  $\alpha$ -Hölder continuous function that we could have used to build our model. In that case, the value of the  $e_i$ -component in  $\hat{f}$  is completely irrelevant for (4.7), so that uniqueness of the “Taylor series” fails. Suppose on the other hand that the underlying space is one-dimensional, that  $\alpha \in (\frac{1}{4}, \frac{1}{2})$ , and that  $W$  is a typical sample path of a Brownian trajectory. In this case it was shown in [27, Thm 3.4] that a bound of the type (4.7) is indeed sufficient to uniquely determine all the coefficients of  $\hat{f}$  (at least for almost all Brownian trajectories).

**Remark 4.4.** The fact that  $\hat{f}$  is uniquely determined by  $f$  in the Brownian case can be interpreted as an analogue to the fact that the Doob-Meyer decomposition of a semimartingale is unique. Since the statement given in [27] is quantitative, it can be interpreted as a deterministic analogue to Norris’s lemma, of which various incarnations can be found in [6, 33, 38].

Consider now a sequence  $W^\varepsilon$  of smooth (random) functions so that  $W^\varepsilon$  converges to Brownian motion in  $C^\alpha$  as  $\varepsilon \rightarrow 0$ . For definiteness, take for  $W_\varepsilon$  piecewise linear interpolations on a grid of size  $\varepsilon$ . Then, if we know *a priori* that we have a bound of the type (4.7) with a proportionality constant of order 1, this determines the coefficients of  $\hat{f}$  “almost uniquely” up to an error of order about  $\varepsilon^{2\alpha - \frac{1}{2}}$ .

What this discussion suggests is that we should really reverse our point of view from what we are used to: instead of fixing a function and asking whether it has a certain Hölder regularity by checking whether it is possible to find a “Taylor expansion” at each point satisfying a bound of the type (4.7), we should take the candidate expansion as our fundamental object and ask under which condition it does indeed approximate one single function / distribution around each point at the prescribed order. More precisely, fix some  $\gamma > 0$  (the order of our “Taylor expansion”) and consider a function  $f: \mathbf{R}^{d+1} \rightarrow T_{<\gamma}$ . Under which assumptions can we find a distribution  $\zeta$  such that  $\zeta$  “looks like” the distribution  $\Pi_z f(z)$  (in a suitable sense) near every point  $z$ ? We claim that the “right” answer is given by the following definition.

**Definition 4.5.** Given a regularity structure  $\mathcal{S}$  and a model  $(\Pi, \Gamma)$  as above, we define  $\mathcal{D}^\gamma$  as the space of functions  $f: \mathbf{R}^{d+1} \rightarrow T_{<\gamma}$  such that the bound

$$\|f(z) - \Gamma_{zz'} f(z')\|_\alpha \lesssim |z - z'|^{\gamma - \alpha} . \tag{4.8}$$

holds for every  $\alpha < \gamma$ , locally uniformly in  $z$  and  $z'$ .

**Remark 4.6.** This definition makes sense and is non-empty even for negative  $\gamma$ , as long as  $\gamma > \inf A$ .

**Remark 4.7.** The notation  $\mathcal{D}^\gamma$  is really an abuse of notation, since even for a given regularity structure there isn’t one single space  $\mathcal{D}^\gamma$ , but a whole collection of them, one for each model  $(\Pi, \Gamma) \in \mathcal{M}$ . More formally, one should really consider the space  $\mathcal{M} \times \mathcal{D}^\gamma$  consisting of

pairs  $((\Pi, \Gamma), f)$  such that  $f$  belongs to the space  $\mathcal{D}^\gamma$  based on the model  $(\Pi, \Gamma)$ . The space  $\mathcal{M} \times \mathcal{D}^\gamma$  also comes with a natural topology.

In the case where  $\mathcal{T}$  is the polynomial regularity structure and  $(\Pi, \Gamma)$  are the usual Taylor polynomials as above, one can see that this definition coincides with the usual definition of  $\mathcal{C}^\gamma$  (except at integer values where  $\mathcal{D}^1$  describes Lipschitz continuous functions, etc). In this case, the component  $f_0(z) = \langle \mathbf{1}, f(z) \rangle$  of  $f(z)$  in  $T_0$  (here we write  $\langle \mathbf{1}, \cdot \rangle$  for the basis element of  $T^*$  dual to  $\mathbf{1}$ ) is the only reasonable candidate for the function represented by  $f$ . Furthermore,  $\langle \mathbf{1}, \Gamma_{zz'} f(z') \rangle$  is nothing but the candidate Taylor expansion of  $f$  around  $z'$ , evaluated at  $z$ . The bound (4.8) with  $\alpha = 0$  is then just a statement of the fact that  $f_0$  is of class  $\mathcal{C}^\gamma$  and that  $f(z)$  is its Taylor series of order  $\gamma$  at  $z$ . The corresponding bounds for  $\alpha > 0$  then follow immediately, since they merely state that the  $\alpha$ th derivative of  $f_0$  is of class  $\mathcal{C}^{\gamma-\alpha}$ .

**4.3. The reconstruction operator.** The situation is much less straightforward when the model space  $T$  contains components of negative homogeneity. In this case, the bounds (4.5) allow the model  $\Pi_z$  to consist of genuine distributions and we do not anymore have an obvious candidate for the distribution represented by  $f$ . The following result shows that such a distribution nevertheless always exists and is unique as soon as  $\gamma > 0$ . This also provides an *a posteriori* justification for our definition of the spaces  $\mathcal{D}^\gamma$ .

**Theorem 4.8.** *Consider a regularity structure  $\mathcal{T} = (A, T, G)$  and fix  $\gamma > r = \inf A$ . Then, there exists a continuous map  $\mathcal{R}: \mathcal{M} \times \mathcal{D}^\gamma \rightarrow \mathcal{S}'$  (the “reconstruction map”) with the property that*

$$|(\mathcal{R}(\Pi, \Gamma, f) - \Pi_z f(z))(\varphi_z^\lambda)| \lesssim \lambda^\gamma, \quad (4.9)$$

*uniformly over  $\lambda \in (0, 1]$  and  $\varphi \in \mathcal{B}_r$ , and locally uniformly over  $z \in \mathbf{R}^{d+1}$ . Furthermore, for any given model  $(\Pi, \Gamma)$ , the map  $f \mapsto \mathcal{R}(\Pi, \Gamma, f)$  is linear. If  $\gamma > 0$ , the map  $\mathcal{R}$  is uniquely specified by the requirement (4.9).*

**Remark 4.9.** In the sequel, we will always consider  $(\Pi, \Gamma)$  as fixed and view  $\mathcal{R}$  as a linear map, writing  $\mathcal{R}f$  instead of  $\mathcal{R}(\Pi, \Gamma, f)$ . The above notation does however make it plain that the full map  $\mathcal{R}$  is not a linear map.

**Remark 4.10.** An important special case is given by situations where  $\Pi_z \tau$  happens to be a continuous function for every  $\tau \in T$  and every  $z$ . Then, it turns out that  $\mathcal{R}f$  is also a continuous function and one simply has

$$(\mathcal{R}f)(z) = (\Pi_z f(z))(z). \quad (4.10)$$

In the general case, this formula makes of course no sense since  $\Pi_z f(z)$  is a distribution and cannot be evaluated at  $z$ .

**Remark 4.11.** We made a slight abuse of notation here since there is really a family of operators  $\mathcal{R}^\gamma$ , one for each regularity. However, this abuse is justified by the following consistency relation. Given  $f \in \mathcal{D}^\gamma$  and  $\tilde{\gamma} < \gamma$ , one can always construct  $\tilde{f}$  by projecting  $f(z)$  onto  $T_{<\tilde{\gamma}}$  for every  $z$ . It turns out that one then necessarily has  $\tilde{f} \in \mathcal{D}^{\tilde{\gamma}}$  and  $\mathcal{R}\tilde{f} = \mathcal{R}f$ , provided that  $\tilde{\gamma} > 0$ . This is also consistent with (4.10) since, if  $\Pi_z \tau$  is a continuous function and the homogeneity of  $\tau$  is strictly positive, then  $(\Pi_z \tau)(z) = 0$ .

We refer to [23, Thm 3.10] for a full proof of Theorem 4.8 and to [22] for a simplified proof that only gives continuity in each “fiber”  $\mathcal{D}^\gamma$ . The main idea is to use a basis of compactly supported wavelets to construct approximations  $\mathcal{R}^n$  in such a way that our definitions can be exploited in a natural way to compare  $\mathcal{R}^{n+1}$  with  $\mathcal{R}^n$  and show that the sequence of approximations is Cauchy in a suitable space of distributions  $\mathcal{C}^\alpha$ . In the most important case when  $\gamma > 0$ , it turns out that while the existence of a map  $\mathcal{R}$  with the required properties is highly non-trivial, its uniqueness is actually quite easy to see. If  $\gamma \leq 0$  on the other hand, it is clear that  $\mathcal{R}$  cannot be uniquely determined by (4.9), since this bound remains unchanged if we add to  $\mathcal{R}$  any distribution in  $\mathcal{C}^\gamma$ . The existence of  $\mathcal{R}$  in the case  $\gamma < 0$  is however still a non-trivial result since in general one has  $\mathcal{R}f \notin \mathcal{C}^\gamma$ !

## 5. Regularity structures for SPDEs

We now return to the problem of providing a robust well-posedness theory for stochastic PDEs of the type (1.2), (1.4), (1.3), or even just (3.3). Our aim is to build a suitable regularity structure for which we can reformulate our SPDE as a fixed point problem in  $\mathcal{D}^\gamma$  for a suitable value of  $\gamma$ .

**Remark 5.1.** Actually, it turns out that since we are interested in Cauchy problems, there will always be some singularity at  $t = 0$ . This introduces additional technical complications which we do not wish to dwell upon.

**5.1. General construction of the model space.** Our first task is to construct the model space  $T$ . Since we certainly want to be able to represent arbitrary smooth functions (for example in order to be able to take into account the contribution of the initial condition), we want  $T$  to contain the space  $\bar{T}$  of abstract polynomials in  $d + 1$  indeterminates endowed with the parabolic grading described in Section 4.1. Since the noise  $\xi$  cannot be adequately represented by polynomials, we furthermore add a basis vector  $\Xi$  to  $T$ , which we postulate to have some homogeneity  $\alpha < 0$  such that  $\xi \in \mathcal{C}^\alpha$ . In the case of space-time white noise, we would choose  $\alpha = -\frac{d}{2} - 1 - \kappa$  for some (typically very small) exponent  $\kappa > 0$ .

At this stage, the discussion following (3.4) suggests that if our structure  $T$  contains a basis vector  $\tau$  of homogeneity  $\beta$  representing some distribution  $\eta$  involved in the description of the right hand side of our equation, then it should also contain a basis vector of homogeneity  $\beta + 2$  (the “2” here comes from the fact that convolution with the heat kernel yields a gain of 2 in regularity) representing the distribution  $K \star \eta$  involved in the description of the solution to the equation. Let us denote this new basis vector by  $\mathcal{I}(\tau)$ , where  $\mathcal{I}$  stands for “integration”. In the special case where  $\tau \in \bar{T}$ , so that it represents an actual polynomial, we do not need any new symbol since  $K$  convolved with a polynomial yields a smooth function. One way of formalising this is to simply postulate that  $\mathcal{I}(X^k) = 0$  for every multiindex  $k$ .

**Remark 5.2.** For consistency, we will also always assume that  $\int K(z)Q(z) dz = 0$  for all polynomials  $Q$  of some fixed, but sufficiently high, degree. Since  $K$  is an essentially arbitrary truncation of the heat kernel, we can do this without loss of generality.

If the right hand side of our equation involves the spatial derivatives of the solution, then, for each basis vector  $\tau$  of homogeneity  $\beta$  representing some distribution  $\eta$  appearing in the description of the solution, we should also have a basis vector  $\mathcal{D}_i\tau$  of homogeneity

$\beta - 1$  representing  $\partial_i \eta$  and appearing in the description of the derivative of the solution in the direction  $x_i$ .

Finally, if the right hand side of our equation involves a product between two terms  $F$  and  $\bar{F}$ , and if basis vectors  $\tau$  and  $\bar{\tau}$  respectively are involved in their description, then we should also have a basis vector  $\tau\bar{\tau}$  which would be involved in the description of the product. If  $\tau$  and  $\bar{\tau}$  represent the distributions  $\eta$  and  $\bar{\eta}$  respectively, then this new basis vector represents the distribution  $\eta\bar{\eta}$ , whatever this actually means. Regarding its homogeneity, by analogy with the case of polynomials, it is natural to impose that the homogeneity of  $\tau\bar{\tau}$  is the sum of the homogeneities of its two factors.

This suggests that we should build  $T$  by taking as its basis vectors some formal expressions built from the symbols  $X$  and  $\Xi$ , together with the operations  $\mathcal{I}(\cdot)$ ,  $\mathcal{D}_i$ , and multiplication. Furthermore, the natural way of computing the homogeneity of a formal expression in view of the above is to associate homogeneity 2 to  $X_0$ , 1 to  $X_i$  for  $i \neq 0$ ,  $\alpha$  to  $\Xi$ , 2 to  $\mathcal{I}(\cdot)$ , and  $-1$  to  $\mathcal{D}_i$ , and to simply add the homogeneities of all symbols appearing in any given expression. Denote by  $\mathcal{F}$  the collection of all formal expressions that can be constructed in this way and denote by  $|\tau|$  the homogeneity of  $\tau \in \mathcal{F}$ , so we have for example

$$|X_i \Xi| = \alpha + 1, \quad |\mathcal{I}(\Xi)^2 \mathcal{I}(X_i \mathcal{D}_j \mathcal{I}(\Xi))| = 3\alpha + 8, \quad \text{etc.}$$

We note however that if we simply took for  $T$  the space of linear combinations of *all* elements in  $\mathcal{F}$  then, since  $\alpha < 0$ , there would be basis vectors of arbitrarily negative homogeneity, which would go against Definition 4.1. What saves us is that most formal expressions are not needed in order to formulate our equations as fixed point problems. For example, the expression  $\Xi^2$  is useless since we would never try to square the driving noise. Similarly, if we consider (1.4a), then  $\mathcal{I}(\Xi)$  is needed for the description of the solution, which implies that  $\mathcal{I}(\Xi)^2$  and  $\mathcal{I}(\Xi)^3$  are needed to describe the right hand side, but we do not need  $\mathcal{I}(\Xi)^4$  for example.

**5.2. Specific model spaces.** This suggests that we should take  $T$  as the linear combinations of only those formal expressions  $\tau \in \mathcal{F}$  that are actually expected to appear in the description of the solution to our equation or its right hand side. Instead of trying to formulate a general construction (see [23, Sec. 8.1] for such an attempt), let us illustrate this by a few examples. We first focus on the case of the KPZ equation (1.3) and we construct subsets  $\mathcal{U}$  and  $\mathcal{V}$  of  $\mathcal{F}$  that are used in the description of the solution and the right hand side of the equation respectively. These are defined as the smallest subsets of  $\mathcal{F}$  with the following properties:

$$\mathcal{T} \subset \mathcal{U} \cap \mathcal{V}, \quad \{\mathcal{I}(\tau) : \tau \in \mathcal{V} \setminus \mathcal{T}\} \subset \mathcal{U}, \quad \{\Xi\} \cup \{\mathcal{D}\tau_1 \cdot \mathcal{D}\tau_2 : \tau_i \in \mathcal{U}\} \subset \mathcal{V}. \quad (5.1)$$

where we used the notation  $\mathcal{T} = \{X^k\}$  with  $k$  running over all multiindices, so that the space of Taylor polynomials  $\bar{T}$  is the linear span of  $\mathcal{T}$ . We then define  $T$  as the space of all linear combinations of elements of  $\mathcal{U} \cup \mathcal{V}$ . We also denote by  $T_{\mathcal{U}}$  the subspace of  $T$  spanned by  $\mathcal{U}$ . This construction is such that if we have any function  $H : \mathbf{R}^{d+1} \rightarrow T_{\mathcal{U}}$ , then we can define in a natural way a function  $\Xi - (\mathcal{D}H)^2 : \mathbf{R}^{d+1} \rightarrow T$  by the last property. Furthermore, by the second property, one has again  $\mathcal{I}(\Xi - (\mathcal{D}H)^2) : \mathbf{R}^{d+1} \rightarrow T_{\mathcal{U}}$ , which suggests that  $T$  is indeed sufficiently rich to formulate a fixed point problem mimicking the mild formulation of (1.3). Furthermore, one has

**Lemma 5.3.** *If  $\mathcal{U}$  and  $\mathcal{V}$  are the smallest subsets of  $\mathcal{F}$  satisfying (5.1) and one has  $|\Xi| > -2$  then, for every  $\gamma > 0$ , the set  $\{\tau \in \mathcal{U} \cup \mathcal{V} : |\tau| < \gamma\}$  is finite.*

The condition  $\alpha > -2$  corresponds to the restriction  $d < 2$ , which makes sense since 2 is the critical dimension for the KPZ equation [32]. The other example we would like to consider is the class of SPDEs (3.3). In this case, the right hand side is not polynomial. However, we can apply the same methodology as above as if the nonlinear functions  $f$  and  $g$  were simply polynomials of arbitrary degree. We thus impose  $\mathcal{T} \subset \mathcal{U} \cap \mathcal{V}$  and  $\{\mathcal{I}(\tau) : \tau \in \mathcal{V} \setminus \mathcal{T}\}$  as before, and then further impose that

$$\left\{ \Xi \prod_{i=1}^m \tau_i : m \geq 1 \ \& \ \tau_i \in \mathcal{U} \right\} \cup \left\{ \prod_{i=1}^m \tau_i : m \geq 1 \ \& \ \tau_i \in \mathcal{U} \right\} \subset \mathcal{V}.$$

Again, we have  $\mathcal{U} \subset \mathcal{V}$  and we define  $T$  as before. Furthermore, it is straightforward to verify that the analogue to Lemma 5.3 holds, provided that  $|\Xi| > -2$ .

**5.3. Construction of the structure group.** Now that we have some idea on how to construct  $T$  for the problems that are of interest to us (with a slightly different construction for each class of models but a clear common thread), we would like to build a corresponding structure group  $G$ . In order to give a motivation for the definition of  $G$ , it is very instructive to simultaneously think about the structure of the corresponding models. Let us first consider some smooth driving noise, which we call  $\xi_\varepsilon$  to distinguish it from the limiting noise  $\xi$ . At this stage however, this should be thought of as simply a fixed smooth function. In view of the discussion of Section 5.1, for each of the model spaces built in Section 5.2, we can associate to  $\xi_\varepsilon$  a linear map  $\mathbf{\Pi} : T \rightarrow \mathcal{C}^\infty(\mathbf{R}^{d+1})$  in the following way. We set

$$(\mathbf{\Pi}X_i)(z) = z_i, \quad (\mathbf{\Pi}\Xi)(z) = \xi_\varepsilon(z), \tag{5.2a}$$

and we then define  $\mathbf{\Pi}$  recursively by

$$\mathbf{\Pi}\mathcal{I}(\tau) = K \star \mathbf{\Pi}\tau, \quad \mathbf{\Pi}\mathcal{D}_i\tau = \partial_i\mathbf{\Pi}\tau, \quad \mathbf{\Pi}(\tau\bar{\tau}) = (\mathbf{\Pi}\tau) \cdot (\mathbf{\Pi}\bar{\tau}), \tag{5.2b}$$

where  $\cdot$  simply denotes the pointwise product between smooth functions. At this stage, it is however not clear how one would build an actual model in the sense of Definition 4.2 associated to  $\xi_\varepsilon$ . It is natural that one would set

$$(\mathbf{\Pi}_z X_i)(z') = z'_i - z_i, \quad (\mathbf{\Pi}_z \Xi)(z') = \xi_\varepsilon(z'), \tag{5.3a}$$

and then

$$\mathbf{\Pi}_z \mathcal{D}_i\tau = \partial_i\mathbf{\Pi}_z\tau, \quad \mathbf{\Pi}_z(\tau\bar{\tau}) = (\mathbf{\Pi}_z\tau) \cdot (\mathbf{\Pi}_z\bar{\tau}). \tag{5.3b}$$

It is less clear *a priori* how to define  $\mathbf{\Pi}_z\mathcal{I}(\tau)$ . The problem is that if we simply set  $\mathbf{\Pi}_z\mathcal{I}(\tau) = K \star \mathbf{\Pi}_z\tau$ , then the bound (4.5) would typically no longer be compatible with the requirement that  $|\mathcal{I}(\tau)| = |\tau| + 2$ . One way to circumvent this problem is to simply subtract the Taylor expansion of  $K \star \mathbf{\Pi}_z\tau$  around  $z$  up to the required order. We therefore set

$$(\mathbf{\Pi}_z\mathcal{I}(\tau))(z') = (K \star \mathbf{\Pi}_z\tau)(z') - \sum_{|k| < |\tau| + 2} \frac{(z' - z)^k}{k!} (D^{(k)}K \star \mathbf{\Pi}_z\tau)(z). \tag{5.3c}$$

It can easily be verified (simply proceed recursively) that if we define  $\mathbf{\Pi}_z$  in this way and  $\mathbf{\Pi}$  as in (5.2) then, for every  $z$ , one can find a linear map  $F_z : T \rightarrow T$  such that  $\mathbf{\Pi}_z = \mathbf{\Pi}F_z$ . In particular, one has  $\mathbf{\Pi}_{z'} = \mathbf{\Pi}_z F_z^{-1} F_{z'}$ . Furthermore,  $F_z$  is ‘‘upper triangular’’

with the identity on the diagonal in the sense of (4.1). It is also easily seen by induction that the matrix elements of  $F_z$  are all given by some polynomials in  $z$  and in the quantities  $(D^{(k)}K \star \Pi_z \tau)(z)$ .

This suggests that we should take for  $G$  the set of all linear maps that can appear in this fashion. It is however not clear in principle how to describe  $G$  more explicitly and it is also not clear that it even forms a group. In order to describe  $G$ , it is natural to introduce a space  $T_+$  which is given by all possible polynomials in  $d+1$  commuting variables  $\{Z_i\}_{i=0}^d$  as well as countably many additional commuting variables  $\{\mathcal{J}_k(\tau) : \tau \in (\mathcal{U} \cup \mathcal{V}) \setminus \mathcal{T} \ \& \ |k| < |\tau| + 2\}$ . One should think of  $Z_i$  as representing  $z_i$  and  $\mathcal{J}_k(\tau)$  as representing  $(D^{(k)}K \star \Pi_z \tau)(z)$ , so that the matrix elements of  $F_z$  are represented by elements of  $T_+$ . There are no relations between these coefficients, which suggests that elements of  $G$  are described by an arbitrary morphism  $f: T_+ \rightarrow \mathbf{R}$ , i.e. an arbitrary linear map which furthermore satisfies  $f(\sigma\bar{\sigma}) = f(\sigma) f(\bar{\sigma})$ , so that it is uniquely determined by  $f(Z_i)$  and  $f(\mathcal{J}_k(\tau))$ .

Given any linear map  $\Delta: T \rightarrow T \otimes T_+$  and a morphism  $f$  as above, one can then define a linear map  $\hat{\Gamma}_f: T \rightarrow T$  by

$$\hat{\Gamma}_f \tau = (I \otimes f) \Delta \tau .$$

(Here we identify  $T$  with  $T \otimes \mathbf{R}$  in the obvious way.) The discussion given above then suggests that it is possible to construct  $\Delta$  in such a way that if we define  $f_z$  by

$$f_z(Z_i) = z_i , \quad f_z(\mathcal{J}_k(\tau)) = (D^{(k)}K \star \Pi_z \tau)(z) , \tag{5.4}$$

then one has  $\hat{\Gamma}_{f_z} = F_z$ . The precise definition of  $\Delta$  is irrelevant for our discussion, but a recursive description of it can easily be recovered simply by comparing (5.3) to (5.2). In particular, it is possible to show that  $\Delta \tau$  is of the form

$$\Delta \tau = \tau \otimes \mathbf{1} + \sum_i c_i^\tau \tau_i \otimes \sigma_i , \tag{5.5}$$

for some expressions  $\tau_i \in T$  with  $|\tau_i| < |\tau|$  and for some non-empty monomials  $\sigma_i \in T_+$  such that  $|\sigma_i| + |\tau_i| = |\tau|$ . Here, we associate a homogeneity to elements in  $T_+$  by setting  $|Z_0| = 2, |Z_i| = 1$  for  $i \neq 0$ , and  $|\mathcal{J}_k(\tau)| = |\tau| + 2 - |k|$ .

In particular, we see that if we let  $e: T_+ \rightarrow \mathbf{R}$  be the trivial morphism for which  $e(Z_i) = e(\mathcal{J}_k(\tau)) = 0$ , so that one only has  $e(\mathbf{1}) = 1$  where  $\mathbf{1}$  is the empty product, then  $\hat{\Gamma}_e \tau = \tau$ . The important fact for our purpose is the following, a proof of which can be found in [23, Sec. 8]. Here, we denote by  $\mathcal{M}: T_+ \otimes T_+ \rightarrow T_+$  the multiplication operator  $\mathcal{M}(\sigma \otimes \bar{\sigma}) = \sigma \bar{\sigma}$  and by  $I$  the identity.

**Theorem 5.4.** *There exists a map  $\Delta^+: T_+ \rightarrow T_+ \otimes T_+$  such that the following identities hold:*

$$\begin{aligned} \Delta^+(\sigma\bar{\sigma}) &= (\Delta^+ \sigma) \cdot (\Delta^+ \bar{\sigma}) , & (\Delta \otimes I) \Delta &= (I \otimes \Delta^+) \Delta , \\ (e \otimes I) \Delta^+ &= (I \otimes e) \Delta^+ = I , & (\Delta^+ \otimes I) \Delta^+ &= (I \otimes \Delta^+) \Delta^+ . \end{aligned} \tag{5.6}$$

Furthermore, there exists a map  $\mathcal{A}: T_+ \rightarrow T_+$  which is multiplicative in the sense that  $\mathcal{A}(\sigma\bar{\sigma}) = (\mathcal{A}\sigma) \cdot (\mathcal{A}\bar{\sigma})$ , and which is such that  $\mathcal{M}(I \otimes \mathcal{A}) \Delta^+ = \mathcal{M}(\mathcal{A} \otimes I) \Delta^+ = e$ , with  $e: T_+ \rightarrow \mathbf{R}$  as above.

**Remark 5.5.** In technical lingo, this lemma states that  $(T_+, \cdot, \Delta^+)$  is a Hopf algebra with antipode  $\mathcal{A}$ , and that  $T$  is a comodule over  $T_+$ .

The importance of this result is that it shows that  $G$  is indeed a group. For any two morphisms  $f$  and  $g$ , we can define a linear map  $f \circ g: T_+ \rightarrow \mathbf{R}$  by  $(f \circ g)(\sigma) = (f \otimes g)\Delta^+ \sigma$ . As a consequence of the first identity in (5.6),  $f \circ g$  is again a morphism on  $T_+$ . As a consequence of the second identity, one has  $\hat{\Gamma}_{f \circ g} = \hat{\Gamma}_f \hat{\Gamma}_g$ . The last identity shows that  $(f_1 \circ f_2) \circ f_3 = f_1 \circ (f_2 \circ f_3)$ , while the properties of  $\mathcal{A}$  ensure that if we set  $f^{-1}(\sigma) = f(\mathcal{A}\sigma)$ , then  $f \circ f^{-1} = f^{-1} \circ f = e$ . Finally, the third identity in (5.6) shows that  $e$  is indeed the identity element, thus turning the set of all morphisms of  $T_+$  into a group under  $\circ$ , acting on  $T$  via  $\hat{\Gamma}$ .

Let us now turn back to our models. Given a smooth function  $\xi_\varepsilon$ , we define  $\Pi_z$  as in (5.3) and  $f_z$  by (5.4). We then also define linear maps  $\Gamma_{zz'}$  by  $\Gamma_{zz'} = \hat{\Gamma}_{\gamma_{zz'}}$  with  $\gamma_{zz'} = f_z^{-1} \circ f_{z'}$ . We then have

**Lemma 5.6.** *For every smooth function  $\xi_\varepsilon$ , the pair  $(\Pi, \Gamma)$  defined above is a model.*

*Proof.* The algebraic constraints (4.4) are satisfied essentially by definition. The first bound of (4.5) can easily be verified recursively by (5.3). The only non-trivial fact is that the matrix elements of  $\Gamma_{zz'}$  satisfy the right bound. If one can show that  $|\gamma_{zz'}(\sigma)| \lesssim |z - z'|^{|\sigma|}$ , this in turn follows from (5.5). This bound is non-trivial and was obtained in [23, Prop. 8.27].  $\square$

**5.4. Admissible models.** Thanks to Lemma 5.6, we now have a large class of models for the regularity structures built in the previous two subsections. However, we do not want to restrict ourselves to this class (or even its closure). The reason is that if we define products in the “naïve” way given by the second identity in (5.3b), then there will typically be some situations where the result diverges as we let  $\varepsilon \rightarrow 0$  in  $\xi_\varepsilon$ . Therefore, we do not impose this relation in general but rather view it as the *definition* of the product, i.e. we interpret it as

$$(\Pi_z \tau) \cdot (\Pi_z \bar{\tau}) := \Pi_z(\tau \bar{\tau}) .$$

However, the remainder of the structure described in (5.3) is required for  $X_i, \mathcal{D}_i$  and  $\mathcal{I}$  to have the correct interpretation. This motivates the following definition.

**Definition 5.7.** Given a regularity structure  $\mathcal{T}$  constructed as in Sections 5.2 and 5.3, we say that a model  $(\Pi, \Gamma)$  is *admissible* if it satisfies  $(\Pi_z X_i)(z') = z'_i - z_i, \Pi_z \mathcal{D}_i \tau = \partial_i \Pi_z \tau$ , as well as (5.3c) and if furthermore  $\Gamma_{zz'} = \hat{\Gamma}_{f_z}^{-1} \hat{\Gamma}_{f_{z'}}$  with  $f_z$  given by (5.4). We will denote the space of all admissible models by  $\mathcal{M}_0 \subset \mathcal{M}$ .

**Remark 5.8.** In the particular case of admissible models for a regularity structure of the type considered here, the data of the single linear map  $\Pi$  as above is sufficient to reconstruct the full model  $(\Pi, \Gamma)$ .

Note that at this stage, it is not clear whether this concept is even well-defined: in general,  $D^{(k)}K \star \Pi_z \tau$  will be a distribution and cannot be evaluated at fixed points, so (5.4) might be meaningless for a general model. It turns out that the definition actually always makes sense, provided that the second identity in (5.4) is interpreted as

$$f_z(\mathcal{J}_k(\tau)) = \sum_{n \geq 0} (D^{(k)}K_n \star \Pi_z \tau)(z) ,$$

where  $K = \sum_{n \geq 0} K_n$  as in (2.1). This is because the bound (2.1), combined with the bound (4.5) and the fact that  $K_n$  is supported in the ball of radius  $2^{-n}$  imply that

$$|(D^{(k)}K_n \star \Pi_z \tau)(z)| \lesssim 2^{(|k| - |\tau| - 2)n} .$$

The condition  $|k| < |\tau| + 2$  appearing in (5.3c) is then precisely what is required to guarantee that this is always summable.

**5.5. Abstract fixed point problem.** We now show how to reformulate a stochastic PDE as a fixed point problem in some space  $\mathcal{D}^\gamma$  based on an admissible model for the regularity structure associated to the SPDE by the construction of Section 5.2. For definiteness, we focus on the example of the KPZ equation (1.3), but all other examples mentioned in the introduction can be treated in virtually the same way. Writing  $P$  for the heat kernel, the mild formulation of (1.3) is given by

$$h = P \star \mathbf{1}_{t>0}((\partial_x h)^2 + \xi) + Ph_0, \quad (5.7)$$

where we write  $Ph_0$  for the harmonic extension of  $h_0$ . (This is just the solution to the heat equation with initial condition  $h_0$ .) In order to formulate this as a fixed point problem in  $\mathcal{D}^\gamma$  for a suitable value of  $\gamma > 0$ , we will make use of the following far-reaching extension of Schauder's theorem.

**Theorem 5.9.** *Fix one of the regularity structures built in the previous section and fix an admissible model. Then, for all but a discrete set of values of  $\gamma > 0$ , there exists a continuous operator  $\mathcal{P}: \mathcal{D}^\gamma \rightarrow \mathcal{D}^{\gamma+2}$  such that the identity*

$$\mathcal{R}\mathcal{P}f = P \star \mathcal{R}f, \quad (5.8)$$

holds for every  $f \in \mathcal{D}^\gamma$ . Furthermore, one has  $(\mathcal{P}f)(z) - \mathcal{I}f(z) \in \bar{T}$ .

**Remark 5.10.** Recall that  $\bar{T} \subset T$  denotes the linear span of the  $X^k$ , which represent the usual Taylor polynomials. Again, while  $\mathcal{P}$  is a linear map when we consider the underlying model as fixed, it can (and should) also be viewed as a continuous nonlinear map from  $\mathcal{M}_0 \times \mathcal{D}^\gamma$  into  $\mathcal{M}_0 \times \mathcal{D}^{\gamma+2}$ . The reason why some values of  $\gamma$  need to be excluded is essentially the same as for the usual Schauder theorem.

For a proof of Theorem 5.9 and a precise description of the operator  $\mathcal{P}$ , see [23, Sec. 5]. With the help of the operator  $\mathcal{P}$ , it is then possible to reformulate (5.7) as the following fixed point problem in  $\mathcal{D}^\gamma$ , provided that we have an admissible model at our disposal:

$$H = \mathcal{P}\mathbf{1}_{t>0}((\mathcal{D}H)^2 + \Xi) + Ph_0. \quad (5.9)$$

Here, the smooth function  $Ph_0$  is interpreted as an element in  $\mathcal{D}^\gamma$  with values in  $\bar{T}$  via its Taylor expansion of order  $\gamma$ . Note that in the context of the regularity structure associated to the KPZ equation in Section 5.2, the right hand side of this equation makes sense for every  $H \in \mathcal{D}^\gamma$ , provided that  $H$  takes values in  $T_{\mathcal{U}}$ . This is an immediate consequence of the property (5.1).

**Remark 5.11.** As already mentioned earlier, we cheat here in the sense that  $\mathcal{D}^\gamma$  should really be replaced by a space  $\mathcal{D}^{\gamma;\eta}$  allowing for a suitable singular behaviour on the hyperplane  $t = 0$ .

It is also possible to show (see [23, Thm 4.7]) that if we set  $|\Xi| = -\frac{3}{2} - \kappa$  for some sufficiently small  $\kappa > 0$ , then one has  $(\mathcal{D}H)^2 \in \mathcal{D}^{\gamma-\frac{3}{2}-\kappa}$  for  $H \in \mathcal{D}^\gamma$ . As a consequence, we expect to be able to find local solutions to the fixed point problem (5.9), provided that we

formulate it in  $\mathcal{D}^\gamma$  for  $\gamma > \frac{3}{2} + \kappa$ . This is indeed the case, and a more general instance of this fact can be found in [23, Thm 7.8]. Furthermore, the local solution is locally Lipschitz continuous as a function of both the initial condition  $h_0$  and the underlying admissible model  $(\Pi, \Gamma) \in \mathcal{M}_0$ .

Now that we have a local solution  $H \in \mathcal{D}^\gamma$  for (5.9), we would like to know how this solution relates to the original problem (1.3). This is given by the following simple fact:

**Proposition 5.12.** *If the underlying model  $(\Pi, \Gamma)$  is built from a smooth function  $\xi_\varepsilon$  as in (5.3) and if  $H$  solves (5.9), then  $\mathcal{R}H$  solves (5.7).*

*Proof.* As a consequence of (5.8), we see that  $\mathcal{R}H$  solves

$$\mathcal{R}H = P \star \mathbf{1}_{t>0}(\mathcal{R}((\mathcal{D}H)^2) + \xi_\varepsilon) + Ph_0 .$$

Combining (5.3b) with (4.10), it is not difficult to see that in this particular case, one has  $\mathcal{R}((\mathcal{D}H)^2) = (\partial_x \mathcal{R}H)^2$ , so that the claim follows.  $\square$

The results of the previous subsection yield a robust solution theory for (5.9) which projects down (via  $\mathcal{R}$ ) to the usual solution theory for (1.3) for smooth driving noise  $\xi_\varepsilon$ . If it were the case that the sequence of models  $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$  associated to the regularised noise  $\xi_\varepsilon$  via (5.3) converges to a limit in  $\mathcal{M}_0$ , then this would essentially conclude our analysis of (1.3).

Unfortunately, this is *not* the case. Indeed, in all of the examples mentioned in the introduction except for (1.2), the sequence of models  $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$  does not converge as  $\varepsilon \rightarrow 0$ . In order to remedy to this situation, the idea is to look for a sequence of “renormalised” models  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  which are also admissible and also satisfy  $\hat{\Pi}_z^{(\varepsilon)} \Xi = \xi_\varepsilon$ , but do converge to a limit as  $\varepsilon \rightarrow 0$ . The last section of this article shows how these renormalised models can be constructed.

**5.6. Renormalisation.** In order to renormalise our model, we will build a very natural group of continuous transformations of  $\mathcal{M}_0$  that build a new admissible model from an old one. The renormalised model will then be the image of the “canonical” model  $(\Pi^{(\varepsilon)}, \Gamma^{(\varepsilon)})$  under a (diverging) sequence of such transformations. Since we want the new model to also be admissible, the only defining property that we are allowed to modify in (5.3) is the definition of the product. In order to describe the renormalised model, it turns out to be more convenient to consider again its representation by a single linear map  $\hat{\Pi}^{(\varepsilon)} : T \rightarrow \mathcal{S}'$  as in (5.3), which is something we can do by Remark 5.8.

At this stage, we do not appear to have much choice: the only “reasonable” way of building  $\hat{\Pi}^{(\varepsilon)}$  from  $\Pi^{(\varepsilon)}$  is to compose it to the right with some fixed linear map  $M_\varepsilon : T \rightarrow T$ :

$$\hat{\Pi}^{(\varepsilon)} = \Pi^{(\varepsilon)} M_\varepsilon . \tag{5.10}$$

If we do this for an arbitrary map  $M_\varepsilon$ , we will of course immediately lose the algebraic and analytical properties that allow to associate an admissible model  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  to the map  $\hat{\Pi}^{(\varepsilon)}$ . As a matter of fact, it is completely unclear *a priori* whether there exists *any* non-trivial map  $M_\varepsilon$  that preserves these properties. Fortunately, these maps do exist and a somewhat indirect characterisation of them can be found in [23, Sec. 8]. Even better, there are sufficiently many of them so that the divergencies of  $\Pi^{(\varepsilon)}$  can be compensated by a judicious choice of  $M_\varepsilon$ .

Let us just illustrate how this plays out in the case of the KPZ equation already studied in the last subsection. In order to simplify notations, we now use the following shorthand graphical notation for elements of  $\mathcal{U} \cup \mathcal{V}$ . For  $\Xi$ , we draw a small circle. The integration map  $\mathcal{I}$  is then represented by a downfacing wavy line and  $\mathcal{DI}$  is represented by a downfacing plain line. The multiplication of symbols is obtained by joining them at the root. For example, we have

$$(\mathcal{DI}(\Xi))^2 = \mathcal{V}, \quad (\mathcal{DI}(\mathcal{DI}(\Xi)^2))^2 = \mathcal{V}\mathcal{V}, \quad \mathcal{I}(\mathcal{DI}(\Xi)^2) = \mathcal{Y}.$$

In the case of the KPZ equation, it turns out that one can exhibit an explicit four-parameter group of matrices  $M$  which preserve admissible models when used in (5.10). These matrices are of the form  $M = \exp(-\sum_{i=0}^3 C_i L_i)$ , where the generators  $L_i$  are determined by the following contraction rules:

$$L_0: \curvearrowright \mapsto \mathbf{1}, \quad L_1: \mathcal{V} \mapsto \mathbf{1}, \quad L_2: \mathcal{V}\mathcal{V} \mapsto \mathbf{1}, \quad L_3: \mathcal{V}\mathcal{Z} \mapsto \mathbf{1}. \tag{5.11}$$

This should be understood in the sense that if  $\tau$  is an arbitrary formal expression, then  $L_0\tau$  is the sum of all formal expressions obtained from  $\tau$  by performing a substitution of the type  $\curvearrowright \mapsto \mathbf{1}$ . For example, one has  $L_0\mathcal{V} = 2\mathbf{1}$ ,  $L_0\mathcal{V}\mathcal{V} = 2\mathcal{V} + \mathcal{Y}$ , etc. The extension of the other operators  $L_i$  to all of  $T$  is given by  $L_i\tau = 0$  for  $i \neq 0$  and every  $\tau$  for which  $L_i$  wasn't already defined in (5.11). We then have the following result, which is a consequence of [23, Sec. 8] and [28] and was implicit in [21]:

**Theorem 5.13.** *Let  $M_\varepsilon$  be given as above, let  $\Pi^{(\varepsilon)}$  be constructed from  $\xi_\varepsilon$  as in (5.2), and let  $\hat{\Pi}^{(\varepsilon)} = \Pi^{(\varepsilon)}M_\varepsilon$ . Then, there exists a unique admissible model  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  such that  $\hat{\Pi}_z^{(\varepsilon)} = \hat{\Pi}^{(\varepsilon)}\hat{F}_z^{(\varepsilon)}$ , where  $\hat{F}_z^{(\varepsilon)}$  relates to  $\hat{\Pi}_z^{(\varepsilon)}$  as in (5.4). Furthermore, one has the identity*

$$(\hat{\Pi}_z^{(\varepsilon)}\tau)(z) = (\Pi_z^{(\varepsilon)}M_\varepsilon\tau)(z). \tag{5.12}$$

Finally, there is a choice of  $M_\varepsilon$  such that  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  converges to a limit  $(\hat{\Pi}, \hat{\Gamma})$  which is universal in that it does not depend on the details of the regularisation procedure.

**Remark 5.14.** Despite (5.12), it is not true in general that  $\hat{\Pi}_z^{(\varepsilon)} = \Pi_z^{(\varepsilon)}M_\varepsilon$ . The point is that (5.12) only holds at the point  $z$  and not at  $z' \neq z$ .

In order to complete our survey of Theorem 1.1, it remains to identify the solution to (5.9) with respect to the renormalised model  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  with the classical solution to some modified partial differential equation. The continuity of the abstract solution map then immediately implies that the solutions to the modified PDE converge to a limit. The fact that the limiting model  $(\hat{\Pi}, \hat{\Gamma})$  is universal also implies that this limit is universal.

**Theorem 5.15.** *Let  $M_\varepsilon = \exp(-\sum_{i=0}^3 C_i^{(\varepsilon)}L_i)$  be as above and let  $(\hat{\Pi}^{(\varepsilon)}, \hat{\Gamma}^{(\varepsilon)})$  be the corresponding renormalised model. Let furthermore  $H$  be the solution to (5.9) with respect to this model. Then, the function  $h(t, x) = (\mathcal{R}H)(t, x)$  solves the equation*

$$\partial_t h = \partial_x^2 h + (\partial_x h)^2 - 4C_0^{(\varepsilon)}\partial_x h + \xi_\varepsilon - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)}). \tag{5.13}$$

**Remark 5.16.** In order to obtain a limit  $(\hat{\Pi}, \hat{\Gamma})$ , the renormalisation constants  $C_i^{(\varepsilon)}$  should be chosen in the following way:

$$C_0^{(\varepsilon)} = 0, \quad C_1^{(\varepsilon)} = \frac{c_1}{\varepsilon}, \quad C_2^{(\varepsilon)} = 4c_2 \log \varepsilon + c_3, \quad C_3^{(\varepsilon)} = -c_2 \log \varepsilon + c_4.$$

Here, the  $c_i$  are constants of order 1 that depend on the details of the regularisation procedure for  $\xi_\varepsilon$ . The fact that  $C_0^{(\varepsilon)} = 0$  explains why the corresponding term does not appear in (1.3). The fact that the diverging parts of  $C_2^{(\varepsilon)}$  and  $C_3^{(\varepsilon)}$  cancel in (5.13) explains why this logarithmic sub-divergence was not observed in [4] for example.

*Proof.* We first note that, as a consequence of Theorem 5.9 and of (5.9), one can write for  $t > 0$

$$H = \mathcal{I}((\mathcal{D}H)^2 + \Xi) + (\dots), \tag{5.14}$$

where  $(\dots)$  denotes some terms belonging to  $\bar{T} \subset T$ .

By repeatedly using this identity, we conclude that any solution  $H \in \mathcal{D}^\gamma$  to (5.9) for  $\gamma$  greater than (but close enough to)  $3/2$  is necessarily of the form

$$H = h \mathbf{1} + \mathfrak{i} + \mathfrak{Y} + h' X_1 + 2\mathfrak{V} + 2h' \mathfrak{Z}, \tag{5.15}$$

for some real-valued functions  $h$  and  $h'$ . Note that  $h'$  is treated as an independent function here, we certainly do not mean to suggest that the function  $h$  is differentiable! Our notation is only by analogy with the classical Taylor expansion. As an immediate consequence,  $\mathcal{D}H$  is given by

$$\mathcal{D}H = \mathfrak{i} + \mathfrak{Y} + h' \mathbf{1} + 2\mathfrak{V} + 2h' \mathfrak{Z}, \tag{5.16}$$

as an element of  $\mathcal{D}^\gamma$  for  $\gamma$  close to  $1/2$ . The right hand side of the equation is then given up to order 0 by

$$(\mathcal{D}H)^2 + \Xi = \Xi + \mathfrak{V} + 2\mathfrak{V} + 2h' \mathfrak{i} + \mathfrak{V} + 4\mathfrak{V} + 2h' \mathfrak{Y} + 4h' \mathfrak{Z} + (h')^2 \mathbf{1}. \tag{5.17}$$

Using the definition of  $M_\varepsilon$ , we conclude that

$$M_\varepsilon \mathcal{D}H = \mathcal{D}H - 4C_0^{(\varepsilon)} \mathfrak{Z},$$

so that, as an element of  $\mathcal{D}^\gamma$  with very small (but positive)  $\gamma$ , one has the identity

$$(M_\varepsilon \mathcal{D}H)^2 = (\mathcal{D}H)^2 - 8C_0^{(\varepsilon)} \mathfrak{Z}.$$

As a consequence, after neglecting again all terms of strictly positive homogeneity, one has the identity

$$M_\varepsilon((\mathcal{D}H)^2 + \Xi) = (M_\varepsilon \mathcal{D}H)^2 + \Xi - 4C_0^{(\varepsilon)} M_\varepsilon \mathcal{D}H - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)}).$$

Combining this with (5.12) and (4.10), we conclude that

$$\mathcal{R}((\mathcal{D}H)^2 + \Xi) = (\partial_x \mathcal{R}H)^2 + \xi_\varepsilon - 4C_0^{(\varepsilon)} \partial_x \mathcal{R}H - (C_1^{(\varepsilon)} + C_2^{(\varepsilon)} + 4C_3^{(\varepsilon)}),$$

from which the claim then follows in the same way as for Proposition 5.12. □

**Remark 5.17.** Ultimately, the reason why the theory mentioned in Section 1.1 (or indeed the theory of controlled rough paths, as originally exploited in [21]) can also be applied in this case is that in (5.15), only *one* basis vector besides those in  $\mathcal{T}$  (i.e. besides  $\mathbf{1}$  and  $X_1$ ) comes with a non-constant coefficient, namely the basis vector  $\mathfrak{Z}$ . The methodology explained in Section 3.1 on the other hand can be applied whenever no basis vector besides those in  $\mathcal{T}$  comes with a non-constant coefficient.

**Acknowledgements.** I am delighted to thank the Institute for Advanced Study for its warm hospitality and the ‘The Fund for Math’ for funding my stay there. This work was supported by the Leverhulme trust through a leadership award, the Royal Society through a Wolfson research award, and the ERC through a consolidator award.

## References

- [1] S. Albeverio and M. Röckner, *Stochastic differential equations in infinite dimensions: solutions via Dirichlet forms*, Probab. Theory Related Fields **89**(3) (1991), 347–386.
- [2] H. Bahouri, J.-Y. Chemin, and R. Danchin, *Fourier analysis and nonlinear partial differential equations*, volume 343 of Grundlehren der Mathematischen Wissenschaften, Springer, Heidelberg, 2011.
- [3] Á. Bényi, D. Maldonado, and V. Naibo, *What is ... a paraproduct?*, Notices Amer. Math. Soc. **57**(7) (2010), 858–860.
- [4] L. Bertini and G. Giacomin, *Stochastic Burgers and KPZ equations from particle systems*, Comm. Math. Phys. **183**(3) (1997), 571–607.
- [5] L. Bertini, E. Presutti, B. Rüdiger, and E. Saada, *Dynamical fluctuations at the critical point: convergence to a nonlinear stochastic PDE*, Teor. Veroyatnost. i Primenen. **38**(4) (1993), 689–741.
- [6] J.-M. Bismut, *Martingales, the Malliavin calculus and hypoellipticity under general Hörmander’s conditions*, Z. Wahrsch. Verw. Gebiete **56**(4) (1981), 469–505.
- [7] J.-M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup. (4) **14**(2) (1981), 209–246.
- [8] R. A. Carmona and S. A. Molchanov *Parabolic Anderson problem and intermittency*, Mem. Amer. Math. Soc. **108**(518) (1994), viii+125.
- [9] R. Catellier and K. Chouk, *Paracontrolled distributions and the 3-dimensional stochastic quantization equation*, ArXiv e-prints, Oct. 2013.
- [10] G. Da Prato and A. Debussche, *Two-dimensional Navier-Stokes equations driven by a space-time white noise*, J. Funct. Anal. **196**(1) (2002), 180–210.
- [11] ———, *Strong solutions to the stochastic quantization equations*, Ann. Probab. **31**(4) (2003), 1900–1916.
- [12] ———, *A modified Kardar-Parisi-Zhang model*, Electron. Comm. Probab. **12** (2007), 442–453 (electronic).
- [13] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, volume 44 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1992.
- [14] R. L. Dobrushin, *Gaussian and their subordinated self-similar random generalized fields*, Ann. Probab. **7**(1) (1979), 1–28.
- [15] P. K. Friz and M. Hairer, *A course on rough paths*, Universitext, Springer, 2014. To appear.
- [16] P. K. Friz and N. B. Victoir, *Multidimensional stochastic processes as rough paths*, volume 120 of Cambridge Studies in Advanced Mathematics, Cambridge University

- Press, Cambridge, 2010. Theory and applications.
- [17] P. Goncalves and M. Jara, *Nonlinear fluctuations of weakly asymmetric interacting particle systems*, Arch. Ration. Mech. Anal. **212**(2) (2014), 597–644.
  - [18] M. Gubinelli, *Controlling rough paths*, J. Funct. Anal. **216**(1) (2004), 86–140.
  - [19] M. Gubinelli, P. Imkeller, and N. Perkowski, *Paraproducts, rough paths and controlled distributions*, ArXiv e-prints, Oct. 2012.
  - [20] M. Hairer, *Rough stochastic PDEs*, Comm. Pure Appl. Math. **64**(11) (2011), 1547–1585.
  - [21] ———, *Solving the KPZ equation*, Ann. of Math. (2) **178**(2) (2013), 559–664.
  - [22] ———, *Introduction to regularity structures*, ArXiv e-prints, Jan. 2014. Braz. J. Prob. Stat., to appear.
  - [23] ———, *A theory of regularity structures*, Invent. Math. **198**(2) (2014), 269–504.
  - [24] M. Hairer and J. Maas, *A spatial version of the Itô-Stratonovich correction*, Ann. Probab. **40**(4) (2012), 1675–1714.
  - [25] M. Hairer, J. Maas, and H. Weber, *Approximating rough stochastic PDEs*, Comm. Pure Appl. Math. **67**(5) (2014), 776–870.
  - [26] M. Hairer, É. Pardoux, and A. Piatnitsky, *A Wong-Zakai theorem for stochastic PDEs*, Work in progress, 2014.
  - [27] M. Hairer and N. S. Pillai, *Regularity of laws and ergodicity of hypoelliptic SDEs driven by rough paths*, Ann. Probab. **41**(4) (2013), 2544–2598.
  - [28] M. Hairer and J. Quastel, *A class of growth models rescaling to KPZ*, Work in progress, 2014.
  - [29] M. Hairer and H. Weber, *Rough Burgers-like equations with multiplicative noise*, Probab. Theory Related Fields **155**(1-2) (2013), 71–126.
  - [30] K. Itô, *Stochastic integral*, Proc. Imp. Acad. Tokyo **20** (1944), 519–524.
  - [31] G. Jona-Lasinio and P. K. Mitter, *On the stochastic quantization of field theory*, Comm. Math. Phys. **101**(3) (1985), 409–436.
  - [32] M. Kardar, G. Parisi, and Y.-C. Zhang, *Dynamic scaling of growing interfaces*, Phys. Rev. Lett., **56**(9) (Mar. 1986), 889–892.
  - [33] S. Kusuoka and D. Stroock, *Applications of the Malliavin calculus. I*, In Stochastic analysis (Katata/Kyoto, 1982), volume 32 of North-Holland Math. Library, pp. 271–306. North-Holland, Amsterdam, 1984.
  - [34] T. J. Lyons, *Differential equations driven by rough signals*, Rev. Mat. Iberoamericana **14**(2) (1998), 215–310.
  - [35] T. J. Lyons, M. Caruana, and T. Lévy, *Differential equations driven by rough paths*, volume 1908 of Lecture Notes in Mathematics, Springer, Berlin, 2007. Lectures from the 34th Summer School on Probability Theory held in Saint-Flour.
  - [36] T. J. Lyons and Z. Qian, *System control and rough paths*, Oxford Mathematical Monographs. Oxford University Press, Oxford, 2002. Oxford Science Publications.
  - [37] P. Malliavin, *Stochastic analysis*, volume 313 of Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin, 1997.
  - [38] J. Norris, *Simplified Malliavin calculus*, In Séminaire de Probabilités, XX, 1984/85,

- volume 1204 of Lecture Notes in Math., pp. 101–130. Springer, Berlin, 1986.
- [39] D. Nualart, *The Malliavin calculus and related topics*, Probability and its Applications (New York). Springer-Verlag, Berlin, second edition, 2006.
- [40] G. Parisi and Y. S. Wu, *Perturbation theory without gauge fixing*, *Sci. Sinica* **24**(4) (1981), 483–496.
- [41] J. Schauder, *Über lineare elliptische Differentialgleichungen zweiter Ordnung*, *Math. Z.* **38**(1) (1934), 257–282.
- [42] L. Simon, *Schauder estimates by scaling*, *Calc. Var. Partial Differential Equations* **5**(5) (1997), 391–407.
- [43] R. L. Stratonovič, *A new form of representing stochastic integrals and equations*, *Vestnik Moskov. Univ. Ser. I Mat. Meh.* **1964**(1) (1964), 3–12.
- [44] E. Wong and M. Zakai, *On the relation between ordinary and stochastic differential equations*, *Internat. J. Engrg. Sci.* **3** (1965), 213–229.

Mathematics Institute, The University of Warwick, U.K.

E-mail: M.Hairer@Warwick.ac.uk



# Hardness of Approximation

Subhash Khot

**Abstract.** This article accompanies the talk given by the author at the International Congress of Mathematicians, 2014. The article sketches some connections between approximability of NP-complete problems, analysis and geometry, and the role played by the Unique Games Conjecture in facilitating these connections. For a more extensive introduction to the topic, the reader is referred to survey articles [39, 40, 64].

**Mathematics Subject Classification (2010).** Primary 68Q17.

**Keywords.** NP-completeness, Approximation algorithms, Inapproximability, Probabilistically Checkable Proofs, Discrete Fourier analysis.

## 1. Introduction

The  $P \neq NP$  hypothesis says that a large class of computational problems known as NP-complete problems do not have efficient algorithms. An algorithm is called efficient if it runs in time polynomial in the size of the input, typically denoted as  $n$ . A natural question is whether one can efficiently compute *approximate* solutions to NP-complete problems and how good an approximation one can achieve. We are interested in both upper and lower bounds: designing algorithms with a guarantee on the quality of approximation (upper bounds) as well as results showing that no efficient algorithm exists that achieves an approximation guarantee beyond a certain threshold (lower bounds). It is the latter question, namely the lower bounds, that is the focus of this article. Such results are known as *inapproximability* or *hardness of approximation* results, proved under a standard complexity theoretic hypothesis such as  $P \neq NP$ .

Let us consider two problems, the Traveling Salesperson (TSP) and the Clique, as illustration. In the (2-dimensional Euclidean version of) TSP problem, we are given a set of  $n$  cities in a plane and the pairwise distances between them and the goal is to find a tour that visits all the cities and has minimum length. In the Clique problem, we are given an  $n$ -vertex graph and the goal is find a clique of maximum size where a clique is a subset of vertices such that all its vertices are pairwise connected by edges. Both the problems are NP-complete<sup>1</sup> and hence one does not hope to efficiently find optimal solutions. Now consider the question of how well one can approximate them. For the TSP problem, for every constant  $\varepsilon > 0$ , Arora and Mitchell [1, 52] designed a polynomial time algorithm that computes a tour with length at most  $1 + \varepsilon$  times the length of the minimum tour. For the Clique problem, Håstad [35] showed that it cannot be approximated at all. Specifically, for every constant  $\varepsilon > 0$ , assuming  $P \neq NP$ , no polynomial time algorithm, given an  $n$ -vertex graph that has a clique of size

---

<sup>1</sup>Proceedings of the International Congress of Mathematicians, Seoul, 2014

at least  $n^{1-\varepsilon}$ , can find a clique of size even  $n^\varepsilon$ . Thus, we know the precise extent to which the TSP and the Clique problems are approximable: the former is approximable as well as one might hope for and the latter is not approximable at all. There are a few more problems for which also we know the precise extent of approximability. In particular, for the 3SAT and the Set Cover problems [30, 36], we know an approximation algorithm that achieves a reasonable (but not too close as TSP) approximation guarantee and we also know that achieving an approximation better than this *threshold* guarantee is an NP-complete problem itself. To emphasize, the last statement implies that an algorithm with approximation guarantee better than the threshold for these problems can then be used to find optimal solutions!

However, for a vast majority of the NP-complete problems of interest, there is (often a big) gap between the quality of the best known approximation algorithms and the known hardness of approximation results. Filling up these gaps, as well as understanding why different NP-complete problems seem to behave differently in terms of their approximability, is largely open. The Unique Games Conjecture was proposed towards making progress on this topic, and in particular towards showing optimal hardness of approximation results, i.e. results that match the quality of the best known approximation algorithms. As it turns out, showing hardness results is closely related to Fourier analysis of boolean functions on a boolean hypercube and to certain problems in geometry, especially related to isoperimetry. This article gives a sketch of some of these connections and cites a couple of open questions towards settling the Unique Games Conjecture. We anticipate that the intended audience of this article is not necessarily familiar with the language and techniques in computer science, so an attempt is made to keep the presentation as self-contained as possible.

## 2. The Unique Games Conjecture

The Unique Games Conjecture [38] states that a certain computational problem called the Unique Game is very hard to approximate. We do state the conjecture here, but we will not really use the statement in the rest of the article.

An instance  $\mathcal{L}$  of the Unique Game problem is a system of linear equations over  $\mathbb{Z}_p$  of a specific form. There are  $n$  variables  $x_1, \dots, x_n$  and  $m$  equations, where  $i^{\text{th}}$  equation is of the form  $x_{i_1} - x_{i_2} = c_i$ . The constants  $c_i \in \mathbb{Z}_p$  may depend on the equation. The goal is to find an assignment to the variables that satisfies a *good* fraction of the equations. Let  $\text{OPT}(\mathcal{L})$  denote the maximum fraction of equations satisfied by any assignment. The Unique Games Conjecture states:

**Conjecture 2.1.** *For every constant  $\delta > 0$ , there is a large enough constant  $p = p(\delta)$ , such that there is no polynomial time algorithm that given an instance of Unique Game over  $\mathbb{Z}_p$  that has an assignment satisfying  $1 - \delta$  fraction of the equations, finds an assignment that satisfies (even)  $\delta$  fraction of the equations.<sup>2</sup>*

<sup>1</sup>There are some subtleties regarding the computational complexity of TSP that we omit here.

<sup>2</sup>The original conjecture is stated in terms of a more general problem and strictly speaking the term Unique Game refers to the general problem. The problem presented here is referred to as Linear Unique Game. It is shown in [41] that the original conjecture is equivalent to the statement here. Also, the problem is conjectured to be NP-complete, rather than just that there is no polynomial time algorithm for it. It is widely believed that NP-complete problems do not have algorithms that run in time  $2^{n^{o(1)}}$ , rather than just in polynomial time. For much of the article, when we say “there is no polynomial time algorithm for a problem”, we really mean “the problem is

A few comments are in order. The term *game* refers to the context of *2-prover-1-round games* where the problem was studied initially. Given an instance of the Unique Game as above, consider the following game between two provers and a verifier: the verifier picks an equation  $x_{i_1} - x_{i_2} = c_i$  at random, sends the variable  $x_{i_1}$  to prover  $P_1$  and the variable  $x_{i_2}$  to prover  $P_2$ . Each prover is supposed to answer with a value in  $\mathbb{Z}_p$ , and the verifier accepts if and only if  $a_1 - a_2 = c_i$  where  $a_1$  and  $a_2$  are the answers of the two provers respectively. The strategies of the provers correspond to assignments  $\sigma_1, \sigma_2 : \{x_1, \dots, x_n\} \mapsto \mathbb{Z}_p$ . The *value* of the game is the maximum over all prover strategies, the probability that the verifier accepts. It is not difficult to show that this value is between  $\text{OPT}(\mathcal{L})$  and  $\max\{1, 4\text{OPT}(\mathcal{L})\}$ . Such games were initially motivated by the study of cryptographic protocols. The term *unique* refers to the property of the equations  $x_{i_1} - x_{i_2} = c_i$  that for every assignment to one variable, there is a unique assignment to the other variable so that the equation is satisfied. Unique Games were studied before in literature, in particular by Feige and Lovász [31] in the context of *parallel repetition*.

The important feature of the Unique Games is that the equations are linear. If one allows equations of arbitrary degree, each equation still depending on two variables, the problem may be referred to as a Non-Unique Game. The statement analogous to Conjecture 2.1 is known to hold for Non-Unique Games (and is very useful). It follows from a combination of the PCP Theorem stated in the next section and the Parallel Repetition Theorem of Raz [61]. For Non-Unique Games, the statement holds even on instances that have an assignment that satisfies all equations, as opposed to only  $1 - \delta$  fraction of the equations. Moreover, one only needs  $p$  to be polynomially large in  $\frac{1}{\delta}$ . For the Unique Games however, if there were an assignment that satisfies all equations, it can be efficiently found (an easy observation). Hence, it is essential in Conjecture 2.1 to consider only  $1 - \delta$  satisfiable instances. Moreover, if the conjecture were correct, it is known that  $p$  would have to be at least exponentially large in  $\frac{1}{\delta}$  [17].

### 3. The Max-Cut Problem, the PCP Theorem, the GW-Algorithm and its Optimality

The Unique Games Conjecture states that the Unique Game problem is hard to approximate. It has been shown that for several other optimization problems of interest, denoting a typical such problem by  $\Pi$ , there is a *reduction* from the Unique Game problem to the problem  $\Pi$  and as a consequence, the problem  $\Pi$  is hard to approximate as well. We sketch one such reduction below and refer the reader to [40] for a list of several reductions of this kind. We note that prior to formulation of the Unique Games Conjecture, researchers had already developed a general framework for similar reductions and techniques to analyze them [2, 13, 27, 35, 36, 61], with some remarkable successes such as Håstad's Clique result mentioned in the introduction. However these prior reductions were from the Non-Unique Game problem. For several problems  $\Pi$  of interest, we do not know how to reduce the Non-Unique Game problem to  $\Pi$ , but we do know how to reduce the Unique Game problem to  $\Pi$ . The Unique Game problem seems to strike a delicate balance: it has a simple enough structure that it is a convenient problem to reduce from and has a complex enough structure that it is plausibly a hard problem.

In this article, we focus on one specific optimization problem, namely the Max-Cut problem, and use it as an illustrative example throughout the article. In this problem, we are given a graph  $G(V, E)$  and the goal is to find a cut, i.e. partition of the vertex set  $V$  into two disjoint sets  $V_1$  and  $V_2$ , so as to maximize the number of edges cut. An edge is said to be cut if its one endpoint is in  $V_1$  and the other endpoint is in  $V_2$ . The size of the cut is the fraction of edges cut. Let  $\text{OPT}(G)$  denote the maximum size of any cut. We will focus on a particular special case of the problem when the graph  $G$  is almost bipartite, i.e. it has a cut that cuts almost all the edges. Let  $\varepsilon > 0$  be a small enough positive constant. The following problem will be the focus of the rest of the article.

**Max-Cut Problem:** Given a graph  $G(V, E)$  such that  $\text{OPT}(G) = 1 - \varepsilon$ . Find (efficiently) a cut of as large size as possible.

We will be interested in the computational complexity of this problem. A couple of observations are immediate. Firstly, the Max-Cut problem is NP-complete and hence one cannot hope to efficiently find a cut of the maximum size, i.e. of size  $1 - \varepsilon$ .<sup>3</sup> Secondly, one can easily find a cut of size  $\frac{1}{2}$ . Simply take a uniformly random cut in the graph; it cuts a fraction  $\frac{1}{2}$  of the edges in expectation and this randomized algorithm, if desired, can be turned into a deterministic algorithm as well. Till early 90's, this was all that was known regarding what is computationally infeasible and what is feasible. Two breakthrough results then led to a significant progress on this question: one from the hardness side, known as the PCP Theorem, and the other from the algorithmic side, namely the Goemans-Williamson's algorithm.

From the hardness side, the PCP Theorem [5, 7, 29] implies that it is not only hard to find a cut of the maximum size, but also hard to find a cut of near-maximum size. Specifically:<sup>4</sup>

**The PCP Theorem:** Assume  $P \neq NP$ . Then there is an absolute constant  $\beta > 1$  such that no polynomial time algorithm, given a graph that has a cut of size  $1 - \varepsilon$ , can find a cut of size  $1 - \beta\varepsilon$ .

The PCP Theorem is stated above as a hardness of approximation result. The acronym PCP stands for *Probabilistically Checkable Proofs* and indeed there is an equivalent formulation of the theorem in terms of *proof checking* (and this is what led to its discovery, as a culmination of much prior work on *interactive proofs*). The theorem states that every NP statement has a polynomial size proof that can be checked by a probabilistic polynomial time verifier by reading only a constant number of bits in the proof! The verifier has the completeness and the soundness property: every correct statement has a proof that is accepted with probability 1 and every proof of an incorrect statement is accepted with only a small probability, say at most 1%. The equivalence between the two viewpoints, namely the hardness viewpoint and the proof checking viewpoint, is simple but illuminating, and has influenced much of the work in this area. In this article, we restrict ourselves to the hardness viewpoint, i.e. the hardness result for the Max-Cut problem as stated above.

From the algorithmic side, Goemans and Williamson [33] designed an efficient algorithm that given a graph  $G(V, E)$  with a cut of size  $1 - \varepsilon$ , finds a cut of size  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon)$ . The latter quantity is approximated as  $1 - \frac{2}{\pi} \sqrt{\varepsilon} - O(\varepsilon^{3/2})$ . We provide a high-level sketch of the

<sup>3</sup>The standard NP-completeness reduction to the Max-Cut problem can be easily modified so that it holds on graphs with maximum cut of size  $1 - \varepsilon$ .

<sup>4</sup>The PCP Theorem actually proves that the stated computational task is NP-complete.

Goemans-Williamson’s algorithm.<sup>5</sup> The algorithm proceeds by computing an embedding  $\phi : V \mapsto \mathbb{S}^{m-1}$  of the set of vertices onto a unit sphere in  $\mathbb{R}^m$ . The dimension  $m$  is unrestricted, but w.l.o.g. can be assumed to be at most  $|V|$ . The embedding is computed by solving a so-called *semi-definite programming (SDP) relaxation* of the problem instance. We omit the description of this step (see [34] for introduction to SDPs and their algorithmic applications), but state the crucial property of the embedding: for most of the edges  $(u, v)$  in the graph, the endpoints  $u, v$  are embedded as points  $\phi(u), \phi(v)$  on the sphere that are nearly antipodal points. Once the embedding has been computed, the algorithm selects a hyperplane  $H$  in  $\mathbb{R}^m$  passing through the origin, uniformly at random from the set of all such hyperplanes. The hyperplane  $H$  cuts the sphere into two parts, which in turn induces a partition of the set  $V$  into two parts, depending on which side of the hyperplane the point  $\phi(v)$  lies, for a vertex  $v \in V$ . This yields the desired cut in the graph. The analysis of the algorithm then shows that the expected size of the cut is at least  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon)$ . Using the approximation cited before, this is at least  $1 - \sqrt{\varepsilon}$  for small enough  $\varepsilon$ .

In spite of the progress offered by the PCP Theorem and Goemans-Williamson’s algorithm, there is still a gap between  $1 - \beta\varepsilon$  and  $1 - \sqrt{\varepsilon}$ , regarding the size of the cut that is infeasible to compute and feasible to compute. Bridging this gap turns out to be an interesting pursuit as we demonstrate in this article. In particular, one could ask whether the Goemans-Williamson algorithm is the best possible algorithm in terms of its approximation guarantee. To the best of author’s information, when the Goemans-Williamson algorithm was discovered, it was viewed as somewhat unnatural and roundabout way of solving a combinatorial problem via a geometric method, and it was believed that a better algorithm would follow soon. However, rather surprisingly, assuming the Unique Games Conjecture, Goemans-Williamson’s algorithm is indeed optimal [41]:

**Theorem 3.1.** *Assume the Unique Games Conjecture. Fix any  $\varepsilon \in (0, \frac{1}{2})$  and let  $\eta > 0$  be an arbitrarily small constant. Then there is no polynomial time algorithm that given a graph with a cut of size at least  $1 - \varepsilon$ , finds a cut of size  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon) + \eta$ .*

Approximating  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon)$  as before, it will be convenient to focus on a (slightly weaker) statement: assuming the Unique Games Conjecture, there is no polynomial time algorithm that given a graph with a cut of size at least  $1 - \varepsilon$ , finds a cut of size  $1 - \frac{1}{2}\sqrt{\varepsilon}$ . Such a statement is proved by reducing the Unique Game problem to the Max-Cut problem. A reduction is a polynomial time procedure that starts with an instance  $\mathcal{L}$  of the Unique Game problem (i.e. a system of linear equations over  $\mathbb{Z}_p$  with two variables per equation) and builds an instance  $G$  of the Max-Cut problem (i.e. a graph) such that finding a large cut in  $G$  amounts to finding a good approximate solution to the system  $\mathcal{L}$ . Since the Unique Games Conjecture states that the latter task is computationally infeasible, so is the former. Specifically, the correctness of such a reduction consists of two statements, referred to as the completeness and the soundness statements: for given  $\varepsilon > 0$ , if  $\delta > 0$  is small enough,

$$\begin{aligned} (\text{Completeness}) \quad \text{OPT}(\mathcal{L}) \geq 1 - \delta &\implies \text{OPT}(G) \geq 1 - \varepsilon. \\ (\text{Soundness}) \quad \text{OPT}(\mathcal{L}) \leq \delta &\implies \text{OPT}(G) \leq 1 - \frac{1}{2}\sqrt{\varepsilon}. \end{aligned}$$

Now, the Unique Games Conjecture states that there is no polynomial time algorithm that given a  $(1 - \delta)$ -satisfiable system  $\mathcal{L}$ , finds a  $\delta$ -satisfying assignment. If the conjecture is

---

<sup>5</sup>Often we are interested in quality of approximation measured as a multiplicative factor. The minimum value of the ratio, over all  $\varepsilon \in (0, 1)$ , between  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon)$  and  $1 - \varepsilon$  is  $\approx 0.878$  and the Goemans-Williamson’s algorithm is often cited as a 0.878-approximation to Max-Cut.

correct, it then follows<sup>6</sup> that there is no polynomial time algorithm that given a graph with a cut of size  $1 - \varepsilon$ , finds a cut of size  $1 - \frac{1}{2}\sqrt{\varepsilon}$ .

We only provide a glimpse of the reduction here. Let  $\mathcal{L}$  be the given linear system over  $\mathbb{Z}_p$ . The reduction constructs, for every variable  $x_i$  in the linear system, a group of  $2^p$  vertices  $C_i$  labeled by boolean strings  $\sigma \in \{-1, 1\}^p$ . For every equation  $x_{i_1} - x_{i_2} = c_i$  in the linear system, there are edges between the group  $C_{i_1}$  and the group  $C_{i_2}$ . Roughly speaking, there is an edge between a vertex  $\sigma$  in group  $C_{i_1}$  and a vertex  $\tau$  in group  $C_{i_2}$  (here both  $\sigma, \tau$  are boolean strings of length  $p$ ) if

$$|\{\ell \in \{1, \dots, p\} \mid \sigma_{\ell+c_i} \neq \tau_\ell\}| \approx (1 - \varepsilon)p.$$

If one considers the special case when the equation is  $x_{i_1} - x_{i_2} = 0$ , then the last condition is same as saying that  $\sigma$  and  $\tau$  have Hamming distance  $\approx (1 - \varepsilon)p$ .

We omit the proofs of the completeness and the soundness properties. The proof of the completeness property is actually immediate from the construction. Proving the soundness property takes some work and though we omit its proof, we describe a key ingredient known as the Majority Is Stablest Theorem. This is a theorem about *noise-stability* of boolean functions on a boolean hypercube, i.e. of functions  $f : \{-1, 1\}^p \mapsto \{-1, 1\}$ . Any such function can be viewed as a cut in the set of vertices  $\{-1, 1\}^p$  and this is how one relates the theorem to the proof of the soundness of the reduction above. We present the Majority Is Stablest Theorem as well as a sketch of its proof, illustrating the connections to probability and Gaussian iso-perimetry.

#### 4. Majority is Stablest and Gaussian Isoperimetry

Suppose  $f : \{-1, 1\}^n \mapsto \{-1, 1\}$  is a boolean function. Every such function can be viewed as a pre-determined rule to decide outcome of an election, also referred to as a *voting scheme*: consider an election with  $n$  voters and two candidates labeled as  $\{-1, 1\}$ . The  $n$  voters vote for either of these candidates, uniformly and independently at random. If  $x_1, \dots, x_n \in \{-1, 1\}$  denotes the sequence of their votes, the winner of the election is declared to be  $f(x_1, \dots, x_n)$ . We focus on a voting scheme  $f$  that is *balanced*, i.e. both the candidates have equal chance of winning the election, and *democratic*, i.e. no individual voter has significant influence on the outcome of the election (formalized below). One example is the majority function  $\text{MAJ}_n = \text{sign}(x_1 + \dots + x_n)$  that corresponds to taking majority vote (say  $n$  is odd). Another example is *majority of majorities* that roughly corresponds to the electoral college system. We desire a voting scheme that is noise-stable, i.e. if a small fraction of votes are corrupted at random, then the probability that the outcome of the election changes is small. The Majority Is Stablest Theorem states that among all balanced and democratic voting schemes, the majority function is the most noise stable (up to a negligible additive error).

Formally, let  $f : \{-1, 1\}^n \mapsto \{-1, 1\}$  be a balanced boolean function, i.e.  $\Pr_x[f(x) = 1] = \Pr_x[f(x) = -1] = \frac{1}{2}$ , where the choice of input  $x$  is uniformly random over  $\{-1, 1\}^n$ . For a co-ordinate  $i \in \{1, 2, \dots, n\}$ , let the influence of the  $i^{\text{th}}$  co-ordinate on the function  $f$

<sup>6</sup>Strictly speaking, this implication is not immediate just from the completeness and the soundness statements. Formally, one shows that a cut of size at least  $1 - \frac{1}{2}\sqrt{\varepsilon}$  in  $G$  can be used, in polynomial time, to find a  $\delta$ -satisfying assignment to  $\mathcal{L}$ . Most reductions are constructive in this sense.

be defined as:

$$\text{Infl}_i(f) := \Pr_x [f(x_1, \dots, x_i, \dots, x_n) \neq f(x_1, \dots, -x_i, \dots, x_n)].$$

This is the probability that the function changes its value when the  $i^{\text{th}}$  co-ordinate is flipped, starting with a uniformly chosen input. A function is democratic if the influence of every co-ordinate is small. Let  $\varepsilon \in (0, \frac{1}{2})$  be a noise parameter. The  $\varepsilon$ -noise stability of the function  $f$  is defined as

$$\text{Stab}_\varepsilon(f) := \Pr_{x,y \sim N_\varepsilon(x)} [f(x) = f(y)], \tag{1}$$

where  $x$  is a uniformly chosen input and  $y$  is chosen from the distribution  $N_\varepsilon(x)$  obtained by flipping every co-ordinate of  $x$  independently with probability  $\varepsilon$  (thus  $y$  is a perturbed or noisy version of  $x$ ). It is known that the noise stability of the majority function  $\text{MAJ}_n$  tends to  $1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon)$  as  $n \rightarrow \infty$ . The Majority Is Stablest Theorem, proved by Mossel, O’Donnell, and Oleszkiewicz [54] (and conjectured in [41]) states that the noise stability of any balanced, democratic function is at most that of the majority function up to a negligible additive error.

**Theorem 4.1.** *Let  $\varepsilon \in (0, \frac{1}{2})$  be a noise parameter and  $\delta > 0$  be an arbitrarily small error parameter. Then for a sufficiently small constant  $\eta > 0$ , any balanced function  $f : \{-1, 1\}^n \mapsto \{-1, 1\}$  such that  $\forall i \in \{1, 2, \dots, n\}, \text{Infl}_i(f) \leq \eta$ , satisfies:*

$$\text{Stab}_\varepsilon(f) \leq 1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon) + \delta.$$

We present a sketch of the proof as it demonstrates the connection to an isoperimetric problem in geometry and its solution by Borell [15]. The proof involves an application of the *invariance principle* [19, 53, 54, 62]. Before we state the invariance principle, we note a few well-known facts. Any function  $f : \{-1, 1\}^n \mapsto \mathbb{R}$  can be represented as a multi-linear polynomial (Fourier or Walsh representation):

$$f(x) = \sum_{S \subseteq \{1, \dots, n\}} \widehat{f}(S) \prod_{i \in S} x_i,$$

where  $\widehat{f}(S) \in \mathbb{R}$  are the Fourier coefficients. When  $f$  is a boolean function, by Parseval’s identity,  $\sum_S \widehat{f}(S)^2 = \mathbb{E}_x [f(x)^2] = 1$ . It is easily proved that

$$\text{Infl}_i(f) = \sum_{i \in S} \widehat{f}(S)^2 \quad \text{and} \quad \text{Stab}_\varepsilon(f) = \frac{1}{2} + \frac{1}{2} \sum_S \widehat{f}(S)^2 (1 - 2\varepsilon)^{|S|}. \tag{2}$$

Using these formulas, the notion of influence and noise-stability can be extended to all multi-linear polynomials (and not just those representing boolean functions). Here is a rough statement of the invariance principle:

**Invariance Principle:** *Suppose  $f$  is a low degree multi-linear polynomial in  $n$  variables and all its variables have small enough influence. Then the distribution of the values of  $f$  is nearly identical when the input is a uniform random point from  $\{-1, 1\}^n$  or a random point from  $\mathbb{R}^n$  with the standard Gaussian measure.*

To motivate the invariance principle, one considers the case when  $f = \sum_{i=1}^n a_i x_i$  is a linear polynomial. Assume w.l.o.g. that  $\sum_{i=1}^n a_i^2 = 1$ . The condition that all variables

have small influence is equivalent to the condition that  $|a_i|$  is small for all  $i \in \{1, \dots, n\}$ . The invariance principle, in this case, states that the distribution of values of  $f(x_1, \dots, x_n)$  where  $x_i$  are i.i.d.  $\{-1, 1\}$  and the distribution of values of  $f(x_1^*, \dots, x_n^*)$  where  $x_i^*$  are i.i.d. standard Gaussian, are nearly identical. Indeed, by the Berry-Esseen Theorem [14, 28], the former distribution is nearly identical to a standard Gaussian and the latter distribution, being an appropriately weighted sum of independent standard Gaussians, is a standard Gaussian itself. The invariance principle is now viewed as a generalization of this special case to *low degree* multi-linear polynomials, with the definition of influences as in Equation (2).

The invariance principle allows us to translate the noise stability problem on boolean hypercube to a similar problem in the Gaussian space and the latter problem has already been solved by Borell [15]! Towards this end, let  $f$  be a boolean function on  $n$ -dimensional hypercube that is balanced and has all influences small enough. We intend to upper bound its  $\varepsilon$ -noise stability. Consider the representation of  $f$  as a multi-linear polynomial:

$$f(x) = \sum_S \widehat{f}(S) \prod_{i \in S} x_i \quad \forall x \in \{-1, 1\}^n.$$

Let  $f^* : \mathbb{R}^n \mapsto \mathbb{R}$  be a function that has the same representation as a multi-linear polynomial as  $f$  (with underlying standard Gaussian measure on  $\mathbb{R}^n$ ):

$$f^*(x^*) = \sum_S \widehat{f}(S) \prod_{i \in S} x_i^* \quad \forall x^* \in \mathbb{R}^n. \tag{3}$$

Assume for the moment that  $f$  has *low degree*. By the invariance principle, the distributions of  $f(x)$  and  $f^*(x^*)$  are nearly identical, and let's assume them to be identical for the sake of convenience. This implies that  $\mathbb{E}[f^*] = \mathbb{E}[f] = 0$  and since  $f$  is boolean, so is  $f^*$ . In other words,  $f^*$  is a partition of  $\mathbb{R}^n$  (with Gaussian measure) into two sets of equal measure. The next observation is that the  $\varepsilon$ -noise stability of  $f$  is same as the  $\varepsilon$ -“Gaussian noise stability” of  $f^* : \mathbb{R}^n \mapsto \{-1, 1\}$ , defined as

$$\text{Stab}_\varepsilon(f^*) := \Pr_{x^*, y^* \sim \mathcal{N}_\varepsilon(x^*)} [f^*(x^*) = f^*(y^*)]. \tag{4}$$

In the definition above,  $x^*$  is chosen from the standard  $n$ -dimensional Gaussian distribution and then  $y^*$  is chosen from the distribution  $\mathcal{N}_\varepsilon(x^*)$ , namely the perturbed or noisy version of  $x^*$ . Formally,  $y^* = (1 - 2\varepsilon)x^* + \sqrt{1 - (1 - 2\varepsilon)^2}z^*$  where  $z^*$  is a standard  $n$ -dimensional Gaussian independent of  $x^*$ . When  $f^*$  is a multi-linear polynomial as in Equation (3), it is easily proved that

$$\text{Stab}_\varepsilon(f^*) = \frac{1}{2} + \frac{1}{2} \sum_S \widehat{f}(S)^2 (1 - 2\varepsilon)^{|S|}.$$

But this expression is same as the  $\varepsilon$ -noise stability of the boolean function  $f$  and thus  $\text{Stab}_\varepsilon(f) = \text{Stab}_\varepsilon(f^*)$ . It is important here that the co-ordinate-wise correlation between the boolean pair  $(x, y)$  is same as the co-ordinate-wise correlation between the Gaussian pair  $(x^*, y^*)$  in Equations (1), (4) defining the boolean and Gaussian noise stability respectively (both correlations equal  $1 - 2\varepsilon$ ). Theorem 4.1 now follows from Borell’s result that upper bounds  $\text{Stab}_\varepsilon(f^*)$ .

**Theorem 4.2.** *If  $g^* : \mathbb{R}^n \mapsto \{-1, 1\}$  is a measurable function with  $\mathbb{E}[g^*] = 0$ , then*

$$\text{Stab}_\varepsilon(g^*) \leq \text{Stab}_\varepsilon(\text{HALF SPACE}) = 1 - \frac{1}{\pi} \arccos(1 - 2\varepsilon),$$

where HALF-SPACE is the partition of  $\mathbb{R}^n$  by a hyperplane through origin.

We note that the error parameter  $\delta$  in the statement of Theorem 4.1 accounts for additive errors involved at multiple places during the argument: firstly, the distributions  $f(x)$  and  $f^*(x^*)$  are only nearly identical. Secondly,  $f$  is not necessarily of low degree, and the invariance principle is not directly applicable. One gets around this issue by *smoothing*  $f$  that *kills* the high degree Fourier coefficients (which are then discarded) and only slightly affects the noise stability. This *truncated* version of  $f$  then has low degree and the invariance principle can be applied. We also note that the statement of Borell's Theorem holds for  $g^*$  that takes values in the interval  $[-1, 1]$  and the noise stability is defined as in Equation (2).

To summarize, an iso-perimetric (type) result in the Gaussian space (e.g. Borell's Theorem) implies a Fourier analytic result on the hypercube (e.g. Majority Is Stablest), which in turn implies correctness of a reduction from the Unique Game problem to an optimization problem  $\Pi$  of interest (e.g. Max-Cut), showing that  $\Pi$  is hard to approximate. It turns out that this scheme applies to several optimization problems  $\Pi$  and not just for Max-Cut. In fact, for a class of problems known as constraint satisfaction problems, Max-Cut being one example, the three components, namely an iso-perimetric type result, a Fourier analytic result and a UGC-based hardness of approximation result, are formally equivalent [58]. The scheme also leads to new iso-perimetric type and Fourier analytic theorems and conjectures, motivated by applications to hardness of approximation (see [39] for examples).

## 5. Counter-examples to Proposed Algorithms

An interesting aspect of the Unique Games Conjecture is that it predicts the existence of *counter-examples* to proposed algorithms and answering whether such counter-examples indeed exist often turns out to be a challenging task with connections to geometry. We briefly explain this scheme and cite one example that leads to non-embeddability results for finite metrics.

Suppose there is a reduction from the Unique Game problem to a computational problem  $\Pi$  (similar to the reduction to the Max-Cut problem described earlier). Thus assuming the Unique Games Conjecture, the problem  $\Pi$  is hard to approximate. Nevertheless, one is free to propose an efficient algorithm  $\mathcal{A}$  towards approximating  $\Pi$  and even a family of efficient algorithms  $\{\mathcal{A}_i\}_{i=1,2,\dots}$  that are increasingly more sophisticated. The Unique Games Conjecture predicts that  $\Pi$  is hard to approximate, and hence each of these proposed algorithms must fail. In particular, there must be a family of counter-examples (i.e. instances of the problem)  $\{\mathcal{C}_i\}_{i=1,2,\dots}$  demonstrating the failure of the corresponding algorithms. Moreover, the more sophisticated the proposed algorithms are, the more sophisticated the counter-examples would need to be. To emphasize, the counter-examples are concrete instances of the problem (e.g. graphs when the Max-Cut problem is considered) with a specific combinatorial or geometric structure.

When this scheme is applied to a problem called Sparsest Cut that is closely related to the Max-Cut problem, the Unique Games Conjecture predicts that there are  $n$ -point finite metrics with non-trivial structural properties. In the Sparsest Cut problem, given a graph, the goal is to cut the graph into two roughly equal sized parts so as to minimize the fraction of edges cut.<sup>7</sup> There is a reduction from the Unique Game problem to the Sparsest Cut problem

---

<sup>7</sup>There are some subtleties regarding the so-called *uniform* and *non-uniform* versions of the problem that we omit here.

[20, 43, 66], so one predicts that the latter problem is hard to approximate.

Nevertheless, since mid-90s, researchers have proposed a family of increasingly sophisticated algorithms based on linear and semi-definite programming relaxation [4, 6, 8, 49–51]. These algorithms *relax* the Sparsest Cut problem to computing a metric on the set of vertices of the given graph that is *well-spread* and minimizes the average distance along the edges of the graph. It is possible to impose increasingly stringent restrictions on the type of metric allowed, leading to increasingly sophisticated algorithms. Cuts in an  $n$ -vertex graph are closely related to  $n$ -point  $\ell_1$  metrics and the approximation quality of the algorithm depends on how *close* the metric happens to be an  $\ell_1$  metric.

However, the Unique Games Conjecture predicts that all these algorithms must fail and hence corresponding counter-examples exist.<sup>8</sup> For some of these algorithms, researchers have already been able to construct such counter-examples (technically known as *integrality gap examples*), which amount to construction of  $n$ -point metrics with increasingly stringent structural properties. Before we state the known results, we introduce a notion of metric embedding.

A metric  $(X, d_X)$  consists of a set of points  $X$  and a distance function  $d_X(\cdot, \cdot)$  on pairs of points that is non-negative, symmetric and satisfies the triangle inequality. An embedding of a metric space  $(X, d_X)$  into another metric space  $(Y, d_Y)$  is a map  $\phi : X \mapsto Y$ . The embedding is said to have distortion  $D$  if distances do not shrink and are not stretched by more than a factor  $D$ , i.e.

$$\forall a, b \in X, \quad d_X(a, b) \leq d_Y(\phi(a), \phi(b)) \leq D \cdot d_X(a, b).$$

An embedding with distortion  $D = 1$  is said to be an isometric embedding. It is easily observed that if  $(X, d_X)$  is a metric, then so is  $(X, \sqrt{d_X})$ , i.e. when the new distances are square root of the original distances. A metric  $(X, d_X)$  is said to be of *negative type* if the metric  $(X, \sqrt{d_X})$  embeds isometrically into  $\ell_2$ . A sub-metric of a metric  $(X, d_X)$  is a subset  $S \subseteq X$  with the same distances between points in  $S$ . We are now ready to state the result predicted by the Unique Games Conjecture and verified by researchers with explicit constructions (some of which precede the prediction).

**Theorem 5.1.** *There are functions  $D(n), t(n) \rightarrow \infty$  as  $n \rightarrow \infty$  and a family of  $n$ -point metrics  $(X, d_X)$  such that*

- *There is no embedding of  $(X, d_X)$  into  $\ell_1$  with distortion  $D(n)$  [8, 16, 51].*
- *The metric  $(X, d_X)$  is of negative type [21–24, 26, 43, 45, 48].*
- *Every sub-metric of  $(X, d_X)$  on  $t(n)$  points embeds isometrically into  $\ell_1$  [42, 59].*

To state the theorem succinctly, there are negative-type metrics that embed isometrically into  $\ell_1$  locally, but do not embed well into  $\ell_1$  globally. The results cited hold for various quantitative settings of the parameters  $D(n), t(n)$ , but we omit these here and refer to [39, 55]. From the algorithmic side, it is possible to impose even more stringent restrictions on the metric (e.g. via the so-called Lasserre SDP relaxation), but then the existence of metrics with these restrictions (on top of those in Theorem 5.1) is open.

---

<sup>8</sup>Here we mean failure to approximate up to a constant multiplicative factor. If the approximation factor is allowed to depend on the size of the graph, the papers cited do indeed give a reasonable approximation.

### 6. Open Problem: Power of Sum-of-Squares Refutation System

In this section and the next, we present two open problems towards settling the Unique Games Conjecture. The first one concerns the power of *refutation systems*. Suppose we have a correct, efficient algorithm for a computational problem  $\Pi$  (computing either exact or approximate solution). Suppose moreover that on some instance  $\mathcal{I}$  of the problem, the algorithm does not find a solution. Since the algorithm is correct, the fact that it does not find a solution, is a proof that no solution exists, and often, a formal proof of infeasibility of a solution can be obtained by examining the execution of the algorithm on the instance  $\mathcal{I}$ . A proof of infeasibility of a solution is referred to as a *refutation*. More specifically, a refutation starts with a false hypothesis that a solution exists and then reaches a contradiction via a sequence of deductions. Naturally, for a refutation derived from the execution of an algorithm, the complexity of the refutation is related to the complexity of the algorithm. Turning this argument around, if on some instance  $\mathcal{I}$  of the problem, if there is no *simple* refutation, this may be considered as evidence that the problem  $\Pi$  has no *simple* or *efficient* algorithm.

This motivates the study of *refutation systems* where a refutation conforms to a given set of rules for deducing successive statements, starting with a hypothesis to be refuted, e.g. a false hypothesis stating that a feasible solution exists when one doesn't. Depending on the kind of deduction rules allowed, one gets different refutation systems and their study is the subject of *proof complexity* (see [11, 12] for surveys). Here we focus on the Lovász-Schrijver, Sherali-Adams and the Lasserre systems. In these systems, each step of the refutation is an inequality and the system specifies how to derive new inequalities from the previous ones. There is a dual, algorithmic view of these systems and from that viewpoint, these systems correspond to LP/SDP relaxations (known as LP/SDP *hierarchies*) that we mentioned before. We refer to [65] for an introduction to and comparison between these systems (hierarchies).

As we said, we wish to show lower bounds for refutation systems, i.e. construct (infeasible) instances  $\mathcal{I}$  such that there is no *simple* refutation within a given system. Showing such lower bounds then corresponds to constructing counter-examples (i.e. *integrality gaps*) for the corresponding LP/SDP relaxation in the dual viewpoint, as discussed in Section 5.

Regarding the Max-Cut problem, reasonable lower bounds are known for the Lovász-Schrijver and Sherali-Adams systems (which are LP based) and also for some basic SDP-based systems [18, 25, 32, 37, 42, 43, 59]. However, showing lower bounds for the Lasserre system (which is SDP based) remains a major challenge and this is our first open problem. The Lasserre system is also known as Sum-of-Squares system and its variants have been studied independently by various authors including Shor, Parrilo, Nesterov, and Lasserre [46, 56, 57, 63]. It is closely related to the Hilbert's 17<sup>th</sup> problem and we refer to [10, 47] for detailed expositions. Here we present the open problem in a self-contained manner.

Let us fix a graph  $G(V = \{1, 2, \dots, n\}, E)$  such that the maximum sized cut in the graph cuts exactly  $(1 - \epsilon)|E|$  edges. We can write down an infeasible set of polynomial equalities and inequalities over reals, denoted  $\mathcal{S}$ , as follows:

$$\mathcal{S} : \quad \forall i \in \{1, \dots, n\}, \quad x_i^2 - 1 = 0 \quad (P_i(x) = 0).$$

$$\sum_{(i,j) \in E} \frac{1 - x_i x_j}{2} - (1 - \epsilon)|E| - 1 \geq 0 \quad (Q(x) \geq 0).$$

The set of equations is written as  $P_i(x) = 0$  and the inequality is written as  $Q(x) \geq 0$  where

$P_i, Q$  are polynomials in  $\mathbb{R}[x_1, \dots, x_n]$  as shown. Let's first see why this set of (in)equalities is infeasible. The equations  $x_i^2 - 1 = 0$  force the variables  $x_i$  to take values in  $\{-1, 1\}$ . Any  $\{-1, 1\}$ -assignment to the variables is viewed as a cut in the graph and then the inequality  $Q \geq 0$  states that the cut cuts at least  $(1 - \varepsilon)|E| + 1$  edges, contradicting the assumption that the maximum sized cut cuts only  $(1 - \varepsilon)|E|$  edges. Indeed the expression  $\frac{1 - x_i x_j}{2}$  equals 1 or 0 depending on whether the edge  $(i, j)$  is cut or not and hence the sum  $\sum_{(i,j) \in E} \frac{1 - x_i x_j}{2}$  equals the number of edges cut.

How could one refute this infeasible set of (in)equalities? One possible way is to come up with polynomials  $\{R_i\}_{i=1}^n, \{S_j, T_j\}_{j=1}^\ell \in \mathbb{R}[x_1, \dots, x_n]$  such that the following polynomial identity holds:

$$\sum_{i=1}^n R_i P_i + (S_1^2 + \dots + S_\ell^2)Q + (T_1^2 + \dots + T_\ell^2) = -1.$$

This would be a contradiction, hence providing a valid refutation. Indeed, since  $P_i = 0$  and  $Q \geq 0$  and the polynomials  $S_j, T_j$  appear only in squared form, the left hand side of the identity is non-negative whereas the right hand side is  $-1$ . The refutation is called a Sum-of-Squares refutation.

It turns out that a Sum-of-Squares refutation always exists and the question is whether there is one that is *simple*. A natural measure of its complexity is the maximum degree of the polynomials  $R_i P_i, S_j^2 Q, T_j^2$  involved, called the degree of the refutation. It is known that a degree  $d$  refutation, if one exists, can be found in time  $n^{O(d)}$ , i.e. in polynomial time for constant  $d$ . Thus it is desirable to have a refutation with constant degree (independent of the size of the graph). From a lower bound perspective, it is known that there are  $n$ -vertex graphs for which any Sum-of-Square refutation requires degree  $\Omega(n)$  (degree  $O(n)$  always suffices).

What if we insist on having a constant degree refutation? One possibility is to start with a hypothesis that is *even more false*. In particular, one can consider the set of (in)equalities:

$$S' : \quad \forall i \in \{1, \dots, n\}, \quad x_i^2 - 1 = 0.$$

$$\sum_{(i,j) \in E} \frac{1 - x_i x_j}{2} - (1 - \varepsilon^2)|E| \geq 0.$$

Note that the inequality hypothesizes that there is a cut that cuts at least  $(1 - \varepsilon^2)|E|$  edges. This hypothesis is much more false than the earlier hypothesis stating that there is a cut that cuts at least  $(1 - \varepsilon)|E| + 1$  edges and thus is plausibly easier to refute. Indeed, for any graph (with maximum cut of size  $(1 - \varepsilon)|E|$ ), the set of (in)equalities  $S'$  has a Sum-of-Squares refutation of degree 2! Such a refutation can be obtained by taking a dual view of the Goemans-Williamson's SDP algorithm for the Max-Cut problem.

These considerations lead to our first open problem: what happens when we use a hypothesis stating that there is a cut that cuts a number of edges that is intermediate between  $(1 - \varepsilon)|E| + 1$  and  $(1 - \varepsilon^2)|E|$ ? Is there always a constant degree refutation (noting that one needs degree  $\Omega(n)$  for some graphs at first extreme and degree 2 always suffices at the second extreme)? Specifically, Let  $\tilde{\varepsilon}$  be any constant such that  $\varepsilon^2 \ll \tilde{\varepsilon} \ll \varepsilon$ . The Unique Games Conjecture predicts, as discussed in Section 3, that no polynomial time algorithm, given a graph with maximum cut of size  $1 - \varepsilon$ , finds a cut of size  $1 - \frac{1}{2}\sqrt{\tilde{\varepsilon}}$ . This prediction, when translated to a prediction regarding lower bounds for the Sum-of-Squares refutation system, states:

**Prediction:** Let  $\varepsilon^2 \ll \tilde{\varepsilon} \ll \varepsilon$ . There are graphs  $G(V = \{1, 2, \dots, n\}, E)$  with the maximum cut of size exactly  $(1 - \varepsilon)|E|$  such that any Sum-of-Squares refutation of the set of (in)equalities:

$$\tilde{S} : \forall i \in \{1, \dots, n\}, \quad x_i^2 - 1 = 0.$$

$$\sum_{(i,j) \in E} \frac{1 - x_i x_j}{2} - (1 - \tilde{\varepsilon})|E| \geq 0.$$

requires a super-constant degree (i.e. tending to  $\infty$  as  $n \rightarrow \infty$ ).

Clearly, constructing graphs that require a super-constant degree refutation for some  $\varepsilon^2 \ll \tilde{\varepsilon} \ll \varepsilon$  would support the Unique Games Conjecture whereas showing that there is always a constant degree refutation for some  $\varepsilon^2 \ll \tilde{\varepsilon} \ll \varepsilon$  would disprove the Unique Games Conjecture.

### 7. Open Problem: Small Set Expander Graphs with Many Large Eigenvalues

The second open problem concerns the existence of *small set expander* graphs with many large eigenvalues. The problem is motivated by the Small Set Expansion Conjecture posed by Raghavendra and Steurer [60]. The conjecture concerns the computational complexity of the *small set expansion* problem which, given a graph, asks for a small (but still of linear size) subset of vertices that does not expand well. The conjecture states that this problem is hard to approximate; a formal statement appears below.

For a  $d$ -regular graph  $G(V, E)$  and a set  $S \subseteq V$ , define the expansion of the set  $S$  as  $\phi(S) := \frac{|E(S, V \setminus S)|}{d \cdot |S|}$ , i.e. the fraction of edges incident on  $S$  that leave  $S$ . Raghavendra and Steurer pose:<sup>9</sup>

**Conjecture 7.1.** *For every constant  $\varepsilon > 0$ , there exists a constant  $\gamma > 0$  such that no polynomial time algorithm, given a regular graph  $G(V, E)$ , can distinguish whether it is a YES Type graph or a NO Type graph as defined below:*

- (YES Type:) *There is a set  $S \subseteq V, |S| = \gamma|V|$  such that  $\phi(S) \leq \varepsilon$ .*
- (NO Type:) *For every set  $S \subseteq V, |S| \approx \gamma|V|, \phi(S) \geq \frac{1}{10}$ .*

As a clarification, we note that a *distinguishing* algorithm takes a graph as input and in polynomial time outputs an answer that is YES if the graph is of YES Type and is NO if the graph is of NO Type. For graphs that are of neither type, the output of the algorithm can be arbitrary. Though the conjecture is phrased as above (as is customary in computer science), the reader may find it more convenient to consider the following version implied by it (computer scientists tend to view the two versions as morally the same):

**Conjecture:** *There is no polynomial time algorithm that, given a graph of the YES Type, meaning one containing a small set (i.e. of relative size  $\approx \gamma$ ) that is almost non-expanding (i.e. has expansion at most  $\varepsilon$ ), finds a small set that is somewhat non-expanding (i.e. has expansion less than  $\frac{1}{10}$ ).*

Finding small non-expanding sets is a natural problem in itself and in addition, Raghavendra and Steurer show that this conjecture implies the Unique Games Conjecture. Therefore,

---

<sup>9</sup>Here  $|S| \approx \gamma|V|$  means that, say,  $|S|$  is between  $\frac{\gamma}{2}|V|$  and  $2\gamma|V|$ .

it is worthwhile to explore this conjecture. As discussed in Section 5, for a computational problem that is predicted to be hard to approximate, the small set expansion problem in this case, one can propose an efficient algorithm and then try to find counter-examples to the proposed algorithm.

It is indeed possible to propose a natural algorithm to find small non-expanding sets [3, 44]. We briefly sketch the algorithm. Let  $A(G)$  be the normalized adjacency matrix of a  $d$ -regular  $n$ -vertex graph  $G(V, E)$ . This is a  $n \times n$  matrix with diagonal entries as 1 and an off-diagonal entry  $(i, j)$  is  $\frac{1}{d}$  if  $(i, j)$  is an edge in the graph and zero otherwise. It is well-known that the eigenvalues of this matrix are in  $[-1, 1]$  and the largest eigenvalue equals 1. Let  $v_1, \dots, v_m \in \mathbb{R}^n$  be the *top eigenvectors*, i.e. those corresponding to eigenvalues that are at least  $1 - O(\varepsilon)$ . For a subset of vertices  $S \subseteq V$ , let  $\mathbf{1}_S \in \mathbb{R}^n$  denote the indicator vector of the subset  $S$ , i.e. its  $i^{\text{th}}$  co-ordinate equals 1 if the  $i^{\text{th}}$  vertex is in  $S$  and zero otherwise. It is easily shown (e.g. [3, Theorem 2.2]) that if the graph has a subset  $S \subseteq V$ ,  $|S| = \gamma n$  such that  $\phi(S) \leq \varepsilon$ , then the indicator vector  $\mathbf{1}_S$  is essentially contained in the linear span of the top eigenvectors  $v_1, \dots, v_m$ . Thus the vector  $\mathbf{1}_S$  and hence the set  $S$  (or rather, an approximation to them) can be found by searching over all the vectors in this  $m$ -dimensional linear span (up to a suitable discretization) and outputting a vector that resembles an indicator vector of a set of size  $\approx \gamma n$ . Let's refer to this algorithm as a *subspace search* algorithm; it runs in time roughly  $2^{O(m)}$ .

Now consider a proposed algorithm to distinguish between the YES Type and NO Type graphs as in the statement of Conjecture 7.1. Compute the eigenvalues and eigenvectors of the matrix  $A(G)$ . If the number of large eigenvalues  $m$  is at most  $n^{o(1)}$ , proceed further and otherwise answer YES. If  $m$  is at most  $n^{o(1)}$ , run the subspace search algorithm and answer YES or NO depending on whether it manages to find a set of size  $\approx \gamma n$  with expansion  $\ll \frac{1}{10}$ . Note that the proposed algorithm always answers YES on a graph of the YES Type.

However, Conjecture 7.1 predicts that every polynomial time algorithm fails in distinguishing between the YES Type and the NO Type graphs. In fact, Raghavendra and Steurer state Conjecture 7.1 in a stronger form, predicting that the task of distinguishing between the YES Type and the NO Type graphs is NP-complete, and every algorithm that runs in time  $2^{n^{o(1)}}$  time fails as well. Considering the proposed algorithm as above, the only reason for it to fail is that it mistakenly answers YES on some graph that is of the NO Type. Thus we are led to the following prediction:

**Prediction:** For every constant  $\varepsilon > 0$ , there exist constants  $\gamma, \delta > 0$  and an infinite family of  $n$ -vertex graphs  $G(V, E)$  of the NO Type, i.e.  $\forall S \subseteq V$ ,  $|S| \approx \gamma n$ ,  $\phi(S) \geq \frac{1}{10}$ , such that the number of its eigenvalues  $\geq 1 - \varepsilon$  is at least  $n^\delta$ .

The open question is whether such graphs exist (see [9] for some progress). It is possible that such graphs do not exist and the Small Set Expansion Conjecture is false (and the Unique Games Conjecture might still be true).

## 8. Conclusion

We have sketched some connections between the Unique Games Conjecture, geometry and analysis. Irrespective of whether the Unique Games Conjecture turns out to be true or false, exploring these connections further, and in particular making progress on the open questions cited, seems worthwhile.

**Acknowledgements.** This work is supported by NSF grants CCF-0832795, 1061938, 1422159, and Simons Collaboration on Algorithms and Geometry grant.

## References

- [1] S. Arora, *Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems*, Journal of the ACM, **45**(5) (1998), 53–782.
- [2] S. Arora, L. Babai, J. Stern, and Z. Sweedyk, *The hardness of approximate optima in lattices, codes, and systems of linear equations*, Journal of Computer and System Sciences, **54**(2) (1997), 317–331.
- [3] S. Arora, B. Barak, and D. Steurer, *Subexponential Algorithms for Unique Games and Related Problems*, In Proc. 51th IEEE Symposium on Foundations of Computer Science, 2010.
- [4] S. Arora, J. Lee, and A. Naor, *Euclidean distortion and the sparsest cut*, In Proc. 37th ACM Symposium on Theory of Computing, pages 553–562, 2005.
- [5] S. Arora, C. Lund, R. Motawani, M. Sudan, and M. Szegedy, *Proof verification and the hardness of approximation problems*, Journal of the ACM, **45**(3) (1998), 501–555.
- [6] S. Arora, S. Rao, and U. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, In Proc. 36th ACM Symposium on Theory of Computing, pages 222–231, 2004.
- [7] S. Arora and S. Safra, *Probabilistic checking of proofs : A new characterization of NP*, Journal of the ACM, **45**(1) (1998), 70–122.
- [8] Y. Aumann and Y. Rabani, *An  $O(\log k)$  approximate min-cut max-flow theorem and approximation algorithm*, SIAM J. Comput., **27**(1) (1998), 291–301.
- [9] B. Barak, P. Gopalan, J. Håstad, R. Meka, P. Raghavendra, and D. Steurer, *Making the Long Code Shorter*, In Proc. IEEE Symposium on Foundations of Computer Science, 2012, 370–379.
- [10] B. Barak and D. Steurer, *Sum-of-Squares Proofs and the Quest toward Optimal Algorithms*, In Proc. of the International Congress of Mathematicians, 2014.
- [11] P. Beame, *Proof complexity*, Computational Complexity Theory, volume 10 of IAS/Park City mathematics series, pages 199–246. American Mathematical Society, 2004.
- [12] P. Beame and T. Pitassi, *Propositional proof complexity: Past, present, and future*, Current trends in theoretical computer science: Entering the 21st century, World Scientific Publishing, 2001, pp. 42–70.
- [13] M. Bellare, O. Goldreich, and M. Sudan, *Free Bits, PCPs, and Nonapproximability—Towards Tight Results*, SIAM Journal on Computing, **27**(3) (1998), 804–915.
- [14] A. Berry, *The Accuracy of the Gaussian Approximation to the Sum of Independent Variates*, Transactions of the American Mathematical Society, **49**(1) (1941), 122–136.
- [15] C. Borell, *Geometric bounds on the Ornstein-Uhlenbeck velocity process*, Z. Wahrsch. Verw. Gebiete, **70**(1) (1985), 1–13.
- [16] J. Bourgain, *On Lipschitz embeddings of finite metric spaces in Hilbert space*, Israel

- Journal of Mathematics, **52** (1985), 46–52.
- [17] M. Charikar, K. Makarychev, and Y. Makarychev, *Near-Optimal Algorithms for Unique Games*, In Proc. Annual ACM Symposium on Theory of Computing, 2006, pp. 205–214.
- [18] ———, *Integrality Gaps for Sherali-Adams Relaxations*, In Proc. ACM Symposium on Theory of Computing, 2009, pp. 283–292.
- [19] S. Chatterjee, *A simple invariance theorem*, arXiv:math/0508213v1, 2005.
- [20] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar, *On the hardness of approximating multicut and sparsest-cut*, In Proc. 20th IEEE Conference on Computational Complexity, pages 144–153, 2005.
- [21] J. Cheeger and B. Kleiner, *Differentiating maps into  $L^1$  and the geometry of BV functions*, Ann. Math., Second Series, **171**, No. 2, 2010.
- [22] ———, *On the differentiation of Lipschitz maps from metric measure spaces to Banach spaces*, Inspired by S.S. Chern, Volume 11 of Nankai Tracts. Math., pages 129–152, 2006.
- [23] J. Cheeger, B. Kleiner, and A. Naor, *Compression bounds for Lipschitz maps from the Heisenberg group to  $L_1$* , Acta Mathematica, **207**(2) (2011), 291–373.
- [24] ———, *A  $(\log n)^{\Omega(1)}$  integrality gap for the sparsest cut SDP*, In Proc. 50th IEEE Symposium on Foundations of Computer Science, 2009.
- [25] W.F. de la Vega and C. Kenyon-Mathieu, *Linear Programming Relaxations of Maxcut*, In Proc. 18th Symposium on Discrete Algorithms, pp. 53–61, 2007.
- [26] N. Devanur, S. Khot, R. Saket, and N. Vishnoi, *Integrality gaps for sparsest cut and minimum linear arrangement problems*, In Proc. 38th ACM Symposium on Theory of Computing, 2006.
- [27] I. Dinur and S. Safra, *The importance of being biased*, In Proc. 34th Annual ACM Symposium on Theory of Computing, 2002.
- [28] C.G. Esseen, *On the Liapunoff limit of error in the theory of probability*, Arkiv for matematik, astronomi och fysik, **A28** (1942), 1–19.
- [29] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy, *Interactive proofs and the hardness of approximating cliques*, Journal of the ACM, **43**(2) (1996), 268–292.
- [30] U. Feige, *Threshold of  $\ln n$  for Approximating Set Cover*, Journal of the ACM, **45**(4) (1998), 634–652.
- [31] U. Feige and L. Lovász, *Two-prover one-round proof systems, their power and their problems*, In Proc. of the ACM Symposium on the Theory of Computing, pages 733–744, 2002.
- [32] U. Feige and G. Schechtman, *On the optimality of the random hyperplane rounding technique for max cut*, Random Struct. Algorithms, **20**(3) (2002), 403–440.
- [33] M. Goemans and D. Williamson, *0.878 approximation algorithms for MAX-CUT and MAX-2SAT*, In Proc. 26th ACM Symposium on Theory of Computing, pages 422–431, 1994.
- [34] M. X. Goemans, *Semidefinite programming in combinatorial optimization*, Math. Program., **79** (1997), 143–161.

- [35] J. Hastad, *Clique is hard to approximate within  $n^{1-\epsilon}$* , Acta Mathematica, **182** (1999), 105–142.
- [36] ———, *Some optimal inapproximability results*, Journal of ACM, **48** (2001), 798–859.
- [37] H. Karloff, *How good is the Goemans-Williamson MAX CUT algorithm?*, In Proc. of the twenty-eighth annual ACM Symposium on Theory of Computing, pages 427–434, 1996.
- [38] S. Khot, *On the power of unique 2-prover 1-round games*, In Proc. 34th ACM Symposium on Theory of Computing, 2002.
- [39] ———, *Inapproximability of NP-complete problems, Discrete Fourier Analysis, and Geometry*, In Proc. of the International Congress of Mathematicians, 2010.
- [40] ———, *On the Unique Games Conjecture*, In Proc. IEEE Conference on Computational Complexity, 2010.
- [41] S. Khot, G. Kindler, E. Mossel, and R. O’Donnell, *Optimal inapproximability results for max-cut and other 2-variable CSPs?*, In Proc. 45th IEEE Symposium on Foundations of Computer Science, pages 146–154, 2004.
- [42] S. Khot and R. Saket, *SDP integrality gaps with local  $\ell_1$ -embeddability*, In Proc. 50th IEEE Symposium on Foundations of Computer Science, pages 565–574, 2009.
- [43] S. Khot and N. Vishnoi, *The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into  $\ell_1$* , In Proc. 46th IEEE Symposium on Foundations of Computer Science, 2005.
- [44] A. Kolla, *Spectral Algorithms for Unique Games*, In Proc. IEEE Conference on Computational Complexity, 2010.
- [45] R. Krauthgamer and Y. Rabani, *Improved lower bounds for embeddings into  $l_1$* , In Proc. ACM-SIAM Symposium on Discrete Algorithms, 2006.
- [46] J.B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM Journal on Optimization, **11** (2001), no. 3, 796–817.
- [47] M. Laurent. *A Comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre Relaxations for 0-1 Programming*, Math. Oper. Res., **28**(3) (2003), 470–496.
- [48] J. R. Lee and A. Naor,  *$l_p$  metrics on the Heisenberg group and the Goemans-Linial conjecture*, In Proc. 47th IEEE Symposium on Foundations of Computer Science, pages 99–108, 2006.
- [49] T. Leighton and S. Rao, *Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms*, Journal of the ACM, Volume 46, Issue 6, pages 787–832, 1999.
- [50] N. Linial, *Finite metric spaces-combinatorics, geometry and algorithms*, In Proc. International Congress of Mathematicians, volume 3, pages 573–586, 2002.
- [51] N. Linial, E. London, and Y. Rabinovich, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, **15**(2) (1995), 215–245.
- [52] J.S.B. Mitchell, *Guillotine subdivisions, approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems*, SIAM Journal on Computing, **28**(4) (1999), 1298–1309.
- [53] E. Mossel, *Gaussian Bounds for Noise Correlation of Functions*, Geometric and

- Functional Analysis, March 2010, Volume 19, Issue 6, pp. 1713–1756.
- [54] E. Mossel, R. O’Donnell, and K. Oleszkiewicz, *Noise stability of functions with low influences: invariance and optimality*, In Proc. 46th IEEE Symposium on Foundations of Computer Science, pages 21–30, 2005.
  - [55] A. Naor,  *$L_1$  embeddings of the Heisenberg group and fast estimation of graph isoperimetry*, In Proc. the International Congress of Mathematicians, 2010.
  - [56] Y. Nesterov, *Squared functional systems and optimization problems*, High performance optimization, **13** (2000), 405–440.
  - [57] P. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Mathematical Programming, **96** (2003), 293–320.
  - [58] P. Raghavendra, *Optimal algorithms and inapproximability results for every CSP?*, In Proc. ACM Symposium on Theory of Computing, 2008, 245–254.
  - [59] P. Raghavendra and D. Steurer, *Integrality gaps for strong SDP relaxations of unique games*, In Proc. 50th IEEE Symposium on Foundations of Computer Science, pages 575–585, 2009.
  - [60] ———, *Graph expansion and the unique games conjecture*, In Proc. 42nd ACM Symposium on Theory of Computing, 2010.
  - [61] R. Raz, *A Parallel Repetition Theorem*, SIAM Journal of Computing, **27**(3) (1998), 763–803.
  - [62] V. Rotar, *Limit theorems for polylinear forms*, J. Multivariate Anal. **9**(4) (1979), 511–530.
  - [63] N.Z. Shor, *An approach to obtaining global extremums in polynomial mathematical programming problems*, Cybernetics and Systems Analysis, **23** (1987), no. 5, 695–700.
  - [64] L. Trevisan, *On Khot’s Unique Games Conjecture*, Bulletin of the AMS, **49**(1) (2012), 91–111.
  - [65] E. Chlamtac and M. Tulsiani, *Convex Relaxations and Integrality Gaps*, Springer Handbook on Semidefinite, Conic and Polynomial Optimization, 2012.
  - [66] M. Tulsiani, P. Raghavendra, and D. Steurer. *Reductions between Expansion Problems*, In Proc. IEEE Conference on Computational Complexity, 2012.

Computer Science Department, New York University, 251 Mercer Street, New York, NY 10012, USA  
E-mail: khot@cs.nyu.edu

# I want to play with mathematics

Adrián Paenza

Good evening, everyone. As you can see there, well, this is the first thing that had to happen. I had to drop my pen. Thank you for coming. It's a great honor for me to be here and I will try to tell you my story. I won't speak of mathematics because I'm not qualified for that. What I'm going to do is tell you my journey... my journey with math, which is an incredible journey... one of the most beautiful things that life has... one of the most beautiful, if not the best, science that we have. And unfortunately, for whatever the reason, people don't perceive it that way.

So I want to tell you my story. Some of the things that I'm going to say are –probably-going to be controversial, if nothing else, because they're going to be my opinions. So they are always open to be debated and that's good. That's what we should do: debate and discuss, always in good faith.

The first thing that happened to me when I got to Seoul, and I got here two or three days ago, was that I was talking to Ona. Ona is a South Korean professor, a friend that I have here in Seoul. I met her when we were members of a jury last year in Providence, Rhode Island in the United States. And the first thing that really surprised me is the way that people here do calculations. And I'm going to write the discussions that I had with Ona.

I was trying to see, how do you people divide here in South Korea. So, I asked her: How do you divide 25 divided by 5? And she told me that 25 divided by 5 here in Korea is 5. And I was surprised because for me, 25 divided by 5 is 14. And I explained it to her. See, here? 1 times 5, that is 5. Then we subtract here, we get 20. Then 14 times 5 is... 4 times 5 is 20. That shows 14. And Ona looked at me saying, "And you're going to give the lecture? Are you sure?" I said: "yes"...

So, Ona was trying to be generous with me. So she says, "Adrian, 5 times 5 is 25." I said, "You're wrong. Look at this. 14 times 5 is... 5 times 4 is 20. 5 times 1 is 5. You add and you get 25." At this point, Ona was totally discouraged. She didn't know what to do. She was about to call the police. She wanted to give me one more chance and so she wrote five times the number 14. "Let me show you something. If you add 5 times 14, you'll see what happens." And I said, "Go ahead." So she started. 4 plus 4 is 8, plus 4 is 12, 16, 20... And I said, "Yes, 21, 22, 23, 24, 25. See? We got it!" So I realized there that we have a difference. Well, anyway, I assume that you understand that we do math anywhere in the world the same way. One of the things that is strange... but look at this. The alphabet... you have a difference right there. I cannot read Korean. But what we have in common are the numbers. Everybody understands numbers. They are a universal way of communicating. No matter what the language is, the numbers are always understood no matter where and what.

And then what I wanted to do is start showing the story of a couple of ironies. Things that happen in life that are very ironic to me. Look at what happens... I'm going to speak about

Irony Number 1. A kid is born and in the first 12 months of his life, or her life, we teach them two main things: how to speak and how to walk. That's the first 12 months. However, the following 12 years, we want them to be quiet and silent. That's crazy! We teach them to do something and then we don't want them to do it!

And then it comes Irony Number 2. In the first 4-5 years of their existence, the kids learn an incredible amount of things. They learn how to speak, they learn how to walk, they learn how to play, they learn how to relate to other people, they learn how to talk to their siblings, how to play with friends, they even learn how to play by themselves... And they learn everything on their own. All of a sudden, some day, very unfortunate for them, we tell them "now you have to go to school." Why? Why do I have to go to school if I learned everything I know without going anywhere? If they could speak, they would say, "I learned everything I know until today, I didn't have to go anywhere. And now, why do you want to take me to school? What is school?" You have to get up at 6 o'clock in the morning, you have to spend there hours and hours of your day, you have homework to do when you get back, you have... Why?

And I feel that at that time, in that particular moment, it would be a spectacular opportunity for our science, for mathematics, because what kids want to do is play. And mathematics has all the tools to show them that they can play, that they can still play, and do what magicians do. What do magicians do with the kids and with the adults? They mesmerize them. They puzzle them. They peak their interest. And kids are in awe. And mathematics has those types of tools too. But what I find is that we show them the wrong door. We don't show them the right way to get to mathematics. We are buried in a lot of technicalities and things that are absolutely boring... it's like saying, for instance, assume that a person has never made a phone call. And before they start making the call, you tell them, "Okay, but before you start making calls, you have to learn all the country codes, all the city codes, all the area codes, and you have to memorize the phone book. Once you know that, then you will make your first phone call."

No! It doesn't work that way. Or I should say, 'it shouldn't be that way'. If nothing else, what math should do, what mathematicians should do, is get more involved in saying, "Wait a second. What we're doing is wrong." Nobody gets into a restaurant through the kitchen. Nobody gets into a house through the bathroom. But naturally, you have to show the kids how beautiful our science is. You have to seduce them, to engage them. How do you do that? You have to show them. Let me tell you here something that I've been thinking and I've been repeating this over the years. This is a story that I hear through Pablo Amster who happens to be now the chair of the Department of Mathematics at the Faculty of Exact and Natural Sciences of the University of Buenos Aires. He told me a story of a group animated cartoons, a series of episodes that were written in the former Czechoslovakia. Listen to the story. There was a little town and there was a king who had a daughter. But the daughter was getting old, was aging, and nobody wanted to marry her. And he was concerned that the daughter could find nobody to marry. So finally the king goes and says to one of his assistants, "Make everybody know that my daughter is going to be waiting for all the pretenders that are going to be in a long line and she will be sitting on a chair and she will see what they can do, their abilities. Show her what they can do. When she finds one that she likes, she's going to give her hand. I'm going to give her hand to whoever is the best of all the candidates." So every episode that was like a little movie, a short movie, a five-minute movie, was showing what each candidate would do. So the first episode, you see a contortionist. There's a long line. The first one... The princess is sitting there and this contortionist moves his body, does a lot

of strange things. The princess shows no emotion, nothing. End of the first episode. Second episode. There's a very wealthy person. He comes with a huge bag with a lot of coins, gold coins. And he displays all the coins on the floor and looks at her like saying, "Here, this could be all for you." Nothing. The princess is unfazed. Third episode. There's a magician and this magician has rabbits and doves, and birds, and colorful scarves. He shows his tricks with cards, he pulls rabbits out of a hat. Beautiful things, but still, nothing. Then there is an acrobat. And the acrobat starts juggling balls, one, two, three, four, five, ten. Nothing. At that point, one starts wondering, "What does she want? What does this princess want?" Nothing seems to touch her. And the line was getting shorter and shorter and shorter until, in the last episode, when there is only one remaining candidate. This candidate was a very short guy. He was having a backpack. When it's his turn, he goes where the princess is, opens his backpack, pulls out a pair of eyeglasses and gives it to the princess. She puts them on and she smiles. And he marries her. So what was the problem? The problem is not that she didn't appreciate what they were doing. She couldn't see!!!!. She couldn't see anything. So how could she react like mesmerized or something if she didn't see anything? Well, that's what we do with math. We have to show people what math is. What we're doing is something wrong. That is not mathematics... It's not that what we show is not part of mathematics. It is. But we shouldn't begin there because we're going to fail. And that is what has happened up til this point. Now, what do we do for that? How do we attack that problem? We already know one thing: people hate mathematics... I mean, the general population. We know this as a fact. But.. why? We know also that there are two groups. The group that consists of the vast majority, who hate math. And on the other side, there seems to be another group, a privileged group, a group of people who really like it. Now, the ones who enjoy math also like to feel... well, you know, because they look at us like, "Those are the nerds. Those are the intelligent ones. Those are the people that are different, those who get it." But nobody says, with more... being more proud, like a badge of honor, "Math is not for me, nothing to do with it." And why? Because in schools and in general, we give answers to the questions that the kids didn't ask. Kids are not dumb. People here, do you know how to drive? I mean, do you know how to drive a car? I assume that most of you know how to drive a car, but at some point, you had to be taught how to do it. Well, when we're sitting in the car for the first, second, third time, and we have someone teaching us, that person, at some point loses his or her patience and starts screaming and abusing us... But why do we tolerate that? We do it because in the end, we understand that we're better off knowing how to drive than not knowing. Although we don't like it, we're willing to pay the prize.

With mathematics, we don't see that. We are told stories about things that we don't know. We're given answers to questions we didn't ask. Not only that, then we have to go home and have to do our homework. Then we get desperate. And we go to our folks, our fathers and mothers say, "Why do I have to be doing this? Why do I have to study this?" And the father and mother, they don't know what to say because they didn't know when it was their time. So what do they do? They say, "You know what? You're going to learn this later. Just wait and you will see." And when does later arrive? Because, usually, I don't know if you have seen it, but there are a lot of people who've invested a lot of time trying to learn something and they have never used it. In life, first, we have problems. Then we look for solutions. In schools, especially in mathematics, not to say in general, I don't know, but in mathematics, we operate the other way around. First we give them solutions, like a theory. And then say, "Well this is where you apply it" or "In what situations do you apply it?" What? What? And we see kids, what they want to do is they would like to play with a video game or with some robots

or encoding a message. Something that has to do with their daily concerns... Things that happen in their every day life... Mathematics has a branch called game theory. We ignore it. I learned about game theory when I was at the university. So that means I was already more than an adolescent. That's wrong. That's wrong. And what happens in school leaves a huge imprint in our lives. And I want to bring two stories to your attention.

The first story ... and it is important if you could walk along this path with me. I used to teach at the University of Buenos Aires at the Faculty of Exact and Natural Sciences. Follow me with this calculation. The classes were attended by 800 students in the average per semester. In a year then, it means 1,600 students. Then, if you think that I've been teaching for over 30 years... say 30. If you multiply 30 by 1,600, you get to 48,000 students. Let's assume that I exaggerated. So, instead of 50,000, let's divide it by two, and we get to 25,000 students.

I think I can say that they are a very large sample: We can conclude that I've seen it all. I've seen people who came to school thinking that they were the stars and then they couldn't progress. And there were others that were just very shy and were not going to last very long, but instead, they ended up being brilliant.

I don't know if you ever thought of this, but society is always looking for the winners. And that is associated with mathematics. Whoever gets it is like someone who's different. But why? Society is always looking to give an award to the guy who arrives first, to the ones who jump higher or the ones who run faster. Ok then. I have a question then: What happens with the second? Or the fifth? Or the twenty-third? This means that we're talking about probably the whole population with a few exceptions. So, it looks that we don't legislate for them, we'll leave them behind. In school, when we introduce math, we should do it through the right doors, and in that case there are certain words that shouldn't be included.

Don't use the word 'no'. Always –as a professor or teacher- always say 'yes'. There's no room for failing. Kids cannot fail. There's no room for the word failing or the concept of failing. "What do you mean failing? I am trying. Encourage me. Help me. Coach me. Show me. Show me where I have my questions".

Mathematicians should help to show that math is about finding patterns, with structures, with puzzles. That's what we should show. Show that there are there and they need to be found. It's like a sea of information, and we have the tools that allow us to sprinkle that sea, and some of the things that we don't see, because they look foggy, they're actually there. They are right there. All of a sudden, in front of our eyes, they emerge and shock us. They were there before, but we couldn't see them. They surface like something different, something new.

But we have to play. We have to allow the kids to play with them. We have to show them that it's good to try, and try, and try even more... until the time that everything will be clearly in front of us.

And I know that this system works. So I'm going to play a three-minute video here to show you what we do in Argentina with some of the kids. Can we play the video for a second? Well, for three minutes, not a second. (Video)

I know it works. I've seen it. I do it on a regular basis. Kids get engaged. And I'm not a very special person. But what we need to do is just show them where the right door is. And I want to just narrate a couple of stories here.

One of the stories is dealing with one of the kids that you just saw on the video. We had arrived with my coworkers on a bus. We parked the bus and I get into the school. You must have seen the kids. Some of them are/were 16 or 17 years old, but some of them are much

younger, say 6, 7, 8 years old.

All of a sudden, I get into one elementary school, I opened the door and I get... as you saw... hundred of kids that come running to me. They all come and converged around me. And they start pulling my sleeves. And one of them asks me: "Listen. How much is a thousand times a thousand?" And the other one from this side says, "Is there a number greater than infinity?" And the third one just looks at me and she says, "Do you ever make a mistake?"

And I will never forget that. This is not a kid whose parents told her to come talk to me because I work on TV. She wasn't awe-struck by, star-struck by someone that she sees on TV. She really cared. She wanted to know if I ever make a mistake because there is this aura that we have.

And that's the message that I want to send. We have to start with "I don't know." Or even "I'm like you, I don't know all the answers, I don't have all the answers". Then, you know what they asked me? I went with them to a class. And they were learning some of the multiplication tables. That makes me suffer. And one of them says, "And do you know the multiplication table of 15?" I said, "No, I don't. Let's figure it out together." And it was so fun, so great when they realized they had the tools to actually go home and tell their parents that they learned the table of 15.

You know that knowledge is power. And when they felt that they had... that they understood where the concept was, it was great for them. They wanted to take pictures with me, but it worked both ways: I wanted to take pictures with them also. I was a better person after I talked with those kids.

Second story. As I told you, I used to teach 800 students at a time, average. Sometimes more. Obviously, it didn't escape them that I worked on TV too. I was their professor at the university, but I was also a sportscaster. I have been working on TV since february of 1972. Soccer, basketball, the NBA..... And by the way, does everybody here know how many balls you can fit in a basket? Have you ever thought of that? Probably if I asked that question, if I conducted a survey here and if I asked you all: "How many balls, basketballs, can fit simultaneously in a basket?". I'm not sure that you know the answer to that question. It really doesn't matter but it's kind of a surprise to learn that you can fit two basketballs simultaneously. Not that we need to know that in order to enjoy a basketball game... But it was very interesting to discover that with them. But I digress. Let me go back to my story. This is what happened. First day of calculus for these kids. I gave my lecture and at the end, remember, it's their first day at college, the first day in that type of class, and picture everybody between say 18 and 20 years old... and at the end, some of them came to the podium, they surround me, they want to talk with me. It's all good. They also know that their professor is a 'celebrity' (so to speak), because they've seen me on TV. Please, I don't want to sound arrogant, it's just a fact. So, when I have 15 or 20 students around me, I started to ask them what career are they trying to follow. I mean, that Calculus class has to be taken by everybody who studies there, no matter if you want to become a mathematician, a physicist, a chemist, a biologist, a geologist, etc.. you get the idea. So, my question to them was: 'what are you going to study?'

One of the students answers the question saying that he was going to study mathematics and computer science. I was surprised because it's not very common that someone is going to follow two careers, let alone one. So, I kind of stopped and looked at him a little bit longer, maybe because I was puzzled by his answer. But I kept going until I asked them all. Then, I had another question to all of them. "What made you come here? What made you start this career?". I guess I want to know if they could point their fingers to some episode in their

lives that made them think: “I want to study math!” Or, “I want to become a chemist”, or something like that.

When I got to the same kid who was going to follow two careers, when it was his turn to answer the second question he goes: “Well, because when I was in grammar school, I saw on TV a person proving that you cannot divide by 0.”

Instead of going to the next kid I stopped, looked at him and I asked him: “What did you say?” He says... He was kind of scared because he didn’t know what did he do that was wrong... he was puzzled. Maybe he said something wrong and he didn’t know. This conversation was taking place in 1996. Seven years prior to that, in 1989, I was hosting a column in the evening news, the most important segment of the day, in prime time. On one of the shows, crazy as I am, I proved that you cannot divide by zero.

So when he answered my question, I told him: “What is your name?” He said, “Christian.” “Christian”, I added, “I need you to come to my house”. Naturally, he was surprised. He didn’t want to come to my house. He didn’t have a car, he had other things to do. So, I said: “Listen. I’m going to take you with my car. I want to show you something and then I’ll bring you back here.” In any event, I didn’t live that far away from school anyway.

But what he didn’t know is that since we have the opportunity to record at home using VCRs and VHS, I have every show where I had worked and/or host in my house. I have them sorted and filed. To make a long story short, he came with me. I went where that particular VHS was, I pulled it out and I played it for him. He wasn’t sure, but let me ask you: “How many people in the prime time news are going to prove that you cannot divide by 0? It HAD to be me.

Now look at it this way. Look at the impact that we have on kids. Look at the impact that we had with that little kid that was asking me if I had ever made a mistake. We have to be very careful with what we do, very, very careful. Because those types of fingerprints, imprints, name it the way you want, are going to stay for a long time. And that is the perception that is very difficult to fight against. That’s why when people say, “I hate math”, it seems to be like it’s a battle that we already lost. And I refuse to give up.

There are a lot of mathematicians here. When we see that something is going wrong, we have to say something. We cannot stay silent like if nothing is happening. Let me bring to your attention something that I saw in New York City. They have like a campaign against terrorism. And so they say if you see something, say something like if you see a package that is suspicious or whatever, say something. So I am going to ‘use’ that sentence and extrapolate here: “If you see something that it’s wrong in math, say it”.

We need to be more involved because otherwise, we keep diagnosing something that we already know. We have failed up til today. In the past, schools didn’t have competitors. When I was growing up, the main source of information was at home and at school. But today, school is just another source of information. The competition is vast: internet, social networks, smart phones, google, etc, etc. School is just one of them, It’s very very important, especially because in school we learn other things. We learn that we are not going to be the king of the house or the queen of the house any more. We have other friends, other classmates. We have to learn how to get frustrated. It’s not always our turn. We are not always the ‘first’. And plus, we learn about structures, discipline. We learn that we’re not alone, we’re not the ‘only ones’ or the ‘preferred ones’. Please, don’t read into what I’m saying that I’m against or opposed to school! I’m not!

But we have to adjust. In the old days, people were mesmerized when there was someone who seemed to have all the knowledge that was out there. Not any more. That’d be impossible.

What we have to do is stimulate the creativity, not accumulating knowledge. These days, if you want to know something, you just google it. You don't even have to be in your house. You can find the answer on your phone, your watch, your laptop, your tablet. "You don't know something, you go find it almost immediately". That's why we have to stimulate creativity. We have to forget the word 'no', or the phrase 'you're wrong!'. There's no room for red marks, or zeros, or an F...There's no room to say that someone has failed. "Failed?". "Failed in what?" "Who are YOU to tell me that I failed?"

And there was one more story that I want to narrate here. And I would like for you to follow this story as closely as you can. This is just a personal favor here, because this was one of the most important moments in my life and I didn't know it when I was going through it. I didn't understand what was happening until...well, you will see as I go along. I learn something new about the story each and every time I narrate it. So follow me with what happened and that's why I ask Ingrid to give me... to let me use her tablet.

Claudio Martinez, who is here sitting in the front row, is the producer of all my TV shows. We work together but he's also my friend. And by the way, I have some friends here that came from all around the world just to be here this moment because it's very important for all of us, not only me. A lot of people here came from the US, from Spain, from Argentina... And they came for this particular moment because we know that we have a task. Something to accomplish, we have to change the perception that people have of math. I'm 65. I won't live forever and I know that there are other people who are going to do a much better job. There are young people ready to do much more than what I/we did. But we know it's boiling, something is happening. Let me go back to the story. Look what happened. Claudio... I told Claudio that I was going to do this thing on the show, on "Altered by Pi". I was going to draw a pizza. And I was going to show a different way to cut a pizza to be shared between two people. "How do you cut a pizza? Say that there are two people to eat." Usually you cut it this way, this way, this way... Do you agree? Don't look at me like as if you have never seen a pizza! This is a pizza, and this is the way that we usually cut a pizza, right?

So I told Claudio: "this is what I want to do on the show. I want to cut it in a different way". "You start here, like in the other case, but you can do this." The second cut, you can make it here instead of here. But you can do this. Can you see this here? So this portion is equal to this portion. This one equals this other one. This one equals this. You get the idea, don't you? So the portions are not the same now but each one still eats the same amount of pizza. In the other way of cutting it, every portion, all the portions are the same size. This is not true anymore. But you can divide the pizza in an even way. So when I suggested this to Claudio, Claudio says, "Why don't we do something?" Let's go to the pizzeria across the street, and we can bring two pizzas and you cut them in front of the camera?

I said, "No, Claudio." "Don't bring the pizzas because I know what's going to happen." I don't cut pizzas on a regular basis. I have only this pair of pants and shirt. I'm going to start cutting the pizza and everything is going to fall on me. My clothes will be ruined and we'll have to stop recording. "I have a better idea," he says. "Why don't we tell Jose," (Jose is the guy who cuts the pizza at the pizzeria and we know him very well because there's where we eat pretty much after every recording). He said: "Why don't we bring Jose here and let him do the cutting?" Again, to make the long story short, Claudio fixes everything and half hour later, Jose shows up with two pizzas, dressed in white like he usually does, and he was even wearing a hat, like a chef. He already had make-up on his face.

He was so happy but, all of a sudden he gets into our studio, and he sees the camera men, people screaming, the lights, the microphones, the cables, the producers, the directors, other

people who run with papers. He froze. I brought him close to me and told him: “Jose, don’t panic! It will be very easy for you. Ignore what surrounds you. Just talk to me. Forget about the cameras, the lights, the microphones, just talk to me. What I need from you is to cut two pizzas. The first one, cut it the same way that you usually do. As far as the second one goes, I will guide you. Don’t worry, this is a recording, so, if we make a mistake we can start it all over again, ok?”. He answered, ‘OK’.

At this point, we start recording this segment of the show. Jose cut the first pizza with no problems (as expected). I didn’t even have time to explain to the audience what I was trying to do. Then, Jose started to cut the second one. He knew what he had to do at the beginning, but he also knew that he had to wait after that. I said, “Jose, now, as the second cut goes, do it above what you usually do, closer to the border”. Jose wanted to do the cut, but although he was cutting it above the middle line, he was doing it in such a way that was hard for anybody to see. I said: “No, Jose, make the cut higher!”. He still didn’t believe me. He wanted to help me. So, he went a little higher, but not AS HIGH AS I WANTED. So, finally, as time was passing by, I tried to push his elbow a little higher to show him what I wanted him to do. At this point, Jose, kind of puzzled with what I was doing... he.. gave up. And he was going to make the second cut diagonally. So, I said: “No, Jose. Perpendicular.”

Let me stop here for a second: when I said perpendicular, he froze. I didn’t understand what happened but he wasn’t moving. So, I said again: “Perpendicular”. As he wasn’t doing anything, I realized that the problem was that... ‘he couldn’t understand me! He didn’t know what the meaning of the word ‘perpendicular’ was. So, as I wanted to keep the recording going, I said: “Jose, 90 degrees!”. Jose wasn’t moving. So, finally, I kind of screamed: “Make a cross.” When I said, “Make a cross”, he understood. It really doesn’t matter what happened after that. Everything went perfect.

Two morals of this story, or two conclusions if you want. As you may imagine, Jose wanted to help me. He really really wanted to do what I asked him to do. However, the problem was that ... my message wasn’t arriving. I mean, I was trying, he was trying, but there was ‘no connection’. He couldn’t understand what I was saying. And how many times in life, not only schools, but in general... With your friends, with your family, with your wife, with your brother, your sister, with your kids... You ask for a favor, the other person wants to help you but they don’t understand what you’re asking them. And the problem gets even worse, because the other person feels embarrassed to say: ‘What? What did you say? What do you want me to do? I just don’t understand what you’re saying!!!’.

And that’s a huge problem that we have. The other person feels that if he (or she) shows that he doesn’t understand, then, he feels that he’s less of a person, a worse person, an ignorant, or something equivalent. And that’s so bad. It’s a shame that we cannot address that problem.

There is like an abuse of power, the power of knowledge. When we know something, it looks like if we abused that power. It’s like we want to make it clear: “I know, you don’t”.

And then I want to draw the second conclusion. When you heard me say, “Perpendicular” and I told you Jose froze, because I realized he didn’t understand the meaning of the word ‘perpendicular’, what happened? You all smiled. You all laughed, as it happened to me the first time that I thought of that story. But, let’s think again about the same situation. Think what that smile/laugh means to the other person. Jose isn’t stupid. People are not stupid. When you smile, although you don’t say it, our body language is what Jose read. He realizes that I knew something that he didn’t know. We have the ‘power’ of knowing what the meaning of the word perpendicular is, and he doesn’t! We have a ‘power’ over them. We know what perpendicular means, and what 90 degrees means, and he doesn’t. They don’t. It’s like if we

made clear to them: 'we're here and you're there' (pointing higher and lower). And that is a huge problem, a huge problem in schools and in life if you want. The other person perceives what you're not saying verbally but your body language betrays you. They know. They smell.

So, the person who has the knowledge has to share that knowledge. This is the only way. We have heard many many times that there's a very unfair distribution of wealth in the world, but there's another type of wealth which is very unfairly distributed: the intellectual wealth. The knowledge!

We need to share the knowledge. If you know something, share it. If the other person doesn't know it don't laugh at him or her. Help her, help him. You're helping yourself and more than that. We need to change the way we communicate that knowledge, make it more accessible and fun. We have to change the way we teach.

I understand that a lot of people are accusing or blaming the teachers, the professors by the way we teach and communicate mathematics. And they're probably right with that argument. But we also have to think that most of the teachers who are teaching today were born in the analogic era, and now, we live in the digital era. Who's going to stop time so that we all make the adjustment? While we train the professors/teachers, we should stop teaching, shouldn't we? You go to a teacher who taught for 30 years in a certain way and now you tell him, "You know what? That's not going to work anymore. Now you need to bring computers, laptops, tablets, notebooks, phones, iPads..." And they're like, "What? What? I don't know how to use them." They're scared. They freeze also. So? What do we do?

We cannot say: "Well, wait for five years or so, until we teach and train and coach our group of teachers and professors and help them adjust to the new times". No, you know that we cannot do that. But how do we solve the problem? Well, with what I call horizontal education. Instead of vertical, where the teacher's here and the students are here... Again, the system of being "I am the one who knows and you are the ones who don't"... Just do it horizontally. We learn together. We have to swallow the pride, like a parent learning with the kids, with their kids. And I've seen that. So we go... and even if you have to alter the chronological order. If there's someone younger who knows something that the older ones don't, he should share what she or he knows. Just come here and show us. That's what we need.

I'm about to finish. I wrote my thesis in 1978, 1979. My, our mentor was Miguel Herrera. There are people in this audience who were also his students. He passed away when he was very young. 44, 45 years old. It really doesn't matter now. He was an expert in Several Complex Variables. Miguel had written a book that was a great summary of everything that was known at the time, and we were reading and studying his book. Every day, he'd come to our office (Nestor Bucari's and my office) to check what had been the progress of the previous day. The story that I want to share is that one day, while Nestor and I were working, we thought that we had found something really new and incredible, but we weren't sure. If what we thought was right, then we had discovered something really very important. Maybe we had made a mistake and we didn't know. We needed Herrera to help us, just by telling us what did he mean with that paragraph that he had written in his book. So, we were anxiously waiting for him to show up that morning. That day, as every day, he'd knock at our door at 8 o'clock in the morning. We thought: "There has to be something wrong here."

When he came, he opened the door and we said, "Miguel, we have found something very important." He says, "Okay, what is it?" So I showed him and he says, "No, that's wrong." I said, "Well, let's see." I said, "What did you write in your book?" So he says, "This is what I meant with what I wrote." No, that's not what you meant. Look here. That's not a fair

conclusion". He says, "No, I see. Let me think about it again". And he went over and over what he had written until a point where he sighed, he sat down on the couch that was right there and said: "You know what guys?" "I don't know what I wrote. I don't understand..."

And those were the key words. When he said: "I don't understand what I wrote", those words stuck with me for ever. He, of all people, the greatest expert in the world in that subject, he... he didn't understand? Not only that: he was saying that he didn't understand what he had written (and we knew that he was right, that the book was right). And what we learned that morning, is that if someone how is that high in everybody's consideration, so highly respected and such a great mathematician, he could say in front of his students, "I don't know."

And what problem do we have in saying, "I don't know?" How many times you see in society, in everyday life, in general, that people are afraid and scared of saying, "I don't understand what you're saying." So, the conclusion is: "Say it! Say it again and again! Don't be embarrassed. It doesn't matter. You're not worse, you're NOT a worse person if you don't understand something. Say it. Let me see, maybe I thought that I understood it and I didn't."

And to finish this, there are a couple of more things I want to say. The education has to be public and free for everybody. There is a huge gap between those who have, like I do, and those who don't. I have pretty much everything I need, and will probably ever need. And there are a lot of people who don't have. There's a problem with that. In order to narrow the gap, we have to spread the knowledge and make education available to everybody. And, in order to do that, we need education to be public and free. Every government has to take charge and is responsible for people's education. Even more: education cannot depend on how much money you (or your parents) have in their bank account. Education has to be free!

I know that I sound like a politician, but so be it. Whomever is ruling a country has to take over and understand that education is a human right. And we all have to help, it's our responsibility.

And especially for us, mathematicians, we need to take a different approach and be more engaged. It's not enough to say that the way we've been doing things so far are bad. We already know that. But we need to be more 'hands on' and participate more. I know that things are going to change. I don't know if I'll see it, but mathematicians and mathematics have the tools to make it change. And if I don't see the changes, at least we'll know that we planted the seeds. And that's all we can ask and care for. Thank you very much.

E-mail: [paenza@gmail.com](mailto:paenza@gmail.com)

## **Panel Discussions**



# How should we teach mathematics better?

Deborah Ball, Bill Barton\*, Jean-Marie Laborde, and Man Keung Siu

## 1. Introduction

Bill Barton introduced the panel, and invited the audience to become participants in an extended discussion after short presentations by each panelist. He asked that the meeting focus on undergraduate mathematics teaching, and to bear in mind the principle of reflexivity: that anything we say about teaching we must be prepared to do ourselves.

## 2. Presentation: Bill Barton

Bill made two points. The first concerned what he calls the “14-year apprenticeship”. That is, that from age 6 to 20 we tell students results and known ways to do mathematics, and get them to practice these things. Only at Masters level do they actually DO mathematics in any authentic sense. He invited the audience to compare this with preparing football players or musicians, and asked where we foster the creativity and imagination that we know is essential to research mathematics?

His second point was about professional development. He noted that it is a hallmark of being a professional—even for those who are the world’s best in their field—that they never stop putting time and energy into getting better. But we are professional researchers, and professional teachers, and he asked how many of us spend half as much time per week improving their teaching as Inbee Park does improving her golf ? (At the time of the Panel Korea’s World No. 3 golfer was doing well in a tournament). Do we spend that amount of time each year?

His suggestions for better teaching, therefore, concern bringing authentic mathematical activity into our undergraduate programmes, and each of us investing time in improving our teaching.

## 3. Presentation: Man Keung Siu

Man Keung began by reframing the title of the panel from “How should mathematics be taught better” to “How can mathematics be taught better” with the reasoning that there are no fixed rules for best teaching. However, we do have some general rules, for example, that a

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

\*Moderator

teacher needs both a brain and a heart, that we must ensure that teachers like both the subject mathematics, and their students. He then emphasized three points.

Less is more: there are not that many concepts that need to be taught in primary and secondary school, and they come up repeatedly. He suggested designing a curriculum around these. He gave an example, the basic idea of which starts in primary arithmetic, and goes via generalisations and beginning algebra, to the rank of linear systems, in particular Toeplitz ones.

The second point was that mathematics is part of culture, not just a tool. As such, the history of its development and its many relationships to other human endeavours should be part of the subject. Of course the history of mathematics is not to be regarded as a panacea to all pedagogical issues in mathematics education, but rather the harmony of mathematics with other intellectual and cultural pursuits that makes it even more worth studying. Furthermore, one should examine a topic from three perspectives: a historical perspective, a mathematical perspective, and a pedagogical perspective. Although the three are related, they are not the same; what happened in history may not be the most suitable way to go about teaching it, and what is best from a mathematical standpoint may not be so in the classroom and is almost always not the same as what happened in history. However, the three perspectives complement and supplement each other.

His final point Man Keung referred to as ‘Mathematics and the Mouse’. In view of the changing learning habits of the younger generation due in part to the widespread use of various technologies, we should consider our old principles and ask some questions:

- (1) How should IT be employed to enable students learn better but not to limit their ability to think critically and in depth?
- (2) How can we ensure that a discovery approach is not to be equated with a hit-and-miss tactic?
- (3) How can we ensure that imaginative thinking is not to be equated with a cavalier attitude, that multi-tasking needs not be identified as sloppy and hasty work, and that the use of IT is not to be identified as following instructions step by step without thinking?

To end, he quoted a passage in the ancient Chinese Book of Rites (禮記·學記) which dated back to more than two thousand years ago (English translation taken from [Legge, 1885]): “Hence it is said, ‘Teaching and learning help each other’ and ‘Teaching is the half of learning.’”

#### **4. Presentation: Jean-Marie Laborde**

Jean-Marie focused on issues connected to the use of technology in the teaching and learning of mathematics today and into the future: What could we wish for, and what can be done to seriously enrich the quality of mathematics education taking advantage of technology? His main points were:

- Technology is not new; nevertheless what we call technology today has a potential that never existed before.
- People are still confused about the effectiveness of technology despite clear evidence of its potential when adequately implemented.

- There is an alarming poverty in the majority of so-called educational mathematics resources, especially on the web. Which raises the question of how to enhance the critical awareness of teachers and policy makers?
- The need for a radical change at political and economic levels to secure a sustainable educational software industry in the same way that textbooks have been sustained for a very long time by an industry coexisting with public institutions.
- The need to start reducing the discrepancy between the kind of mathematical knowledge students can achieve in a computer based environment (which will be their future working environment) and the still very conservative style of teaching, almost everywhere.

## **5. Points from Discussion**

### **Opinions offered**

- Attempts to incorporate literature and history in mathematics courses can also be refreshing.
- In general, employers seem to want problem-solving skills above nearly anything else.
- Different countries are using different technologies: blackboards, PowerPoint etc.
- Different countries or institutions also have different access to technologies.
- There are no universal answers to whether the use of any technology is good or bad: appropriate use is what matters.
- If you use computers to teach, you should use computers to evaluate. This causes logistic challenges, but is very motivating.

### **Experiences reported**

- Two people reported being “marked down” for using teaching technologies in their university classes.
- One noted that “flipped classes” are working at a pre-calculus level, but a lot of technology was needed to do it.
- One reported using local examples to provide effective teaching in the absence of technology.
- One reported using Maple in a calculus class, and finding that it helped with letting students do a range of examples.

### **Unanswered questions posed during discussion**

- Should we teach the geometry-inspired students differently than the algebra-inspired ones? Comments included noting that in the time of Descartes, “geometer” meant “mathematician”, and that mathematics education theory talks about “versatility”—the idea that students should learn to approach problems in a variety of ways by a variety of means.
- How should we balance “understanding” with “getting through the curriculum”.

- Should we change our teaching behavior because of what students prefer?

In summary, Jean-Marie noted that Blackboard or whiteboards are merely variants on the theme of presentation, but that active software is very different. Many systems are designed by and for mathematicians, yet learners of mathematics have different needs. There is also a problem of cost and resources. Politicians are not aware of these issues, and mathematicians are not aware of the wider uses of mathematics. As a final comment, Bill asked the audience to consider using the ideas expressed (and other new thinking) to teach in new ways as well as just trying to teach better in old ways.

Deborah Ball, University of Michigan, USA

E-mail: [dball@umich.edu](mailto:dball@umich.edu)

Bill Barton, University of Auckland, New Zealand

E-mail: [b.barton@auckland.ac.nz](mailto:b.barton@auckland.ac.nz)

Jean-Marie Laborde, Université Joseph Fourier, France

E-mail: [jean-marie.laborde@cabri.com](mailto:jean-marie.laborde@cabri.com)

Man Keung Siu, University of Hong Kong, Hong Kong

E-mail: [mathsiu@hku.hk](mailto:mathsiu@hku.hk)

# Mathematical Massive Open Online Courses (MOOCs): Report of a Panel Discussion

James H. Davenport

**Abstract.** The author moderated a Panel consisting of Bill Barton, Robert Ghrist, Matti Pauna and Ángel Ruiz at the 2014 International Congress of Mathematicians. This paper contains the initial panel brief, the author’s summary of the Panel statements, the question-and-answer session, and some conclusions.

## 1. Panel Brief

This panel (The live panel can be seen at <https://www.youtube.com/watch?v=bRjkbmuCm20>.) had been arranged by the Committee on Electronic Information and Communication (CEIC) of the International Mathematical Union (IMU). The title was carefully chosen, not “MOOCs in general”. Though the panel could have had a long, and interesting to some, debate on general questions about MOOCs, it was part of the the ICM and therefore the focus was on Mathematics. What might make Mathematics a special subject for MOOCs? CEIC’s initial thoughts, circulated to the audience, were as follows.

1. The highly sequential nature of mathematics: it is little use trying Analysis II until one has done Analysis I, and so on. In a given university, Directors of Studies (or their equivalent) carefully plan an *ordered* curriculum, and write a list of pre-requisites etc. An *individual* MOOC provider might do the same (though there is little evidence of this so far), but there is currently no evidence of a general catalogue/list of prerequisites.



This is not helped by the language of mathematics: “Elementary Proofs of the Prime Number Theorem” is unlikely to be “elementary” in the usual sense (OED 7a: “Rudimentary, introductory”) of the word.

2. The notation. Many Virtual Learning Environments (VLEs) do not support the display of mathematical notation well, and certainly not the construction of mathematical notation “on the fly”. Equally, entering mathematics is often difficult for students. Notation is also not as universal as is commonly believed, and a student who learned arithmetic the Spanish way, even though fluent in English in general, may well be baffled by Anglo-Saxon long division of polynomials (or *vice versa*).

**However** if this barrier can be crossed, MOOCs way well be a means of getting advanced mathematics to those who would not otherwise have the opportunity (as closed online courses already do).

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

3. The fact that mathematics is a practical subject: one learns by doing. It is not clear how MOOCs can support the sort of routine exercise that is a vital part of learning mathematics. Multiple Choice Questions do not fit the bill here.

There are solutions to some of these problems, but they are not widely known, and not in the “mainstream” VLE/MOOC systems.

## 2. Panelists

**JHD** *James Davenport, University of Bath, U.K. (Moderator)*

He is a CEIC member, and knows something about MOOCs as the University of Bath has been involved in MOOCs as part of the U.K.’s FutureLearn consortium, which is led by the U.K.’s Open University, which many people would argue has been involved in MOOCs since it opened its doors in 1971. The Open University had advised FutureLearn against early involvement in Mathematics, since OU experience had been that Mathematics was one of the more difficult subjects to do well. The second wave of FutureLearn MOOCs will include a statistics MOOC.

JHD uses quite a bit of “MOOC technology” in the large “face-to-face” courses he currently teaches. His initial opinion is that there is not a simple either/or between traditional teaching and MOOCs.

**BB** *Bill Barton, University of Auckland, New Zealand*

He works in the Mathematics Department at the University of Auckland, but is a mathematics educator. He does teach undergraduate mathematics to large classes. Given his rôle in the International Commission on Mathematical Instruction (ICMI), he wrote the MOOC report for CEIC in 2012/13 — a long time ago in this subject. Auckland is also a member of FutureLearn, and is developing a Statistics MOOC.

**MP** *Matti Pauna, University of Helsinki, Finland*

He has been involved in many projects regarding online teaching and assessment of mathematics.

**AR** *Ángel Ruiz, Universidad de Costa Rica, Costa Rica*

He is a Vice President of the International Commission on Mathematical Instruction (ICMI). He is the Director General of the Mathematics Education Reform in Costa Rica that elaborates pedagogical resources and carries on several courses: face-to-face, blended and a collection of 20 MOOCs for 2014-2015. The MOOCs are oriented to in-service teachers preparation to deploy the new curriculum.

**RG** *Robert Ghrist, University of Pennsylvania, USA*

He wrote and produced a calculus course in the Coursera framework, available at <https://class.coursera.org/calcsing-005/lecture/preview>, and there’s a trailer at <http://www.youtube.com/watch?v=BKpBbzYYXrk>.

## 3. Panelists Statements

**3.1. Bill Barton.** I have two main points, and various sub-points.

1. MOOCs, in general, and Mathematical MOOCs in particular, are still in a very early stage of development, and they are only one part of the impact of transformative technology on mathematics education. It will change things, but we do not yet know how. We mathematicians must stay involved and treat it seriously and rationally use these technologies and research their impact.
2. I personally do not see MOOCs *replacing* face-to-face education of various kinds (although it will change it). My understanding of the educational process, and particularly the mathematical educational process, is that working together **both** at a distance **and** face-to-face is required. Indeed, one of my optimisms about MOOCs is that they will enable us to *enhance* face-to-face teaching and learning. In particular, it may allow us more focus on mentoring authentic mathematical activity at undergraduate level.
  - We need to maintain a research stance on MOOCs. That is, we need to keep an open mind on whether MOOCs, or future manifestations of MOOCs, or MOOC technology, are positive or negative for mathematics education in each of its many contexts — and we must make evidence-based decisions on these matters.
  - It is essential, as with any change, that we have people who *dream* what can be; we have people who are pushing for things that cannot be done now; and we have realists who worry about issues of cost, impact on lecturers and students.
  - We must be partially guided by the students' world — else we risk becoming passed by, to the detriment of mathematics. When I was a student I remember mathematics teachers resisting the introduction of scientific calculators because we would not then experience logarithms properly through using slide-rules.
  - My prediction is that MOOC technology is likely to become integrated more and more into conventional programmes, both in supportive ways, substantive ways, and as main components – to the benefit of everyone. (Of course stand alone MOOCs are likely to exist).
  - Our fear of MOOCs changing mathematics as we think it should be is, for me, simply an argument for staying strongly involved. Or we will be passed by: think of the rise of testing, how that has been cornered by governments and private testing institutions because we did not hang on to the rational assessment of our own subject.
  - Assessment: the first time a pilot flies an actual passenger jet it is full of passengers. And he or she has certainly been competently assessed by the flight simulator as competent. Of course, the (human) system will still have taken great precautions against personation etc., so our legitimate worry is not about electronic assessment *per se*, but against the human problems that might go with it.

**3.2. Matti Pauna.** His complete slides are at <http://blog.wias-berlin.de/imu-icm-panel-moocs/files/2014/08/ICM-Pauna.pdf>.

I have been involved in many projects regarding online teaching and assessment of mathematics over the last ten years. This has formed my belief that you cannot learn mathematics without doing it and you really need feedback as well (from peers, computers or instructors, and the last becomes difficult in a MOOC context). We have a substantial emphasis on assessment, but this is assessment *for* learning, not just assessment *of* learning.

Our main educational vehicle is World Education Portals (see <http://myweeps.com>), which provides free education material. There is a mixture of slides and animations. Students

Using the method of integration by substitution, find the following integral: Run the question tests...

$$\int \frac{\sin(4x)}{\cos(4x)+3} dx.$$

Your last answer was interpreted as follows:  $-\log(\cos(4x)+3)/4$   
 This answer is invalid.  
 You seem to be missing \* characters. Perhaps you meant to type  $-\log(\cos(4*x)+3)/4$ .

---

Using the method of integration by substitution, find the following integral:

$$\int \frac{\sin(4x)}{\cos(4x)+3} dx.$$

Your last answer was interpreted as follows:  

$$\frac{-\ln(\cos(4.x)+3)}{4}$$

STACK assists with the correct input of formulas

Figure 1. Automatic Assessment (STACK) Example

get instant (computer generated) feedback online, and can monitor their own progress. This means that students do need to learn syntax of the system but that usually happens within a week. Our vehicle for this is the Automatic Assessment tool STACK created by Chris Sangwin (now at Loughborough): see [2], and Figure 1 for an example, which also shows how the system responds to syntax errors without penalising the student. This provides:

- diagnostic tests at the start of course so instructors know where they stand;
- continuous learning by practising and getting constructive feedback.

We also make use of peer assessment, where the process for a workshop module in Moodle works as follows.

1. Student are given homeworks that are to be submitted by Wednesday: an example is in Figure 2.
2. After the submission deadline, an example solution (or several solutions) is provided.
3. According to the model solution and assessment criteria, students have to grade and give constructive feedback to five randomly selected students by Sunday.
4. A student's own grade from this assignment is the average of the five grades.
5. The teacher's role is to monitor and support.

This forces students to study the model solution in order to mark (as well as study the solutions of others). Students learn how to communicate mathematics. This creates interaction between students, which is particularly important in an online course. The instructors must make it clear that the aim is to help others to learn, rather than to become a grader.

One effect of this whole approach is that a large amount of data is collected. We are only beginning to answer the question “what can we learn from these data” — questions such as “which parts of the course helped the most”.

## Problem given in peer assignment

**Complex Limit Problem**

In this workshop you need to compute the limit

$$\lim_{x \rightarrow \infty} \frac{\sin^2(\sqrt{x+1} - \sqrt{x})}{1 - \cos^2 \frac{1}{x}}.$$

Show all the steps of the computation.

Figure 2. Peer Assessment Example

**3.3. Ángel Ruiz.** His complete slides are at <http://blog.wias-berlin.de/imu-icm-panel-moocs/files/2014/08/Angel-Ruiz-on-MOOCs-ICM-2014-short.pptx>. There's also a supporting document: <http://blog.wias-berlin.de/imu-icm-panel-moocs/files/2014/08/MOOCs-in-the-reform-of-Mathematics-Education-in-Costa-Rica-final.docx>.

MOOCs are in an initial stage so a wide and flexible perspective is necessary, hence I am offering this one, from a small developing country. It shows the use of this e-learning strategy for very specific objectives.

In May 2012, a major reform in the mathematics curriculum of all primary and secondary education was adopted in Costa Rica. For its implementation the most important activities are courses for in-service teachers face-to-face, blended and virtual. I will concentrate now of course on the virtual courses. These courses are associated with the new curriculum, thus the nature is a little different from university courses. The content is not mathematics, not general pedagogy but a specific pedagogy of mathematics.

Why did we go for MOOCs? They are dynamic through videos, so the teachers can make contact with prestigious researchers who elaborated the new curriculum and conduct its implementation. Also we thought it would be easier but we are no longer so sure of that: to elaborate these courses has taken us more effort than we expected! We are using, basically, Powerpoint presentations, and we always have a person talking to you, so there is clear eye contact with the professor of the course.

We expect a greater completion rate than other MOOCs have experienced for two reasons:

1. the courses are very specific;
2. the new curriculum must be implemented.

There is a maximum number of participants because of the lack of resources. Note that there are some voluntary face to face activities involved as well: these are conducted by Ministry of Education officials. So we do have open courses, but not so massive, and online but with an external support within the different regions of this country.

**3.4. Robert Ghrist.** I'm a research mathematician and engineer at the University of Pennsylvania, and I want to tell you a little bit about my experience with MOOCs, more specifically my Coursera Calculus course. Only half the people signed up do anything and only a tiny

proportion of those make it to the end<sup>1</sup> of these fourteen weeks. The number that matters isn't how many people make it to the end but how much content is transferred. For example, over 1.6 million 15-minute videos were viewed — that's a lot of calculus watched! The course contains lectures that evolve over time, and after each lecture there is a homework that takes one, two, maybe several, hours. There is a lot of information at personal webpage (<http://www.math.upenn.edu/~ghrist/>). The main idea is that video is key, and indeed very well suited to mathematics. Our subject has a real advantage over other subjects in the ways in which video etc. can enhance the presentation. I've put a lot of effort into taking advantage of what the medium can do. Small clips, slides, comic book style, all draw the student in. Most of all it allows us to do what we as mathematicians want to do: show students what we see in our heads. MOOCs allows us to increase the bandwidth of our communication. It is far easier to show the beauty of mathematics in this medium than on a chalkboard.

Mathematics as we know is a very subtle art. It takes years of discipline and training to appreciate.

So we are at a relative disadvantage with respect to other fields in terms of this. I see a future where we have a lot more mathematics majors, because we can communicate better, because we can show many higher and more beautiful truths.

We complain that our students think that our rules are tricks that you follow like a robot. We know it is not that true, but we struggle to show that to our students. It is doable on a chalkboard, but not that easy. It will be easy to communicate, not only to our students, but to the rest of the world.

I would like to see a future where we have a lot more mathematicians making content. It will take much more work on our behalf to do this but the result will be worth it.

## 4. Discussion

**4.1. A (post-doc) speaker from NYU.** I hear a lot about MOOCs, and wonder how they will affect me. But my real question is “What problem are MOOCs solving?”

**JHD** In UK it allows for a critical mass of graduate students from different universities to attend courses that no individual university would find it economic to run.

**BB** Engagement with students perhaps. But then, what problem did evolution solve? The world is changing and we need to adapt.

**MP** No specific problem. Technology helps with visualisation of mathematical concepts.

**RG** Students can't pause or rewind live lectures. Lecturers do not give great lectures every time, nor can they, but you can get perfect content on MOOCs.

**AR** The mathematical community must understand MOOCs broadly and use them in different ways as appropriate and when they are effective. For the Costa Rica example it was needed for the flexibility as the teachers were all over the country (rather than being in one university) and needed to study in the evenings.

**4.2. Ingrid Daubechies, President IMU.** People in developing countries like the idea of MOOCs. But if you need to stream them online then it can be difficult without a good internet connection. Surely they need to be downloadable or local.

---

<sup>1</sup>The actual numbers were roughly 150,000, 70,000 and 3,000.

**AR** Some regions in Costa Rica have much better internet than others. So different regions are treated differently in terms of the logistics — sometimes material gets sent out physically. Also, one can take steps like adjusting resolution to help with worse connection. Finally, note that the internet in developing countries may be very different in 10 years time.

**BB** We did not stop developing radio when most of the world could not get it. If MOOCs prove effective the infrastructure will follow.

**4.3. Marie Farge (ENS Paris).** Do you make your MOOCs free to all? Note that this does not mean just to download but to also reuse, to take a component, to re-use in one's own teaching. This requires an appropriate licence, as the default position does not permit re-use.

**AR** Our courses are free online. Were intended only for Costa Rican teachers but have been used internationally already! Work not licensed for reuse yet but there is intention to do so.

**4.4. Thorsten Koch, TU Berlin.** There was recently a major article in *The Economist* [3, 4, a linked pair] stating that online courses are much cheaper. Does traditional education no longer pay? Will you charge for them?

**RG** There is nothing wrong with free/low-cost material as long as the quality is high. Even if this technology only works for first year courses (and students can take them in high school) that then reduces the time to degree, and would save a lot of money (in the U.S., often \$50–60,000), which is a positive thing.

**BB** Innovation is driven by those who dream the impossible. But we also need the realists, and do need to worry about cost and the impact on academia. If the change is just driven by economics then it won't be high quality and the change won't last.

**JHD** There's nothing specific about mathematics here, so can we move on?<sup>2</sup>

**4.5. Ingrid Daubechies, President IMU.** Being specifically mathematical, how does homework work on the Coursera platform? How do you grade 3000 students?

**RG** It is multiple choice (due to constraints on the technology and platform) with some automated feedback on wrong answer. Not many people will agree that multiple choice is the best way to assess mathematics. This is the main shortcoming of our course, but it will evolve and get better. Some constraints with the technology, and I preferred to experiment elsewhere.

**MP** Important to allow students to come up with their own solutions. Automatically generate problems. Some students take a quiz multiple times, getting a different question each time. Gameify it and they practice more.

**JHD** I have used the system MP was speaking about, and it can take me over an hour to write a quiz question and design all the alternative feedback mechanisms: it is certainly not free to the author! There is other software available but each piece tends to be tailored to a specific system and not well supported by the generic MOOC software that institutions tend to buy.

---

<sup>2</sup>This was in accordance with the remit of the panel. It does not mean that the issue isn't of wider interest.

**4.6. László Lóvasz, Past President IMU.** We seem to have a choice between face to face with an average instructor versus online with an excellent instructor. Are there any studies on which is better?

**JHD** MP has done some studies.

**MP** We have some limited studies (tens of students). Our experience at Helsinki shows that the MOOC-supported class is slightly better. Mika Seppälä in Florida State University is also using this material to support students in his sections of Calculus, and these students did a grade better, on average

**BB** It is not an either/or situation. MOOC technology will become integrated into face to face courses. We cannot test a new education idea like a drug trial, but that doesn't mean we shouldn't develop new ideas. We should be thinking about evaluation throughout.

**RG** In chess, we can set up competitions between human and computer, to find out "which is better". The best by far though is the human player augmented by software. I believe that the same will happen in education. Why not rely on the well-produced videos to motivate and excite students then get face to face time to work out the details with students.

**4.7. Jean-Marie Laborde, Grenoble.** I agree with the last speaker. Intelligent tutoring systems (ITS) have been around for a while and they have generally over-claimed. To make progress we need to understand how students learn, and what is the best way to support this.

**All** General nodding in agreement throughout this speech

**BB** Agree but the current situation is not well understood and does not examine aspects such as idea exploration and concept formation properly either. We may be expecting more of MOOCs than we are currently getting from conventional teaching.

**4.8. A speaker from Mexico.** What real world problems did you have implementing this? Technology, graphic design etc.

**MP** I am a fan of this technology of course. But it of course requires a lot of IT skills which need to be developed. Getting software provided throughout the university is hard. This is a new subject and we are experimenting. You need to be ready to make mistakes and errors and react to them. Students are quick to complain and flag up problems with the software, but tend to be happy if it is fixed quickly.

**JHD** At my university it took 6 months to get IT guys to install STACK in a system that could be used by undergraduates due to their quality assurance procedures.

**RG** It took an enormous amount of time and there was a huge learning curve. A good analogy is the video game industry. At the start one could make a good game with two people and two months. Now, in the very big video game industry, you can (and have to) have an enormous team with a huge budget. We could be on a similar trajectory.

**4.9. Mina Teicher, Bar-Ilan University.** How will MOOCs effect the job market, in the United States and in the rest of the world?

**BB** A more interesting question is how will it change the mathematics learned?

**4.10. A speaker from Canada.** I am very much in favour of these technological developments, having been the first person on my university's web site. Even if you can see their benefits, how do you implement this organisationally (office hours, workload management, credit for working on these etc.)? As has been stated, these courses take a great deal of time to develop. If you have thousands of students, how do you tutor them? My Teaching Assistants have strict working conditions, not taking more than fifty students etc. Teaching credit is independent of class size in my university.

**RG** I am not a trade union negotiator. I did not get any teaching credit for developing my MOOC course, so I took a risk. It paid off through all the positive feedback, but this is not a model that will work generally. But universities must find a better model. As well as credit and compensation issues, there are also ownership issues. That's probably why my university's effort is headed by someone from the Law School.

**4.11. An unknown speaker.** How do we use this technology to motivate students, especially the more able ones?

**BB** I think these technologies offer the opportunity to really motivate the more able student, as illustrated by RG's presentation.

**JHD** But it won't do that automatically: we the authors have to put the effort in.

**AR** I am dealing with teachers, not university students. But we are dealing, in general, with a new generation of students: much more visual, much more multi-tasking. They don't need a manual to operate a device.

A separate question for us is accreditation. Our courses are free (which is good for the students), but the process of accreditation (at least in Costa Rica) has not caught up with these changes. I look forward to discussing this in Brazil in four year's time.

**4.12. An unknown speaker.** Even if all the courses are not (yet) available worldwide, can we get a copy of the presentations?

**JHD** Yes: at the website (<http://blog.wias-berlin.de/imu-icm-panel-moocs/>), and under a CCBY licence.

**4.13. A speaker from Spain.** Will MOOCs replace face-to-face teaching? In face-to-face, we can see if we are losing the class, or some members of it, and we get interaction. These are valuable things I would be sorry to lose. but in Spain the economic point of view is very powerful.

**BB** Some things can only be done face-to-face. Others can be done better by MOOCs. I therefore expect the two to become more integrated. We have a fear of changing the nature of teaching and mathematics. Some of these fears are genuine. That fear is an argument for us to be **more** involved, to ensure the changes, which are coming anyway, are positive. We need to have a rational voice in these changes. We have lost control of mathematics testing in schools (to government or private companies). This was detrimental and so we must keep in control of MOOCs to avoid similar loss.

**RG** If mathematicians do not build good online content, and demonstrate how we can improve our teaching *by augmenting* our teaching with these technologies, then non-mathematicians will build bad courses.

**4.14. An unknown speaker from Argentina.** How does one promote this technology, if it is good?

Also, how should we use the errors of the students to improve the class with MOOCs?

**BB** Students already have their communication methods, mostly digitally enhanced and through social media<sup>3</sup>. Similarly, we need to ensure technology enhances learning. Once again, it is not an either/or question.

**JHD** I put my face-to-face lectures online for students who miss out. These can contain my errors, either mathematical or pedagogical. The students present have already debugged the lecture. Without students present at the time of recording there would be many more uncaught errors.

**BB** I leave the errors in: they promote more reaction!

**4.15. Marie Farge (ENS Paris).** The example of the computer game industry shows the size of the industrial investments involved. Therefore shouldn't everything (the content and the software etc.) be open source, so that good ideas can be widely adopted?

**RG** There is plenty of technology around to do media delivery. The hardest gap is the last half-metre [between the technology and the student]. Therefore we need good software tools for course development. The technology to send it around the world is already here. But currently my content is made in Powerpoint. I need a better platform.

**MP** WEPS is completely open source and free. Our base system, Moodle, is under heavy development by a big community, who are very conscious of their responsibility for production software.

The content I have under myweps.com is also free: the only restriction is that if you modify the content you must put it back up on the server.

**4.16. A speaker from Africa.** If you do not use technology, the students will — they will video lectures and distribute them etc., as I have seen in Cameroon. We are very happy to use cash machines and book airline tickets online: education has to benefit from this, and support the interaction so necessary in mathematics. We can, and do, make use of computer algebra (WIRIS) to do computations supporting our students work.

**All** General nods of approval.

**JHD** I am afraid we are out of time: my thanks to panellists and audience. MOOCs are *part* of the solution to a range of problems, not the be all and end all. Discussion will continue online (<http://blog.wias-berlin.de/imu-icm-panel-moocs/>).

## 5. Moderator's Conclusions

1. One common theme is that the question is not "MOOCs or face-to-face". The question is how best to take advantage of the strengths of both. "MOOC" is in fact a marketing phrase encompassing a range of technologies, and the real issue is which of these technologies can we use to enhance our teaching, and how.

---

<sup>3</sup>Author's note: though not much studied in mathematics, social media are ubiquitous among students. See [1] for one example.

2. Assessment was a significant topic: see Panel Brief point 1 and section 4.5. RG was using multiple choice because it was available, and admitted to its weaknesses, MP demonstrated a much more advanced piece of technology, but it's non-trivial for the course author to use well. MP's points about the integrality of assessment to the learning process are important, and his phrase *assessment for learning*, rather than just assessment of learning, is worth remembering.
3. The Intellectual property/licensing issues, raised in sections 4.3 and 4.15 are interesting. Most universities have not addressed this systematically (section 4.10).
4. The question of infrastructure (Internet bandwidth etc.) was raised in section 4.2, but is well-answered there, and we should not let infrastructure deter us, though we need to be conscious of it.
5. Evaluation of the technologies is difficult (even blind experiments are practically impossible, and double blind completely so) and poorly understood: some experiences are reported in section 4.6.

**Acknowledgements.** The Moderator is grateful to Matthew England and Ravi Vakil for taking notes, and the ICM Team for recording the Panel session. Above all, he is grateful to the Panel members, before, during and after the ICM, but any errors or misattributions are his fault alone. His own attendance was supported by EPSRC under grant EP/J003247/1, and by a National Teaching Fellowship from the U.K.'s Higher Education Academy.

## References

- [1] V. Perišić, F. Harvey, and T. Smith, Exploring use of social media in teaching analysis. <http://scotland.heacademy.ac.uk/assets/documents/stem-conference/Conference-proceedings-2014/MSOR/MSOR-227-O.pdf>, 2014.
- [2] C.J. Sangwin, *Computer-Aided Assessment of Mathematics*. Oxford University Press, 2013.
- [3] The Economist, Creative destruction. <http://www.economist.com/news/leaders/21605906-cost-crisis-changing-labour-markets-and-new-technology-will-turn-old-institution-its>, 2014.
- [4] The Economist, Creative destruction and Briefing: the future of universities. <http://www.economist.com/news/briefing/21605899-staid-higher-education-business-about-experience-welcome-earthquake-digital>, 2014.

University of Bath, UK

E-mail: J.H.Davenport@bath.ac.uk



# IMAGINARY PANEL: Math communication for the future – A Vision Slam

Carla Cederbaum, Alicia Dickenstein, Gert-Martin Greuel\*, David Grünberg, Hyungju Park, and Cédric Villani

**Abstract.** The IMAGINARY panel held on August 20, 2014, consisted of a “Vision Slam” of ideas on mathematics communication. We give an account of the expositions and we highlight the ideas and history of the IMAGINARY project.

**Mathematics Subject Classification (2010).** 00A09, 97A80, 97U99, 97-02, 97-04.

**Keywords.** Mathematical popularization, Mathematics communication, Mathematics education.

## 1. Introduction : Alicia Dickenstein

The name of the IMAGINARY Panel on Math Communication, was inspired on the “Poetry Slam” competitions, at which poets read or recite original work and are judged by the audience. A “Theme Slam” is one in which all performances must conform to a specific theme. In the IMAGINARY panel, the theme was a vision on the communication of mathematics for the general public.

The first panelist was Gert-Martin Greuel, together with Andreas Matt the heart and soul of the IMAGINARY project, created during his term (2002-2013) as director of the Mathematisches Forschungsinstitut Oberwolfach (MFO), Germany. Greuel is a Professor at the University of Kaiserslautern, Germany, with an impressive record of service and editorial responsibilities. He is a recognized specialist in Singularity Theory and Computer Algebra.

The second speaker of the panel was Cédric Villani, who has applied mathematical analysis to various areas of partial differential equations, probability theory, statistical physics and differential geometry. Among his many prestigious awards is the Fields Medal which he received at the ICM 2010 in Hyderabad, India. Villani is also renowned for his numerous activities in mathematics outreach and communication.

The next speaker was David Grünberg, who pleaded for more involvement of mathematicians in communicating their research to the younger generation. Grünberg is a teacher of mathematics and theory of knowledge (so far, in Costa Rica, England, Tanzania, Austria, Togo and Switzerland) and holds an Engineering degree. He is currently the head of the department of mathematics at the International School of Lausanne.

The fourth panelist was Carla Cederbaum, a young and very active mathematician. She got her PhD in 2011 in Berlin, Germany, in the area of differential geometry and geometric

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

\*Moderator

analysis. She held an Assistant Research Position at Duke University, USA, and is now a postdoc at the University of Tübingen and at the MFO, Germany. She is the author of a popular science book in mathematics which has just been translated from the original German to Korean, and the Senior Editor of the “Snapshots of modern mathematics from Oberwolfach”.

The last panelist was Hyungju Park, the chairperson of the successful ICM Seoul 2014. He is a professor at Pohang University of Science and Technology, Korea, and the director of the National Institute for Mathematical Sciences (NIMS), Korea. He has been elected as one of the Members-at-Large of the Executive Committee of the International Mathematical Union (IMU) for the period 2015-2018.

A Slam has winners. In this case, the winner was the audience. The “prize” was offered right after the panel. There were tours guided by volunteers through the fantastic NIMS-IMAGINARY exhibition, produced by the NIMS Institute from Korea in collaboration with the ICM committee and the Mathematisches Forschungsinstitut Oberwolfach (MFO) from Germany, and hosted by Hyungju Park. Some of the guides have produced material for the exhibition themselves. We would like to acknowledge the inspiration and support for this exhibit and for the panel from Andreas Matt, creative Project Manager of the IMAGINARY Project.

### **The IMAGINARY experience in Argentina**

I was the moderator of the panel. I am a Professor in the Department of Mathematics of the School of Exact and Natural Sciences at the University of Buenos Aires, Argentina. I have been elected as one of the Vice-presidents of the IMU for the period 2015-2018. My area of research includes different aspects of algebraic geometry and its applications, including effective methods.

I first heard about IMAGINARY during a visit to MFO in November 2007. I was fascinated by the possibilities of the software Surfer that was been developed for the year of Mathematics in Germany in 2008. It can be freely downloaded from <http://imaginary.org/program/surfer>. It is now one of many free software available at the Imaginary web page <http://imaginary.org/programs>.

Surfer allows to visualize 3d images of algebraic surfaces, that is, the (real) solutions of a polynomial equation  $\{f(x, y, z) = 0\}$  in 3 variables. The simplest such surfaces are a plane (where  $f$  is a linear polynomial) or a sphere of radius  $r$  centered at a point  $(x_0, y_0, z_0)$  (where  $f$  has the form  $(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 - r^2$ ). The incredible diversity of real algebraic surfaces is astonishing! Most familiar objects can be approximated by unions of them. But what I found more interesting: it is very easy just to play with Surfer choosing one of the many images in the Gallery and changing the parameters, and . . . almost anything one does is BEAUTIFUL. No deep knowledge is needed, one can just play and enjoy the beauty. On the other side, if one wants to get a particular shape, then it is necessary to pause to think and use mathematical concepts.

So, the first main effect is that the mathematical equations which are non tempting (and non beautiful) for most of the people, give immediately rise to objects that are in general not related to mathematics at all and that look pleasant to everybody. And secondly, the educational possibilities of this interaction between formulas and forms, between mathematics and art, are big.

A mathematical exhibit using Surfer and another interactive beautiful software called Morenaments (<http://imaginary.org/program/morenamentals>) was first held in Argentina dur-

ing the 2012 edition of the huge Science and Technology fair “Tecnópolis” organized by the Argentinian Government. Millions of visitors attended this fair and were enthralled by the display.

Together with my colleagues Gabriela Jeronimo, Santiago Laplagne and Ursula Molter we presented an outreach project to the University of Buenos Aires, which was funded. During 2012 and 2013 we visited several public high schools in the city of Buenos Aires and worked with the students for around 90 minutes. These students come in general from low income families, which do not necessarily have a computer at home, but they do have in general small netbooks distributed by the national or the city governments. Even with these technical limitations, the activities were very successful. The students worked passionately and produced in this small amount of time interesting forms, that we posted in our website <http://moebius.dm.uba.ar/> together with manuals for students and teachers (in Spanish). We also played with another basic free software called Britney created by Santiago Laplagne (available at: <http://moebius.dm.uba.ar/page.php?code=1>), which allows to visualize fractals.

Our conclusions are that having mathematics mediated by computers made the introduction of mathematical concepts easier. More importantly, we hope that revealing the beauty of mathematical objects to the general public will allow them to have a friendlier view of mathematics and will open the way for them to enjoy it while enhancing their mathematical thinking.

## **2. IMAGINARY – Mathematical creations and experiences : Gert-Martin Greuel**

IMAGINARY is the name of a collaborative mathematics outreach project that aims to improve the image and understanding of mathematics and in this way awakes an interest and fuels passion for the subject in children and adults. This goal is achieved in different ways: on the one hand by showing the beauty and art in mathematics and on the other hand through surprising applications.

IMAGINARY is a project of the Mathematisches Forschungsinstitut Oberwolfach (MFO) and it was born in conjunction with the Year of Mathematics in 2008 in Germany. It started with the travelling exhibition “IMAGINARY through the eyes of mathematics”, shown in many cities in Germany.

### **Exhibitions**

Exhibitions are the IMAGINARY way to reach out to a broad public in real life. They are shown in galleries, at museums, in schools, banks, universities, parks or train stations. Exhibitions are diverse: they can include images, interactive programs, sculptures, puzzles, games, text boards, etc.. Visitors can take print-outs of their creations home and everybody can easily stage an own exhibition. In fact, many of the exhibitions were self-organized.

Let me give an impression of the original travelling exhibition. Since 2008, it has been shown in over 60 cities in Germany alone. But it has also travelled further afield to 4 continents, 29 countries and over 120 cities with more than 1 million visitors in total. In Europe, IMAGINARY has been presented in 17 countries with talks, workshops, media activities and, in most cases, exhibitions.

What made the exhibition unique from the beginning, is its highly interactive and intuitive nature and its open access and open source philosophy. This is also reflected in the many positive comments left in the guest book by visitors having experienced the unexpected beauty and the “joy of comprehension”: - *This already beautiful exhibition is obtaining a special liveliness by excellent leadership.* - *Super, especially that you can also use the program in the school.* - *A wonderful exhibition. I have spent much time here and met many beautiful things, it had to take place more often and actually as a permanent event!* - *Thank you and keep it up!* - *It is a fantastically beautiful exhibition.* - *The magic world of mathematics is not easy to understand. But you can bring them closer.* - *We were again there, because it was so fascinating.* - *I should have perhaps studied math ....* - *Simply gorgeous, cool programs.* - *Mathematics makes happy.*



Figure 1. Exhibition at the Leibniz-University Hannover, 2008



Figure 2. Cedric Villani inaugurating the exhibition in Paris, 2010

### **SURFER Creations**

One of the main attractions of an IMAGINARY exhibition is the SURFER, a program that calculates and displays algebraic surfaces in real time. Visitors can enter and change polynomial equations on a large touchscreen with their fingers, shift parameters, determine the colours of the surfaces and turn the figures as they like. The great thing about SURFER is that you don't have to understand the underlying mathematics (algebraic geometry) a priori, you can experiment, try, follow your intuition and creativity and this way learn mathematics and create unique art work like pictures or animations.

SURFER was developed by the MFO in collaboration with the Martin Luther University Halle-Wittenberg and the University of Kaiserslautern, mainly by Christian Stussak. Many visitors of an IMAGINARY exhibition downloaded the SURFER and created their own algebraic surfaces, with sometimes really surprising results.

For example, Valentina Galata started in 2008 when she was a 17 years old high school student to remodel 'real world objects' based on algebraic surfaces with the SURFER. Other users of the SURFER created really artistic pictures of algebraic surfaces, like the beautiful Sunflower image by Torolf Sauermann.

### **Relation to Research**

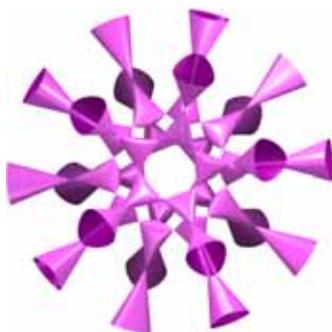
The origin of the SURFER is very closely linked with current mathematical research. A first version goes back to Stephan Endrass, a student of the mathematician Wolf Barth, who



Figure 3. “Cappuccino” by Valentina Galata



Figure 4. “Sunflower” by Torolf Sauer-  
mann



discover the “Barth sextic”. The Barth sextic’ is a beautiful surface of degree 6 with the symmetry of an icosahedron (and with a terrible complicated equation). It holds the world record with 65 simple nodes, the maximum possible number of singularities. From degree  $d = 7$  on, the maximum number of singularities on a surface of degree  $d$  is unknown.

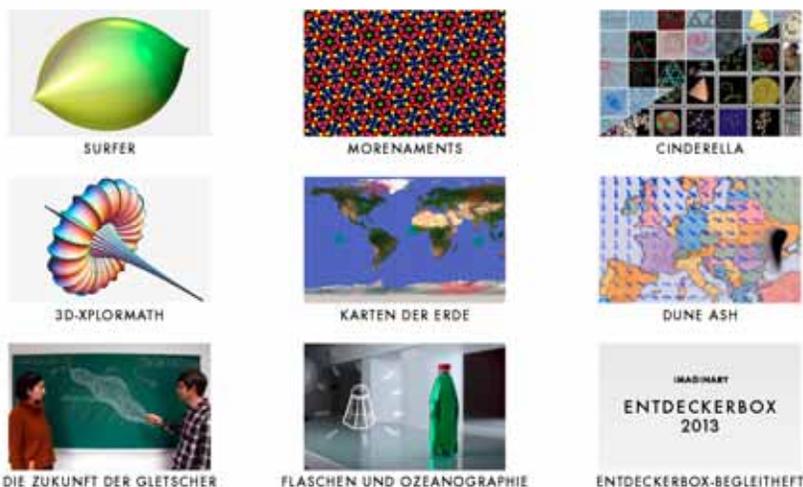
A new IMAGINARY project is to connect modern mathematics and current research to outreach. Mathematicians visiting the MFO are asked to write about their current work but for a general public. These so-called “snapshots of modern mathematics” are then reviewed and edited and distributed through the project. For details see the overview by Carla Cederbaum in this article.

**IMAGINARY in schools and classrooms**

First schools started to copy the exhibition or parts of it, for example the pictures or the programmes by a high school in Saarbrücken. Also first self-organized exhibition were held e.g. in Kiev and IMAGINARY competitions were organized e.g. by a newspapers in Greece. The Girl’s Day at the TU Berlin was a one day event to attract school girls to study mathematics by using the SURFER. Especially the programme SURFER is ideal to be used in the “school context”, for example a 4-days workshop for school students aged 12-14 in Vienna, called “Kinder-Uni-Kunst”. The idea was to create mathematical animations with music and while making the films with SURFER learning basic underlying concepts of algebraic geometry. See also the personal point of view by the mathematics teacher David Grünberg about IMAGINARY in school in this article.

A collection of IMAGINARY worksheets of different levels of difficulty has been developed for school children aged between 5 and 17 years. IMAGINARY booklets with questions and explanations are used during exhibitions for guided school tours or at special workshops. The so-called “Entdeckerbox” (discovery box) is primarily aimed at use

in the classroom and provides resources for teachers in order to make mathematics lessons more interactive and interesting for the pupils. It contains 3D-sculptures, nine programs and films and, as a special highlight, the booklet “Problems for children from 5 to 150” by V.I. Arnold. This text has been translated into 6 languages and may be downloaded at [imaginary.org/search/node/arnold](http://imaginary.org/search/node/arnold).



### The mathematics of planet earth

However, the original exhibition was not enough; it focused on a very beautiful yet small part of mathematics. The project needed to grow further and the Mathematics of Planet Earth Year 2013 (MPE) presented a good opportunity to do so. A competition for virtual exhibition modules themed around MPE was announced, and IMAGINARY provided the required web infrastructure in order to make the modules of the competition available online. At the launch of the MPE year in Europe at the UNESCO in Paris, the web interface to IMAGINARY - open mathematics ([imaginary.org](http://imaginary.org)) went live, displaying entries for the competition and, of course, the winners.

At the same time, a complete MPE exhibition is also available, consisting of a series of modules with a more applied mathematics focus, such as a program that calculates the displacement of volcanic ash clouds (Dune Ash) or a film discussing how mathematical modelling of glacial movement works in order to predict the future behaviour of glaciers.

### International spreading

The “travelling exhibition” IMAGINARY developed into a “spreading exhibition” through many partners who independently started to stage it and further expand it. IMAGINARY exhibitions were shown in 4 continents, 29 countries and over 120 cities with more than 1 million visitors. An example is the RSME (the Royal Spanish Mathematical Society) who took the exhibition at the occasion of its Centennial, added new texts and translations and staged it in more than 13 cities. Another wonderful example is the cooperation with National Institute of Mathematical sciences (NIMS) and the ICM 2014 in Korea that was made possible mainly by Hyungju Park. The exhibition NIMS IMAGINARY during the ICM was visited by about 12.000 visitors, among them many school classes, and attracted a lot of media coverage.



The exhibits were also installed and shown in science and mathematics museums. For example the MiMa Museum for Minerals and Mathematics in Oberwolfach, the new Mo-Math in New York, the “Mathematisches Kabinett” in the Deutsches Museum in Munich, the Forms & Formulas exhibition in the National Museum of Natural Sciences and History in Lisbon, and the CosmoCaixa museums in Barcelona and Madrid.

In 2014, many new exhibitions have been launched around the world. In particular, IMAGINARY has started a collaboration with the African Institute for Mathematical Sciences (AIMS) and, in association with AIMS, an interactive IMAGINARY event was organised for the first time in Africa at the 10th anniversary of the pi-day celebrations in Dar es Salaam, Tanzania. In November 2014, a workshop and exhibition will be organised in Cape Town to plan future mathematics communication activities with partners on the African continent. IMAGINARY exhibitions are currently on tour or planned in Germany, Russia, Spain, Norway, Portugal, and Hungary, and new projects in France and in Turkey are on the way.

### Who stands behind IMAGINARY?

IMAGINARY is a project by the MFO, accounted by its director Gerhard Huisken, with funding from the Klaus Tschira Stiftung. It is maintained by a committed core team (mathematicians, software engineers, graphic designers, etc.), who run the project, develop the Internet platform and give advice on how to coordinate exhibitions, but also dream up new ventures of where IMAGINARY will go in the future. The excellent achievements and the impact of the project was acknowledged in November 2013 by the Deutsche Mathematiker Vereinigung (DMV) when the German Media Prize for Mathematics was awarded to Gert-Martin Greuel, the former director of MFO and scientific advisor of IMAGINARY, and An-



dreas Daniel Matt, the curator and project manager of IMAGINARY.

Besides the core team, and most importantly, IMAGINARY it is a community driven project by and for the community. This means that anyone who has an interesting piece of software, film or other type of interactive material can upload this to the website and make it available to the rest of the community. Of course, anyone can just use the material and create a mathematics event, exhibition or workshop. In this way, the community becomes an integral part in the communication process by not only experiencing but also creating content and thus advancing mathematics communication to the 21st century. We hope that many institutions make use of the content and infrastructure of IMAGINARY, and take an active part in shaping its future.

*Parts of this text have been published in Gert-Martin Greuel, Andreas D. Matt & Antonia S. J. S. Mey: IMAGINARY- Mathematics Communication for the 21st Century, in EMS Newsletter 92, June 2014.*

### **3. Mathematicians, journalists and the general public : Cédric Villani, notes by Bianca Violet and Severina Klaus**

I appreciate very much that the ICM here in Seoul takes care of these global issues about not just doing mathematics but also communicating about mathematics. I think it is so important.

In the past years I have been involved in various activities at institutional level and at personal level. I already have been working in the project of a math museum at the Institute Poincaré in Paris for some years as well.

But in this slam I will only talk about my personal experience of communication as a mathematician, not about my institutional experience.

My first encounter with a journalist from the outside world, not a scientific journalist, was about 10 years ago and it was a disaster! After the interview, the guy went back to his boss and said “I met a crazy guy; I did not understand a single word”. He had not dared to tell me that he did not understand anything, so I kept on talking. It was a complete disaster and we had to rearrange things and so on. Thinking of the first encounter I can say that mathematical communication is not something that you are born with, it is not something that is natural, it is something that you train.

My second encounter was in fact a training operation in 2007 - Etienne Ghys had recommended that I attend this training session, so I went. Four persons of our laboratory attended the training, and we were all delighted with this experience.

This is my first advice: If you want to do some serious math communication - get some training with an inspiring guide!

And this guy who trained us, he was a media person and an expert in communication; and in particular he explained about the psychology and constraints of journalism. How you have to put him with you on your side, what are his difficulties, his expectations, what is he afraid of, what is the margin, etc. It was really interesting.

And years later when I was invited by a school of scientific journalists to give a lecture I reversed the sides. Trying to put them into the brain of a mathematician and explain how we have difficulties when we face a journalist and how the contact can be difficult. It is important to be aware of this.

And this changes everything. We already have enough trouble communicating with each other and explaining mathematical research, so why should we bother to communicate to

general non-mathematicians? There are various reasons for this and it is good to have all of them in mind.

First, which we are very sensitive to, is making sure that the young generations are interested in this and know that these are good jobs, inspiring jobs. Maybe these are not the jobs in which you make the most money, but these are jobs in which you feel good - and you may recall, by the way, that in 2009 mathematicians were ranked as the number one job in the world in terms of how rewarding it is by the Wall Street Journal.

This is the goal we think of in most countries, but that is not the only one. Another goal is to feel good about the way people look at us. You don't want people thinking that these are crazy nerds doing their stuff and we don't know what these mathematicians are doing. People say that it is good to give them money, but who knows what the hell they are doing with that money. It is important that people have a good opinion about us as a profession and so on. And just to be heard. It may be sad, but nowadays if you don't remind people that you exist as a job, as a community, people completely forget. Then one day you lose your funding, one day people will say I don't think this is a good job and so on. And we don't want this to happen.

Another important goal is to maintain the link and coherence of society. All pieces of society are important, this we know; and we have to recall how important we are to other people, the same as we need the engineers, we need finance people, we need artist people, we need everybody to make the world run, and we are part of this. And it is known, by the way, that the part of mathematics and mathematical research in the GDP is much higher than one would guess and it is increasing year after year.

And a final thing is that sometimes many people, people who take decisions, who run things, are very much in need of our advice on many things. Not really about explaining mathematics, but over the past years I got plenty of invitations for instance from clubs with people running companies or administrations etc. looking for general guidance on how to approach complex problems, on how you do it in your research work. And all the time they say it reminds them of some expertise this is so inspirational and so on. The whole world is full of people with difficult decisions to make and they don't know whom to ask advice of. We as researchers always have decisions to make which are complicated; and they are in demand of our advice in this.

What does a general audience want? This is a mixture of many things. Some of them are at a university. Some of them are thinking "will this be a good job for my kids?" Many of them think "I was so bad in math, I am angry by this, at least maybe I have a chance to understand at last, prove myself that I was not so dumb". Many people at the end of talks, they come and tell me "ah, if only I had a teacher like you! I would not have been so bad at math." I really think if they had a teacher like me it would probably have changed nothing. But it is good that they have this feeling, that the guilt was not on their side, that they were not intrinsically dumb and so on. And also some people, they just need to inquire about the world. They heard that mathematics has part in finances, in the economy, in space exploration, in whatever. And they wonder what do they do, these mathematicians? It is curiosity. And even those people who are very bad at math, they have the right to understand what we are doing, in the same way as people who are bad at writing for instance have the right to follow and see what is going on in literature and get informed about the trends in culture. Mathematics is technology and science but it is also part of the culture, and many people are interested in this.

Now, with that in mind, we see there are very different audiences we need to reach. And

you have to make sure, if you communicate, that part of what you say will be interesting for this or that audience. The best is if you can mix a little bit of everything in your talk; you cannot always do it, but sometimes it works.

I have tried a number of different forms of communication. Some are good for this and others are good for that. After the ICM of 2010 I received a lot of invitations. And after each talk you have to think: what did work? what did not work? what lesson can I draw for the next speech?

I did radio. Radio is good because you can say things and people listen to you. I did television. Television is good because people see the attitude that you have; but don't expect that they listen to you. They are not interested in what you are saying when you are on TV, but it is very efficient to reach people.

I was in some movies, I prepared some articles for some newspapers, I also did some exercises in which they ask you for some text which is a mixture of something like poetry or literature and something like mathematics. I did a lot of public lectures, for example in High Schools, for little kids, 10 year olds, for older kids, for students, for politicians, CEO's, workers - all kinds of possible things. I never went on Facebook or twitter out of a lack of time though.

There are many obstacles. And it happened more than once that after something went wrong I thought I should stop and did it again, and in the end it was a big success.

To single out just one experience which was the most life-changing in this, it was the book. Seriously speaking though, this book, *Théorème Vivant*, came by accident because I met some editor in a dinner. He understood I wanted to do math communication, and he had no interest whatsoever in explaining me the mathematical concepts. All he wanted to know was how I work in my daily life, what I do, what I think, what is my life and so on. It was very embarrassing. And in the end I decided I would go for a concept that described our theorem but not its meaning. Just how we made it. It was a long work and as in all long works it was full of unexpected things, two years and a half full of ups and downs. And I just put in the mathematical equations, the mathematical words with no explanation, and this was contrary to any classical work of communication. It just gives you an impression of what it is to be in the brain of a mathematician.

And the first day it was out I was worried about what people will say? What will the colleagues say? But it worked beautifully. And then I started to receive these comments. Hundreds of comments, comments like "Your book changed my life"- and nothing can make you feel prouder than that.

It's a thing we all have to learn. There are many people out in the world that are looking at us mathematicians for inspiration. There are people somewhere who would be happy to listen to what we do and our fears and anxieties and how we overcome them and so on. So share with them; that's the most important thing.

#### **4. Communicating math research to the younger generation : David Grünberg**

- **Open Mathematics - what it is and why you should do it.**
- **Mathematics teachers are in the "trenches" fighting for the mathematics of tomorrow. We need Mathematicians to help us.**

I like to believe in the multiverse theory! The reason is that, in a multiverse, maybe

there's a region of space-time in which I, after school, continued with pure Mathematics! This is our reality, though, and in this one, this, is what happened: I learnt the same mathematics in school as you, but then our paths diverged: you guys kept going with the equations, and you are here at the ICM with your bright lemmas and elegant proofs. Whereas I ... I went travelling and started teaching Mathematics in schools to help my finances along the way. I taught Mathematics simply because at that time, this was the only subject I might conceivably teach. Oh, and just before that, I got myself a degree in Engineering, but I can't say that too loud, otherwise someone will say: "An engineer was let into the ICM ?! - someone call security!"

But don't get me wrong. I have learnt to love Mathematics! As a math teacher, I do math every day, I learn math, I teach math... And I don't even have to publish anything. I get to do maths with fabulous people - my students! Looked at it like this, I rather like my corner of the multiverse... I am here to talk to you about something that reunites, us, though, in this very real universe: Mathematics Communication. Mathematics Communication is a network of bridges spanning the divide between Mathematics research and the wider population, which includes those young people who, we are hoping, will one day continue what you guys are doing. I'm here to give you my take on one of these bridges: the IMAGINARY project. And my aim is to convince you to take some steps on that bridge, should you not have tried yet.

IMAGINARY is something that can reunite you, the mathematician, with school pupils. My first contact with IMAGINARY was when I followed a link to the web platform. "*Open mathematics*" mmh... - *I'm wondering what that might mean! Woa, these pictures look really attractive!...* Ah, *algebraic surfaces (Picture 4)*, mmh... *oh that looks like a 3D version of what my students do with functions. I've always thought it's a good idea to introduce area via 3D shapes and surface area - our surrounding is 3 dimensional, after all. Perhaps something similar can be said about functions: work in 3D seems so much more real than in the coordinate plane....*

Such thinking got me going with IMAGINARY. Note the importance of the aesthetic appeal here, something I'm particularly sensitive to. I dreamt of organising an IMAGINARY exhibit in my school. After all, our art department does exhibits as a matter of course - why not the Mathematics department? No - wait! Why not the mathematics department TOGETHER with the art department? I started to understand what 'open mathematics' means.

"Open Maths" is all about exchanges. Once I heard a TED-talk in Paris by a mathematician on the "ingredients of good ideas". And one of these ingredients was "EXCHANGES". You know who was talking? A certain Cedric Villani.

The very idea of IMAGINARY is that people can exchange and participate. Our slogan, "OPEN Mathematics", says IMAGINARY wants to be an ingredient in the Villani-"Good-ideas" formula. 'Open' means something like "open house"- you are free to come and go, there are no locked cupboards. You can take things out or bring things. An open house becomes YOUR house; I know a family with 5 adolescent children. Their parents are very much 'open house', and so the kids bring their friends home all the time. I can tell you: that house is alive! You've got the core of my message right here: An OPEN house makes for a house that is ALIVE!

Open Mathematics lives through participation. You upload your maths or download someone else's. The main point is not THAT someone downloads your work, but it is WHAT that person will do with it - given some quality control, people will start to use stuff you put out there in ways you couldn't imagine. Other mathematicians will use it; Museums will

use it (for instance the “FormulaMorph” installation at the MoMath, New York, based on IMAGINARY material); schools will use it (for example: me!); artists will use it (as a visit at the IMAGINARY exhibition will make obvious); even chefs will use the material! (a chef cook actually got inspired after visiting an IMAGINARY exhibit in Spain!)

It’s amazing what people will do! They will even translate texts into Korean for you! Right now this summer, I have a student doing a project based on some research material made available on the platform. Point is: At IMAGINARY we have witnessed that process of creative transformation of content into new ideas that NO ONE could anticipate.

In my town in Switzerland, there’s periodically a “free market”. People bring things that they want to pass on - for free. And you can come and take anything you want - for free, you cannot pay. Even the food is free! IMAGINARY is often asked: where can you buy the exhibit? Now you know the answer: There’s nothing to buy at IMAGINARY - it’s based on **participation**.

Let me tell you a little more about my job: You know, Mathematics teachers worldwide are facing some tough challenges with modern developments in mathematics. First, there is the splintering of mathematics research into many highly specialised nooks and corners. This makes the transfer of new mathematics into school curricula difficult. But at the same time, technology, like computer algebra systems for example, questions whether schools can teach mathematics for much longer the way it is still mostly done today. And then, as you know, our subject is under a lot of pressure and scrutiny, squeezed somewhere between PISA evaluations and back-to-basics prophets.

I’ve heard someone say: “Mathematics could one day disappear, like the classics did (Latin and Greek)” I wish I could have taken that person to Villani’s talk at ‘Bridges’ on Monday, where he explained that the movie industry is one of the biggest consumers of mathematics. I say: We could well see too few students choosing to continue their study of mathematics if school mathematics doesn’t somehow stay connected to maths research and change the way the subject itself evolves. And we need YOU to help schools adapt!

My students need you. In order to keep mathematics irrigated with young talent, we DEFINITELY need the cooperation between those involved in mathematics research, mathematics communication and mathematics education. It is one of the ambitions of IMAGINARY to provide a platform for such collaboration.

You might think: why should I spend effort making my mathematics approachable to wider crowds? Why should my mathematics be “open”? My students, and my colleagues and myself at school, and teachers worldwide are one of the reasons why I’d like you to consider communicating your work as much as possible to the wider community: We need you. We maths teachers are at the forefront of teaching YOUR subject, so if you aim to stay in touch with us, it will be a lot easier for us to connect with mathematics research.

With my class of Grade 10, we did a unit on algebraic surfaces, something I would not have dreamt of doing had not a mathematician gone through the trouble of producing a quality and user friendly software to experiment with. It was an interesting experience. In the end, we did a competition about who can produce the most fanciful snowman, using a single formula. What I’m trying to say is this: I can guarantee you one thing: If you can spot an area of your work that can be somehow transmitted at a lower level, create high quality, user-friendly material about it, make it available for free, then there’s a maths teacher somewhere in the world who will use it, but probably much more than one, because we are talking to each other.

You know what: In my school there is a large magazine stand in front of the library. In

envy the science teachers: they have several magazines they can read to keep up-to-date with what's happening in science (New Scientist, Nature. . .). There's no magazine that tells me with simple language and cool graphics what's going on in mathematics today because no one is writing one for our sort of audience - A colleague joked that I can always read the numbers in the science magazine...

I find initiatives like IMAGINARY useful in helping me keep in touch with what's happening in mathematics TODAY.

This session here is about VISIONS FOR MATHEMATICS COMMUNICATION. Here is my vision:

In the near future, there will be more sharing between mathematicians, more "open mathematics", and an increased focus by mathematicians on communicating their work.

## 5. Snapshots of modern Mathematics : Carla Cederbaum

In the last year, a team of mathematicians and mathematics communicators at the Mathematisches Forschungsinstitut Oberwolfach (MFO) has developed a new scheme to communicate modern mathematics and mathematics research in writing to a wide audience. The resulting texts are called "snapshots of modern mathematics from Oberwolfach" and have been collected and made available for free via [www.imaginary.org](http://www.imaginary.org) since January 2014 as part of the project *Oberwolfach meets IMAGINARY*, funded by the Klaus Tschira Foundation and the Oberwolfach Foundation.

Our goals for the snapshot project are to show that mathematics

- can be understood,
- is diverse,
- has surprising practical applications,
- is fun, elegant, and creative,
- ...

and that mathematicians

- are individual people,
- are diverse in personality etc.,
- have different motivations,
- are approachable,
- ...

In particular, we want to encourage the readers to be curious about modern mathematics and mathematical research.

**Visions.** I am the senior editor of the snapshots; my vision is that the snapshots will be widely read by a diverse worldwide community, used in or for secondary and tertiary education and as a source of inspiration by journalists, and that other institutions adopt the format we have developed to contribute their own snapshot series on the IMAGINARY platform for

the worldwide community. Furthermore, it would be great if volunteers from all around the world would take the time to translate their favorite snapshot(s) into their native language and upload the translation on the IMAGINARY platform so that others can benefit and read a snapshot in a language that might be easier for them than English.

In the following, I will provide more information on the idea of snapshots in general and some of the specifics about the snapshots from Oberwolfach.

**What is a snapshot?** A “snapshot of modern mathematics” is a short text

- written by a (group of) mathematician(s),
- edited by mathematics communicators,
- peer-reviewed by specialists,
- possibly illustrated by a designer, and
- distributed for free via the IMAGINARY platform [www.imaginary.org](http://www.imaginary.org)
- under a Creative Commons license.

See Figure 5 for excerpts of some examples and [www.imaginary.org](http://www.imaginary.org) for all snapshots that are currently available.

Snapshot of modern mathematics  
from Oberwolfach
№ 11/2014

### Arrangements of lines

---

Brian Harbourne • Tomasz Szemberg

We discuss certain open problems in the context of arrangements of lines in the plane.

#### 1 Introduction

Imagine a finite set of lines in the plane. These lines may intersect each other in certain points – at some points only two lines might meet, at other places three or even more lines might meet. Such a set of lines is called an arrangement  $A$ . The arrangement may, for example, just consist of three lines joining vertices  $P$ ,  $Q$  and  $R$  of a triangle, see Figure 1.

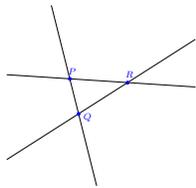


Figure 1: An arrangement of three lines

One can calculate the *self-crossing count*  $c$  of the arrangement by counting the number of lines through each intersection point, squaring it, and adding

1



Figure 1: Euler and Goldbach's letter. It seems non-trivial to find a genuine portrait of Goldbach.

value of  $C$  was

$$C = e^{3100} \approx 2 \cdot 10^{1346}$$

(Liu-Wang [8]), which was way too large. We simply cannot hope to check the first  $10^{1346}$  cases by computer – in fact, it is highly doubtful that any earthly or alien civilization that will ever exist could ever check, say,  $10^{120}$  cases of any conceivable statement one by one: the number of picoseconds since the beginning of the universe is less than  $10^{30}$ , whereas the number of protons in the observable universe is currently estimated at  $\sim 10^{80}$ , meaning that even parallel computing and galactic dictatorship wouldn't be enough.

I managed to bring  $C$  down to  $10^{27}$ . The binary Goldbach conjecture had already been checked by computers up to  $4 \cdot 10^{18}$  [9]; using that fact, one can check the ternary Goldbach conjecture up to  $10^{27}$  in a few hours on a modern desktop computer. (In fact, D. Platt and I [6] had already checked it up to  $8.8 \cdot 10^{30}$  on parallel computers.) This means the ternary (that is, weak) Goldbach conjecture is now proven for all (odd) integers.

It is clear why a brute-force computation can check a conjecture such as Goldbach's only for  $n$  smaller than some constant  $C$ : a computation has to be finite. But why would a mathematical proof ever give a bound valid only for  $n$  larger than a constant  $C$ ?

Figure 5. Sample excerpts from different snapshots. The snapshot on the right is by Harald Helfgott.

Snapshots are aimed at a wide audience, including

- secondary school teachers and instructors of undergraduates,

- science journalists,
- secondary school and undergraduate students, or
- just anyone with an interest in modern mathematics and mathematical research.

For example, teachers and instructors might use some snapshots to demonstrate to their students that mathematics is still an active research field: What kinds of questions do mathematicians research? What other fields are intertwined with mathematics? Are notions we teach in secondary school and in the undergraduate curriculum relevant for or at least used in mathematical research?

Science journalists who are searching for research results and stories that are of interest for their readers/listeners/viewers. In mathematics, this is particularly hard as our research publication are very difficult to read for outsiders. Instead of doing this hard work or going by word of mouth, journalists could use the snapshots to identify topics they would like to report on. At the same time, a snapshot's reference list and its authors might be a good place to start if a journalist would like to find out more about the topic.

Students, on the other hand, may find the snapshots helpful when trying to decide if they would like to pursue an undergraduate or graduate degree in mathematics.

All snapshots are assigned to categories specifying their “mathematical subject(s)” and possible “connections to other fields”. This allows readers to learn about the different areas of mathematical research and their interconnections. At the same time, they can get a first impression of the diverse applications of mathematics in other areas of research as well as of the influences other fields have and have had on the genesis of mathematics as a discipline.

The first snapshots have been collected in the Fall of 2013 at the MFO. However, the idea of producing such snapshots and distributing them via the IMAGINARY platform is by no means protected by the MFO. To the contrary, we are happy to support and/or advise other institutions (institutes, national or international societies) who are interested in setting up their own snapshot scheme along the above lines. We are in the process of producing hands on guidelines for institutions who are considering to do so.

**Snapshots of modern mathematics – from Oberwolfach.** The snapshot project started in January 2014 at the MFO. We find the snapshot **authors** among the participants of the scientific programs of the institute. They are volunteers identified by the scientific organizers of the respective program. To facilitate writing for such a general audience, we provide writing guidelines including hints how to make a text more accessible, see <http://mfo.de/math-in-public/snapshots>. Moreover, we have prepared a  $\LaTeX$  class and template that support authors in giving adequate copyright credit etc.

Our editors suggest editorial changes to the authors, thus making the snapshots more accessible and understandable and generally easier to read. To do so, they make thoughts present in the snapshots very explicit, bridge gaps to the secondary school curriculum and its scope (as far as possible), introduce redundancy, illustrate ideas, formulate questions to the reader, check language and grammar, insert cross-references between the snapshots, etc. When a snapshot is finalized, the organizer(s) of the scientific program at MFO who selected the author(s) also act as reviewers (“communicated by”). The editors then make approved snapshots available on the IMAGINARY platform.

Our team consists of currently 2 junior editors – mathematics graduate students who are also mathematics communicators, Sophia Jahns and Lea Renner. We get tremendous support from the IMAGINARY team, in particular from Christoph Knoth (web design), An-

dreas Matt (coordination), Antonia Mey (author contacts), Konrad Renner (design and web design), Christian Stussak (IT), and Bianca Violet (author contacts) as well as from the MFO staff.

**Getting involved.** If you are interested in joining our team, contributing material for schools, translating snapshots into other languages, or setting up a snapshot series at your institution, reporting on the snapshot project or using snapshots in other publications, please feel free to get in touch with me via [cederbaum@mfo.de](mailto:cederbaum@mfo.de).

## 6. The KAOS initiative – Knowledge Awake On Stage : Hyungju Park

Many accomplishments of modern mathematics can often be explained quite clearly to the public when they are presented in connection with their implications in other intellectual areas. In this regard, KAOS (Knowledge Awake On Stage) started aiming to be a long series of conversations with the audience about the intricate web of mathematical structures that run through the fabric of contemporary civilization. It features lectures enhanced by stage effects, given by experts in various sectors of academia and society including natural scientists, social scientists, writers, critics, musicians, and artists, each of which are followed by a conversation with two host mathematicians (Minhyong Kim of Oxford and Hyungju Park of POSTECH).

General Vision: The guests will have varying degrees of interest in mathematics and science, and one goal of the conversation will exactly be to uncover the connection and relevance of mathematics to their work through a careful combination of questions and interactions. The series is aimed at the educated public. Efforts will be made to keep the lecture and the entire conversation at a level accessible to any educated person or student with a serious interest. Technical portions will be supplemented by explanatory material to be supplied by the two hosts

KAOS 1. November 28, 2012

Title: Mathematics of Match Making

The lecture was given by one of the hosts, Minhyong Kim of Oxford University, on the Nobel Economics Prize winning work of Lloyd Shapely and Alvin Roth. The subsequent talk-show part was presided by the other host, Hyungju Park of POSTECH, with the participation of two guests.

KAOS 2. May 29, 2013

Title: Geometry, Topology and Matters\*

The lecture was given by a renowned physicist, Philip Kim of Columbia University (now at Harvard University), on Graphene Physics. The subsequent talk show was conducted by the speaker and two host mathematicians.

KAOS 3. October 5, 2013

Title: Fantasy of Music and Mathematics\*

The master of ceremony was a well-known Korean pop singer, Lucid Fall. After a 20 minute long introductory lecture given by Hyungju Park, the main lecture was given by a renowned Berlin-based opera composer Eunsuk Chin on Contemporary Music. Minhyong Kim gave a concluding lecture on various aspects of mathematics and music. The talk show part was led by Lucid Fall with the participation of the main speaker and the two host mathematicians.

#### KAOS 2014

A yearlong program consisting of five lectures by renowned Korean mathematicians are planned in celebration of Seoul ICM 2014 on the theme of “Essence of Mathematics”

1. Number by Minhyong Kim of Oxford University, Mar 2014
2. Function by Seungyeol Ha of Seoul National University, May 2014
3. Structure by Seok-Jin Kang of Seoul National University, in the fall, 2014
4. Shape by Jun-Muk Hwang of Korea Institute for Advanced Study, in the fall, 2014
5. Counting by Jeong-Han Kim, Korea Institute for Advanced Study, in the fall, 2014

**Acknowledgements.** Alicia Dickenstein wants to thank the ICWM TOGETHER Project from the ICM Korea 2014 Support Program, for the possibility of attending ICM 2014.

Carla Cederbaum, MFO and University of Tübingen, Germany  
E-mail: cederbaum@mfo.de

Alicia Dickenstein, Universidad de Buenos Aires, Argentina  
E-mail: alidick@dm.uba.ar

Gert-Martin Greuel, University of Kaiserslautern, Germany  
E-mail: greuel@mathematik.uni-kl.de

David Grünberg, International School of Lausanne, Switzerland  
E-mail: David.Grunberg@isl.ch

Hyungju Park, POSTECH & NIMS, Republic of Korea  
E-mail: alanpark@postech.ac.kr

Cédric Villani, Institut Henri Poincaré, France  
E-mail: villani@ihp.fr



# The World Digital Mathematics Library: Report of a Panel Discussion

Peter J. Olver

**Abstract.** A summary of the Panel Discussion of the World Digital Mathematics Library held at the 2014 International Congress of Mathematicians in Seoul, South Korea, on August 20, 2014.

## 1. Introduction

The increasing ubiquity of the World Wide Web in the waning years of the twentieth century inspired the vision of a World Digital Mathematics Library (WDML), containing digitized versions of the entire corpus of mathematical research literature, both contemporary and historical, in a distributed system of interlinked repositories. The unique attributes of mathematics, including the eternal validity of mathematical results and constructions, make the WDML especially compelling. More than just a collection of digitized research papers and books, the WDML will include the abilities to search, link, annotate, index, classify, mine, compute, etc., that will form a wide ranging toolbox of applications that incorporate the desirable features but go beyond the current capabilities of MathSciNet, zbMATH, Google Scholar, Wolfram Alpha, etc., and thereby foster the next generation of mathematical research and its manifold applications. Moreover, a commitment to openness, ensuring that the WDML is freely accessible throughout the worldwide research and education communities, lies at the heart of this vision.

The WDML vision was codified by the General Assembly (GA) of the International Mathematical Union (IMU) who, in 2006, endorsed a statement, *Digital Mathematics Library: A Vision for the Future*, [7], of the Committee on Electronic Information and Communication (CEIC) that "... endorses this vision of a distributed collection of past mathematical scholarship that serves the needs of all science, and encourages mathematicians and publishers of mathematics to join together in implementing this vision."

While digitization projects gained momentum and scope in the intervening years [21], while a number of "local" initiatives, such as the European Digital Mathematics Library (EuDML) [8], Math-Net.Ru [16], and several country-based DML's (e.g., DML-CZ [5], DML-PL [20], NUMDAM [17]) have demonstrated proof of (at least some aspects of) the concept, and while some of the required software tools are under active development by a number of groups, both academic and commercial, the overall implementation of a truly Global Digital Mathematics Library has remained tantalizingly out of reach. Nevertheless, several recent developments have rekindled expectations that we may at last have both the means and the will to realize the WDML within the near future. These developments include:

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

On June 1-3, 2012, the CEIC organized a Symposium on *The Future World Heritage Digital Mathematics Library* that was held at the U.S. National Academy of Sciences and involved over 50 participants from throughout the world. The meeting was supported by a grant from the Alfred P. Sloan Foundation. Participants, keynote talks, position statements, panel discussions, breakout sessions and more can all be found on the Conference Wiki [24].

In conjunction with the NAS Symposium, the Sloan Foundation further funded a broad-based committee to write the report *Developing a 21st Century Global Library for Mathematics Research*, that explores the practical mechanisms, challenges, and capabilities that are required for the realization of the WDML. The report was published by the US National Research Council in March, 2014, [6].

On August 17, 2014, in conjunction with the International Congress of Mathematicians in Seoul, South Korea, the IMU and CEIC hosted a meeting of a select group of 21 experts to plan the next practical steps towards the construction of the Global Digital Mathematical Library (GDML). As a result, a smaller eight person working group (WG) under the sponsorship of the IMU, was created. The WG members are Patrick Ion, chair (USA), Thierry Bouche (France), Bruno Buchberger (Austria), Michael Kohlhase (Germany), Jim Pitman (USA), Olaf Teschke (Germany), Stephen Watt (Canada), and Eric Weisstein (USA). The WG is charged with the tasks of designing a road map for the practical next steps towards the GDML, determining its organizational structure, prioritizing the different requirements for its implementation, estimating an incremental budget, both start-up and sustaining funds, and fostering the writing of proposals to funding organizations, with the goal that these next steps will be realized before the end of 2014.

## 2. Statements by Panelists

As an outgrowth of the preceding developments, a Panel Discussion on the World Digital Mathematics Library was organized by the CEIC, and held at the 2014 International Congress of Mathematicians in Seoul, South Korea, on August 20, 2014. The invited panelists were:

- Thierry Bouche, Université Joseph Fourier, Grenoble, France
- Ingrid Daubechies, Duke University, USA
- Gert-Martin Greuel, University of Kaiserslautern, Germany
- Patrick Ion, Mathematical Reviews, USA
- Rajeeva Karandikar, Chennai Mathematical Institute, India
- June Zhang, Peking University, China

The moderator was:

- Peter Olver, University of Minnesota, Minneapolis, USA

To focus the panel discussion, the following questions were circulated in advance to the invited panelists and to the audience.

1. How would you define the WDML?

2. What is the user base that should be targeted?
3. What are the main features expected from a WDML?
4. In existing systems, what would you identify as good practices (respectively bad practices)?
5. How can the community engage with existing stakeholders (publishers, societies, universities, funders) so as to keep the central WDML entity lean enough to be sustainable?

The ensuing discussion is summarized below. The proceedings began with short presentations by each panelist in turn, which were then followed by questions and comments from the audience and responses by the panel.

*Peter Olver* initiated the discussion with some remarks on the vision and history of the WDML, and where we currently stand, summarizing the points outlined in the introduction to the present article. The World Digital Mathematics Library is a vision which started gathering steam in the late 1990's. The IMU signed on as an early proponent of this vision, and a document supporting the WDML prepared by the CEIC was endorsed by the Executive Committee and the General Assembly in 2006, [7]. The basis of the WDML is to have access to the world's mathematical literature, in a *searchable, linked, computable, indexed*, etc., form. A variety of relevant resources and background materials can be found on the CEIC portion of the IMU website, [11]. However, there have been some significant advances and initiatives in the last couple of years, and the intent is that this Panel will help set in motion the next steps necessary to the realization of the WDML.

Various regional initiatives exist, and thus prove, to some extent, the overall concept. However, a truly global digital library remains to be realized. In 2012, the CEIC organized a meeting at the U.S. National Academy of Sciences funded by the Sloan Foundation, [24], which was followed by the writing of a report, issued by the US National Research Council in March, 2014 [6]. The meeting on Sunday at ICM 2014 and formation of a Working Group mentioned above was summarized; see also Patrick Ion's contribution below. The support and publicizing of this initiative by the mathematical world at large is needed, so that we will be able to go to funding agencies, the community, private foundations, etc. to seek the required support. Indeed, the stars seem to be aligning now, and it would be a shame not to take advantage of the moment.

*Ingrid Daubechies* began the presentations by commenting that she is very enthusiastic about the prospects for a WDML, and, while not an expert herself, in the remaining months of her Presidency will continue to foster the realization of the WDML vision.

*Rajeeva Karandikar* then addressed the Panel Briefing questions as follows:

1. The WDML should aim to be a one-stop virtual location for all the needs of a mathematician as far as literature is concerned — where they can search and retrieve archival material, find researchers currently working on similar subjects, or in other areas in mathematics that use research on a theme of interest.
2. Research mathematicians and Ph.D. students in mathematics form the initial target user base. Eventually, this will include scientists/engineers using mathematical results for research and development, students, historians, the educated citizen, etc.
3. The main features expected from a WDML are:

- Search and retrieval of original content *permanently* available.
  - Provide access at either no cost or minimal cost to the end user/institution.
  - Include journals and research monographs; maybe later other books.
  - Provide, to a user, names of people working in an area or who use work in an area.
4. Not enough exposure to comment, presuming that the question is about existing DML's. There is a plethora of bad practices in the publishing world, but we all know them already.
  5. The goal is to enlist mathematics departments to participate in the setting up of a distributed WDML, and thereby avoid major investment in hardware. Mathematicians need to impress on publishers that things are going to change, whether they like it or not. In particular, publishers will have to adjust to much lower revenues. Once the project gets underway and funding agencies see the benefits, we can then persuade them to help fund the WDML activity going forward.

For this to work, we have to digitize older material and bring it into the digital library. We also need to ensure that henceforth the research work funded by public funds does not end up under the copyright restrictions imposed by commercial publishers.

Further remarks:

- I endorse the statement in yesterday's Publishing Panel, calling on all publishers to allow open access to all papers after at most a five-year window, and hand over the material to the WDML to make it accessible to the community. The IMU should start by setting a deadline (e.g., December 31, 2014) for implementation, after which the community will not co-operate with recalcitrant publishers by declining to submit manuscripts, serve on editorial boards, referee papers, etc.
- The publishing metadata should be openly licensed, and not proprietary to the Math-SciNet/Zentralblatt systems as it is at present. The mathematical societies must be convinced to join in the enterprise, and we do not want to lose their support.
- I agree with Ingrid Daubechies' blog post of 2012 that Editorial Boards should become independent societies that then sign licensing agreements with publishers to assume ownership of their journals. It will be hard for one or two individual boards to do this, but acting collectively could put significant pressure on publishers to agree to this new arrangement.
- All new research work should appear on an open archive, under a Creative Commons license. In this way, publishers will no longer have an exclusive control over the material.
- Most of the discussion has been centered on journals, but we should also consider books, especially research monographs. Most authors do not write research books for money and would be happy to put their works in the public domain.

To quote Peter Olver, [18]: "The time is ripe for a radical rethinking of the traditional academic model for scholarly communication within mathematics. . . . If we are not properly engaged, the future will be decided for us and, almost certainly, will not be to our liking."

*June Zhang* introduced herself as a librarian not a mathematician. She gave an introduction to the China Academic Library and Information System (CALIS), a public information

system for higher level teaching and research, which has been funded by the Chinese Government since 1998. Phase 3 (the current phase) is funded with 200M RMB in government support. There are currently 1200+ member libraries and 250+M resources. The principal goals of CALIS are :

- To promote, maintain and improve resource sharing;
- To organize libraries to build the Chinese Academic Digital Library (eduChina);
- To provide information services at a high academic level for teaching and research;
- To extend cooperation nationally and internationally.

CALIS provides seven primary services:

1. Academic search engine (eDu);
2. Inter Library Loan and Document Delivery Services (eDe);
3. Reference Service (eWen);
4. Union Catalogue Service;
5. Foreign Language Journals Service;
6. Imported Database Management;
7. Training and Certification.

The CALIS digital library system structure has four basic layers:

1. Basic infrastructure: hosts, servers, database, storage, security, mid-level software;
2. Resources: knowledge base, databases, reports, courses, bibliographic information, navigation;
3. Applications: including support applications such as data management and service applications such as document delivery, interlibrary loan, full text service, communication service, etc.;
4. Digital Library Portal: designed for the end users, who in turn help decide on the desired portal functions.

Standards run through the whole process of constructing the Digital Library, including meta-data schema, technical schema, and many interoperability standards. Maintenance of data and systems is the most important aspect once the library is constructed. Organization and management is another essential long-term component. All service and development of the Digital Library are governed by the CALIS management group in the administrative center in Peking University. CALIS currently contains:

- 140+M books, including 0.25M Chinese language mathematics and 0.1M other language mathematics;
- 228 Chinese Mathematics journals, 1130 other mathematical journals, and 604 open access journals;
- 46M dissertations, including 32,763 Chinese language mathematics and 16,929 other language mathematics.

CALIS members contribute all their metadata and holdings information. CALIS cooperates with the Korean Education and Research Information Service (KERIS) and with the Japanese National Center for Science Information System (NACSIS).

*Gert–Martin Greuel* began by stating that this is a very important panel for the community, and it needs to address difficult details. The panel should talk about mathematical knowledge, which is a very complex system, but still *primarily* based on scientific publications, although software, databases, etc. also exist. Mathematical knowledge may be considered as a continually growing huge building in which it is necessary that each floor is reached and no stone is lost. Thus, unlike other sciences, in mathematical research the literature plays a very important role due to the timeliness and validity of its achievements, that are mainly preserved in the scientific literature. Mathematical knowledge does not become obsolete, so the entire literature must be available to the research mathematician and user of mathematics. The published research literature is the principal component and should be the starting point of a WDML, but this is not all and we should be open to other types of resources in the future.

Further points:

- Much of the literature has been digitized, but this has been done by different sources, commercial and non-commercial, to different standards, and with different licensing conditions.
- We need to think of the architecture for the WDML, which is not just a repository of digitized mathematics, but also semantic tools will be needed later on in more and more sophisticated ways.
- Stephen Wolfram noted that this is a \$100M project, and this cannot be done by volunteers. We have to come to grips with the issues of long term maintenance.
- We also have to think about content, e.g. border sciences such as biology, computer science, etc.
- Searching, computing, indexing, linking, and the like will be important.
- There are many commercial copyright interests, who will not give their content for free. We have to engage them, and the best we can hope for is a reasonable moving wall.
- Archiving requires longer term evaluation and design, as well as long-term funding.
- We have to cope with growth of the literature: zbMATH now records 120K items/year; arXiv has an even steeper growth rate with no flattening seen yet.
- High quality and specific metadata, with mathematical search options is important, semantic content analysis, author disambiguation and author profiles.

Recall the IMU/CEIC requirement: Each article should include a separate list of references with links to the indexing databases Mathematical Reviews and Zentralblatt. Mathematical Reference Databases have several advantages:

- provide identifiers for the indexed mathematical literature;
- ensure completeness of the mathematical literature;
- have high quality and well-structured metadata, as well as math-specific search options;

- are restricted to the mathematical literature, and hence have little extraneous noise;
- include semantic content analysis (MSC, keywords, abstract, reviews);
- provide linking of information, e.g., full texts if available, references, etc.;
- provide author disambiguation and author profiles.

Mathematical Reference Databases are also engaged in the development of necessary tools for the WDML and hence can provide core services for the WDML.

- deciding continuously what to index, i.e., what is math literature?
- development of metadata schemes for mathematical publications;
- maintenance of the Mathematics Subject Classification (MSC);
- form pilot partners for the use of the methods for publishing and presenting mathematical knowledge, e.g., use of MathML as presentation format.

*Thierry Bouche* began with a quote by Jean–Pierre Serre:

Mathematicians just make their results available to everyone as if they were on shelves waiting to be fetched.

Note that digital libraries don't currently have the good features (well-organized, permanent, etc.) of paper libraries. Digital power tools can, and should, be opening new paths for research and serendipity.

Mathematical validated literature never becomes obsolete. (Old results are not superseded by newer ones: they are their foundation, full proofs are sometimes never written twice!) The mathematical literature is valid only as a whole, building a wide network of references, and useful to other sciences in an asynchronous fashion. The mathematical corpus is the set of all (potentially) referenceable published works. It must be carefully archived, indexed, and preserved, and must be widely accessible over the long term.

We thus need a reference library, which should be

- comprehensive
- up-to-date
- well organized
- long lasting
- widely open
- easy to use for non-mathematicians
- and digital, with power tools opening new paths for research and serendipity!

The European Digital Mathematics Library (EuDML) has finished a three-year 1.6M euro funded project, which has now become a follow-up unfunded consortium under the auspices of the European Mathematical Society (EMS).

In a nutshell, we produced

- A critical mass in content, approximately 6% of the mathematical corpus.
- A cooperation network.

- A math-savvy fully functional digital library, with MathML metadata, math mining, MSC, links to/from math databases.
- A good looking Web site with unique navigation tools adapted to our user community.
- Internal and external deep interlinking, MSC browsing, reference lookup.
- A number of productivity and interoperability devices enabling the main service, some production ready, some more experimental.
- EuDML initiated organizational model and policies, under the strong control of science through EMS.

The three-point EuDML policy is that the content be:

1. Scientifically validated, and published in final form;
2. Physically hosted at one of the partner institutions;
3. Openly accessible after a reasonable moving wall: 0–5 years.

We now understand the basic layer, and scaling from 6% to 30% of Mathematical Corpus is at hand. But the most pressing demand from mathematicians is 100% of the Corpus. The gaps are items that are not digitized, or not professionally digitized, missing item-level metadata, inability to harvest existing metadata, or metadata that the content owner does not agree to provide. All these are surmountable, but require some effort — technical, legal, political — and support from the community [4].

To ensure stability, we need a distributed and replicated physical archive. Correctness matters to us, so it is of high importance to let intelligent agents generate derived mathematical knowledge that would be machine readable for enhancing service (spontaneous crowdsourcing, OCR, structure/semantics recognition, . . .) but this process should be transparent to the users and never hide the original, unmodified sources. We are talking of an infrastructure for research that will be the daily working tool of mathematicians worldwide. What is really needed is long-term institutional support just like your university or department library today!

*Patrick Ion* just retired from near 30 years at Mathematical Reviews. He was heavily involved in  $\text{\TeX}$ , MSC, etc., and served as co-chair of the W3C Math WG that produced the standard MathML, that is now a recognized part of HTML5.

At the first ICM in 1897 there was a session under the chairmanship of Peano concerned with questions of how to encode mathematical knowledge, [22]. Indeed it was in connection with such efforts that Peano developed his axioms for the natural numbers. At the 1928 ICM in Bologna there was active discussion of how to provide comprehensive bibliographic resources for mathematics to everyone [1]. Now the IMU sees the possibility of realizing the current dream of a Global Digital Mathematical Library or World Digital Mathematical Library.

The adjective digital is important here as it is the new digital technologies that allow better access to the resources of mathematical knowledge than ever before. We are in the presence of a transformative technology, and we can capitalize on it to everyone's benefit. I can imagine that in 16th century Europe, or even earlier in 14th century Korea [10], when printing from metal type was a brand-new technology, people saw the possibilities of the new forms of book for the recording and dissemination of knowledge. That they were right we all now know. That sort of opportunity is open to us again now.

The adjective global is important too. We all think the truths of our subject to be global, independent of location in this world. We think of ourselves as a world-wide community. This is well demonstrated by our being gathered at the ICM from over 120 countries. A GDML can have a global reach as a result of the digital technology mentioned, particularly the internet. It will be a shared global good. We can hope for global support for the idea and expect that there can be contributions to a GDML from all over the world. It will provide benefits all over the world. The earliest mathematical artifact some think to be from Ishango in Congo, [9]; perhaps this GDML can be a help in Africa. But as Adrian Paenza emphasized in his Leelavati Prize lecture at ICM 2014, [12], the main goal has to be to offer solutions to problems that the people you serve want solved. We think of ourselves as a world-wide community, so a GDML can have a global reach, thanks to digital technology.

IMU President Ingrid Daubechies and Chair Peter Olver of the IMU's CEIC took the initiative to work toward a WDML or GDML through consultations with a broad expert group. This culminated in comprehensive report from a Workshop at the US National Academy of Sciences, [6]. Now a small working group of eight persons, which I am to chair, has been given the task of making, by the end of this year, concrete proposals for work setting up a GDML. Then resources can be found, so to speak, to virtually break ground on building a GDML.

The GDML WG represents a variety of backgrounds and interests and is about as international as 8 people can be, if where their careers have carried them is taken into account. They are united by a belief that there are opportunities for building a GDML to serve the mathematical community and disseminate mathematical knowledge as widely as it is needed, and by a wish to make that happen starting now. The WG will of course be calling upon the expertise of the community, about the square of 8 in size, that Ingrid Daubechies and Peter Olver have been consulting, as well as on many others. The WG's activities will be reported on through the IMU's CEIC web site and we, of course, will be happy to hear from the community of ideas for services a GDML may provide and what problems it may solve. We expect that realizing a GDML will naturally involve both the academic and industrial mathematical communities and collaboration with those who have served it well for a long time — very importantly the publishing business world-wide. The WG's goal is to get GDML projects defined and started in comparatively short order.

I see essentially four facets to the GDML initiative:

- Community aspects;
- Literature aspects — relatively well-understood after EuDML's efforts, but essential;
- Knowledge management aspects — less well-understood;
- Administrative aspects — June Zhang has described the scale.

They are all discussed in the NRC report. The WG is to make concrete what's suggested there on all four fronts.

Some parts of a GDML require work that is understood, or already done in part, but that just takes much time and effort to complete. Other parts require serious investigation and prototyping which also takes time, even nowadays, although the general ideas may seem clear. The WG is made up of members who think now is the time to realize the new opportunities for a GDML. We all believe that now is the time to *do* this.

### 3. Audience Questions and Comments

Following the presentations by the panelists, the audience was given the opportunity to ask questions and make comments.

Marie Farge, France, began by stating that we have given our copyright away. There are different traditions under a variety of national laws: copyright, author right, etc.. The kind of copyright agreements we sign with some publishers are illegal under French law. So we should talk to good lawyers on an international level about this issue, and in this way we can put pressure on the publishers to release the copyrights. Ingrid Daubechies answered that this was duly noted, and that we should involve a very good international lawyers early on.

Mina Teicher, Israel, asked how far back are you going to go — 19th century, 18th, 16th? Several on the Panel answered in principle as far as we can. We should start with the more easily accessed material, e.g. older uncopyrighted material should be more readily available.

Alexey Ustinov, Russia, remarked that many of the digital libraries in Russia, which are very good, are illegal! Gert-Martin Greuel emphasized that everything the WDML does has to be legal.

Gerhard Paseman, USA, asked about the political implications of a WDML. The publishers are very good at lobbying the government, and might influence the funding of mathematics were such a digital library to affect their business. Ingrid Daubechies replied that in some countries, including US, there are legal obligations to ensure that publicly-funded research is publicly-accessible. Countries are passing laws to ensure this. Paseman replied that that doesn't address the issue. Suppose we had such a library — how would that affect government funding in the future? If the AMS sees a decrease in their revenue as a result, they may ask that their government only funds those researchers who published with them. Daubechies said the publishers have not seen a decrease in revenue from implementing a moving wall, and Gert-Martin Greuel said that publishers may well profit from implementing a moving wall. Moreover, once a digital library is in place with structured metadata, which is the hard part, and basic services, publishers will be able to make use of that to enhance their own web pages, offer additional services, etc. Peter Olver added that representatives from the US National Science Foundation (NSF) and the German Deutsche Forschungsgemeinschaft (DFG) attended the 2012 NAS Workshop, and were very supportive of the effort.

Gizem Karaali, USA, said that the Mathematical Association of America (MAA), as the result of an initiative by the educational directorate of NSF, created a mathematics digital library, Math-DL, as a component of the National Science Digital Library. Her question was if the WDML is mainly geared towards research, how do you see your work combining with educational needs? Rajeeva Karandikar replied that in the long term the WDML would include educational components, but in the short term the emphasis should be on research. Gert-Martin Greuel brought up the issue of digital identifiers, such as DOI. Lack of digital identifiers for everything makes extending the WDML into educational material much more of a challenge. Marie Farge remarked that something along these lines has already been started in the US.

Thomas Banchoff, USA, former president of the MAA, stated that the MAA has placed all of its publications on JSTOR with a moving wall. JSTOR treats mathematics education as seriously as mathematics research. He then asked how JSTOR fits into the WDML initiative. Gerhard Paseman noted that he has to pay for JSTOR. Peter Olver said that a danger is that JSTOR recopyrights material in that it charges users to access its corpus. Thierry Bouche said that, given the extent of its holdings, if it weren't subscription-based, JSTOR would be

close to what we want. On the other hand, it has no mathematics specificity, so that all the enhancements we envision related to mathematical knowledge do not exist on JSTOR. He also remarked that, since the tragic 2013 death of Aaron Swartz, JSTOR has modified its copyright procedures. Patrick Ion added that JSTOR is very successful at many things. The original mission of JSTOR was to avoid having to construct more library buildings, and the Mellon Foundation decided that the \$90 million it provided in funding was the cheaper alternative to the library shelf space people were demanding. However, the mission of JSTOR is very different from the WDML, and the mathematics that they have, while very good, is rather inaccessible, even to those with good subscriptions. While they have a good business model, they aren't a simple model for a GDML. Banchoff asked if this is going to complicate the WDML plans. Ion replied that he didn't think so because JSTOR journals can be redigitized or access can be negotiated with JSTOR or publishers.

Marie Farge asked that, since you want to make it free, what's your business model? Peter Olver answered that, as Stephen Wolfram said in Sunday's meeting, the WDML is a public good, not a business, and so there is no business model. We do, however, need a sustainability model. Thierry Bouche asked what is the business model of your institution's library? There is a lot of money going to university libraries that could go towards the WDML. Gert-Martin Greuel added that this is a difficult question. Having accurate and complete metadata that is secure, archived, etc., could be of interest to commercial publishers, who could help support the WDML consortium to help enhance their own services. Governments could use it to evaluate their own researchers. But we do not want to require Universities to buy subscriptions. Ingrid Daubechies said that there is currently a lot of money going into assembling and providing access to literature through libraries, but there are major challenges for the administrative and community aspects. There are certainly money streams that could be used when it makes sense to everyone. Marie Farge then added that having free access will boost industry, and their demand for mathematics and mathematicians, as well as aiding retired mathematicians and young students.

An unidentified questioner asked whether you can look at other models that exist, iTunes, Napster etc., which are successful in their own way. Patrick Ion agreed and pointed to music indexing services, such as MusicBrainz.org. Peter Olver added that, while the discussion has concentrated on getting access to the mathematical literature, the WDML is more than just a collection of papers, books, etc. and will include applications allowing one to search for theorems and ideas, determine whether something you found is related to results in an old paper in a different area, various indices of concepts, theorems, formulas, and so on.

Another unidentified questioner then asked how do we follow up on today's Panel. Peter Olver announced that a blog concerning the WDML will be hosted by the CEIC, starting soon after the end of the Congress, and available on the CEIC website [14]. The initiative will require community input and community buy-in. Indeed, without the support of the mathematics community, the WDML will not succeed.

A questioner from Mumbai, India, asked whether the WDML will have some form of quality control. Many open access journals take money from authors to publish and are sheer junk and will publish anything people pay for. Mathematicians shouldn't have to pay for publishing since it brings corruption into the system. The WDML should set up some form of standards. Thierry Bouche replied that this is not an easy question. In EuDML, there is an advisory board linked to the European Mathematical Society which makes an effort to enforce quality standards, but this is not necessarily perfect. Gert-Martin Greuel added that zbMATH and Math Reviews do check for quality every year when deciding which journals

to index. This is difficult, but needs to be done. The questioner then asked whether we can put a stop to author-pays journals. Greuel said that zbMath stops indexing substandard journals. Peter Olver recommended the IMU/CEIC-produced document on best practices for journals [3]. This concluded the proceedings.

#### 4. Conclusion

As a result of the Meeting and the Panel, the IMU has initiated a WDMML blog [14], to provide a forum for ongoing discussion of the emerging WDMML. A key conclusion of the panel discussion, coupled with contemporaneous developments, is that there is now real potential for significant near-term progress on the realization of the WDMML.

**Acknowledgements.** The moderator is grateful to James Davenport for taking notes, to Thierry Bouche and Patrick Ion for help with the text, and to the ICM 2014 Team for recording the panel session, which can be found on YouTube [13]. A permanently archived version of this video will appear subsequently on the IMU website.

#### References

- [1] R.C. Archibald, *Plans for reviving Bibliotheca Mathematica*, in: *Atti del Congresso Internazionale dei Matematici*, Bologna 3-10 Settembre 1928, vol. 6, pp. 473–474; see <http://www.mathunion.org/ICM/ICM1928.6/>
- [2] J. Ball and J. Borwein, *ACCESS: Who gets what access, when and how?*, MSRI Digitizing Mathematics Workshop, 2005. <http://www.mathunion.org/fileadmin/CEIC/Publications/MSRI.pdf>
- [3] Best Current Practices for Journals, International Mathematical Union, 2010. <http://www.mathunion.org/fileadmin/CEIC/bestpractice/bpfinal.pdf>
- [4] T. Bouche, *The Digital Mathematics Library as of 2014*, *Notices Amer. Math. Soc.* **61** (2014) 1085–1088.
- [5] Czech Digital Mathematics Library, <http://dml.cz>
- [6] *Developing a 21st Century Global Library for Mathematics Research*, The National Academies Press, 2014.
- [7] Digital Mathematics Library: A Vision for the Future, International Mathematical Union, 2006, [http://www.mathunion.org/fileadmin/CEIC/Publications/dml\\_vision.pdf](http://www.mathunion.org/fileadmin/CEIC/Publications/dml_vision.pdf)
- [8] The European Digital Mathematics Library, <https://eudml.org>
- [9] [http://en.wikipedia.org/wiki/Ishango\\_bone](http://en.wikipedia.org/wiki/Ishango_bone)
- [10] <http://en.wikipedia.org/wiki/Jikji>
- [11] <http://www.mathunion.org/ceic/resources/icm-2014-panels/>
- [12] <http://www.youtube.com/watch?v=Te4iO6jkuG0>
- [13] <https://www.youtube.com/watch?v=OERXmv2oIyU>
- [14] The IMU WDMML Blog, 2014–, <http://blog.wias-berlin.de/imu-icm-panel-wdml/>

- [15] A. Jackson, *The Digital Mathematics Library*, Notices Amer. Math. Soc. **50** (2003) 918–923.
- [16] Math-Net.Ru, <http://www.mathnet.ru>
- [17] NUMDAM: Recherche et téléchargement d’archives de revues mathématiques numérisées, <http://www.numdam.org>
- [18] P. J. Olver, *Journals in Flux*, Notices Amer. Math. Soc. **58** (2011), 1124–1126.
- [19] J. Pitman and C. Lynch, *Planning a 21st Century Global Library for Mathematics Research*, Notices Amer. Math. Soc. **61** (2014) 776–777.
- [20] The Polish Digital Mathematics Library, <http://pldml.icm.edu.pl>
- [21] U. Rehmann, DML: Digital Mathematics Library, [http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml\\_links.html](http://www.mathematik.uni-bielefeld.de/~rehmann/DML/dml_links.html)
- [22] E. Schröder, Über Pasigraphie, ihren gegenwärtigen Stand und die pasigraphische Bewegung in Italien, in: *Verhandlungen des ersten Internationalen Mathematiker-Kongresses in Zürich vom 9. bis 11. August 1897*, pp. 147–162; see <http://www.mathunion.org/ICM/ICM1897/>
- [23] Some Best Practices for Retrodigitization, International Mathematical Union, 2006. [http://www.mathunion.org/fileadmin/CEIC/Publications/retro\\_bestpractices.pdf](http://www.mathunion.org/fileadmin/CEIC/Publications/retro_bestpractices.pdf)
- [24] Symposium Wiki for The Future World Heritage Digital Mathematics Library, held at National Academy of Sciences, 2012. <http://www.wias-berlin.de/imu/archive/WDML/>

University of Minnesota, USA  
E-mail: [olver@umn.edu](mailto:olver@umn.edu)



# Why STEM (Science, Technology, Engineering and Mathematics)?

Jean-Pierre Bourguignon, Ingrid Daubechies, Myung-Hwan Kim, and Youngah Park\*

## 1. Introduction

This is a brief note of the Invited Panels “Why STEM” held on the 18th of August, 2014. The meeting was chaired by Youngah Park, President of Korea Institute of Science & Technology Evaluation and Planning (KISTEP). The panelists were:

- Ingrid Daubechies: President of the International Mathematical Union (IMU)
- Myung-Hwan Kim: President of the Korean Mathematical Society (KMS)
- Jean-Pierre Bourguignon: President of the European Research Council (ERC)

## 2. Why STEM?: Introductory Presentation

Youngah Park gave an overview of the session and raised topics for discussion. Written below is the introductory presentation she gave, which covers the importance of STEM education, global interest in STEM education, and so on.

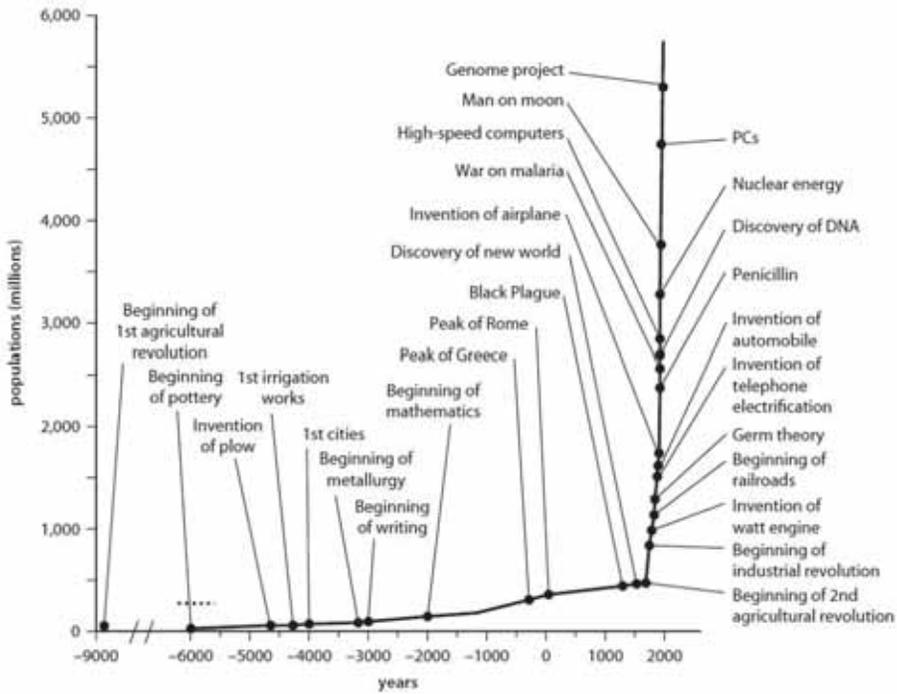
Human history has been advancing with the progress in the areas of science and technology as well as in mathematics. For millennia, humans were unable to escape from the Malthusian Trap, limited by crop yields in increasing the population. Then, the emergence of incredible ideas and inventions in science and technology triggered rapid improvement in economic productivity, providing freedom from the Malthusian Trap and allowing the realization of human development. The era of concentrated human life improvement after the 1800s corresponds with tremendous progress in modern science and technology. The consideration of these events shows how crucial developments in STEM were to the immense improvements in living standards.

South Korea is experiencing similar progress. Korea was one of the poorest countries in the world 60 years ago. With the end of the Korean War in the 1950s, Korea did not possess any S&T knowledge or infrastructure. However, starting from the light industry exports, such as wigs and shoes, Korea has acquired enough knowledge to develop heavy industry and ICT industries. Today, we are finally at the forefront of high-tech industries. Also, Korea has been able to act as a developed nation within the international community, hosting the ICM and

---

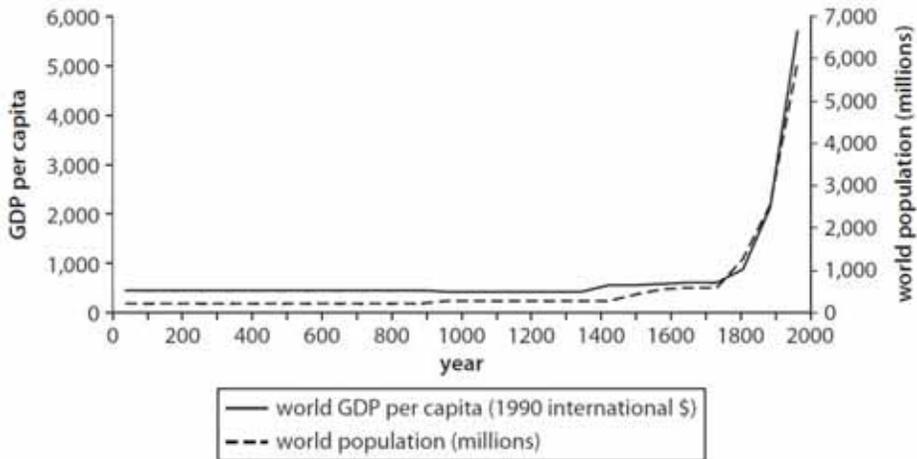
■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

\*Moderator



Source: Commission on Growth and Development, 2008

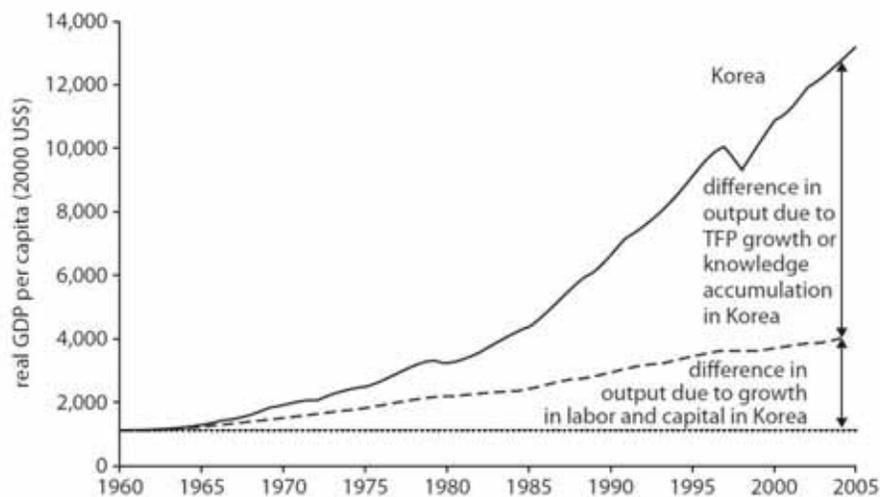
Figure 1. World Population Growth and Major Technology Events, 9000BC to Present



Source: Maddison, 2006

Figure 2. Growth in Population and GDP per Capita in the Past 2,000 Years

inviting many prestigious mathematicians. In allowing such developments, much research attests to the central role of investment and knowledge acquisition in science, technology, and engineering. Figure 3 also reveals the fact.



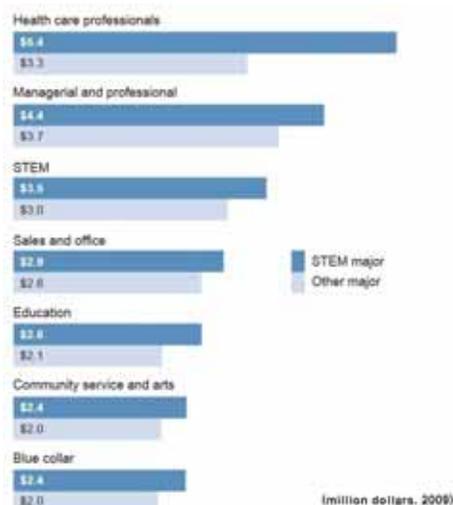
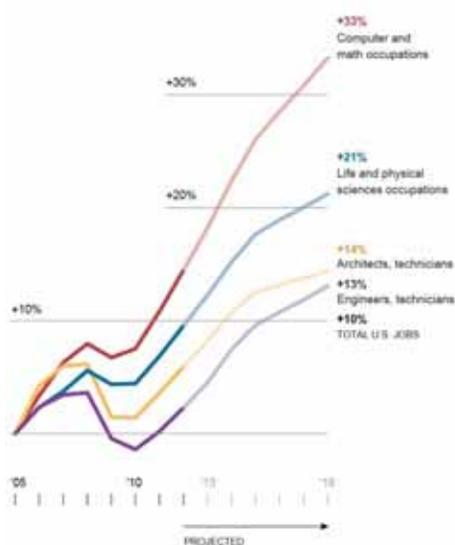
Source: World Bank, 2007

Figure 3. How Innovation Contributes to Growth in Republic of Korea, 1960-2005

Specifically, how important is mathematics in light of contemporary social value? Mathematics is perhaps perceived as distant from our everyday lives, and it may be difficult to observe its economic and innovative value. However, mathematics, without doubt, creates diverse opportunities for industrial advancement. For instance, Captain Davy Jones, from the “Pirates of the Caribbean” film series, comes to life through computer graphics, rather than just special effects and makeup. With this cutting-edge technology, the “Pirates of the Caribbean” has grossed almost USD 5.4 billion worldwide from 2003 to 2013, which is an amazing achievement because it is so close to Samsung Electronics’ gross sales in Q1 2014. Without this type of mathematical tool, the “Pirates of the Caribbean” would not have been created, and we would not be able to enjoy the fruits of the modern movie industry.

We cannot overlook the story of Prof. Stanley Joel Osher. He was the winner of Carl Friedrich Gauss Prize in ICM 2014. His achievements could explain how useful mathematical applications are in our daily lives. Prof. Osher has outstanding achievements in shock capturing, level set methods, PDE (Partial Differential Equation)-based methods in computer vision, and image processing. These have diverse applications in MRI scans, medical image analysis, computer chip design, weather forecasting, and so on. Also, he is particularly famous for catching criminals using image restoration technology. Mathematics contributes not just to the growth of the economy but also to the improvement of quality of life and the resolution of social problems.

The economic progress and societal gains from STEM developments are the reason why we need to emphasize STEM education. In addition, STEM education can ensure the better lives for future generations. As an example, the United States has high expectations for STEM job growth in the fields of computing, math, life science, physics, engineering, and so on. In terms of percentage change in jobs, computer- and math-related occupations are expected to grow by 33% and life and physical sciences-related occupations has increased by 21% from 2005 to 2012, and projections through 2018. This is really strong growth compared with the total U.S. jobs, which is estimated at 10%, because the demand for specialized labor in such



Source: The Promise of Science and Technology Work, The New York Times, December 7, 2013

Figure 4. High Expectation for STEM Job Growth test Figure 5. High Income Expectation for STEM Major

fields will increase according to the direction of societal development. Furthermore, STEM majors have been found to earn much higher incomes than other majors. In case of health care professionals, employee who majored in STEM area earns USD 5.4 million in average during a whole lifetime while other major earns USD 3.3 million. This is because of the higher value creation ability and innovative capacity of STEM major workers. To be precise, they are likely to have high labor productivity, logical ability, and analytic skill.

As the world realizes the significance of STEM education, each nation is beginning to establish relevant policies and promote investment in STEM education. For example, to reaffirm America's role as the world's engine of scientific discovery and technological innovation, President Obama is placing an emphasis on investments for STEM education. In 2009, the U.S. also started the "Educate to Innovate Campaign." In the launching event, President Obama made a speech that states:

*Reaffirming and strengthening America's role as the world's engine of scientific discovery and technological innovation is essential to meeting the challenges of this century. That's why I am committed to making the improvement of STEM education over the next decade a national priority.*

*President Obama. November 23, 2009*

This campaign has garnered over USD 700 million in public-private partnerships. The campaign has leveraged the unique capacities of the private sector, cultivated effective STEM instructors, bolstered federal investment in STEM, and broadened participation to inspire a more diverse STEM talent pool. More specifically, the U.S has developed a lot of STEM education programs, which aims to raise the efficacy of STEM education, such as Transforming Undergraduate Education in STEM (TUES) of National Science Foundation (NSF),

ITEEA EbD (Engineering by Design), Curriculum of International Technology and Engineering Educators Association (ITEEA), Project Lead The Way (PLTW) Program. Another representative nation that emphasize on STEM education is the United Kingdom. The United Kingdom is supporting STEM education using the Science & Innovation Investment Framework to boost investments. The government's overall ambitions include improvements in the quality of science teachers in all education institutions, student performance in science, numbers of students choosing STEM subjects in higher education, and the proportion of qualified students who pursue careers in the R & D sector. They also have developed STEM education programs, such as the National HE STEM Program. It started since 2009 with GBP-21 million budget and supports higher education institution in the exploration of new approaches to recruit students and deliver programs of study within the STEM disciplines. In Korea, the government is currently promoting the creative economy as a principal national policy agenda and STEM education is receiving much attention as the means of fostering a creative labor force. President Park of Korea is emphasizing the importance of cultivating creative human resources in realizing the creative economy, and thus announced "The 2nd Basic Plan—Fostering and Supporting S&T Human Resources." There are some relevant issues in STEM education with Korea's situation. In Table 1, according to the Program for International Student Assessment (PISA), Korean students demonstrate high performance in math and science. In math, they ranked first out of all OECD countries and placed between second to fourth in sciences in 2012. Korean students invest a significant amount of time in their studies, thus achieving remarkable levels of academic achievement in PISA.

Despite the highest level of performance in math and science, Korean students have weak motivation to perform mathematical activities purely for the joy gained from the activity itself. They don't believe that they can successfully perform in mathematics, given academic tasks at designated levels, and possess a high level of anxiety about their performance in math exams. These findings have important implications for the direction of future STEM education. Although the Korean educational system has succeeded in raising the performance of students by teaching advanced math and science, it has decreased student enjoyment.

The Ministry of Education of Korea is presently in the process of reforming the high school curriculum under the philosophy of integrating liberal arts and science track for convergence education. According to the reform plan, the credit points of science subjects will

Table 1. OECD Program for International Student Assessment (PISA) Results: 2000-2012

		PISA 2000 (41)	PISA 2003 (40)	PISA 2006 (57)	PISA 2009 (75)	PISA 2012 (65)	
Math	Average	547	542	547	546	554	
	Rank	OECD	2	2	1~2	1~2	1
		Total	3	3	1~4	3~6	3~5
Science	Average	552	538	522	538	538	
	Rank	OECD	1	3	5~9	2~4	2~4
		Total	1	4	7~13	4~7	5~8

\* (): number of countries / Source: OECD

Table 2. PISA Index of Affective Characteristics out of 35 OECD Countries (selected areas)

Index		Rank of Republic of Korea
Motivation to Learn Mathematics	Intrinsic Motivation	29
	Instrumental Motivation	32
Mathematics Self-belief	Mathematics Self-efficacy	34
	Mathematics Anxiety	4

Source: OECD

be reduced to 10 credits points from 15 credits points. Also, the proportion of science classes is reduced to 10.8% from the current 15.1% of all class hours. Moreover, considered that science classes consist of physics, chemistry, biology, and earth science, this reform plan poses a threat to the logical thinking abilities of our students and the creativity of our society. To cover the diversity of STEM subjects, students spend a lot of time learning to reflect the contents, but the Korean government is pursuing reforms in the opposite direction. The most important thing is to look at these issues as a whole and to strive for general and acceptable ideas regarding the STEM education policy.

The topics for discussion were as follows.

1. Impact of mathematics on the advancement of science & technology and national competitiveness
2. Importance of STEM education to middle and high school students
3. Policy to develop students' problem-solving ability

### 3. Panel Discussion 1: Ingrid Daubechies

If you look at the achievement of student and the economic growth in a country, there is a very high correlation. Also, I believe that the investment on education through which students would learn mathematical concepts and mathematical knowledge is extremely profitable for a country. I don't have direct logical corollary between them. However, if you start losing math, you will lose some of the benefits. I think that Korea has been one of the best cases where we see enormous growth based on very solid math and science education. Korea would be a country where skeptical question would come up much less than elsewhere, so the situation Dr. Park showed us surprises me.

In the U.S., students in high school have enormous freedom to organize their curriculum. We try very hard to convince students who are smart, especially those who come from maybe less advantaged conditions or don't already have a lot advantages from their family background, to take science and mathematics education, which is one of the best ways for social mobility. My son is a high school teacher in an institute. He is delighted when he manages to convince smart students to take more mathematics, which is very useful for them. If they have enough knowledge in mathematics, they could have many choices in college and they could consider many options in terms of careers afterward. I'm not even talking about the choice of doing mathematics and research.

Mathematicians and researchers constitute a minority of the population. However, because we have knowledge that can preserve and raise our society, it is essential to communicate with the next generation.

#### 4. Panel Discussion 2: Myung-Hwan Kim

Korean War broke out in 1950 and lasted for three years. During the three-year span, the tragic war thoroughly devastated the Korean peninsula. In 1961, eight years after the ceasefire, the GDP per capita was roughly 80USD - less than a quarter per day! Korea at the time was among the poorest nations in the world.

From the late 1960's or early 70's, Korea started transforming itself from an agricultural country into an industrial country. By 2012, Korea became the seventh member of the 20-50 Club with GDP over 20,000USD per capita and population exceeding 50 million. This success is referred as the 'Miracle of the Han River'.

Many experts say that Korean's emphasis on education is behind this amazing economic growth.<sup>1</sup> Even during the war, tent schools were operational, which symbolizes the Koreans' passion on education. Korean parents believe that education makes their children's lives successful. They make every effort to provide for their children's education, which sometimes is criticized as obsession for their children. This fervor for education, however, is admired and praised by many people, particularly from outside Korea.

Korean kids spend considerable length of time on studying mathematics. The reason is simple. Mathematics is one of the most important subjects for high school students preparing university entrance exams. Students who are not good at mathematics have little chance to pass entrance examinations for top universities in Korea. As a result, the average math-skill of Koreans is among the highest in the world. This is proved by the results in the PISA Test, in which Korea ranked in the top five every time.

In the International Mathematical Olympiad which is an annual competition among mathematically talented kids from all over the world, Korea also ranked in the top five most of the time. This is an indication that Korean kids perform extremely well also in the highest-level competitions in problem-solving mathematics. Identifying and training of these mathematically gifted kids is conducted by the Korean Mathematical Society funded by the government.

In Korea, many talented students today seek to study mathematics instead of ever-popular medicine or engineering. Korean mathematics community is excited with this encouraging news and at the same time is faced with an important task and a noble duty of guiding these students toward a bright future through mathematics. This is the bright side of mathematics education of Korea.

On the other side, however, there are lots of problems in mathematics education as well.<sup>2</sup> Some examples are:

1. Most students study mathematics only for the purpose of entering universities.
2. Students are exposed only to those problems that can be solved in three minutes. Speed counts the most.

---

<sup>1</sup>Professor Hanushek, a renowned economist at Stanford University, affirmed at the first MENAO Conference that there is indeed a close relation between the economic growth rate and the average years of schooling.

<sup>2</sup>Science education of Korea suffer more or less the same problems.

3. Since Korean SAT is a multiple-choice test, students learn how to exclude wrong answers rather than to find right ones, and keep practicing to reduce mistakes.
4. It is one of big social concerns that parents should pay extra educational expense to send their children to private tutoring schools.
5. Many students give up and hate mathematics in the early stage of schooling and later become antagonists of mathematics.

The worst problem appears to be the frequent change of high school curricula. The Ministry of Education keeps changing the curricula with big changes in every five or so years and small changes nearly every year. In this year again, the ministry the ministry of Education prepares a reforming plan for high school curricula. The science and engineering community is concerned with the anticipated plan, since the ministry is trying to reduce the number of credit hours for compulsory mathematics and sciences courses. The science and engineering community in Korea oppose the plan, since it firmly believes that reducing the credit hours for math and science can never be an effective solution but results in much worse problems. What should be taught in high school should be taught there.

The Ministry of Education together with mathematics community should make every conscientious effort to find ways of reducing the number of students giving up mathematics in their early stage of education. STEM and/or STEAM seem to be an excellent means of achieving this goal.

### **5. Panel Discussion 3: Jean-Pierre Bourguignon**

The teaching of mathematics and science at the school level has an irreplaceable role in exposing students to problem solving, as well as making them autonomous in thinking. This requires time and a personal engagement by students so that they feel at home with the approach. Teachers have to be trained to make such an appropriation happen smoothly. A key issue is of course the sequential nature of the teaching of these disciplines, which is stricter than in other subjects taught in school. One must indeed base what is learned in a given year on what has been already learned, and hopefully properly understood, the years before.

The use of advanced technology in the classroom is inevitable as it is so wide spread in the society. It needs to be seconded with appropriate pedagogy to help students with calculations and more generally mathematical exercises, in which guessing and estimating a priori results plays an important role to check what is proposed by the machine. It also offers new dimensions that allow for experimentation with some well-chosen mathematical problems, something that would be very time consuming or just impossible without such tools. This can bring a more stimulating and certainly less repetitive environments for the study of mathematics in class, and is likely to appeal to the students favouring a less formal approach to the learning of the discipline. To be widely successful, such an approach will also require teachers to be trained accordingly.

One must also take the information provided by international tests such as PISA for what they are. For now a number of years, Korean students have been doing very well at these tests, being regularly at the top of this international benchmarking, showing the efficient job done by teachers in Korean schools. From what I understand some concern is now raised by the fact that, if many Korean students are doing extremely well at these tests, the number of those doing less well and also having developed some apprehension with the learning

of the disciplines may be increasing. Such a situation, that should indeed be given serious considerations (my own country, France, has neglected such a message for too long), needs to be addressed, but this can certainly not be done not by lowering the standards which will affect the performance of all Korean students, including the very good ones, but by giving proper attention to the students that need to be accompanied appropriately. It is likely that proposing various approaches to the learning of the discipline, without diminishing its ambition, will help in dealing with this question.

Last but not least, in modern societies where the role of information and knowledge is increasing, mathematics contributes in many more ways than 20 years ago to the economic development in industry AND the services. Actually, this point is even overly underestimated by mathematicians themselves. One of the consequences of this fact is that, in the work force, the number of people who have to feel comfortable with mathematics to deliver what is expected from them in their job has grown considerably. This is due to the central position taken by structuring and dealing with information and data in general in many different areas of economy. This is especially true in the context of the « creative economy » that has become a major objective for the Korean government. This new perspective gives one more reason not to diminish the role and place given to the teaching of science and mathematics in schools. A reduction could have serious negative effects on the competitive edge that the Korean economy has built in the last 20 years, that is based on a very competent and dedicated work force.

## 6. Contributions from the Floor and Answer from the Panelists

**Audience 1.** I was very interested in proposed reforms in Korea. I have seen similar reforms in several countries in Europe. I also think that these reforms are probably a very bad metaphor. The reality is that if we decrease the difficulty of students in STEM education, we can increase the number of students who are willing to do it. However, the knowledge will remain constant as a whole. But this is wrong and I have figured out that sometimes, decreasing in the level of education make decreasing a number of students who choose STEM as their main subject. As such, we must be very careful to do this reform, and I believe that it doesn't work. The second thing that wasn't presented here is about the open choice for curriculum. Is there any reason why people think open in choices is a good idea? In some cases, it makes very bad social results because some students come from good environment and they know very well what to choose. On the other hand, many students don't know how to make consistent education. Being open for choice could affect the difference among diverse social classes.

**Audience 2.** The content and the level of mathematics education are going down, and this could be very special examples of a few countries, including Korea and France. However, this is a very striking example because of the history in those countries, and it's clear that we want to change the situation. The panelists gave information that 30% of the recent advancement of economy is based on science and technology and mathematics. This is a very good argument for STEM. However, let's consider the goal of the Ministry of Education and the Ministry of Finance or policy makers in general. If we say STEM is so important, then they would say "Okay, mathematics is important and investment is needed on it, but we don't have to create everything here. Also, not every country has to do everything for the economy. We can get

fruits of science and technology from other countries, like medicine. Medicine is found and developed in the United States, and it is being used all over the world. It could be the same for mathematics. STEM is so important, but we don't necessarily have to own it. If it is elsewhere, we can use it.

**Jean-Pierre Bourguignon.** If you look at the economy worldwide, several major companies have formed in the last 20 years. These are a completely new kind of company. Here in Korea is a good example for the situation. Most of Korean global companies suddenly developed very high-level technology and they are extremely successful. Samsung was not a huge company. Now, it's really one of the dominant companies in the world. And the way it keeps this position is just by hiring the best people. They have done very well. Korea needs to have the potential of having a significant role in the next important new areas. Nobody is able to predict where these areas will come from.

Actually, the remark I have made to journalists to whom we spoke before coming here was about a recent conversation I had with the person-in-charge of research in one of the great multinational companies dealing with transportations, real physical networks such as water distribution and so on. He mentioned that, globally 8% of the engineers working in the company are what you can call mathematical engineers and the target for the company in 2025 was that this percentage should reach 20%. They believe that the new developments will come from managing data and services nurtured by data, leading to advising states, regions and many organizations. Who are the people going to collect and analyze these data? Basically, mathematicians and statisticians. There's no physics in it. There's no chemistry in it. It's a different kind of activity and this new kind of activity will become a very dynamic economic sector.

I am closing with this mention. We all have to try and know where the new things would come from and who is going to deal with them. It's not clear which country is going to do it. Maybe it would be a completely different country. But there is great need to be reactive, to have trained the right people, waiting to be involved in new profiles jobs.

**Ingrid Daubechies.** I'd like to add one more thing to the mention what Dr. Bourguignon made. In the U.S., it has been very successful in attracting young people from elsewhere. More and more, U.S.-born students or students from second or third generation in the U.S. are opting not to go into science and technology but a lot of science and technology-trained students from elsewhere come to the U.S. and stay. If this reverses last, I think it will be catastrophic for the U.S. because they will not have their own reserves. You can't follow the path of the U.S.

**Youngah Park.** Yes. In Korea, we used to go abroad, but now we are making creative people by ourselves. Also, we are thinking about attracting many foreign scientists and engineers from the outside, especially from Asia. It's because we believe that our future will depend almost entirely on creative people like scientists and engineers.

**Audience 3.** I want to ask for the panelist's opinion on effect of smart media on STEM education. Here in Korea, when you go to the metro, you can see everybody has a smartphone in his hand at a young age, and I was wondering whether there is any correlation between familiarity with smart media and other soft skills. This development in information technology contributes to personal ability or education? In Europe, we eventually discussed whether

it's good or bad for people to have so much information. It seems that they never concentrate seriously on longer problems because they are used to read short articles and get simple information.

**Myung-Hwan Kim.** I think we have to find some way to help education sector to use these information technologies. I don't know if there is any successful system in other countries, but I think it's a good idea to use this as an instrument for teaching. It would be the best harmonious way that we should bring.

## 7. Concluding Remarks

We believe that "Invited Panel: Why STEM?" provided a lively forum for networking and discussion for all participants from mathematics society, industries, and policy area. We can say that this event was a good occasion for fruitful discussions on the STEM education. It gave us many implications: correlation among mathematics, science & technology, national competitiveness, importance of STEM education to secondary education, policy for students' problem-solving ability, and so on. All of the comments from panelists and audiences will be helpful to make future education policies and the STEM education itself.

The President of KISTEP, Youngah Park, invites those who had not chance to express opinion to send a letter to [ypark@kistep.re.kr](mailto:ypark@kistep.re.kr). She is gratified to all the panelists and organizers and hopes that there would be more places for discussion in the near future.

Jean-Pierre Bourguignon, Institut des Hautes Études Scientifiques, France

E-mail: [jpb@ihes.fr](mailto:jpb@ihes.fr)

Ingrid Daubechies, Duke University, USA

E-mail: [ingrid@math.duke.edu](mailto:ingrid@math.duke.edu)

Myung-Hwan Kim, Seoul National University, Republic of Korea

E-mail: [mhkimath@snu.ac.kr](mailto:mhkimath@snu.ac.kr)

Youngah Park, KISTEP, Republic of Korea

E-mail: [ypark@kistep.re.kr](mailto:ypark@kistep.re.kr)



# Mathematics is everywhere

Eduardo Colli, Fidel R. Nemenzo, Konrad Polthier, and Christiane Rousseau\*

**Abstract.** “Mathematics is everywhere” was the title for a panel at ICM 2014. The four panelists discussed what can be put under this title, what are the messages that can be passed to the public, and how to pass these messages. To most mathematicians, it seems obvious that mathematics is everywhere, and a living discipline within science and technology. Yet, how many of them are able to convey the message? And, when most people look around, they do not see mathematics, they do not know about the mathematics underlying the technology, they know very little about the role of mathematics in the scientific venture. Can we help building a powerful message? Can we unite forces for better passing it?

**Mathematics Subject Classification (2010).** Primary 00A05; Secondary 00B10.

**Keywords.** Popularization of mathematics, Mathematics a living discipline, Mathematics is everywhere.

## 1. Introduction

*“Mathematics is everywhere.”*

Suppose you are a mathematician and you are put in front of this statement. Are you convinced? If you are in front of your classroom, are you able to explain the statement? And if you are in front of the public, or of a journalist, what examples will you choose to illustrate the statement? We, the panelists, have the impression that many of our mathematician colleagues are convinced. Yet, many of us lack good examples to pass the message. Indeed, the message should first please us before we decide to transmit it.

Let us now go to the schools. The teachers all know in principle that mathematical education is important but how can they answer the question “What is mathematics useful for, nowadays that calculators can do the computations, for us and that software solves the problems we used to learn to solve by hand?” How many teachers can take you to a tour of the city and show you the maths in all modern gadgets that you use, from a parabolic antenna, to a GPS, to the architecture of a building and the synchronization of traffic lights?

If we now go to the public and claim that mathematics is everywhere, we could expect many skeptical faces. . .

Mathematics and its creative role in science, technology and society deserves to be better known. On the mathematical community side, this could result in more support of the society to mathematics, more interest of the kids in schools for their mathematics courses, and more

---

■ Proceedings of the International Congress of Mathematicians, Seoul, 2014

\*Moderator

interesting jobs for mathematically trained graduates. On the other side, everyone benefits from the contribution of mathematics to a better organization of our society, including public health system, management of resources, organization of transports. And mathematical breakthroughs in technology contribute to the creation of high technology companies.

This brings us to the gaps within our communities: gaps between mathematicians and other scientists, between pure and applied mathematicians, between researchers in mathematical sciences and mathematics educators. Perhaps the deeper gap lies between what mathematics is and how the public, including politicians and policy makers, sees it.

A challenge for the mathematical community is to convey the beauty and value of mathematics, that “mathematics is everywhere”. Technology — used by people everyday — is always an opportunity to explain to the public the power of mathematics. More difficult is to convey the idea that abstract ideas too are beautiful and important and that a brilliant idea can make a breakthrough. At the most basic level, there is the need to restore everyone’s (especially young people’s) sense of awe and wonder.

Mathematicians should also be at the forefront of efforts to communicate mathematics, and bridge the gaps within the math community and society. International collaboration can make a significant difference. Preparing the right material for communicating mathematics requires energy and is time consuming, especially when it is hands-on. Also, anyone of us is always limited by his(her) own taste of mathematics. Putting the material on line allows sharing resources and using material on many different topics. Translating existing material can enrich significantly the material accessible in a given country. But it does not suffice that material be on line for it to be used. . .

## 2. Mathematics is everywhere : Christiane Rousseau

This title is an extraordinary slogan. If you are not convinced, let us give a few examples.

Mathematics is everywhere in technology. Without mathematics, there would be no CT Scan. Indeed, a CT Scan only gives a series of numbers, namely the quantity of energy absorbed along the different rays through the body inside a plane, i.e. the Radon transform of the image, and the inverse Radon transform allows recovering a 2D image from these scattering data. Now, open your computer: Google’s algorithm, which is so efficient, relies on the stationary distribution of a Markov chain: a clever idea created an empire. The small files for the images that you see on the Web have been compressed using Fourier transform in JPEG format, or wavelets in JPEG2000. Sensitive data is encrypted using number theory or algebraic geometry. And your computer is built with transistors that are sophisticated switches: building software ultimately comes to decomposing any operation into parallel sequences of elementary operations on 0 and 1. More examples in [11].

Mathematics is everywhere in science. Already for more than two thousand years, mathematics has evolved closely with physics, finding in physics a source of problems and providing solutions to physical problems. More recently, mathematics increased relationships with other sciences, and especially biology. The spectrum of applications in biology is immense, from the functioning of the cells or groups of cells and organs including the brain, to the functioning of the full body, with all the potential medical applications. Other types of applications include interactions of living populations, spreading and control of infectious diseases, ecology and ecosystems, and how biodiversity is organized on the planet.

The international year *Mathematics of Planet Earth 2013* (MPE2013) had a very impor-

tant outreach component with the goal of showing the many applications of mathematics to, on one hand, discovering, understanding and managing our planet and, on the other hand, helping facing the planetary challenges of climate change and sustainability. The unprecedented collaboration around this international year is certainly explained by the increasing awareness among the public and the scientists that the planet is in real trouble, and that mathematics has a role to play in this issues. The theme is much wider than climate change and sustainability. Putting mathematical “glasses” we can discover the interior of the Earth by analyzing seismic waves generated by large earthquakes. Studying the planetary motions inside the solar system allows explaining the past climates of the Earth, and also the chaotic behavior of the inner planets, from which we cannot exclude, either a collision between two planets or expelling one planet from the solar system. The MPE2013 website ([www.mpe2013.org](http://www.mpe2013.org)) provides resources for enriching the curriculum and for outreach activities. This makes it easy for any teacher or professor to address such themes in secondary school, or even in undergraduate education. It is remarkable that MPE2013 occurred with almost no budget. This highlights how collaboration can significantly increase the impact of our outreach activities.

As mathematicians we work too much in isolation, and we should join forces with scientists to pass the message. I went in December 2013 to the Fall meeting of the American Geophysical Union and attended an education session. I was very impressed by the nice material that was presented, including numerical simulations: simulation of the formation of the Gran Canyon, simulation of a climate model in which the user could change the parameters, etc. Such material could easily enrich the curriculum in several of mathematics courses at the undergraduate level.

Having led MPE2013 since its inception in 2009, I was surprised by the number of mathematicians and teachers of mathematics who looked first excited by the theme and then disoriented if they had to produce examples. I would start listing a few applications simple to explain: how to calculate the length of the day depending on the season and the latitude, how a sundial works, provided that you correct time through the equation of time, how to draw a map of the Earth, how the GPS works, how to model the spread of an epidemic, etc. After a while, my vis-a-vis could continue the game and provide new examples I had not thought of. This means that even the convinced people need help to find good answers to questions like “What is mathematics useful for?”, and “Has everything been found in mathematics?” Yet, these very important questions deserve significant answers. An answer could start with “Mathematics is everywhere.”, then list a few applications where mathematics are hidden, before you continue with explaining the mathematics of your favorite application.

**An example** I am a passionate of popularization of mathematics. I like to discuss examples and present strong scientific messages out of them. I love powerful ideas which are unifying in science. One of them comes from Turing’s seminal paper “The chemical basis of morphogenesis”[14]. It is the idea that the loss of stability of an equilibrium through diffusion creates patterns. Take a flat stretch of dry sand and let the wind blow: there will always be a small irregularity that will stop some grains of sand, starting the beginning of a dune. Since the wind blows regularly, sand deserts are never flat, but rather covered with dunes. The same occurs with waves on the lakes and oceans, and with snow sastrugies in Antarctica. Turing idea’s was that morphogenesis in biology has a chemical origin, with reaction diffusion phenomena involving several chemical reactants. Initially, the embryo has spherical symmetry, and the loss of stability of this equilibrium through diffusion leads to the

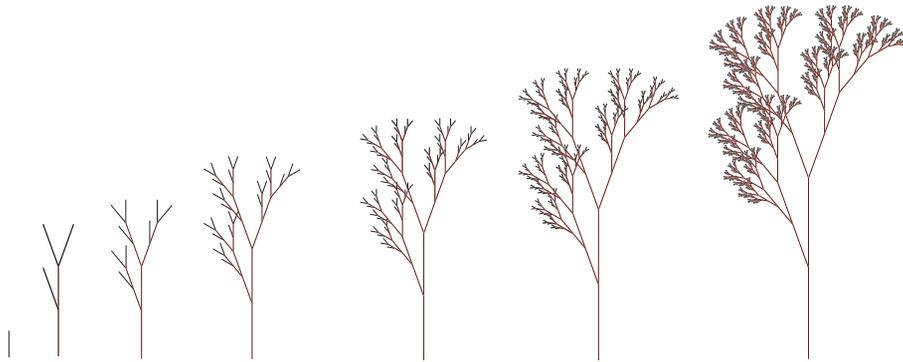


Figure 1. An example of  $L$ -system with two rules. On the figure, appears the initial condition, a segment of type  $F$ , and the first six iterations. The symbol  $S$  corresponds to a stalk piece, and the symbol  $F$  to a terminal branch. The first rule corresponds to replacing a terminal branch by a stalk of length 2 and three terminal branches. It is given by:  $F \mapsto S + [F] - S + [S] - -[S]$ . The second rule doubles the length of the stalk:  $S \mapsto SS$ .  $\pm$  corresponds to a rotation of a given angle to the left or to the right, and the notation [branch] means that we must come back to the beginning of the branch before starting the next instruction.

formation of limbs, nose, ears, etc. The same model has been introduced in many areas from phyllotaxy when leaves appear along the stalk to the growth of plants modeled by  $L$ -systems ([7]). The  $L$ -systems have been introduced by the biologist, Aristid Lindenmayer. They model the growth of complex plants by iteration of a small number of operations, similar to iterations of a few instructions of a cellular automata (see Figure 1).

Reaction-diffusion models have also been proposed for modeling the fractal patterns appearing on some seashells ([8]), and for the patterns of animal coatings (see for instance [10]). A discrete version of the reaction-diffusion model can be given for the pattern of the *Cymbiola Innexa* REEVE (see Figure 2), which resembles a lot the Sierpinski carpet. The pattern is generated one line at a time, similar to the ridges of a real shell. There are two reactants: the activator (A) is colored and given the value 1, while the inhibitor (I) is white and given the value 0. If one divides the image into pixels, then the pattern is formed iteratively row by row. The color of a pixel is the sum modulo 2 of the two pixels that touch it by the corner on the preceding row.

The reaction-diffusion model used to describe the patterns of animal coatings is especially interesting. The same model allows four different types of patterns (see Figure 3):

- spots;
- labyrinths;
- gaps;
- stripes.

For animal coatings, the patterns which appear depend only on the size and shape of the surface at the time of the pattern formation. In particular, stripes usually occur on thin tubular regions. It is especially striking that these four types of patterns are exactly the ones observed in vegetation patterns. Vegetation patterns occur when there is not enough water for full vegetation cover. Above a certain threshold of moisture, vegetation can survive.

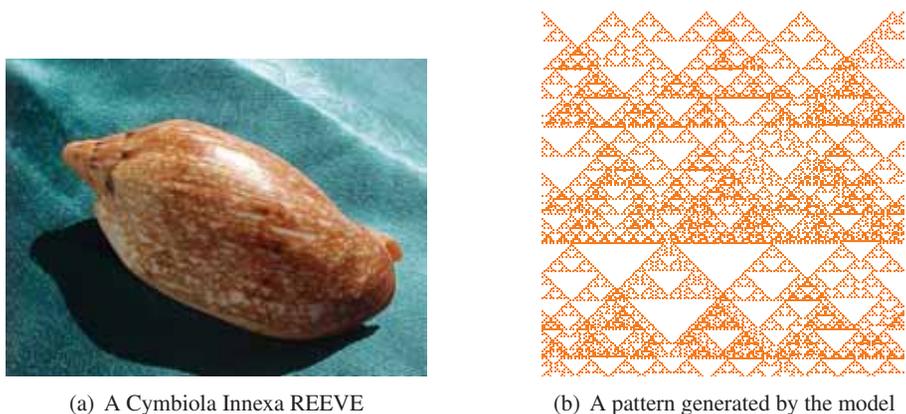


Figure 2. A *Cymbiola Innexa* REEVE (Photo credit: Ian Holden, Schooner Specimen Shells), and a pattern generated by computer, using the discrete reaction-diffusion model described.

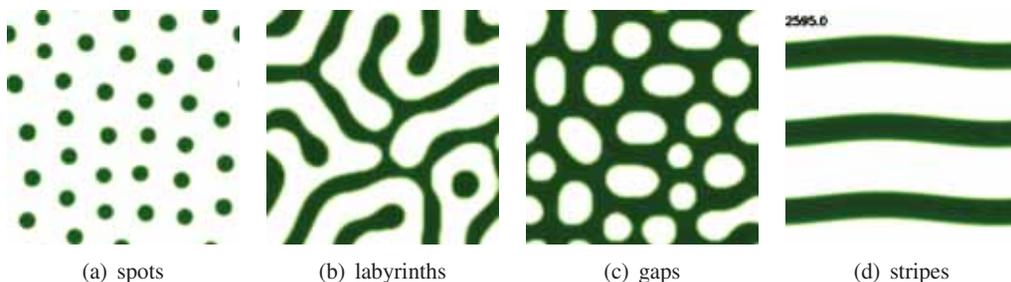


Figure 3. The four types of patterns (images provided by A. Provenzale from papers [3] and [4]). The stripes are perpendicular to the slope.

Spots are observed when the moisture is minimum, then labyrinths, then gaps. Stripes occur on slopes.

In chemistry, the most famous example is the Belousov-Zhabotinsky reaction-diffusion, an oscillating reaction creating regular patterns before moving to a chaotic behavior.

More recently, the model has also been used in the context of ecological invasions, spread of epidemics, tumor growth and wound healing. James Murray has been particularly active in all these applications.

### 3. Using math-glasses in Seoul – a walk in the city with eyes and mind tuned in on mathematics : Eduardo Colli

The slogan “math is everywhere” is commonly used in math popularization and is generally accepted and understood by those who teach mathematics. But when one is faced with the task of convincing someone that this assertion is true some philosophical issues arise: what exactly is meant by “everywhere”? What does it mean “is”?

I think that it is more or less a common sense that “everywhere” in real life is something near the concept of a dense set in mathematics. We cannot escape, in our urban life, of seeing

matters which are closed interrelated with mathematical concepts. But on the other hand it would be rather pretentious to claim that every aspect of life or of human life is related to mathematics: think about a conversation between two old friends, telling each other how about their lives are in that day. Or think in a film, or in a romance, or in a play. Can we say there is math there? I would answer “no, at least based on what we know in the present days”.

Moreover, what do we mean by saying “there is math here or there”? In fact, there are some different ways of relating mathematics to what we see in real life. We give three.

1. One is by means of an application, understood as some technique which is developed by human beings in order to solve some problem. The GPS technology is an often cited example: it is based on the mathematical assertion that in three-dimensional Euclidean space if we know our distance to four points (the satellites), and these points are in generic position (in this case, there is no three of them lying in the same line and the four do not lie in the same plane) then our position is determined. In fact, three satellites suffice if the second solution can be discarded by other means. The implementation of this principle also uses more advanced mathematics, since relativity theory must be taken into account for precise measurement of these distances.
2. Another one is by means of relating a phenomenon to a mathematical model, like a body in free fall. Nature is full of examples of patterns arising from simpler rules that generate them. Even if this hypothesis is controversial, I assume that mathematics *is* in these patterns.
3. A third one is when some human being gets inspiration in mathematical concepts to bring some kind of delight to real life. This is common in architecture, design and art – see for example the Brazilia Cathedral, which is a hyperboloid of revolution, or the catenaries present in Gaudí’s work, but it may also be the case in the creation of games – like Hex – and puzzles – like Rubik’s cube. In these cases math appears purposely but with no intention of solving a problem.

These three categories are not completely disjoint. For example, an object of design may seem to be inspired in mathematics, but the designer itself created it without being aware of its mathematical relations. So the mathematics of the object appear as a model of the object, as it was created by nature. Other example is when a good mathematical model of a phenomenon can prove itself useful to applications only much later than its discovery – for example Newton laws of mechanics and gravity to explain the movement of planets and much later the launch of satellites and spaceships. Here the mathematics that we attach to the phenomenon brings the seed of a technical application.

At the end, mathematics is both language and science, it is both tool and inspiration, it is application and abstraction. It is inherent to the human being and that is why it is everywhere.

**The walk** It is endowed with this way of seeing – that we call here “math-glasses”, a term already used elsewhere – that we will take a quick walk through Seoul. The reader will easily recognize the three above mentioned ways of seeing “mathematics everywhere” and I invite him/her to wear math-glasses from now on in every situation possible.

I won’t reproduce pictures of the visited places, but rather indicate in the footnotes the internet sources where they can be found. One may also use keyword searches using the

names of places and buildings. It should be clear that this text was written several months before ICM, so I will be honest to say that in fact I did not do a true walk, but an ‘internet walk’ in Seoul.

**Modern architecture** Our first stop is in front of the (new) Seoul City Hall.<sup>1</sup> This building was opened in 2012 and was conceived by Yoo Kerl. It seems that mathematics have inspired the artistic realization of Kerl’s motivations – see the composition with triangles, the approximation of an oval surface by polygons and the curved surface in the facade, resembling a graph of a two variable function,  $z = f(x, y)$ , with  $z$  in the horizontal direction, perpendicular to the facade.

But apart from the direct mathematical inspiration of this particular building, it calls our attention to the incredibly powerful tool that mathematics, together with computers, brought to architects, designers, engineers and other with CAD – Computer Aided Design. These softwares are the uttermost modern application of geometry, particularly analytic geometry, linear algebra and curve/surface approximation and interpolation techniques. Nowadays it is impossible to walk in an urban area, indoors or outdoors, without seeing buildings and objects that were first drawn in some software like these.

**East Asia ancient architecture** Our next stop is at the hall of Changdeokgung Palace, a typical example of East Asia ancient architecture.<sup>2</sup> I was astonished to find that mathematics is also used to master the conception of new buildings in this style, also taking into account the differences in style of Japanese, Chinese and Korean cultures. In [13] and references therein it is discussed the procedural modeling of this kind of construction through CAD softwares where the user simply chooses some parameters and the basic structure of walls and roofs is automatically drawn following a prescribed algorithm.

In this case, human beings are modeling the creation of other human beings and, as a result, establishing a path to preserving cultural heritage.

**New branches of geometry thinking: packing** If we have children with us, why not spending some time in a candy store, with its plastic cylinders fulfilled with colored candies? Those who wear math-glasses easily recognize matters related to sphere packing. Although children are mostly interested on colors, tastes and textures, the seller could be particularly wondering whether one of his/her packing cylinders lets less or more empty space accordingly to the size of the balls inside.

Reasoning in simple terms shows that we may not expect great differences as a function of the ball sizes, since locally the ratio between empty and filled space is the same, independently of the size, as long as the balls have the same size. Of course we have to neglect the effects of the wall, but they are small when the balls are much smaller than the cylinder radius.

The sphere packing problem in infinite three-dimensional Euclidean space has challenged mathematicians for three centuries with the so called Kepler conjecture, stating that  $\frac{\pi}{\sqrt{18}}$  is the best density possible for equal spheres. A computer-assisted proof has been provided by Hales and Ferguson, and they are working on a proof that could be checked by an

---

<sup>1</sup>I like the picture of Minseok Kim’s Blog, at [linkwind.blogspot.com.br/2012/11/new-city-hall-of-seoul.html](http://linkwind.blogspot.com.br/2012/11/new-city-hall-of-seoul.html)

<sup>2</sup>There is a nice picture at [en.wikipedia.org/wiki/Changdeokgung](http://en.wikipedia.org/wiki/Changdeokgung)

automatic proof checker program.<sup>3</sup>

**Parabolas or not?** Night is falling and we go take a look at the Moonlight Rainbow Fountain, at Banpo Bridge.<sup>4</sup> We immediately identify the parabolic shapes of the jets. Are they really parabolic? If there was no air resistance, they would be.

But suddenly each jet goes out with different slope.<sup>5</sup> What would we see if we were placed at the beginning of the bridge, looking sideways at the jets? (see Figure 3). We assume that despite the different slopes, the water goes out always with the same velocity. The jets fulfill a portion of the sight plane, which is bounded by a curve: the envelope. What curve does the envelope give? The answer is surprising: it is also a parabola!

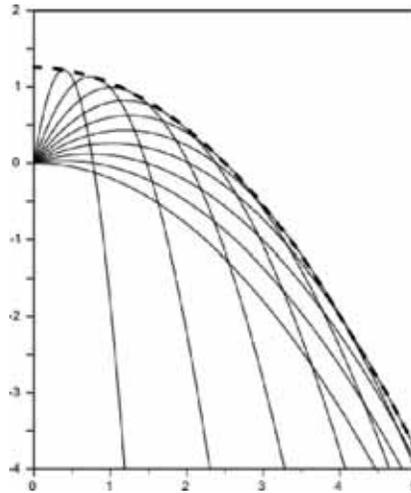


Figure 4. Parabolic fountain jets viewed from sideways. Each jet is launched at the same velocity  $V_0$ . The dotted line shows the envelope

The introduction of air resistance complicates things a little bit and trajectories are no longer parabolic. If it was a thrown object, a good model for air resistance is a force proportional to the square of the absolute velocity and opposite to the velocity vector. Therefore the parametrization of the trajectory would be the solution of the second order ordinary differential equation  $(\ddot{x}, \ddot{y}) = (0, -mg) - \alpha \sqrt{\dot{x}^2 + \dot{y}^2} (\dot{x}, \dot{y})$ .

Notice also that each jet can have its own color, an effect that can be obtained injecting light aligned with the water outlets. Then light is imprisoned in the jet, which is a very interesting effect of total reflection, when the angle of the incident ray is so small that there is no refraction, a simple consequence of Snell's law. This principle is used in the construction of optical fibers.

**Traffic jam** Now it is time to go back to the hotel, but traffic conditions are not favorable. Why are there traffic jams even in open roads without any traffic signals or car accidents?

<sup>3</sup>See [en.wikipedia.org/wiki/Kepler\\_conjecture](http://en.wikipedia.org/wiki/Kepler_conjecture)

<sup>4</sup>[en.wikipedia.org/wiki/Banpo\\_Bridge](http://en.wikipedia.org/wiki/Banpo_Bridge)

<sup>5</sup>In the link [spreadsosomeawesome.com/2011/09/29/han-gang-river-cruise-in-south-korea/](http://spreadsosomeawesome.com/2011/09/29/han-gang-river-cruise-in-south-korea/) one can see a picture where the jet slopes vary sinusoidally

The flow on a road can be modeled in many ways and one of them is supposing it is a continuous (see [1], for example). This continuous model has three time-and-position-dependent relevant variables: flow  $q = q(x, t)$ , in cars per second, concentration  $k = k(x, t)$ , in cars per meter, and speed  $v = v(x, t)$ , in meters per second. These three variables are related by two simple equations, in such a way that it suffices to study only one of them. The space and time evolution of this remaining variable is then ruled by a partial differential equation, that can explain wave effects of the traffic. This equation is an example of conservative laws, which deserves a whole mathematical area on its own.

A group in Japan [12] shows experiment and analysis with real cars in a circular lane (a video recording can also be seen). From the video it is clear that, at least for a high concentration, constant concentration is an unstable state.

**Subway queues** Giving up to get a taxi drive in the traffic jam, maybe the subway is a good alternative. But subways have queues in the platforms [15]. People arrive at the platform of train line obeying to some probabilistic distribution law, in general something like a Poisson distribution, that says the probability of  $k$  people arrive between two train departures. On the other hand, there is a matrix assigning probabilities of getting out at a given station given that the passenger got in at some other station.

With enough data the subway administration could then calculate the probability of having a queue exceeding the capacity of the platform, avoiding dangerous situations. Or they could carry on a careful study about the ideal time lapse between two trains.

**A hands-on 3-sphere** I am tired and not willing to face these queues. Better spending time at a coffee shop, playing with a *souvenir* that I have found in some trinket shop, the magical folding cube (Figure 3).<sup>6</sup>



Figure 5. Magic folding cube

But I am still wearing my math-glasses. I start wondering what if this object was a kind of spacecraft of 8 chambers (the smaller cubes), each face of the chamber with a door that opens only when there is another chamber at the other side of the wall. This situation is achieved by a suitable articulation of the spaceship that makes the walls touch. There is no door to the outside, only doors that connect one chamber to the other. An astronaut in this spacecraft, where does he/she live?

<sup>6</sup>Taken from [www.gyroscope.com](http://www.gyroscope.com)

The answer is: he/she lives in a 3-sphere! The 3-sphere is the only compact orientable simply connected manifold of dimension three, but this we know only now, after G. Perelman has proven a conjecture of Poincaré that stood unanswered for more than one hundred years!

**Conclusion** In this one day walk we saw geometry in its pure form or related to modern aspects like interpolation and packing; ordinary and partial differential equations; statistics and probability; and topology. Further exploration can go much further than I did in the above paragraphs. The examples could be not only directly used in classroom, but also be a source of inspiration in the task of seeking out day-by-day examples.

#### 4. Bridging gaps and communicating mathematics : Fidel Nemenzo

In the late 50s, the British intellectual C.P. Snow, in his famous lecture ‘The Two Cultures’, lamented the fragmentation of the learning in the academe and the widening gap between the humanities/social sciences and the sciences. His lecture drew attention to the failure of specialists to communicate their ideas across the boundaries of their disciplines.

There are gaps too within our mathematical community, brought about by specialization and the different ‘languages’ we speak. There is the gap between ‘pure’ and ‘applied’ mathematics, between research mathematicians and mathematics educators.

But perhaps the deeper gap lies between those who do and teach mathematics on one hand, and those who use mathematics— everyone! This is a gap between what mathematics is and how the public sees it. The lack of understanding and appreciation of mathematics on the part of the public has grave implications on our students’ education, our school systems and government policy. In some countries, funding support for mathematics and the basic sciences is dwindling, in favor of the applied disciplines which are seen to have more ‘direct’ benefits to industry and society. This underscores the need to communicate the beauty and power of mathematics to the broadest possible audience and convey the message that mathematics is part of almost every aspect of our lives.

Like ordinary language, mathematics allows one to represent and communicate ideas and meanings. It has been described as the language for the study of patterns about quantity, space and shape, and structure. Mathematics is abstract, but because of its precision, it is the language of science, helping us model and understand the natural and physical world and providing the ideas that power modern technology. In fact, it is also increasingly becoming part of the language for understanding and modeling social phenomena in a diverse range of disciplines such as economics and sociology.

Technology — ubiquitous and used by people everyday — gives us an opportunity to explain to the public the powers of mathematics.

Take for example, error-correction codes. Human beings are equipped with the ability to detect and correct errors, and thus we are able to read and correct the following corrupted message: “*you can raed a taotl mses wouthit a porbelm. This is bcuseae the human mnid deos not raed ervey lteter by istlef, but the word as a wlohe*”. Word processing programs too have error correction: MS Word can convert the mistyped word ‘*mathemaitsc*’ to the correctly spelt ‘mathematics’, by identifying the corrupted word with the ‘closest’ item in its collection of legitimate words. This implies the use of some notion of ‘distance’ between words.

It is a safe assumption that errors occur whenever we transmit data or information across

‘noisy’ channels. Data can be in the form of text messages, digital images, sound or movie files, etc. Coding theory is the mathematics behind the packaging of information so that we are able to efficiently transmit the information, and detect and correct the errors. The idea is to construct abstract mathematical objects called ‘codes’ which are used to represent the data one wishes to transmit. Good codes are equipped with algebraic structure and some notion of ‘distance’ that allows efficient error detection and correction. The traditional codes are constructed as vector spaces over finite fields. But in the last two decades there has been growing interest in codes over finite rings. There is also a class of codes built from algebraic curves over finite fields.

Originating from the works of Shannon and Hamming during the mid-20th century, and clearly motivated by the requirements of engineering and information technology, research in coding theory draws ideas from many fields of mathematics, such as number theory, ring and field theory, linear algebra, combinatorics and geometry. Coding theory is the mathematics working behind the scene in many of the gadgets and machines we use everyday — such as mobile phones, CD and DVD players, etc.

Prime numbers — the building blocks of the integers — are both fascinating and mysterious. They have pleasing properties that are a source of delight for school children, amateur number enthusiasts and professional mathematicians. They also baffle, and give rise to many open problems in mathematics. One particular problem is computational: given an integer, decompose it as a product of prime factors (prime factorization). There is no known efficient algorithm for this, and the difficulty of prime factorization serves as the basis for the security of well-known public key cryptosystems, such as the RSA cryptosystem. Cryptosystems are methods of encrypting (and decrypting) information for secure transmission.

Another mathematical object used in cryptography is an elliptic curve, whose properties can be used to construct cryptosystems with higher security and shorter keys. Elliptic curves are smooth cubic curves whose points are endowed with some neat algebraic structure (over the rationals, a finitely generated abelian group) and arithmetic. Although they have been studied as abstract objects for the past 150 years, elliptic curves have some surprising applications. Andrew Wiles’ proof of Fermat’s Last Theorem (1995) is based on the modularity of elliptic curves. About thirty years ago, elliptic curve cryptosystems were introduced as an alternative to the standard RSA-type methods ([6], [9]). A point on a curve (defined over finite fields) can easily be multiplied by an integer, but it is very difficult to compute the number, given the original point and the result. The security of elliptic curve cryptosystems is based on the difficulty of this computational problem.

Everyday, most people use and enjoy the benefits of technology such as their mobile phones, digital cameras and the internet, unaware these run on the power of mathematical ideas.

More than 50 years ago, the mathematician G.H. Hardy once rejoiced in the ‘uselessness’ of number theory, “whose very remoteness from ordinary human activities”, he said, “should keep it gentle and clean.” [5] Today, number theory — the study of numbers, such as prime numbers — is no longer seen as ‘useless’. Like coding theory, it works behind the scenes in the internet, securing our email and financial transactions, authenticating sources of data, and ensuring the safe passage of information. This is the irreversible trend in mathematics — the divisions between the pure and the applied are breaking down. Abstract ideas in mathematics, developed for their own sake, are now finding new applications. And conversely, problems in physics, IT and engineering, the physical and biological sciences, are stimulating new research in mathematics.

While the main task of mathematicians is to do mathematics, they should share, with educators, the responsibility of promoting the right attitude towards mathematics, among students, the popular media, our governments, and the broader public. A well-informed public is necessary for a public culture that is supportive of mathematics. There are too many biases, fears and misconceptions out there about our discipline that should be dispelled in all possible arenas, including social media. Mathematicians should be at the forefront of communicating the delight, beauty and power of mathematics as both language and tool, and dispelling the stereotype of a mathematician as a performer of mental acrobatics, cut off from society. An educated public need not have a grasp of equations and formulas, but should understand the role of mathematics in shaping our world.

## 5. Conclusion

The different contributions have highlighted the challenges facing our mathematical community. On the one hand, the suggested title of this panel, “*Mathematics is everywhere*”, is a fascinating slogan that would deserve to be exploited more often, both in terms of strong messages in science and technology, but also as a game to be played when we look around us and dismantle what we see to discover the mathematics hidden in so many objects or phenomena around us. On the other hand, the image of our discipline still needs improvement, and our community is facing challenges mainly in terms of communication, collaboration between communities, and preparation of messages, resources and material to communicate. We should join forces to improve the situation, especially considering the fact that, with the web, it is easier than ever to share resources and spread the message to the largest public possible.

**Acknowledgements.** Eduardo Colli was supported by Grant #2014/08628-1, São Paulo Research Foundation (FAPESP). Christiane Rousseau was supported by NSERC in Canada.

## References

- [1] John A. Adam, *X and the City. Modeling aspects of urban life*, Princeton and Oxford: Princeton University Press, 2012.
- [2] Robert B. Banks, *Towing icebergs, falling dominoes, and other adventures in applied mathematics*, Princeton, New Jersey: Princeton University Press, 1998.
- [3] Gilad E., von Hardenberg J., Provenzale A., Shachak M., and Meron E., *Ecosystem engineers: from pattern formation to habitat creation*, Physical Review Letters, **93** (2004), 098105 1–4.
- [4] ———, *A mathematical model of plants as ecosystem engineers*, J. of Theoretical Biology **244** (2007), 680–691.
- [5] G.H. Hardy, *A Mathematician’s Apology*, Cambridge University Press, 1940.
- [6] Neal Koblitz, *Elliptic curve cryptosystems*, Mathematics of Computation **48** (177) (1987), 203–209.
- [7] Lindenmayer, A. and Prusinkiewicz, P., *The algorithmic beauty of plants*, Springer-

- Verlag, 1990 (pdf bversion can be downloaded for free at <http://algorithmicbotany.org/papers/#abop>).
- [8] Meinhardt, H., *The algorithmic beauty of sea shells*, Springer-Verlag Berlin Heidelberg, 2009 (fourth edition).
  - [9] Victor Miller, *The uses of elliptic curves in cryptography*, CRYPTO **85** (1985), 417–426.
  - [10] Murray, J.D., *Mathematical Biology*, Springer Interdisciplinary Applied Mathematics, New York, 2002-2003.
  - [11] Rousseau C. and Saint-Aubin Y., *Mathematics and Technology*. Springer Undergraduate Series in Mathematics and Technology, Springer, NY, 2008.
  - [12] Yuki Sugiyama, Minoru Fukui, Macoto Kikuchi, Katsuya Hasebe, Akihiro Nakayama, Katsuhiko Nishinari, Shin-ichi Tadaki, and Satoshi Yukawa, *Traffic jams without bottlenecks: experimental evidence for the physical mechanism of the formation of a jam*, New J. Phys. **10** (2008), 033001. doi:10.1088/1367-2630/10/3/033001
  - [13] Soon Tee Teoh, Generalized descriptions for the procedural modeling of ancient East Asian buildings, DOI:10.2312/COMPAESTH/COMPAESTH09/017-024. In proceedings of: Computational Aesthetics 2009: Eurographics Workshop on Computational Aesthetics, Victoria, British Columbia, Canada, 2009.
  - [14] Turing A., *The chemical basis of morphogenesis*, Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences **237** (1952), 37–72.
  - [15] Xin-yue Xu, Jun Liu, Hai-ying Li, Yan-fang Zhou, *Probabilistic model for remain passenger queues at subway station platform*, J. Cent. South Univ. **20** (2013), 837–844.

Eduardo Colli, Universidade de São Paulo, Brazil

E-mail: colli@ime.usp.br

Fidel R. Nemenzo, University of the Philippines Diliman, Philippines

E-mail: fidel@math.upd.edu.ph

Konrad Polthier, Frei Universität Berlin, Germany

E-mail: Konrad.Polthier@fu-berlin.de

Christiane Rousseau, University of Montreal, Canada

E-mail: rousseac@dms.umontreal.ca



## Other Activities

The Organizing Committee strived to bring changes in the public's awareness and perception of mathematics. To that end, various cultural programs were organized such as public lectures, public exhibitions, Baduk events and Math Movie Screening. Over 23,000 of the general public enjoyed the benefit of these programs.

Two public lectures were held: one on the 13th and the other on the 20th of August. The speaker of the first public lecture in the evening of the Opening Ceremony was James Simons, the Founder and CEO of Renaissance Technologies, who delivered his lecture under the title, *My Life in Mathematics*. The speaker of the second public lecture was the 2014 Leelavati Prize Winner, Adrián Paenza. Both lectures together drew a large audience of about 5,500, a considerable portion of which was secondary and college students. Korean subtitles were provided during both lectures to assist these young students.

Exhibitions were open to the general public during the days of the Congress, where 33 organizations participated to provide visitors with informative and unique experience. A special exhibition also featured at Seoul ICM called IMAGINARY, which is an experience-centered exhibition developed by Mathematisches Forschungsinstitut Oberwolfach of Germany and jointly managed by the National Institute for Mathematical Sciences (NIMS) of Korea. Baduk programs were organized to provide participants with a glimpse of this traditional strategic game, also known as Go. Five world-renowned professional players, including Changhyuk Yoo and Changho Lee, played simultaneous games with 25 participants, selected through an application-and-review process. Public Baduk lectures were also delivered as part of this cultural program.

Over 1,600 participated in Math Movie Screening to watch the French mathematics documentary film, *How I Came to Hate Maths (Comment J'ai Détesté Les Maths)*. One of the actual characters of the movie, Cédric Villani, a 2010 Fields Medalist, personally attended the event and also held a question-and-answer session along with Jean-Pierre Bourguignon and Gert-Martin Greuel. The event was jointly hosted by the Embassy of France in Korea and the Organizing Committee.

All congress ceremonies, public lectures and scientific programs, including prize winner lectures, plenary and invited lectures, special lectures and panel discussions, were recorded and open to the public online. Official Seoul ICM application was also developed for both Android and iOS to further ease participants' access to all aspects of the congress. Over 3,000 downloaded the application.

The Organizing Committee, together with Springer, produced Seoul Intelligencer as per the recent tradition of the congress. Daily newspaper, Math&Presso, was distributed on each day of the Congress, to deliver onsite excitements and valuable information through interview articles and event coverages.

# List of Participants

Aalipour Hafshejani, Ghodratollah (Iran)  
Abara, Ma Nerissa (Philippines)  
Abate, Marco (Italy)  
Abbaspour, Hossein (France)  
Abdollahi, Alireza (Iran)  
Abdujabbarov, Ahmadjon (Uzbekistan)  
Abdulkhalikov, Kanat (UAE)  
Abdulov, Alisher (Kazakhstan)  
Abdyldaeva, Elmira (Kyrgyzstan)  
Abdymanapov, Sarsengali (Kazakhstan)  
Aberkane, Idriss (France)  
Abgrall, Rémi (Switzerland)  
Abiev, Nurlan (Kazakhstan)  
Abikenova, Sholpan (Kazakhstan)  
Abouzaid, Mohammed (USA)  
Abrahamsson, Leif (Sweden)  
Abreu, Eduardo (Brazil)  
Acharya, Saraswati (Nepal)  
Adamczewski, Boris (France)  
Addawe, Joel (Philippines)  
Adem, Alejandro (Canada)  
Adewumi, Sunday (Nigeria)  
Adhikari, Avishek (India)  
Adi Kusumo, Fajar (Indonesia)  
Adj, Sriwulan (Indonesia)  
Agarwal, Praveen (India)  
Agbor, Agbor Dieudonne (Cameroon)  
Agol, Ian (USA)  
Agora, Elona (Argentina)  
Agrawal, Mamta (India)  
Ahmad, Sk Safique (India)  
Ahmad, Sarfraz (Pakistan)  
Ahmedov, Anvarjon (Malaysia)  
Ahmedov, Bobomurat (Uzbekistan)  
Ahn, Chang Min (Republic of Korea)  
Ahn, Chi Young (Republic of Korea)  
Ahn, Danbee (Republic of Korea)  
Ahn, Dong Jin (Republic of Korea)  
Ahn, Donghyun (Republic of Korea)  
Ahn, Dongkyun (Republic of Korea)  
Ahn, Eunkyung (Republic of Korea)  
Ahn, Heungju (Republic of Korea)  
Ahn, Hyewon (Republic of Korea)  
Ahn, Jae Sol (Republic of Korea)  
Ahn, Jaewook (Republic of Korea)  
Ahn, Jeaman (Republic of Korea)  
Ahn, Jeoung-Hwan (Republic of Korea)  
Ahn, Ji Su (Republic of Korea)  
Ahn, Jin Hoo (Republic of Korea)  
Ahn, Jiweon (Republic of Korea)  
Ahn, Joong Hyun (Republic of Korea)  
Ahn, Juneyeong (Republic of Korea)  
Ahn, Keon Hui (Republic of Korea)  
Ahn, Soyoung (Republic of Korea)  
Ahn, Su Bin (Republic of Korea)  
Ahn, Sungjoon (Republic of Korea)  
Ahn, Y.K (Republic of Korea)  
Ahn, Yonghyun (Republic of Korea)  
Ahn, Young Joon (Republic of Korea)  
Aidos, Yerkara (Kazakhstan)  
Aistleitner, Christoph (Japan)  
Akhbari, Roghayeh (Iran)  
Akin, Hasan (Turkey)  
Akram, Saima (Pakistan)  
Akyildiz, Ersan (Turkey)  
Akyildiz, Yilmaz (Turkey)  
Al Salti, Nasser (Oman)  
Alahmadi, Adel (Saudi Arabia)  
Alarcon, Maria Del Mar (Spain)  
Alberdi Celaya, Elisabete (Spain)  
Albinsky, Irina (Israel)  
Albuquerque, Rui (Portugal)  
Aleksandrov, Alexander (Russia)  
Alekseev, Gennady (Russia)  
Alekseev, Anton (Switzerland)  
Alexeeva, Alexandra (Switzerland)  
Algarni, Said (Saudi Arabia)  
Al-Ghassani, Asma (Oman)  
Ali, Istkhari (India)  
Ali, Rosihan (Malaysia)  
Allaire, Gregoire (France)  
Al-Mdallal, Qasem (UAE)  
Almocera, Alexis Erich (Philippines)  
Al-Mohy, Awad (Saudi Arabia)  
Al-Mosawi, Riyadh (Iraq)  
Alonso Moron, Manuel (Spain)  
Alrashed, Maryam (Kuwait)  
Al-Sharawi, Ziyad (Oman)

Alsulami, Hamed (Saudi Arabia)  
Altug, Ali (USA)  
Alvarez, María Alejandra (Chile)  
Alvarez-Vazquez, Lino (Spain)  
Alves, Manuel Joaquim (Mozambique)  
Alves Dos Santos, Luiz (Belgium)  
Al-Yasry, Ahmad (Iraq)  
Alymkulov, Keldibay (Kyrgyzstan)  
Ambat, Vijayakumar (India)  
Ambethkar, Vusala (India)  
Amenta, Alex (Australia)  
Amini, Massoud (Iran)  
An, Congpei (P.R. China)  
An, Yanbin (P.R. China)  
An, Bokyoung (Republic of Korea)  
An, Byunghee (Republic of Korea)  
An, Hong-Min (Republic of Korea)  
An, Huibeom (Republic of Korea)  
An, Il Ju (Republic of Korea)  
An, Jaehyun (Republic of Korea)  
An, Jeongsun (Republic of Korea)  
An, Minjung (Republic of Korea)  
An, Youngjoo (Republic of Korea)  
An, Phan Thanh (Vietnam)  
Anak Agung Gede, Ngurah (Indonesia)  
Ananchuen, Nawarat (Thailand)  
Ananchuen, Watcharaphong (Thailand)  
Andami Ovono, Armel (Gabon)  
Anderson, Jean (USA)  
Andjiga, Nicolas Gabriel (Cameroon)  
Ando, Kazunori (Japan)  
Andreescu, Titu (USA)  
Andreev, Pavel (Russia)  
Andres, Sebastian (Germany)  
Andruskiewitsch, Nicolás (Argentina)  
Angiono, Ivan Ezequiel (Argentina)  
Aniversario, Imelda (Philippines)  
Anona, Frederic Manelo (Madagascar)  
Antonyan, Natella (Mexico)  
Antonyan, Sergey (Mexico)  
Anwar, Matloob (Pakistan)  
Aoki, Miho (Japan)  
Araujo, Carolina (Brazil)  
Arbieto, Alexander (Brazil)  
Arceo, Carlene Perpetua (Philippines)  
Archibald, Thomas (Canada)  
Ardakov, Konstantin (UK)  
Arenas, Manuel (Chile)  
Areum, Lee (Republic of Korea)  
Arguz, Nuromur Hulya (Germany)  
Arias, Jeanine Concepcion (Philippines)  
Arizmendi, Octavio (Mexico)  
Arjunwadkar, Himanee (India)  
Arkut, Ibrahim Cahit (Turkey)  
Arnaud, Marie-Claude (France)  
Arnoux, Cédric (France)  
Arnoux, Chloé (France)  
Arnoux, Pierre (France)  
Arora, Ashish (India)  
Arora, Sanjeev (USA)  
Arthur, Dorothy (Canada)  
Arthur, James (Canada)  
Aru, Juhan (France)  
Arzarello, Ferdinando (Italy)  
Asaoka, Masayuki (Japan)  
Ash, Andrew (USA)  
Ash, J. Marshall (USA)  
Ashrafi Ghomroodi, Seyed Ali Reza (Iran)  
Asmuss, Svetlana (Latvia)  
Assal, Miloud (Tunisia)  
Assim, Jilali (Morocco)  
Astashkin, Sergey (Russia)  
Astashova, Irina (Russia)  
Astrakova, Anna (Russia)  
Asuega, Tasi (USA)  
Atan, Kamel (Malaysia)  
Atindogbe, Cyriaque (Benin)  
Atkinson, Christopher (USA)  
Attar, Akram (Iraq)  
Attiya, Adel (Egypt)  
Avetisyan, Karen (Armenia)  
Avetisyan, Lilit (Armenia)  
Avila, Artur (France)  
Avramidi, Ivan (USA)  
Avramidi, Valentina (USA)  
Awanou, Gerard (USA)  
Axtell, Jonathan (USA)  
Ayoub, Joseph (Switzerland)  
Ayupov, Shavkat (Uzbekistan)  
Ayupova, Fakhriya (Uzbekistan)  
Ayzenberg, Anton (Japan)  
Bacani, Jerico (Philippines)  
Bach, Volker (Germany)  
Bachmann, Tom (Germany)  
Backman, Spencer (USA)  
Bae, Eunok (Republic of Korea)  
Bae, Gi Chan (Republic of Korea)  
Bae, Gyujong (Republic of Korea)  
Bae, Hanwool (Republic of Korea)  
Bae, Hyeongjin (Republic of Korea)  
Bae, Hyeong-Ohk (Republic of Korea)  
Bae, Hyo Min (Republic of Korea)  
Bae, Jaegug (Republic of Korea)  
Bae, Jae-Hyeong (Republic of Korea)

- Bae, Jeongwoo (Republic of Korea)  
 Bae, Jin Ho (Republic of Korea)  
 Bae, Junsik (Republic of Korea)  
 Bae, Myoungjean (Republic of Korea)  
 Bae, Sang Hyeon (Republic of Korea)  
 Bae, Soohyun (Republic of Korea)  
 Bae, Sookyung (Republic of Korea)  
 Bae, Sunghan (Republic of Korea)  
 Bae, Sunghoon (Republic of Korea)  
 Bae, Taeyeon (Republic of Korea)  
 Bae, Yeon Uck (Republic of Korea)  
 Bae, Yeongjin (Republic of Korea)  
 Bae, Yeonho (Republic of Korea)  
 Bae, Yongju (Republic of Korea)  
 Bae, Younggon (Republic of Korea)  
 Bae, Younghan (Republic of Korea)  
 Baek, Chung Hun (Republic of Korea)  
 Baek, Eun Ha (Republic of Korea)  
 Baek, Hunki (Republic of Korea)  
 Baek, Jongyeon (Republic of Korea)  
 Baek, Sanghoon (Republic of Korea)  
 Baek, Seung Jin (Republic of Korea)  
 Baek, Sueng (Republic of Korea)  
 Bagdasar, Ovidiu (UK)  
 Bagherzad Sessary, Seyed Mohammad Reza (Iran)  
 Bahayou, Mohamed Amine (Algeria)  
 Bai, Yao (Sweden)  
 Baik, Sangwon (Republic of Korea)  
 Bailey, Kathie (USA)  
 Bak, Jisoo (Republic of Korea)  
 Bak, Jong-Hyeok (Republic of Korea)  
 Bak, Sejong (Republic of Korea)  
 Bak, Soyoon (Republic of Korea)  
 Balachandran, Krishnan (India)  
 Baladi, Viviane (France)  
 Balakrishnan, Jennifer (UK)  
 Balashchenko, Vitaly (Belarus)  
 Balilescu, Loredana (Romania)  
 Ball, John (UK)  
 Ball, Palden (UK)  
 Ball, Sedhar (UK)  
 Ball, Tenzin (UK)  
 Ballard, Matthew (USA)  
 Balleier, Carsten (Germany)  
 Balmaceda, Jose Maria (Philippines)  
 Baltaeva, Umida (Uzbekistan)  
 Ban, Kornheng (Cambodia)  
 Banchoff, Thomas (USA)  
 Bandara, Lashi (Australia)  
 Bandari, Somayeh (Iran)  
 Bandeira, Afonso (USA)  
 Bandyopadhyay, Antar (India)  
 Bang, Guk-Yeong (Republic of Korea)  
 Bang, Je-Hyeon (Republic of Korea)  
 Bang, Sejeong (Republic of Korea)  
 Bang, Sekwon (Republic of Korea)  
 Bang, Seung Jin (Republic of Korea)  
 Bang, Suk Hee (Republic of Korea)  
 Banisch, Ralf (Germany)  
 Bao, Yuanyuan (Japan)  
 Bao, Weizhu (Singapore)  
 Baowan, Duangkamon (Thailand)  
 Barak, Boaz (USA)  
 Baranek, Alejandro (Argentina)  
 Barany, Imre (Hungary)  
 Barik, Sasmita (India)  
 Bark, Junoh (Republic of Korea)  
 Barmak, Jonathan (Argentina)  
 Barrack, Duncan (UK)  
 Barrett, David (USA)  
 Barrow-Green, June (Norway)  
 Barsbold, Bazarragchaa (Mongolia)  
 Barton, Bill (New Zealand)  
 Bartosiewicz, Zbigniew (Poland)  
 Bartoszynski, Tomek (USA)  
 Barua, Rana (India)  
 Basak, Biplab (India)  
 Basilla, Julius (Philippines)  
 Baskoro, Edy Tri (Indonesia)  
 Basmajian, Ara (USA)  
 Basse, Unanaowo (Nigeria)  
 Bauer, Ingrid (Germany)  
 Bayati, Shamila (Iran)  
 Begum, Shamsun Naher (Bangladesh)  
 Behforooz, Hossein (USA)  
 Behle, Markus (Germany)  
 Behn, Antonio (Chile)  
 Behrend, Kai (Canada)  
 Bell, Renee (USA)  
 Belolipetsky, Mikhail (Brazil)  
 Benis Sinaceur, Hourya (France)  
 Benitez, Julius (Philippines)  
 Benito, Angelica (USA)  
 Benkart, Georgia (USA)  
 Bennis, Driss (Morocco)  
 Benoist, Yves (France)  
 Berczi, Gergely (UK)  
 Berdinsky, Dmitry (New Zealand)  
 Berdyshev, Abdumauvlen (Kazakhstan)  
 Berezina, Miryam (Israel)  
 Bergeaud, Jean-Pierre (France)  
 Betty, Rowena Alma (Philippines)  
 Beyaz, Ahmet (Turkey)

Bhadra Man, Tuladhar (Nepal)  
Bhak, Zachary (USA)  
Bhanushe, Mandar (India)  
Bhargava, Divakar (India)  
Bhargava, Madhu (India)  
Bhargava, Brij (USA)  
Bhargava, Manjul (USA)  
Bhargava, Mira (USA)  
Bhargava, Mudita (USA)  
Bhargava, Nalin (USA)  
Bhat, B V Rajarama (India)  
Bhatia, Rajendra (India)  
Bhatt, Abhay Gopal (India)  
Bhatta, Chet Raj (Nepal)  
Bhattacharjee, Debashis (India)  
Bhatti, Faqir (Pakistan)  
Bhowmik, Bappaditya (India)  
Bibi, Nargis (UK)  
Biembengut, Maria Salett (Brazil)  
Biquard, Olivier (France)  
Birkenfeld, Judith (Spain)  
Bisson, Gaetan (Polynesia)  
Blaavand, Jakob (UK)  
Blocki, Zbigniew (Poland)  
Blumberg, Andrew (USA)  
Bohman, Thomas (USA)  
Bojanowska-Jackowska, Agnieszka (Poland)  
Bok, Younggyu (Republic of Korea)  
Bokayev, Nurzhan (Kazakhstan)  
Boldt, Sebastian (Germany)  
Bollman, Dorothy (Puerto Rico)  
Bonforte, Matteo (Spain)  
Bonifant, Araceli (USA)  
Bonnans, Joseph Frederic (France)  
Bonzi, Bernard Kaka (Burkina Faso)  
Boo, Changwoo (Republic of Korea)  
Boo, Deok Hoon (Republic of Korea)  
Boote, Yumi (UK)  
Borg, James (Malta)  
Borodin, Alexei (USA)  
Bossoto, Basile Guy Richard (Congo)  
Botirov, Golibjon (Uzbekistan)  
Bouayad, Alexandre (France)  
Bouayad, Sebastien (France)  
Bouche, Thierry (France)  
Boudaoud, Fatima (Algeria)  
Boussaid, Omar (Algeria)  
Bozicevic, Mladen (Croatia)  
Braha, Naim (Albania)  
Braides, Andrea (Italy)  
Braverman, Mark (USA)  
Braz E Silva, Pablo (Brazil)  
Breschi, Giancarlo (Spain)  
Bressan, Juliana (Brazil)  
Breuillard, Emmanuel (France)  
Brezzi, Franco (Italy)  
Brooke-Taylor, Andrew (UK)  
Broughan, Jacqueline (New Zealand)  
Broughan, Kevin (New Zealand)  
Brown, Francis (France)  
Brown, Jeff (Republic of Korea)  
Brundan, Jonathan (USA)  
Bu, Sunyoung (Republic of Korea)  
Bucher-Karlsson, Michelle (Switzerland)  
Bucur, Alina (USA)  
Buczynski, Jaroslaw (Poland)  
Budzynski, Piotr (Poland)  
Buffa, Annalisa (Italy)  
Bui, Xuan Hai (Vietnam)  
Bui Thanh, Tu (Vietnam)  
Bui Van, Dinh (Vietnam)  
Bujalance, Emilio (Spain)  
Bujalance Garcia, Jose Antonio (Spain)  
Bulatov, Andrei (Canada)  
Bulboaca, Teodor (Romania)  
Bulca, Betul (Turkey)  
Bulut, Aynur (USA)  
Bunwong, Kornkanok (Thailand)  
Buot, Jude (Philippines)  
Burrill, Sophie (Canada)  
Bursztyn, Henrique (Brazil)  
Burton, Benjamin (Australia)  
Butler, Steve (USA)  
Buyukboduk, Kazim (Turkey)  
Buzzi, Claudio Aguinaldo (Brazil)  
Byeon, Dongho (Republic of Korea)  
Byeon, Jeayoung (Republic of Korea)  
Byun, Gi Hyun (Republic of Korea)  
Byun, Jisoo (Republic of Korea)  
Byun, Jungi (Republic of Korea)  
Byun, Sangho (Republic of Korea)  
Byun, Seok Hyun (Republic of Korea)  
Byun, Seongmin (Republic of Korea)  
Byun, Sungsoo (Republic of Korea)  
Byun, Sun-Sig (Republic of Korea)  
C R, Saranya (India)  
Cabarrubias, Bituin (Philippines)  
Cabral, Emmanuel (Philippines)  
Cabrera, Alejandro (Brazil)  
Caceres, Luis (Puerto Rico)  
Cai, Wenxiang (Canada)  
Cai, Jihua (P.R. China)  
Cai, Yong (P.R. China)  
Calderbank, Authur (USA)

- Camacho, Cesar (Brazil)  
 Camia, Federico (Netherlands)  
 Campana Ramia, Maria (Brazil)  
 Campillo, Antonio (Spain)  
 Cancès, Eric (France)  
 Candes, Emmanuel (USA)  
 Cangul, Ismail Naci (Turkey)  
 Cangul, Oya (Turkey)  
 Cannarsa, Piermarco (Italy)  
 Canoy, Sergio Jr. (Philippines)  
 Cao, Chongguang (P.R. China)  
 Cao, Xueyun (P.R. China)  
 Cardoso, Isolda (Argentina)  
 Carlini, Enrico (Australia)  
 Carlos Augusto Cardoso, Gorito (Brazil)  
 Carmona, Juan (Colombia)  
 Carnahan, Scott (Japan)  
 Carneiro, Emanuel (Brazil)  
 Carnielli Abranches Ramos, Ariana (Brazil)  
 Carocca, Angel (Chile)  
 Carpio, Kristine Joy (Philippines)  
 Casals, Roger (Spain)  
 Casals-Ruiz, Montserrat (UK)  
 Casas, Oscar (Colombia)  
 Cassiano, Anna Grazia Vincenza (Italy)  
 Cassy, Bhangy (Mozambique)  
 Castro-Jimenez, Francisco-Jesus (Spain)  
 Catanese, Fabrizio (Germany)  
 Catinas, Emil (Romania)  
 Catinas, Teodora (Romania)  
 Catoiu, Stefan (USA)  
 Cavalcante, Marcos Petrucio (Brazil)  
 Cavalcante, Pedro Paulo (Brazil)  
 Cederbaum, Carla (Germany)  
 Celebi, A. Okay (Turkey)  
 Celebi, Gulay (Turkey)  
 Celeste, Raquel (Philippines)  
 Celeste, Richell (Philippines)  
 Cervo, Manuela (Brazil)  
 Cesnavicius, Kestutis (USA)  
 Cha, Daseul (Republic of Korea)  
 Cha, Jaehoon (Republic of Korea)  
 Cha, Sanghyeon (Republic of Korea)  
 Cha, Seokbin (Republic of Korea)  
 Cha, Seong Jun (Republic of Korea)  
 Cha, Soon Hyuk (Republic of Korea)  
 Cha, Sun Hee (Republic of Korea)  
 Cha, Ye Sle (Republic of Korea)  
 Cha, Young Kwang (Republic of Korea)  
 Chae, Gab Byung (Republic of Korea)  
 Chae, Jiseok (Republic of Korea)  
 Chae, Kuem-Seon (Republic of Korea)  
 Chae, Moon Kook (Republic of Korea)  
 Chae, Myeongju (Republic of Korea)  
 Chae, Seok Joo (Republic of Korea)  
 Chaichi, Mohamad (Iran)  
 Chakraborty, Partha Sarathi (India)  
 Chaleyat-Maurel, Mireille (France)  
 Chalishajar, Dimplekumar (USA)  
 Challa, Durga Prasad (India)  
 Chamorro, Diego (France)  
 Chan, Tony F (Hong Kong)  
 Chan, Sony (Republic of Korea)  
 Chan, Tsz On Mario (Republic of Korea)  
 Chandee, Vorrapan (Thailand)  
 Chang, Gyu Whan (Republic of Korea)  
 Chang, Hoon (Republic of Korea)  
 Chang, Hyonman (Republic of Korea)  
 Chang, Jaewon (Republic of Korea)  
 Chang, Jeongwook (Republic of Korea)  
 Chang, Joo Sup (Republic of Korea)  
 Chang, Keehoon (Republic of Korea)  
 Chang, Kun Soo (Republic of Korea)  
 Chang, Kyung Yoon (Republic of Korea)  
 Chang, Seunghwan (Republic of Korea)  
 Chang, Wan Uk (Republic of Korea)  
 Chang, Yeonsu (Republic of Korea)  
 Chang, Younhea (Republic of Korea)  
 Chang, Fei-Huang (Taiwan)  
 Chang, Koukung Alex (Taiwan)  
 Chang, Ting-Pang (Taiwan)  
 Changkyu, Park (Republic of Korea)  
 Chanty, Piseth (Cambodia)  
 Charney, Ruth (USA)  
 Chatterjee, Pralay (India)  
 Chatterjee, Sourav (USA)  
 Chattopadhyay, Arup (India)  
 Chatzidakis, Zoé (France)  
 Chebotarev, Alexander (Russia)  
 Chebotarev, Vladimir (Russia)  
 Chebotareva, Liudmila (Russia)  
 Chemla, Karine (France)  
 Chen, Guang (P.R. China)  
 Chen, Jianmin (P.R. China)  
 Chen, Minghao (P.R. China)  
 Chen, Peide (P.R. China)  
 Chen, Xian (P.R. China)  
 Chen, Yin (P.R. China)  
 Chen, Xinxin (France)  
 Chen, Louis (Singapore)  
 Chen, Chin-Yun (Taiwan)  
 Chen, Jein-Shan (Taiwan)  
 Chen, Jiun Cheng (Taiwan)  
 Chen, Jungkai (Taiwan)

Chen, Shyanshiou (Taiwan)  
 Chen, Dawei (USA)  
 Cheng, Chongqing (P.R. China)  
 Cheng, Cong (P.R. China)  
 Cheng, Wei (P.R. China)  
 Cheng, Yan-Hsiou (Taiwan)  
 Cheon, Dong Hwan (Republic of Korea)  
 Cheon, Dong Wook (Republic of Korea)  
 Cheon, Gi-Sang (Republic of Korea)  
 Cheon, Jeungeun (Republic of Korea)  
 Cheon, Jung Hee (Republic of Korea)  
 Cheon, Sung Ho (Republic of Korea)  
 Cheon, Younhwan (Republic of Korea)  
 Cheong, Daewoong (Republic of Korea)  
 Cheong, Minseok (Republic of Korea)  
 Cheraku, Venkata Ganapathi Narasimha Kumar (India)  
 Cherinda, Marcos (Mozambique)  
 Cherinda, Nilsa Adelaide Issufo Enoque Pondja (Mozambique)  
 Chernikov, Artem (France)  
 Cheung, Rex (USA)  
 Chi, Dongpyo (Republic of Korea)  
 Chi, Soo Hyun (Republic of Korea)  
 Chiang, Yuan-Jen (USA)  
 Chiang-Hsieh, Hung-Jen (Taiwan)  
 Chiba, Takashi (Japan)  
 Chierchia, Luigi (Italy)  
 Chin, Yohan (Republic of Korea)  
 Chintala, Vineeth (India)  
 Chiu, Faye (USA)  
 Chiyonobu, Taizo (Japan)  
 Cho, Ae-Kyoung (Republic of Korea)  
 Cho, Allen Myungsin (Republic of Korea)  
 Cho, Beongeun (Republic of Korea)  
 Cho, Bumkyu (Republic of Korea)  
 Cho, Cheol Hyun (Republic of Korea)  
 Cho, Cheong Ho (Republic of Korea)  
 Cho, Chuhee (Republic of Korea)  
 Cho, Dahye (Republic of Korea)  
 Cho, Dong Hyun (Republic of Korea)  
 Cho, Durkbin (Republic of Korea)  
 Cho, Eun Ho (Republic of Korea)  
 Cho, Eun Young (Republic of Korea)  
 Cho, Gyeong-Mi (Republic of Korea)  
 Cho, Hong Seok (Republic of Korea)  
 Cho, Hye Ran (Republic of Korea)  
 Cho, Hyeon Jun (Republic of Korea)  
 Cho, Hyun Woo (Republic of Korea)  
 Cho, Hyun Woong (Republic of Korea)  
 Cho, Hyung Chan (Republic of Korea)  
 Cho, Hyungchan (Republic of Korea)  
 Cho, Hyunsoo (Republic of Korea)  
 Cho, Hyun-Woo (Republic of Korea)  
 Cho, Jeon Rim (Republic of Korea)  
 Cho, Jeong Suk (Republic of Korea)  
 Cho, Jin-Hwan (Republic of Korea)  
 Cho, Jinseok (Republic of Korea)  
 Cho, Junghee (Republic of Korea)  
 Cho, Jun-Mo (Republic of Korea)  
 Cho, Kyenghwan (Republic of Korea)  
 Cho, Kyu Han (Republic of Korea)  
 Cho, Kyung (Republic of Korea)  
 Cho, Kyunghyun (Republic of Korea)  
 Cho, Min Jun (Republic of Korea)  
 Cho, Minshik (Republic of Korea)  
 Cho, Nak Eun (Republic of Korea)  
 Cho, Raesang (Republic of Korea)  
 Cho, Sang Wook (Republic of Korea)  
 Cho, Sangbum (Republic of Korea)  
 Cho, Sangheum (Republic of Korea)  
 Cho, Sanghyun (Republic of Korea)  
 Cho, Seok Ho (Republic of Korea)  
 Cho, Seong Yun (Republic of Korea)  
 Cho, Soojin (Republic of Korea)  
 Cho, Suhngkeun (Republic of Korea)  
 Cho, Sung Hae (Republic of Korea)  
 Cho, Sung Hyun (Republic of Korea)  
 Cho, Sung Jae (Republic of Korea)  
 Cho, Sung Je (Republic of Korea)  
 Cho, Sung Vin (Republic of Korea)  
 Cho, Sungjoo (Republic of Korea)  
 Cho, Sungwon (Republic of Korea)  
 Cho, Sungyoon (Republic of Korea)  
 Cho, Won Sik (Republic of Korea)  
 Cho, Won-Seo (Republic of Korea)  
 Cho, Yeol Je (Republic of Korea)  
 Cho, Yeong Gyeong (Republic of Korea)  
 Cho, Yong Jin (Republic of Korea)  
 Cho, Yong Joong (Republic of Korea)  
 Cho, Yong Soo (Republic of Korea)  
 Cho, Yonggeun (Republic of Korea)  
 Cho, Yonghwa (Republic of Korea)  
 Cho, Yong-Kum (Republic of Korea)  
 Cho, Yongseung (Republic of Korea)  
 Cho, Yoonjoo (Republic of Korea)  
 Cho, Yoonki (Republic of Korea)  
 Cho, Young Hyun (Republic of Korea)  
 Cho, Young Kyoung (Republic of Korea)  
 Cho, Youngae (Republic of Korea)  
 Cho, Youngjin (Republic of Korea)  
 Cho, Youngmin (Republic of Korea)  
 Cho, You-Young (Republic of Korea)  
 Cho, Yu Jeong (Republic of Korea)

Cho, Yumi (Republic of Korea)  
 Cho, Yunsun (Republic of Korea)  
 Choe, Boo Rim (Republic of Korea)  
 Choe, Dong Heon (Republic of Korea)  
 Choe, Hyeongmin (Republic of Korea)  
 Choe, Insong (Republic of Korea)  
 Choe, Jaigyoung (Republic of Korea)  
 Choi, Beomjun (Republic of Korea)  
 Choi, Bokyoung (Republic of Korea)  
 Choi, Chan Young (Republic of Korea)  
 Choi, Chanhyuk (Republic of Korea)  
 Choi, Daebeom (Republic of Korea)  
 Choi, Dahyeon (Republic of Korea)  
 Choi, Do Hoon (Republic of Korea)  
 Choi, Do Young (Republic of Korea)  
 Choi, Dong Hyeok (Republic of Korea)  
 Choi, Dong June (Republic of Korea)  
 Choi, Eun Hye (Republic of Korea)  
 Choi, Eunmi (Republic of Korea)  
 Choi, Gun Hee (Republic of Korea)  
 Choi, Gwanghyeon (Republic of Korea)  
 Choi, Ha Nul (Republic of Korea)  
 Choi, Hagyun (Republic of Korea)  
 Choi, Hakho (Republic of Korea)  
 Choi, Hang Chul (Republic of Korea)  
 Choi, Heesun (Republic of Korea)  
 Choi, Heungsu (Republic of Korea)  
 Choi, Hojin (Republic of Korea)  
 Choi, Howon (Republic of Korea)  
 Choi, Hyangchul (Republic of Korea)  
 Choi, Hyeonseok (Republic of Korea)  
 Choi, Hyesung (Republic of Korea)  
 Choi, Hyounggyu (Republic of Korea)  
 Choi, Hyun Jung (Republic of Korea)  
 Choi, Hyung Tae (Republic of Korea)  
 Choi, Hyungjun (Republic of Korea)  
 Choi, Hyungun (Republic of Korea)  
 Choi, I Sol (Republic of Korea)  
 Choi, Ik-Han (Republic of Korea)  
 Choi, Ilkyoo (Republic of Korea)  
 Choi, In Han (Republic of Korea)  
 Choi, Jae Seok (Republic of Korea)  
 Choi, Jae Sung (Republic of Korea)  
 Choi, Jaeyong (Republic of Korea)  
 Choi, Jeong-Ok (Republic of Korea)  
 Choi, Jieun (Republic of Korea)  
 Choi, Jihoon (Republic of Korea)  
 Choi, Jihoon (Republic of Korea)  
 Choi, Jihyun (Republic of Korea)  
 Choi, Jin Ho (Republic of Korea)  
 Choi, Jin Hyuck (Republic of Korea)  
 Choi, Jinwon (Republic of Korea)  
 Choi, Jiyun (Republic of Korea)  
 Choi, Jong Geon (Republic of Korea)  
 Choi, Jonghyeon (Republic of Korea)  
 Choi, Jongkeun (Republic of Korea)  
 Choi, Joon Young (Republic of Korea)  
 Choi, Jun (Republic of Korea)  
 Choi, Jun Hyung (Republic of Korea)  
 Choi, Junesang (Republic of Korea)  
 Choi, Jungwoo (Republic of Korea)  
 Choi, Junho (Republic of Korea)  
 Choi, Junhwa (Republic of Korea)  
 Choi, Junjae (Republic of Korea)  
 Choi, Jun-Won (Republic of Korea)  
 Choi, Kwang Ju (Republic of Korea)  
 Choi, Kyeongsu (Republic of Korea)  
 Choi, Minki (Republic of Korea)  
 Choi, Minnseok (Republic of Korea)  
 Choi, Min-Suk (Republic of Korea)  
 Choi, Minyeop (Republic of Korea)  
 Choi, Moo Jin (Republic of Korea)  
 Choi, Myung Hwa (Republic of Korea)  
 Choi, Myung Jae (Republic of Korea)  
 Choi, Myung-Jun (Republic of Korea)  
 Choi, Seo Jong (Republic of Korea)  
 Choi, Seok Youn (Republic of Korea)  
 Choi, Seonmi (Republic of Korea)  
 Choi, Seung Ho (Republic of Korea)  
 Choi, Seung Ho (Republic of Korea)  
 Choi, Seunghoe (Republic of Korea)  
 Choi, Seung-II (Republic of Korea)  
 Choi, So Yeon (Republic of Korea)  
 Choi, Soeun (Republic of Korea)  
 Choi, Sooyeon (Republic of Korea)  
 Choi, Soyeon (Republic of Korea)  
 Choi, Suh Hyun (Republic of Korea)  
 Choi, Suhyoung (Republic of Korea)  
 Choi, Su-Jeong (Republic of Korea)  
 Choi, Sung Rak (Republic of Korea)  
 Choi, Sung Woo (Republic of Korea)  
 Choi, Sun-Ho (Republic of Korea)  
 Choi, Sunhwa (Republic of Korea)  
 Choi, Sunyong (Republic of Korea)  
 Choi, Suyoung (Republic of Korea)  
 Choi, Taeyoung (Republic of Korea)  
 Choi, Whanhyuk (Republic of Korea)  
 Choi, Won Chang (Republic of Korea)  
 Choi, Wonkyu (Republic of Korea)  
 Choi, Woocheol (Republic of Korea)  
 Choi, Yong Seok (Republic of Korea)  
 Choi, Yongho (Republic of Korea)  
 Choi, Yoonho (Republic of Korea)  
 Choi, Young Won (Republic of Korea)

- Choi, Young Yun (Republic of Korea)  
 Choi, Younggi (Republic of Korea)  
 Choi, Youngji (Republic of Korea)  
 Choi, Young-Jun (Republic of Korea)  
 Choi, Young-Kwang (Republic of Korea)  
 Choi, Youngook (Republic of Korea)  
 Choi, Youngwoo (Republic of Korea)  
 Choi, Youn-Seo (Republic of Korea)  
 Choi, Yun Sung (Republic of Korea)  
 Choi, Young-Pil (UK)  
 Choi, Kyudong (USA)  
 Choie, Youngju (Republic of Korea)  
 Chon, Inheung (Republic of Korea)  
 Chong, Chi Tat (Singapore)  
 Choo, Jungsun (Republic of Korea)  
 Chooi, Wai Leong (Malaysia)  
 Chowdhury, Mohammad Showkat Rahim (Australia)  
 Choy, Jaeyoo (Republic of Korea)  
 Christodoulou, Demetrios (Greece)  
 Chu, Hahng-Yun (Republic of Korea)  
 Chu, May (USA)  
 Chuang, Chih-Sheng (Taiwan)  
 Chudnovsky, Maria (USA)  
 Chugh, Renu (India)  
 Chuluun, Delgermurun (Mongolia)  
 Chumley, Timothy (USA)  
 Chun, Ga Yoon (Republic of Korea)  
 Chun, Jinhee (Republic of Korea)  
 Chun, Jonghee (Republic of Korea)  
 Chun, Minhyeok (Republic of Korea)  
 Chun, Seogbeom (Republic of Korea)  
 Chun, Sun Hyang (Republic of Korea)  
 Chun, Yongjin (Republic of Korea)  
 Chung, Shun Wai (P.R. China)  
 Chung, Ping Ngai (Hong Kong)  
 Chung, Da Woon (Republic of Korea)  
 Chung, Daewon (Republic of Korea)  
 Chung, Dago (Republic of Korea)  
 Chung, Dong Myung (Republic of Korea)  
 Chung, Hong (Republic of Korea)  
 Chung, In Jae (Republic of Korea)  
 Chung, Jaeyoung (Republic of Korea)  
 Chung, Ji Su (Republic of Korea)  
 Chung, Jin Il (Republic of Korea)  
 Chung, Kiryong (Republic of Korea)  
 Chung, Kwangman (Republic of Korea)  
 Chung, Kyungmi (Republic of Korea)  
 Chung, Min Young (Republic of Korea)  
 Chung, Sangwon (Republic of Korea)  
 Chung, Seungjoon (Republic of Korea)  
 Chung, Soon-Yeong (Republic of Korea)  
 Chung, Yoonwon (Republic of Korea)  
 Chung, Youngbok (Republic of Korea)  
 Chung, Youngjin (Republic of Korea)  
 Chuzhoy, Julia (USA)  
 Ciliberto, Ciro (Italy)  
 Ciocan-Fontanine, Ionut (USA)  
 Cirilo, Patricia (Brazil)  
 Clemens, Charles Herbert (USA)  
 Clutterbuck, Julie (Australia)  
 Cohen, Henri (France)  
 Collera, Juancho (Philippines)  
 Colli, Eduardo (Brazil)  
 Colon-Reyes, Omar (Puerto Rico)  
 Conlon, David (UK)  
 Conrad, Eric Roger (USA)  
 Conrad, Kirsten Griffiths (USA)  
 Conrad, Nicholas Leighton (USA)  
 Conrad, Phillip Alexander (USA)  
 Contreras, Gonzalo (Mexico)  
 Coons, Michael (Australia)  
 Corcino, Cristina (Philippines)  
 Corcino, Roberto (Philippines)  
 Corsman, Jens (Sweden)  
 Cortes, Victor (Chile)  
 Cortez, Maria Isabel (Chile)  
 Cortinas, Guillermo (Argentina)  
 Corwin, Ivan (USA)  
 Cowen, Carl (USA)  
 Craig, Walter (Canada)  
 Crispin Quinonez, Veronica (Sweden)  
 Crovisier, Sylvain (France)  
 Crucella, Julián (Argentina)  
 Cuenca, Jose Antonio (Spain)  
 Cui, Chengri (P.R. China)  
 Curbelo, Jezabel (Spain)  
 Curbera, Guillermo (Spain)  
 Curtis, Marcia (USA)  
 Da Silva, Lía (Chile)  
 Dafermos, Mihalis (UK)  
 Dafni, Galia (Canada)  
 Dai, Wanyang (P.R. China)  
 Dai, Wen-Rong (P.R. China)  
 Dalawat, Chandan Singh (India)  
 Dalitz, Wolfgang (Germany)  
 Damanik, David (USA)  
 Damian, Florin (Moldova)  
 Dan, Yuya (Japan)  
 D'Andrea, Carlos (Spain)  
 Dang, Anh Tuan (Vietnam)  
 Dao, Thi Thu Ha (Vietnam)  
 Dao Van, Dung (Vietnam)  
 Darafsheh, Mohammad Reza (Iran)

- Darby, Alastair (UK)  
 Dark, Chan Ho (Republic of Korea)  
 Darkhovsky, Boris (Russia)  
 Darmon, Henri (Canada)  
 Darmon, Maia (Canada)  
 Das, Paritosh Chandra (India)  
 Das, Soumya (India)  
 Daskalopoulos, Panagiota (USA)  
 Datta, Basudeb (India)  
 Datta, Mrinmoy (India)  
 Daubechies, Ingrid (USA)  
 Daus, Leonard (UAE)  
 Davenport, James (UK)  
 David, Sinnou (France)  
 Davidoff, Giuliana (USA)  
 Davis, Diana (USA)  
 Davvaz, Bijan (Iran)  
 De Bont, Petronella (Netherlands)  
 De Guzman, Nino Jose (Philippines)  
 De Klerk, Ben-Eben (South Africa)  
 De La Cruz, Ralph (Philippines)  
 De León, Manuel (Spain)  
 De Los Reyes, Juan Carlos (Ecuador)  
 De Melo, Welington (Brazil)  
 De Sousa Ribeiro Junior, Ernani (Brazil)  
 De Young, Gregg (Egypt)  
 Deb, Biswajit (India)  
 Debbi, Latifa (Algeria)  
 Debinska-Bielak, Teresa (Poland)  
 Debinska-Nagorska, Anna (Poland)  
 Debnath, Joyati (USA)  
 Debnath, Narayan (USA)  
 Degos, Jean Guy (France)  
 Degos, Jean-Yves (France)  
 Del Barco, Viviana (Argentina)  
 Del Pobil, Angel P. (Spain)  
 Dela Cruz, Laarni (Philippines)  
 Dela Cruz, Romar (Philippines)  
 Dela Rosa, Kennett (Philippines)  
 Deligero, Eveyth (Philippines)  
 Delp, Kelly (USA)  
 Deltell, Juan Carlos (Spain)  
 Demailly, Jean-Pierre (France)  
 Demidenko, Gennady (Russia)  
 Demirci, Musa (Turkey)  
 Dencker, Nils (Sweden)  
 Deng, Shijin (P.R. China)  
 Depablo, Arturo (Spain)  
 Desquith, Etienne (Ivory Coast)  
 Devaney, Robert (USA)  
 Devkota, Jyoti (Nepal)  
 Devos, Matt (Canada)  
 Dewan, Kum Kum (India)  
 Dhall, Sakshi (India)  
 Dharmatti, Sheetal (India)  
 Di Teodoro, Antonio (Venezuela)  
 Diallo, Abdoul Salam (Senegal)  
 Dias Moreira, Antonio Marcus (Brazil)  
 Díaz Rojas, Daniela (Chile)  
 Diazchavez Pacheco, Maria De Lourdes (Mexico)  
 Dickenstein, Alicia (Argentina)  
 Didenko, Victor (Brunei Darussalam)  
 Diemer, Colin (USA)  
 Dierkes, Ulrich (Germany)  
 Dilman, Valery (Russia)  
 Dinar, Yassir (Sudan)  
 Ding, Ling (Australia)  
 Ding, Hao (P.R. China)  
 Ding, Zhiyuan (USA)  
 Dinger, Ulla (Sweden)  
 Dinh, Dung (Vietnam)  
 Dinh, Nho-Hao (Vietnam)  
 Dinh, Trung (Vietnam)  
 Dinh Thanh, Duc (Vietnam)  
 Diop, Mamadou Abdoul (Senegal)  
 Djitte, Ngalla (Senegal)  
 Djoric, Mirjana (Serbia)  
 Djurdjevac Conrad, Natasa (Germany)  
 Dmytryshyn, Andrii (Sweden)  
 Do, Eui Hyeon (Republic of Korea)  
 Do, Genahee (Republic of Korea)  
 Do, Geonho (Republic of Korea)  
 Do, Jeong-Hyeok (Republic of Korea)  
 Do, Sang Woo (Republic of Korea)  
 Do, Young Hoo (Republic of Korea)  
 Doan, The Hieu (Vietnam)  
 Doan Thi, Kim Quy (Vietnam)  
 Dobbins, Michael Gene (Republic of Korea)  
 Dobrinen, Natasha (USA)  
 Dogra, Netan (UK)  
 Donaldson, James (USA)  
 Dong, Yujun (P.R. China)  
 Dontchev, Assen (USA)  
 Dosmagulova, Karlygash (Kazakhstan)  
 Dotti, Isabel (Argentina)  
 Douillet, Pierre (France)  
 Dowla, Arif (Bangladesh)  
 Drach, Kostiantyn (Ukraine)  
 Drouard, Genevieve (France)  
 Drugeon, Jean-Pierre (France)  
 Drummond-Cole, Gabriel C. (USA)  
 Du, Chengyong (P.R. China)  
 Du, Wenxue (P.R. China)  
 Du, Yuling (P.R. China)

Duan, Ben (P.R. China)  
 Dubey, Ritesh (India)  
 Dubey, Shruti (India)  
 Duchin, Moon (USA)  
 Duchnak, Martin (USA)  
 Dudek, Andrzej (USA)  
 Duduchava, Roland (Georgia)  
 Duk, Yong (Republic of Korea)  
 Dumitrescu, Sorin (France)  
 Dunbayev, Yerkin (Kazakhstan)  
 Dunham, Douglas (USA)  
 Dunne, Edward (USA)  
 Duong, Duc (Vietnam)  
 Duplantier, Bertrand (France)  
 Dyatlov, Semen (USA)  
 Dzhililov, Akhtam (Uzbekistan)  
 Dzhuraev, Abubakir (Kyrgyzstan)  
 Eberhard, Sean (UK)  
 Eden, Richard (Philippines)  
 Eegunjobi, Adetayo (Namibia)  
 Efendiev, Yalchin (USA)  
 Eisenbrand, Friedrich (Switzerland)  
 Eisenbud, David (USA)  
 El Dhaba, Amr (Egypt)  
 El Haouma, Siham (Morocco)  
 El Tom, Mohamed (Sudan)  
 El Yacoubi, Nouzha (Morocco)  
 Elduque, Alberto (Spain)  
 El-Fayez, Faiza (Saudi Arabia)  
 Elgarem, Noha (Egypt)  
 Elizalde, Emilio (Spain)  
 Elizalde, Sergi (USA)  
 Elizar, Elizar (Australia)  
 El-Khatib, Youssef (UAE)  
 Elsabaa, Fawzy (Egypt)  
 Emerton, Helena (USA)  
 Emerton, Matthew (USA)  
 Emerton, Nicholas (USA)  
 Endam, Joemar (Philippines)  
 Eneya, Levis (Malawi)  
 England, Matthew (UK)  
 Engler, Tina (Germany)  
 Enomoto, Kazuyuki (Japan)  
 Enomoto, Kyoko (Japan)  
 Entov, Michael (Israel)  
 Eom, Dohyeon (Republic of Korea)  
 Eom, Jieun (Republic of Korea)  
 Eom, Jihee (Republic of Korea)  
 Eom, Junyong (Republic of Korea)  
 Eom, Seong Sik (Republic of Korea)  
 Eom, Seungjae (Republic of Korea)  
 Eom, Soo Kyung (Republic of Korea)  
 Eom, Tae Kang (Republic of Korea)  
 Ephremidze, Lasha (Georgia)  
 Erdős, Laszlo (Austria)  
 Ergenc, Hulya (Turkey)  
 Erkursun, Nazife (Turkey)  
 Erlacher, Evelina (Austria)  
 Erokhovets, Nikolay (Russia)  
 Esenturk, Emre (Turkey)  
 Esnault, Helene (Germany)  
 Espinar, Jose (Brazil)  
 Essel, Emmanuel Kwame (Ghana)  
 Estrada, Yuriria (Mexico)  
 Esztergombi, Katalin (Hungary)  
 Etukudo, Udobia (Nigeria)  
 Euh, Yunhee (Republic of Korea)  
 Eum, Ick Sun (Republic of Korea)  
 Eun, Kyoungsoon (Republic of Korea)  
 Eun, Nam Hyun (Republic of Korea)  
 Exner, Pavel (Czech Republic)  
 Exnerova, Jana (Czech Republic)  
 Eynard, Bertrand (France)  
 Ezome, Tony (Gabon)  
 Facchini, Alberto (Italy)  
 Fairag, Faisal (Saudi Arabia)  
 Fairweather, Carol (USA)  
 Fairweather, Graeme (USA)  
 Faminskii, Andrei (Russia)  
 Fan, Jiuyu (P.R. China)  
 Fang, Daoyuan (P.R. China)  
 Fang, Fuquan (P.R. China)  
 Fang, Jinhui (P.R. China)  
 Farag, Omima (Egypt)  
 Farah, Ilijas (Canada)  
 Farb, Beatrice (USA)  
 Farb, Benson (USA)  
 Farb, Felix (USA)  
 Farge, Marie (France)  
 Farhoodi, Roozbeh (Iran)  
 Fariborzi Araghi, Mohammad Ali (Iran)  
 Farjudian, Amin (P.R. China)  
 Farr, Graham (Australia)  
 Fathi, Albert (France)  
 Fathi, Max (France)  
 Favoco, David (Canada)  
 Fedoseev, Alexey (Russia)  
 Felshtyn, Alexander (Poland)  
 Fenecios, Jonald (Philippines)  
 Feng, Yanquan (P.R. China)  
 Ferenczi, Sebastien (France)  
 Fermanian Kammerer, Clotilde (France)  
 Fernandez, David (Spain)  
 Ferrari, Ana María (Uruguay)

Figalli, Alessio (USA)  
 Filinovskiy, Alexey (Russia)  
 Filip, Simion (USA)  
 Fintushel, Ronald (USA)  
 Fisher, Tom (UK)  
 Flores Espinoza, Ruben (Mexico)  
 Flores-Pena, Andrea (USA)  
 Fock, Vladimir (France)  
 Fong, Chamberlain (USA)  
 Foondun, Mohammud (UK)  
 Foulon, Corinne (France)  
 Foulon, Patrick (France)  
 Foupouagnigni, Mama (Cameroon)  
 Fox, Jacob (USA)  
 Fragoulopoulou, Maria (Greece)  
 Freitag, James (USA)  
 Freitag, James Earl (USA)  
 Fu, Sunyu (P.R. China)  
 Fu, Yongqiang (P.R. China)  
 Fujita, Hajime (Japan)  
 Fujita, Yasutsugu (Japan)  
 Fujiwara, Takashi (Japan)  
 Fukuda, Shigetaka (Japan)  
 Funaki, Tadahisa (Japan)  
 Furman, Alexander (USA)  
 Furuya, Junko (Japan)  
 Furuya, Yasuo (Japan)  
 Futaki, Akito (Japan)  
 Gabai, David (USA)  
 Gagnon, Ludovick (France)  
 Gaidhani, Yogeshri (India)  
 Gaiko, Valery (Belarus)  
 Gajjar, Pravina (Sweden)  
 Galatius, Soren (USA)  
 Galeano Penaloza, Jeanneth (Colombia)  
 Galiya, Taugynbayeva (Kazakhstan)  
 Gallagher, Isabelle (France)  
 Gallian, Joseph (USA)  
 Gan, Wee Teck (Singapore)  
 Ganesan, Arthi (India)  
 Gang, Ju Hyeon (Republic of Korea)  
 Gang, Yeonghun (Republic of Korea)  
 Gao, Jianmin (P.R. China)  
 Gao, Qin (P.R. China)  
 Garcia, Gaston Andres (Argentina)  
 Garcia, Mauro (Colombia)  
 Garcia-Colin, Natalia (Mexico)  
 Garcia Huidobro, Marta (Chile)  
 Garcia Iglesias, Agustin (Argentina)  
 Garcia Martinez, Sandra Carolina (Brazil)  
 Garcia Martinez, Xabier (Spain)  
 Garcia Ramos, Yboon Victoria (Peru)  
 Garciano, Agnes (Philippines)  
 Garg, Shelly (India)  
 Garrisi, Daniele (Italy)  
 Garroni, Adriana (Italy)  
 Gashi, Qendrim (Albania)  
 Gathogo, Margaret (Kenya)  
 Gatsinzi, Jean Baptiste (Namibia)  
 Geisser, Thomas (Japan)  
 Geldhauser, Carina (Germany)  
 Gelfand, Sergei (USA)  
 Gelfreich, Vasily (UK)  
 Gentry, Craig (USA)  
 Geon, Hyukjun (Republic of Korea)  
 George, Santhosh (India)  
 Gerasimov, Anton (Russia)  
 Gerasymova, Tetiana (Ukraine)  
 Getzler, Ezra (USA)  
 Geum, Young Hee (Republic of Korea)  
 Gevorkyan, Ashot (Armenia)  
 Ghahramani, Lili (USA)  
 Ghahramani, Saeed (USA)  
 Ghang, Whan (Republic of Korea)  
 Gharge, Sanjeevani (India)  
 Ghimire, Ram (Nepal)  
 Ghisa, Dorin (Canada)  
 Ghisa, Olivia (Canada)  
 Gholam, Roqia (Pakistan)  
 Gholami, Mohammad (Iran)  
 Ghorpade, Sudhir (India)  
 Ghrist, Robert (USA)  
 Ghys, Étienne (France)  
 Gi Yong, Kwon (Republic of Korea)  
 Gilbert, Anna (USA)  
 Gim, Dong Gyun (Republic of Korea)  
 Gim, Geunho (USA)  
 Ginestet, Cedric (USA)  
 Gjergji, Rexhep (Albania)  
 Gladkov, Alexander (Belarus)  
 Glazyrin, Alexey (USA)  
 Go, Hyunju (Republic of Korea)  
 Godoy Mesquita, Jaqueline (Brazil)  
 Góes Mesquita, Luís (Brazil)  
 Goetze, Friedrich (Germany)  
 Goh, Do Won (Republic of Korea)  
 Goldston, Daniel (USA)  
 Gomes Soares, Marcio (Brazil)  
 Gon, Yasuro (Japan)  
 Goncharov, Maxim (Russia)  
 Gong, Gwang Joe (Republic of Korea)  
 Gong, Woosik (Republic of Korea)  
 Gong, Jianhua (UAE)  
 Gonzales, Salome (Peru)

Gonzalez, Jose (Canada)  
Gonzalez Jimenez, Santos (Spain)  
Goo, Jung Seo (Republic of Korea)  
Gorin, Vadim (Russia)  
Gottlieb, Daniel (USA)  
Gouin-Lamourette, Etienne (France)  
Govindankuttymenon, Sajith (India)  
Gowrisankaran, Chandra (Canada)  
Gowrisankaran, Kohur (Canada)  
Granario, Daryl (Philippines)  
Gray, Jeremy (UK)  
Greaves, Gary (UK)  
Green, Ben (UK)  
Green, Kathryn Kert (USA)  
Green, Mark (USA)  
Greenberg, Ralph (USA)  
Greenwell, Christopher (UK)  
Greuel, Gert-Martin (Germany)  
Grey, Matthias (Denmark)  
Griffiths, Marian (USA)  
Griffiths, Phillip (USA)  
Grimmett, Geoffrey (UK)  
Groenewald, Elizabeth (South Africa)  
Groenewald, Nico (South Africa)  
Gross, Mark (UK)  
Gross, Benedict (USA)  
Grötschel, Martin (Germany)  
Grover, Priyanka (India)  
Groves, Daniel (USA)  
Grünberg, David (Switzerland)  
Gu, Chang Gyu (Republic of Korea)  
Gu, Gun Wong (Republic of Korea)  
Gu, Yoon Hoe (Republic of Korea)  
Guayjarernpanishk, Panat (Thailand)  
Gudapati, Nishanth (Germany)  
Gudapati, Venkata Krishna (Germany)  
Gueye Diagne, Salimata (Senegal)  
Guezane-Lakoud, Assia (Algeria)  
Gui, Luying (P.R. China)  
Guo, Baina (P.R. China)  
Guo, Jinghua (P.R. China)  
Guo, Wei (P.R. China)  
Guo, Zhengguang (P.R. China)  
Guon, Yong Sik (Republic of Korea)  
Gupta, Arvind (India)  
Gupta, Shiv (USA)  
Guralnick, Robert (USA)  
Gürkan, Yasmin (Germany)  
Gutu, Valeriu (Moldova)  
Guven, Busra (Turkey)  
Gwak, Hyeon Gi (Republic of Korea)  
Gwan Young, Park (Republic of Korea)  
Gyllenberg, Mats (Finland)  
Ha, Sang-Wook (Austria)  
Ha, Chang Woo (Republic of Korea)  
Ha, Hun Soung (Republic of Korea)  
Ha, Hyun Jeong (Republic of Korea)  
Ha, Jeong Ho (Republic of Korea)  
Ha, Jin Won (Republic of Korea)  
Ha, Jiyoun (Republic of Korea)  
Ha, Junhong (Republic of Korea)  
Ha, Junsoo (Republic of Korea)  
Ha, Ki Sik (Republic of Korea)  
Ha, Sangwon (Republic of Korea)  
Ha, Seok Min (Republic of Korea)  
Ha, Seung-Yeal (Republic of Korea)  
Ha, Taehun (Republic of Korea)  
Ha, Taeyoung (Republic of Korea)  
Ha, Ye Rim (Republic of Korea)  
Haagerup, Uffe (Denmark)  
Haan, Jaeho (Republic of Korea)  
Haboush, William (USA)  
Hadi, Setiawan (Indonesia)  
Hadid, Samir (UAE)  
Haeng Beom, Shin (Republic of Korea)  
Hagelstein, Paul (USA)  
Haider, Azeem (Saudi Arabia)  
Hailu, Habtu Zegeye (Botswana)  
Hairer, Martin (UK)  
Hairer, Xue-Mei (UK)  
Hales, Thomas (USA)  
Hallnas, Erik (Sweden)  
Hallnas, Eva (Sweden)  
Hallnas, Hanna (Sweden)  
Hallnas, Martin (UK)  
Ham, Dong Hyung (Republic of Korea)  
Ham, Jiyoung (Republic of Korea)  
Ham, Seheon (Republic of Korea)  
Hamada, Tatsuyoshi (Japan)  
Hamenstädt, Ursula (Germany)  
Hamidou, Toure (Burkina Faso)  
Han, Shuyang (Australia)  
Han, Qi (P.R. China)  
Han, Areum (Republic of Korea)  
Han, Chang Ho (Republic of Korea)  
Han, Chang Hui (Republic of Korea)  
Han, Chong-Kyu (Republic of Korea)  
Han, Chulho (Republic of Korea)  
Han, Dong Geun (Republic of Korea)  
Han, Donghun (Republic of Korea)  
Han, Dongkyu (Republic of Korea)  
Han, Huijun (Republic of Korea)  
Han, Hyeon Woo (Republic of Korea)  
Han, Hyun Koo (Republic of Korea)

Han, Hyun Soo (Republic of Korea)  
 Han, Hyun-Gu (Republic of Korea)  
 Han, Jae Woo (Republic of Korea)  
 Han, Jeongyeop (Republic of Korea)  
 Han, Jinsu (Republic of Korea)  
 Han, Jiwoo (Republic of Korea)  
 Han, Jiyoung (Republic of Korea)  
 Han, Jongmin (Republic of Korea)  
 Han, Joon Ho (Republic of Korea)  
 Han, Joon Tack (Republic of Korea)  
 Han, Joon Woo (Republic of Korea)  
 Han, Joung Woo (Republic of Korea)  
 Han, Ju Hyun (Republic of Korea)  
 Han, Jung Won (Republic of Korea)  
 Han, Kangjin (Republic of Korea)  
 Han, Ki Hyun (Republic of Korea)  
 Han, Ki-Reem (Republic of Korea)  
 Han, Kyuwool (Republic of Korea)  
 Han, Mingi (Republic of Korea)  
 Han, Myung Hun (Republic of Korea)  
 Han, Sang-Eon (Republic of Korea)  
 Han, Seok Gun (Republic of Korea)  
 Han, Seok Won (Republic of Korea)  
 Han, Seong Hyeok (Republic of Korea)  
 Han, Seung Min (Republic of Korea)  
 Han, Sol (Republic of Korea)  
 Han, Soo-Ho (Republic of Korea)  
 Han, Soo-Yun (Republic of Korea)  
 Han, Suhyun (Republic of Korea)  
 Han, Sumin (Republic of Korea)  
 Han, Sunhyuk (Republic of Korea)  
 Han, Yonggu (Republic of Korea)  
 Han(Hahn), Sang Geun (Republic of Korea)  
 Hansen, Brittany (USA)  
 Hao, Li (P.R. China)  
 Harada, Shinya (Japan)  
 Harington, Robert (USA)  
 Harris, Michael (France)  
 Harsh, Harsh Vardhan (India)  
 Harutyunyan, Tigran (Armenia)  
 Hashemi, Amir (Iran)  
 Hashimoto, Kenji (Japan)  
 Hashimoto, Takashi (Japan)  
 Hattori, Tae (Japan)  
 Hawksley, Andrea (USA)  
 He, Tian-Xiao (USA)  
 He, Yanxiang (USA)  
 Hee Sang, Ann (Republic of Korea)  
 Hegarty, Peter (Sweden)  
 Hegde, Suresh (India)  
 Helfgott, Harald (France)  
 Hemakul, Wanida (Thailand)  
 Henstridge, John (Australia)  
 Heo, Giseon (Canada)  
 Heo, Eun Ji (Republic of Korea)  
 Heo, Gyong Min (Republic of Korea)  
 Heo, Hye-Rim (Republic of Korea)  
 Heo, Jaeseong (Republic of Korea)  
 Heo, Ji Won (Republic of Korea)  
 Heo, Jihye (Republic of Korea)  
 Heo, Joon (Republic of Korea)  
 Heo, Jun Young (Republic of Korea)  
 Heo, Jung Won (Republic of Korea)  
 Heo, Shin Wook (Republic of Korea)  
 Heo, Sungyeon (Republic of Korea)  
 Heo, Taehyeok (Republic of Korea)  
 Heo, Yeongjae (Republic of Korea)  
 Heo, Yun (Republic of Korea)  
 Hernandez, Daniel (USA)  
 Hernandez-Hernandez, Daniel (Mexico)  
 Hessari, Peyman (Republic of Korea)  
 Hill, Michael (USA)  
 Hingston, Nancy (USA)  
 Hinz, Andreas (Germany)  
 Hinz, Christine (Germany)  
 Hirabayashi, Izumi (Japan)  
 Hirachi, Kengo (Japan)  
 Hirasaka, Mitsugu (Republic of Korea)  
 Hirasawa, Go (Japan)  
 Hirasawa, So (Japan)  
 Hittmeyer, Stefanie (New Zealand)  
 Hlavac, Adam (Czech Republic)  
 Ho, Toan (Vietnam)  
 Ho Dang, Phuc (Vietnam)  
 Ho Young, Kim (Republic of Korea)  
 Hoa, Ta Thi Phuong (Vietnam)  
 Hoang, Truong (Vietnam)  
 Hodgson, Bernard (Canada)  
 Hoegenhaven, Amalie (Denmark)  
 Hogben, Leslie (USA)  
 Hoh, Mi Kyeoung (Republic of Korea)  
 Holden, Helge (Norway)  
 Holdener, Judy (USA)  
 Hollanti, Camilla (Finland)  
 Holmsen, Andreas (Norway)  
 Homma, Masaaki (Japan)  
 Hong, Soonjo (Chile)  
 Hong, Chae Young (Republic of Korea)  
 Hong, Chan Yong (Republic of Korea)  
 Hong, Geomseul (Republic of Korea)  
 Hong, Gilbert (Republic of Korea)  
 Hong, Hansol (Republic of Korea)  
 Hong, Hyeok-Pyo (Republic of Korea)  
 Hong, Hyerim (Republic of Korea)

Hong, Hyunsil (Republic of Korea)  
 Hong, Hyunsook (Republic of Korea)  
 Hong, Kisuk (Republic of Korea)  
 Hong, Kyungpyo (Republic of Korea)  
 Hong, Kyusik (Republic of Korea)  
 Hong, Sangsoo (Republic of Korea)  
 Hong, Seo Jun (Republic of Korea)  
 Hong, Seokhyun (Republic of Korea)  
 Hong, Seoung Wook (Republic of Korea)  
 Hong, Seung Min (Republic of Korea)  
 Hong, Song Woo (Republic of Korea)  
 Hong, Soon Ki (Republic of Korea)  
 Hong, Suehui (Republic of Korea)  
 Hong, Sung Sa (Republic of Korea)  
 Hong, Sunghyun (Republic of Korea)  
 Hong, Won Eui (Republic of Korea)  
 Hong, Yeong Beom (Republic of Korea)  
 Hong, Young Hee (Republic of Korea)  
 Hong, Guixiang (Spain)  
 Hooda, Neha (India)  
 Hoovers, Ingrid (Netherlands)  
 Horita, Vanderlei (Brazil)  
 Horiuchi, Toshio (Japan)  
 Horiuchi, Yukari (Japan)  
 Horoldagva, Batmend (Mongolia)  
 Hoshi, Akinari (Japan)  
 Hou, Haoling (P.R. China)  
 Hounkonnou, Mahouton Norbert (Benin)  
 Hryn, Aliaksandr (Belarus)  
 Hsiao, Chin-Yu (Taiwan)  
 Hu, Hanwen (P.R. China)  
 Hu, Huiying (P.R. China)  
 Hu, Ze-Chun (P.R. China)  
 Hu, Wen-Guei (Taiwan)  
 Hu, Weiwei (USA)  
 Huang, Chengming (P.R. China)  
 Huang, Rung-Tzung (Taiwan)  
 Huashu, Zhan (P.R. China)  
 Huggett, Stephen (UK)  
 Hughes, Kenneth (South Africa)  
 Huh, Joon Suk (Republic of Korea)  
 Huh, Sang Eun (Republic of Korea)  
 Huh, Sukmoon (Republic of Korea)  
 Hui, Kin Ming (Taiwan)  
 Hui-Jae, Heo (Republic of Korea)  
 Hung, Kuo-Chih (Taiwan)  
 Hunt, Susan (Canada)  
 Hunt, John (South Africa)  
 Hunt, Vivien (UK)  
 Hunziker, Markus (USA)  
 Huo, Jinna (P.R. China)  
 Huo, Yongliang (P.R. China)  
 Hur, Joon Hwan (Republic of Korea)  
 Hur, Namook (Republic of Korea)  
 Hur, Youngmi (Republic of Korea)  
 Husniah, Hennie (Indonesia)  
 Huu Du, Nguyen (Vietnam)  
 Huynh, Vu (Vietnam)  
 Hwang, Chi-Ok (Republic of Korea)  
 Hwang, Chul Hyen (Republic of Korea)  
 Hwang, Dongseon (Republic of Korea)  
 Hwang, Eun Seo (Republic of Korea)  
 Hwang, Gyeongha (Republic of Korea)  
 Hwang, Hojun (Republic of Korea)  
 Hwang, Hongtaek (Republic of Korea)  
 Hwang, In Tak (Republic of Korea)  
 Hwang, In Woo (Republic of Korea)  
 Hwang, Iueseong (Republic of Korea)  
 Hwang, Ji-Hun (Republic of Korea)  
 Hwang, Jihye (Republic of Korea)  
 Hwang, Jina (Republic of Korea)  
 Hwang, Jiwon (Republic of Korea)  
 Hwang, Ji-Won (Republic of Korea)  
 Hwang, Jong Yun (Republic of Korea)  
 Hwang, Junha (Republic of Korea)  
 Hwang, Junhyeong (Republic of Korea)  
 Hwang, Jun-Muk (Republic of Korea)  
 Hwang, Kyuchan (Republic of Korea)  
 Hwang, Kyungwon (Republic of Korea)  
 Hwang, Nam Young (Republic of Korea)  
 Hwang, Namhyeok (Republic of Korea)  
 Hwang, Sarang (Republic of Korea)  
 Hwang, Se Min (Republic of Korea)  
 Hwang, Se Won (Republic of Korea)  
 Hwang, Seong Cheol (Republic of Korea)  
 Hwang, Seongha (Republic of Korea)  
 Hwang, Seoyeon (Republic of Korea)  
 Hwang, Seungsu (Republic of Korea)  
 Hwang, Soon Gyu (Republic of Korea)  
 Hwang, Sun Hong (Republic of Korea)  
 Hwang, Sunwook (Republic of Korea)  
 Hwang, Taekgyu (Republic of Korea)  
 Hwang, Woonjae (Republic of Korea)  
 Hwang, Yoon Tae (Republic of Korea)  
 Hwang, Sukjung (UK)  
 Hyeon, David (Republic of Korea)  
 Hyeon, Kim (Republic of Korea)  
 Hyeryeon, Lee (Republic of Korea)  
 Hyodo, Fumitake (Japan)  
 Hyon, Yun Kyong (Republic of Korea)  
 Hytönen, Tuomas (Finland)  
 Hyun, Jihoon (Republic of Korea)  
 Hyun, Jong Yoon (Republic of Korea)  
 Hyun, Sue Yeon (Republic of Korea)

Hyun, Woosik (Republic of Korea)  
 Hyun, Yongwoo (Republic of Korea)  
 Hyun, Yoonsuk (Republic of Korea)  
 Ibrahim, Noor (Malaysia)  
 Ichihara, Kazuhiro (Japan)  
 Ignatov, Ventsislav (Bulgaria)  
 Igodt, Paul (Belgium)  
 Ih, Su-Ion (USA)  
 Itaka, Shigeru (Japan)  
 Ikeda, Koichiro (Japan)  
 Ikeda, Ilhan (Turkey)  
 Illarionov, Andrei (Russia)  
 Ilolov, Mamadsho (Tadjikistan)  
 Im, Bo-Hae (Republic of Korea)  
 Im, Bokhee (Republic of Korea)  
 Im, Dongmin (Republic of Korea)  
 Im, Hyun Jae (Republic of Korea)  
 Im, Jaekyeong (Republic of Korea)  
 Im, Miju (Republic of Korea)  
 Im, Seonghyuk (Republic of Korea)  
 Im, Youngho (Republic of Korea)  
 Indrati, Christiana Rini (Indonesia)  
 Ion, Bonita (USA)  
 Ion, Patrick (USA)  
 Ip, Ivan (Japan)  
 Iranmanesh, Ali (Iran)  
 Ishi, Hideyuki (Japan)  
 Ishida, Atsuhide (Japan)  
 Ishii, Daisuke (Japan)  
 Ishii, Shihoko (Japan)  
 Ismail, Fudziah (Malaysia)  
 Ismail, Zuhaila (Malaysia)  
 Iswadi, Hazrul (Indonesia)  
 Ito, Hidekazu (Japan)  
 Itoh, Jin-Ichi (Japan)  
 Itoh, Mitsuhiro (Japan)  
 Iturriaga, Renato (Mexico)  
 Ivanov, Georgy (Norway)  
 Ivanovici, Danela Oana (France)  
 Iyer, Jaya (India)  
 Izumi, Mariko (Japan)  
 Izumi, Masaki (Japan)  
 Jackowski, Stefan (Poland)  
 Jadala, Venkat Ramana Reddy (India)  
 Jafari Rad, Nader (Iran)  
 Jain, Subit Kumar (India)  
 Jain, Tanvi (India)  
 Jain, Surender (USA)  
 Jambu, Madeleine (France)  
 Jambu, Michel (France)  
 Jampana, Phanindra Varma (India)  
 Jamsranjav, Davaadulam (Mongolia)  
 Jana, Purbita (India)  
 Jang, Alim (Republic of Korea)  
 Jang, Bongsoo (Republic of Korea)  
 Jang, Boreum (Republic of Korea)  
 Jang, Busik (Republic of Korea)  
 Jang, Chang Geun (Republic of Korea)  
 Jang, Changrim (Republic of Korea)  
 Jang, Chanho (Republic of Korea)  
 Jang, Chungjae (Republic of Korea)  
 Jang, Deok-Kyu (Republic of Korea)  
 Jang, Hyeokju (Republic of Korea)  
 Jang, Hyo Seok (Republic of Korea)  
 Jang, Jaeduck (Republic of Korea)  
 Jang, Jiwoong (Republic of Korea)  
 Jang, Jooho (Republic of Korea)  
 Jang, Jun Hyuk (Republic of Korea)  
 Jang, Jun-Ha (Republic of Korea)  
 Jang, Jun-Hyuk (Republic of Korea)  
 Jang, Junmyeong (Republic of Korea)  
 Jang, Junyoung (Republic of Korea)  
 Jang, Kihoon (Republic of Korea)  
 Jang, Kyoungseok (Republic of Korea)  
 Jang, Kyung Min (Republic of Korea)  
 Jang, Ryoung Woo (Republic of Korea)  
 Jang, Seon-Myeong (Republic of Korea)  
 Jang, Seung Uk (Republic of Korea)  
 Jang, Sohyuen (Republic of Korea)  
 Jang, Sun Young (Republic of Korea)  
 Jang, Won Yong (Republic of Korea)  
 Jang, Woong Bi (Republic of Korea)  
 Jang, Yeonhui (Republic of Korea)  
 Jang, Young Seok (Republic of Korea)  
 Jang, Youngjae (Republic of Korea)  
 Jang, Yu Seon (Republic of Korea)  
 Januszkiewicz, Tadeusz (Poland)  
 Jee, Young Myong (Republic of Korea)  
 Jeltsch, Marianne Regula (Switzerland)  
 Jeltsch, Rolf (Switzerland)  
 Jena, Susil Kumar (India)  
 Jenaliyev, Muvasharkhan (Kazakhstan)  
 Jensen, Tommy (Republic of Korea)  
 Jeon, Byungyun (Republic of Korea)  
 Jeon, Daeyeol (Republic of Korea)  
 Jeon, Hansol (Republic of Korea)  
 Jeon, Hanul (Republic of Korea)  
 Jeon, Hong Chan (Republic of Korea)  
 Jeon, Hyeong Jin (Republic of Korea)  
 Jeon, Hyun Ji (Republic of Korea)  
 Jeon, Hyun Soo (Republic of Korea)  
 Jeon, Inho (Republic of Korea)  
 Jeon, Intae (Republic of Korea)  
 Jeon, Jin-Hwan (Republic of Korea)

Jeon, Jiwon (Republic of Korea)  
 Jeon, Jong Duek (Republic of Korea)  
 Jeon, Kangsan (Republic of Korea)  
 Jeon, Kiwan (Republic of Korea)  
 Jeon, Min Gyu (Republic of Korea)  
 Jeon, Myeong Jae (Republic of Korea)  
 Jeon, Sang Wook (Republic of Korea)  
 Jeon, Seongmin (Republic of Korea)  
 Jeon, Sungweon (Republic of Korea)  
 Jeon, Wonju (Republic of Korea)  
 Jeon, Young Ju (Republic of Korea)  
 Jeon, Youngmok (Republic of Korea)  
 Jeong, Byeongin (Republic of Korea)  
 Jeong, Chang Hyeon (Republic of Korea)  
 Jeong, Chan-Hyuck (Republic of Korea)  
 Jeong, Chan-Woo (Republic of Korea)  
 Jeong, Chanyeong (Republic of Korea)  
 Jeong, Chulhwan (Republic of Korea)  
 Jeong, Daegyem (Republic of Korea)  
 Jeong, Daehyeon (Republic of Korea)  
 Jeong, Dahee (Republic of Korea)  
 Jeong, Deokhyun (Republic of Korea)  
 Jeong, Eunhee (Republic of Korea)  
 Jeong, Eunji (Republic of Korea)  
 Jeong, Eunju (Republic of Korea)  
 Jeong, Ga Ram (Republic of Korea)  
 Jeong, Ha Seung (Republic of Korea)  
 Jeong, Han-Young (Republic of Korea)  
 Jeong, Hyeon Wi (Republic of Korea)  
 Jeong, Imsoon (Republic of Korea)  
 Jeong, In Chan (Republic of Korea)  
 Jeong, Ja A (Republic of Korea)  
 Jeong, Jae Yoon (Republic of Korea)  
 Jeong, Jaehyun (Republic of Korea)  
 Jeong, Jeongmi (Republic of Korea)  
 Jeong, Ji In (Republic of Korea)  
 Jeong, Jiin (Republic of Korea)  
 Jeong, Jinhyuck (Republic of Korea)  
 Jeong, Jinyoung (Republic of Korea)  
 Jeong, Jisu (Republic of Korea)  
 Jeong, Jiye (Republic of Korea)  
 Jeong, Jooheon (Republic of Korea)  
 Jeong, Jun Sik (Republic of Korea)  
 Jeong, Keunyoung (Republic of Korea)  
 Jeong, Kwang-Woo (Republic of Korea)  
 Jeong, Kyeonghoon (Republic of Korea)  
 Jeong, Kyung Chul (Republic of Korea)  
 Jeong, Min Hee (Republic of Korea)  
 Jeong, Min Su (Republic of Korea)  
 Jeong, Moonja (Republic of Korea)  
 Jeong, Myeongjin (Republic of Korea)  
 Jeong, Myeong-Ju (Republic of Korea)  
 Jeong, Naehyeok (Republic of Korea)  
 Jeong, Piljun (Republic of Korea)  
 Jeong, Sangtae (Republic of Korea)  
 Jeong, Se Hun (Republic of Korea)  
 Jeong, Seok Hyeon (Republic of Korea)  
 Jeong, Seonggu (Republic of Korea)  
 Jeong, Seonghee (Republic of Korea)  
 Jeong, Seongjin (Republic of Korea)  
 Jeong, Seulgi (Republic of Korea)  
 Jeong, Seung Hoon (Republic of Korea)  
 Jeong, Seungwon (Republic of Korea)  
 Jeong, Sihyun (Republic of Korea)  
 Jeong, Soyeong (Republic of Korea)  
 Jeong, Su Min (Republic of Korea)  
 Jeong, Sung-Jo (Republic of Korea)  
 Jeong, Sunny (Republic of Korea)  
 Jeong, Ui-Hyeon (Republic of Korea)  
 Jeong, Wonbo (Republic of Korea)  
 Jeong, Wonje (Republic of Korea)  
 Jeong, Woo Seock (Republic of Korea)  
 Jeong, Yewon (Republic of Korea)  
 Jeong, In-Jee (USA)  
 Jeronimo, Gabriela Tali (Argentina)  
 Jerrard, Robert (Canada)  
 Jha, Kanhaiya (Nepal)  
 Ji, Shaolin (P.R. China)  
 Ji, Sehyun (Republic of Korea)  
 Ji, Si Yun (Republic of Korea)  
 Ji, Un Cig (Republic of Korea)  
 Ji, Yonggwon (Republic of Korea)  
 Ji, Yun-Seong (Republic of Korea)  
 Ji, Lizhen (USA)  
 Jia, Yanhe (P.R. China)  
 Jiang, Jinxi (P.R. China)  
 Jiang, Yunping (USA)  
 Jill, Mesirov (USA)  
 Jin, Hailan (P.R. China)  
 Jin, Kai (P.R. China)  
 Jin, Yinglie (P.R. China)  
 Jin, Yuanfeng (P.R. China)  
 Jin, Zhezhi (P.R. China)  
 Jin, Gyo Taek (Republic of Korea)  
 Jin, Hong Sung (Republic of Korea)  
 Jin, Hyeokjun (Republic of Korea)  
 Jin, Hyeong-Jun (Republic of Korea)  
 Jin, Jin-Hee (Republic of Korea)  
 Jin, Jonguhn (Republic of Korea)  
 Jin, Joohwan (Republic of Korea)  
 Jin, Ju Chan (Republic of Korea)  
 Jin, Sangdon (Republic of Korea)  
 Jin, Seokho (Republic of Korea)  
 Jin, Seokjo (Republic of Korea)

- Jin, Seong Hoon (Republic of Korea)  
 Jin, Sun Sook (Republic of Korea)  
 Jin, Yongtak (Republic of Korea)  
 Jing, Naihuan (USA)  
 Jitman, Somphong (Thailand)  
 Jo, Baek Gun (Republic of Korea)  
 Jo, Changhun (Republic of Korea)  
 Jo, Chang-Seong (Republic of Korea)  
 Jo, Hailey (Republic of Korea)  
 Jo, Han Seong (Republic of Korea)  
 Jo, Hong-Kwon (Republic of Korea)  
 Jo, Hyung-Rok (Republic of Korea)  
 Jo, Hyunju (Republic of Korea)  
 Jo, Janghyun (Republic of Korea)  
 Jo, Jun Hong (Republic of Korea)  
 Jo, Ki-Yeong (Republic of Korea)  
 Jo, Minseo (Republic of Korea)  
 Jo, Myung Sang (Republic of Korea)  
 Jo, Seonggeun (Republic of Korea)  
 Jo, Sihun (Republic of Korea)  
 Jo, Sung Bin (Republic of Korea)  
 Jo, Young-Hun (Republic of Korea)  
 Joachin, Heinze (Germany)  
 Joe, Albert (Republic of Korea)  
 Joeng, Eunhee (Republic of Korea)  
 Joh, Hyun Jun (Republic of Korea)  
 Joh, Jae Wan (Republic of Korea)  
 Joh, Seong Eun (Republic of Korea)  
 John, Cotrina (Peru)  
 Johnson, Sam (India)  
 Johnston, Desmond (UK)  
 Jones, Martha (USA)  
 Jones, Vaughan (USA)  
 Joo, Geon (Republic of Korea)  
 Joo, Hayeon (Republic of Korea)  
 Joo, Jaeun (Republic of Korea)  
 Joo, Jeongeun (Republic of Korea)  
 Joon, Kwon (Republic of Korea)  
 Jooseop, Shin (Republic of Korea)  
 Joshi, Nalini (Australia)  
 Joswig, Michael (Germany)  
 Joung, Yewon (Republic of Korea)  
 Joung, Yong Sig (Republic of Korea)  
 Joussaume, Aurélie (France)  
 Ju, Hyeong-Kwan (Republic of Korea)  
 Ju, Jeonghoon (Republic of Korea)  
 Ju, Jung Hun (Republic of Korea)  
 Judge, Jonathan (USA)  
 Jueun, Kang (Republic of Korea)  
 Jumakhayeva, Gulbarshin (Kazakhstan)  
 Jun, Hyunmi (Republic of Korea)  
 Jun, Jae Uk (Republic of Korea)  
 Jun, Jongha (Republic of Korea)  
 Jun, Sang Geun (Republic of Korea)  
 Jun, Sook Heui (Republic of Korea)  
 Jun, Sung Chan (Republic of Korea)  
 Jung, Bo Young (Republic of Korea)  
 Jung, Chul Yoon (Republic of Korea)  
 Jung, Daewon (Republic of Korea)  
 Jung, Eunok (Republic of Korea)  
 Jung, Hae Won (Republic of Korea)  
 Jung, Han Sol (Republic of Korea)  
 Jung, Hongtaek (Republic of Korea)  
 Jung, Hyeyoung (Republic of Korea)  
 Jung, Hyojin (Republic of Korea)  
 Jung, Hyun Sik (Republic of Korea)  
 Jung, Hyunmyung (Republic of Korea)  
 Jung, Jaewon (Republic of Korea)  
 Jung, Jang Yong (Republic of Korea)  
 Jung, Ji Hye (Republic of Korea)  
 Jung, Ji Hyung (Republic of Korea)  
 Jung, Ji Yoon (Republic of Korea)  
 Jung, Jin Wook (Republic of Korea)  
 Jung, Jong Soo (Republic of Korea)  
 Jung, Joon Hyuk (Republic of Korea)  
 Jung, Jun Young (Republic of Korea)  
 Jung, Junehyuk (Republic of Korea)  
 Jung, Junhwa (Republic of Korea)  
 Jung, Keehoon (Republic of Korea)  
 Jung, Kwang Man (Republic of Korea)  
 Jung, Kyomin (Republic of Korea)  
 Jung, Lee Soo (Republic of Korea)  
 Jung, Miyoung (Republic of Korea)  
 Jung, Seoung Dal (Republic of Korea)  
 Jung, Soim (Republic of Korea)  
 Jung, Soon-Mo (Republic of Korea)  
 Jung, Soyeun (Republic of Korea)  
 Jung, Sung Min (Republic of Korea)  
 Jung, Sungeun (Republic of Korea)  
 Jung, Tacksun (Republic of Korea)  
 Jung, Taeup (Republic of Korea)  
 Jung, Uijin (Republic of Korea)  
 Jung, Won Seok (Republic of Korea)  
 Jung, Woo Ki (Republic of Korea)  
 Jung, Woochul (Republic of Korea)  
 Jung, Woojung (Republic of Korea)  
 Jung, Woo-Seok (Republic of Korea)  
 Jung, Woosung (Republic of Korea)  
 Jung, Yoon Mo (Republic of Korea)  
 Jung, Young-Han (Republic of Korea)  
 Jung, Younghoon (Republic of Korea)  
 Jung, Joeun (USA)  
 Jung, Paul (USA)  
 Just, Andrzej (Poland)

Juyumaya, Jesus (Chile)  
 K.C., Gokul (Nepal)  
 Kabulov, Armanbek (Kazakhstan)  
 Kagunda, Josephine (Kenya)  
 Kahn, Jeremy (USA)  
 Kahng, Byeong Hoon (Republic of Korea)  
 Kalaj, David (Montenegro)  
 Kalelkar, Tejas (India)  
 Kalidass, Mathiyalagan (India)  
 Kalimoldayev, Maksat (Kazakhstan)  
 Kalmenov, Tynysbek (Kazakhstan)  
 Kalmenova, Antonina (Kazakhstan)  
 Kamaku, Peter Waweru (Kenya)  
 Kamalakkannan, Kalaivani (India)  
 Kameda, Masumi (Japan)  
 Kamiyoshi, Tomohiro (Japan)  
 Kamran, Tayyab (Pakistan)  
 Kanas, Stanislaw (Poland)  
 Kang, Mihyun (Austria)  
 Kang, Sung Joo (Canada)  
 Kang, Shuhua (P.R. China)  
 Kang, Sooran (New Zealand)  
 Kang, B.G. (Republic of Korea)  
 Kang, Bong Gwan (Republic of Korea)  
 Kang, Bowon (Republic of Korea)  
 Kang, Bumtle (Republic of Korea)  
 Kang, Byungsoo (Republic of Korea)  
 Kang, Chan Hee (Republic of Korea)  
 Kang, Changmin (Republic of Korea)  
 Kang, Chulmin (Republic of Korea)  
 Kang, Chunghyuk (Republic of Korea)  
 Kang, Deok Hun (Republic of Korea)  
 Kang, Deokhoon (Republic of Korea)  
 Kang, Dong Yeap (Republic of Korea)  
 Kang, Du Ho (Republic of Korea)  
 Kang, Eun Ji (Republic of Korea)  
 Kang, Geon (Republic of Korea)  
 Kang, Hee-Won (Republic of Korea)  
 Kang, Ho Jin (Republic of Korea)  
 Kang, Ho Joon (Republic of Korea)  
 Kang, Hyeonbae (Republic of Korea)  
 Kang, Hyeongwoo (Republic of Korea)  
 Kang, Hyosang (Republic of Korea)  
 Kang, Hyun Doo (Republic of Korea)  
 Kang, Hyungcheol (Republic of Korea)  
 Kang, Hyunsuk (Republic of Korea)  
 Kang, Il Se (Republic of Korea)  
 Kang, Jeang Min (Republic of Korea)  
 Kang, Jee Yeon (Republic of Korea)  
 Kang, Jeong Hoon (Republic of Korea)  
 Kang, Ji Hyun (Republic of Korea)  
 Kang, Jung Hun (Republic of Korea)  
 Kang, Junoh (Republic of Korea)  
 Kang, Kyubong (Republic of Korea)  
 Kang, Kyungkeun (Republic of Korea)  
 Kang, Mi Yeong (Republic of Korea)  
 Kang, Moon Seok (Republic of Korea)  
 Kang, Myeongmin (Republic of Korea)  
 Kang, Myungjoo (Republic of Korea)  
 Kang, Naehun (Republic of Korea)  
 Kang, Nam-Gyu (Republic of Korea)  
 Kang, Pricillia (Republic of Korea)  
 Kang, Pyung-Lyun (Republic of Korea)  
 Kang, Seok-Jin (Republic of Korea)  
 Kang, Seong Mo (Republic of Korea)  
 Kang, Seong Su (Republic of Korea)  
 Kang, Seonghyeon (Republic of Korea)  
 Kang, Seongku (Republic of Korea)  
 Kang, Seungkoo (Republic of Korea)  
 Kang, Sinuk (Republic of Korea)  
 Kang, Soonja (Republic of Korea)  
 Kang, Soon-Yi (Republic of Korea)  
 Kang, Sun Jong (Republic of Korea)  
 Kang, Sunbu (Republic of Korea)  
 Kang, Sung Hee (Republic of Korea)  
 Kang, Sunghee (Republic of Korea)  
 Kang, Sunggho (Republic of Korea)  
 Kang, Sunghyun (Republic of Korea)  
 Kang, Sungkyung (Republic of Korea)  
 Kang, Sungmo (Republic of Korea)  
 Kang, Wanmo (Republic of Korea)  
 Kang, Wonwoo (Republic of Korea)  
 Kang, Yeonghun (Republic of Korea)  
 Kang, Yeongsook (Republic of Korea)  
 Kang, Youn Seung (Republic of Korea)  
 Kang, Young Chan (Republic of Korea)  
 Kang, Yun Seok (Republic of Korea)  
 Kang, Dae Han (USA)  
 Kania-Bartoszynska, Joanna (USA)  
 Kapustka, Grzegorz (Poland)  
 Kara Hansen, Ayse (Turkey)  
 Karaali, Gizem (USA)  
 Karabash, Illia (Ukraine)  
 Karandikar, Rajeeva (India)  
 Karim, Ham (Cambodia)  
 Karimjanov, Ikboljon (Uzbekistan)  
 Karimov, Umed (Tadjikistan)  
 Karimov, Erkinjon (Uzbekistan)  
 Karimov, Jasurbek (Uzbekistan)  
 Karjanto, Natanael (Indonesia)  
 Karmanova, Maria (Russia)  
 Kartiko, Sri Haryatmi (Indonesia)  
 Kasickova, Linda (Czech Republic)  
 Kasimov, Vagif (Azerbaijan)

- Kassa, Semu Mitiku (Ethiopia)  
 Kassabov, Martin (USA)  
 Kassel, Fanny (France)  
 Katagi, Nagaraj (India)  
 Kathuria, Leetika (India)  
 Kato, Keiichi (Japan)  
 Katona, Gyula O.H. (Hungary)  
 Katz, Nets (USA)  
 Kaur, Harpreet (India)  
 Kawamata, Yujiro (Japan)  
 Kayar, Zeynep (Turkey)  
 Kayvanfar, Saeed (Iran)  
 Kazachkov, Ilya (UK)  
 Kearsley, Anthony (USA)  
 Kedem, Rinat (USA)  
 Kedlaya, Kiran (USA)  
 Kedukodi, Babushri Srinivas (India)  
 Keller, Mitchel (USA)  
 Keller, Thomas (USA)  
 Kemoklidze, Tariel (Georgia)  
 Kempainen, Antti (Finland)  
 Kenig, Carlos (USA)  
 Ker Hsin, Ong (Malaysia)  
 Kerckhoff, Steve (USA)  
 Kerimbekov, Akylbek (Kyrgyzstan)  
 Kerschner, Stephanie (Germany)  
 Keum, Jeon Young (Republic of Korea)  
 Keum, Jihoon (Republic of Korea)  
 Keum, Jonghae (Republic of Korea)  
 Keune, Frans (Netherlands)  
 Keyfitz, Barbara (USA)  
 Khabelashvili, Albert (Russia)  
 Khaldi, Rabah (Algeria)  
 Khammash, Ahmed A. (Saudi Arabia)  
 Khan, Arshad (India)  
 Khan, Kamran (India)  
 Khan, Nadia (Pakistan)  
 Khanduja, Sudesh (India)  
 Khanevsky, Michael (USA)  
 Kharat, Archana (India)  
 Kharat, Vilas (India)  
 Kharchenko, Vladislav (Mexico)  
 Kharlampovich, Olga (Canada)  
 Khelemskiy, Alexander (Russia)  
 Khim, Dongho (USA)  
 Kho, Dong Yeong (Republic of Korea)  
 Khosrovshahi, Gholamreza B. (Iran)  
 Khot, Amol (India)  
 Khot, Jayashree (India)  
 Khot, Subhash (India)  
 Khot, Neev (USA)  
 Khots, Boris (USA)  
 Khudoyberdiyev, Abror (Uzbekistan)  
 Khumbah, Nkem (USA)  
 Ki, Dohyeong (Republic of Korea)  
 Ki, Nohyun (Republic of Korea)  
 Ki, Sungock (Republic of Korea)  
 Kiefer, Frank (Germany)  
 Kiem, Young-Hoon (Republic of Korea)  
 Kifle, Yirgalem Tsegaye (Ethiopia)  
 Kikuchi, Keiichi (Japan)  
 Kil, Seung Ho (Republic of Korea)  
 Kilicman, Adem (Malaysia)  
 Kim, Hansol (Canada)  
 Kim, Young-Heon (Canada)  
 Kim, A Young (Republic of Korea)  
 Kim, Au Jin (Republic of Korea)  
 Kim, Bara (Republic of Korea)  
 Kim, Beom Jin (Republic of Korea)  
 Kim, Beomseok (Republic of Korea)  
 Kim, Bo Kyoung (Republic of Korea)  
 Kim, Bohyun (Republic of Korea)  
 Kim, Bokki (Republic of Korea)  
 Kim, Boran (Republic of Korea)  
 Kim, Boseong (Republic of Korea)  
 Kim, Bum Soo (Republic of Korea)  
 Kim, Bum Soo (Republic of Korea)  
 Kim, Bum Su (Republic of Korea)  
 Kim, Bumsig (Republic of Korea)  
 Kim, Byeong Chan (Republic of Korea)  
 Kim, Byeorhi (Republic of Korea)  
 Kim, Byoung Soo (Republic of Korea)  
 Kim, Byoung-II (Republic of Korea)  
 Kim, Byung Chan (Republic of Korea)  
 Kim, Byung Chun (Republic of Korea)  
 Kim, Byung Hak (Republic of Korea)  
 Kim, Byunghan (Republic of Korea)  
 Kim, Byunghoon (Republic of Korea)  
 Kim, Catherine Eun-Young (Republic of Korea)  
 Kim, Chae Lin (Republic of Korea)  
 Kim, Chan Jin (Republic of Korea)  
 Kim, Chan Kyo (Republic of Korea)  
 Kim, Chang Heon (Republic of Korea)  
 Kim, Chang Hun (Republic of Korea)  
 Kim, Chang Hyun (Republic of Korea)  
 Kim, Chang Ik (Republic of Korea)  
 Kim, Changjoon (Republic of Korea)  
 Kim, Chan-Gyun (Republic of Korea)  
 Kim, Cheolhyeong (Republic of Korea)  
 Kim, Chong Woo (Republic of Korea)  
 Kim, Chul Ho (Republic of Korea)  
 Kim, Chul Jun (Republic of Korea)  
 Kim, Dae Il (Republic of Korea)  
 Kim, Dae June (Republic of Korea)

Kim, Dae San (Republic of Korea)  
Kim, Dae Young (Republic of Korea)  
Kim, Daehwan (Republic of Korea)  
Kim, Daeyeoul (Republic of Korea)  
Kim, Dai-Sik (Republic of Korea)  
Kim, Dano (Republic of Korea)  
Kim, Davin (Republic of Korea)  
Kim, Deok Hyeon (Republic of Korea)  
Kim, Do Hyeon (Republic of Korea)  
Kim, Do Hyeon (Republic of Korea)  
Kim, Do Hyun (Republic of Korea)  
Kim, Do Hyun (Republic of Korea)  
Kim, Do Sang (Republic of Korea)  
Kim, Do Young (Republic of Korea)  
Kim, Dohan (Republic of Korea)  
Kim, Dohhoon (Republic of Korea)  
Kim, Dohyeong (Republic of Korea)  
Kim, Do-Hyung (Republic of Korea)  
Kim, Do-Hyung (Republic of Korea)  
Kim, Dokyoung (Republic of Korea)  
Kim, Dong Han (Republic of Korea)  
Kim, Dong Hoon (Republic of Korea)  
Kim, Dong Hoon (Republic of Korea)  
Kim, Dong Hwan (Republic of Korea)  
Kim, Dong Hyeon (Republic of Korea)  
Kim, Dong Kyu (Republic of Korea)  
Kim, Dong Seo (Republic of Korea)  
Kim, Dong Wook (Republic of Korea)  
Kim, Donggeun (Republic of Korea)  
Kim, Dongguen (Republic of Korea)  
Kim, Donggun (Republic of Korea)  
Kim, Dongha (Republic of Korea)  
Kim, Dongho (Republic of Korea)  
Kim, Donghyeon (Republic of Korea)  
Kim, Donghyun (Republic of Korea)  
Kim, Donghyun (Republic of Korea)  
Kim, Dongjun (Republic of Korea)  
Kim, Dong-Min (Republic of Korea)  
Kim, Dong-Ryul (Republic of Korea)  
Kim, Dong-Soo (Republic of Korea)  
Kim, Dongsu (Republic of Korea)  
Kim, Donguk (Republic of Korea)  
Kim, Donguk (Republic of Korea)  
Kim, Dongwoo (Republic of Korea)  
Kim, Dongyung (Republic of Korea)  
Kim, Down-Woon (Republic of Korea)  
Kim, Doyeong (Republic of Korea)  
Kim, Doyoung (Republic of Korea)  
Kim, Doyun (Republic of Korea)  
Kim, Du Gyu (Republic of Korea)  
Kim, Edward (Republic of Korea)  
Kim, Eunjung (Republic of Korea)  
Kim, Eun-Kyung (Republic of Korea)  
Kim, Eunmi (Republic of Korea)  
Kim, Gang Chan (Republic of Korea)  
Kim, Geon Ha (Republic of Korea)  
Kim, Geun Ho (Republic of Korea)  
Kim, Gi Su (Republic of Korea)  
Kim, Gi Yong (Republic of Korea)  
Kim, Gi Yong (Republic of Korea)  
Kim, Gunwoo (Republic of Korea)  
Kim, Gwang Hui (Republic of Korea)  
Kim, Gyu Jong (Republic of Korea)  
Kim, Gyu Yeol (Republic of Korea)  
Kim, Ha Eun (Republic of Korea)  
Kim, Ha Young (Republic of Korea)  
Kim, Han (Republic of Korea)  
Kim, Hanbyul (Republic of Korea)  
Kim, Hansol (Republic of Korea)  
Kim, Harim (Republic of Korea)  
Kim, Hark-Mahn (Republic of Korea)  
Kim, Hawoon (Republic of Korea)  
Kim, Hayan (Republic of Korea)  
Kim, Heejin (Republic of Korea)  
Kim, Heonnam (Republic of Korea)  
Kim, Heung Gyu (Republic of Korea)  
Kim, Hoil (Republic of Korea)  
Kim, Hong Kyun (Republic of Korea)  
Kim, Hongin (Republic of Korea)  
Kim, Hong-Jong (Republic of Korea)  
Kim, Hongsook (Republic of Korea)  
Kim, Hong-Suk (Republic of Korea)  
Kim, Hoonjoo (Republic of Korea)  
Kim, Hosung (Republic of Korea)  
Kim, Hun (Republic of Korea)  
Kim, Hun-Jae (Republic of Korea)  
Kim, Hunnam (Republic of Korea)  
Kim, Hwajoon (Republic of Korea)  
Kim, Hwankoo (Republic of Korea)  
Kim, Hwi-Dong (Republic of Korea)  
Kim, Hye Ji (Republic of Korea)  
Kim, Hye Ryeong (Republic of Korea)  
Kim, Hyeji (Republic of Korea)  
Kim, Hyeon Jin (Republic of Korea)  
Kim, Hyesun (Republic of Korea)  
Kim, Hyoeun (Republic of Korea)  
Kim, Hyoung Min (Republic of Korea)  
Kim, Hyuk (Republic of Korea)  
Kim, Hyun (Republic of Korea)  
Kim, Hyun Ah (Republic of Korea)  
Kim, Hyun Dong (Republic of Korea)  
Kim, Hyun Jin (Republic of Korea)  
Kim, Hyun Jin (Republic of Korea)  
Kim, Hyun Joong (Republic of Korea)

Kim, Hyun Jung (Republic of Korea)  
 Kim, Hyun Ki (Republic of Korea)  
 Kim, Hyun Kyu (Republic of Korea)  
 Kim, Hyun Moon (Republic of Korea)  
 Kim, Hyun Woo (Republic of Korea)  
 Kim, Hyung Jeon (Republic of Korea)  
 Kim, Hyung Seop (Republic of Korea)  
 Kim, Hyung-Jun (Republic of Korea)  
 Kim, Hyungyoon (Republic of Korea)  
 Kim, Hyun-Min (Republic of Korea)  
 Kim, Hyunseok (Republic of Korea)  
 Kim, Hyuntae (Republic of Korea)  
 Kim, Hyunyeon (Republic of Korea)  
 Kim, Ihn Sue (Republic of Korea)  
 Kim, Ilgirn (Republic of Korea)  
 Kim, In-Kyun (Republic of Korea)  
 Kim, Inseo (Republic of Korea)  
 Kim, Insu (Republic of Korea)  
 Kim, Jae Deok (Republic of Korea)  
 Kim, Jae Hee (Republic of Korea)  
 Kim, Jae Hyeon (Republic of Korea)  
 Kim, Jae Min (Republic of Korea)  
 Kim, Jae Uk (Republic of Korea)  
 Kim, Jaeho (Republic of Korea)  
 Kim, Jaehoon (Republic of Korea)  
 Kim, Jaehyun (Republic of Korea)  
 Kim, Jaehyung (Republic of Korea)  
 Kim, Jaewoo (Republic of Korea)  
 Kim, Jaeyoung (Republic of Korea)  
 Kim, Jaeyoung (Republic of Korea)  
 Kim, Jang Soo (Republic of Korea)  
 Kim, Jeasoo (Republic of Korea)  
 Kim, Jeewook (Republic of Korea)  
 Kim, Jeong Ah (Republic of Korea)  
 Kim, Jeong Han (Republic of Korea)  
 Kim, Jeong Ho (Republic of Korea)  
 Kim, Jeong In (Republic of Korea)  
 Kim, Jeong Ok (Republic of Korea)  
 Kim, Jeong San (Republic of Korea)  
 Kim, Jeong-Gyoo (Republic of Korea)  
 Kim, Jeongho (Republic of Korea)  
 Kim, Jeonghun (Republic of Korea)  
 Kim, Jeongju (Republic of Korea)  
 Kim, Jeongook (Republic of Korea)  
 Kim, Jeong-Rae (Republic of Korea)  
 Kim, Jeongseop (Republic of Korea)  
 Kim, Jeongsu (Republic of Korea)  
 Kim, Jeungyoon (Republic of Korea)  
 Kim, Ji Hun (Republic of Korea)  
 Kim, Ji Su (Republic of Korea)  
 Kim, Ji Sun (Republic of Korea)  
 Kim, Ji Woo (Republic of Korea)  
 Kim, Ji Yeon (Republic of Korea)  
 Kim, Ji Young (Republic of Korea)  
 Kim, Jieon (Republic of Korea)  
 Kim, Jigu (Republic of Korea)  
 Kim, Jihwan (Republic of Korea)  
 Kim, Ji-Hye (Republic of Korea)  
 Kim, Jin Hee (Republic of Korea)  
 Kim, Jinha (Republic of Korea)  
 Kim, Jinhwa (Republic of Korea)  
 Kim, Jinhung (Republic of Korea)  
 Kim, Jinsu (Republic of Korea)  
 Kim, Jinsu (Republic of Korea)  
 Kim, Jinuk (Republic of Korea)  
 Kim, Jiseong (Republic of Korea)  
 Kim, Jisu (Republic of Korea)  
 Kim, Jiwon (Republic of Korea)  
 Kim, Jiwon (Republic of Korea)  
 Kim, Jiwon (Republic of Korea)  
 Kim, Jiyeon (Republic of Korea)  
 Kim, Jiyun (Republic of Korea)  
 Kim, Jjun (Republic of Korea)  
 Kim, Jong Jin (Republic of Korea)  
 Kim, Jong Kyu (Republic of Korea)  
 Kim, Jong Ryul (Republic of Korea)  
 Kim, Jong Suk (Republic of Korea)  
 Kim, Jong Won (Republic of Korea)  
 Kim, Jongchan (Republic of Korea)  
 Kim, Jongeun (Republic of Korea)  
 Kim, Jongmin (Republic of Korea)  
 Kim, Jongsu (Republic of Korea)  
 Kim, Jongtae (Republic of Korea)  
 Kim, Jon-Lark (Republic of Korea)  
 Kim, Joo Young (Republic of Korea)  
 Kim, Joo-Hyun (Republic of Korea)  
 Kim, Joon Oh (Republic of Korea)  
 Kim, Joon Pyo (Republic of Korea)  
 Kim, Joonha (Republic of Korea)  
 Kim, Joonhee (Republic of Korea)  
 Kim, Joonhyung (Republic of Korea)  
 Kim, Joonil (Republic of Korea)  
 Kim, Joowan (Republic of Korea)  
 Kim, Joseph (Republic of Korea)  
 Kim, Joung-Mi (Republic of Korea)  
 Kim, Joy (Republic of Korea)  
 Kim, Ju Hong (Republic of Korea)  
 Kim, Ju Hyung (Republic of Korea)  
 Kim, Ju Hyung (Republic of Korea)  
 Kim, Jueun (Republic of Korea)  
 Kim, Jueun (Republic of Korea)  
 Kim, Jun (Republic of Korea)  
 Kim, Jun Hyeok (Republic of Korea)  
 Kim, Jun Young (Republic of Korea)

Kim, Junbeom (Republic of Korea)  
Kim, Jung Han (Republic of Korea)  
Kim, Jung Ho (Republic of Korea)  
Kim, Jung Woo (Republic of Korea)  
Kim, Jung-A (Republic of Korea)  
Kim, Jungeun (Republic of Korea)  
Kim, Jungsoo (Republic of Korea)  
Kim, Junseong (Republic of Korea)  
Kim, Junsu (Republic of Korea)  
Kim, Juntae (Republic of Korea)  
Kim, Kee Tack (Republic of Korea)  
Kim, Keun Yong (Republic of Korea)  
Kim, Keun-Young (Republic of Korea)  
Kim, Ki Won (Republic of Korea)  
Kim, Kitae (Republic of Korea)  
Kim, Kunwoo (Republic of Korea)  
Kim, Kwang-Seob (Republic of Korea)  
Kim, Kyeong Hi (Republic of Korea)  
Kim, Kyeong Min (Republic of Korea)  
Kim, Kyeong Seok (Republic of Korea)  
Kim, Kyeong Yeob (Republic of Korea)  
Kim, Kyeong Yeop (Republic of Korea)  
Kim, Kyeonghun (Republic of Korea)  
Kim, Kyoung Hee (Republic of Korea)  
Kim, Kyoung Min (Republic of Korea)  
Kim, Kyoung Mo (Republic of Korea)  
Kim, Kyoung-Hoon (Republic of Korea)  
Kim, Kyoungsun (Republic of Korea)  
Kim, Kyoung-Tark (Republic of Korea)  
Kim, Kyu Sang (Republic of Korea)  
Kim, Kyu Sik (Republic of Korea)  
Kim, Kyung Man (Republic of Korea)  
Kim, Kyung Min (Republic of Korea)  
Kim, Kyung Seo (Republic of Korea)  
Kim, Kyung Soo (Republic of Korea)  
Kim, Kyung-Hwa (Republic of Korea)  
Kim, Kyunghwan (Republic of Korea)  
Kim, Kyungmin (Republic of Korea)  
Kim, Kyung-Won (Republic of Korea)  
Kim, Lami (Republic of Korea)  
Kim, Luke (Republic of Korea)  
Kim, Man-Joong (Republic of Korea)  
Kim, Meehye (Republic of Korea)  
Kim, Mee-Kyoung (Republic of Korea)  
Kim, Mi Young (Republic of Korea)  
Kim, Mi Young (Republic of Korea)  
Kim, Min Ah (Republic of Korea)  
Kim, Min Chul (Republic of Korea)  
Kim, Min Do (Republic of Korea)  
Kim, Min Gyeong (Republic of Korea)  
Kim, Min Jae (Republic of Korea)  
Kim, Min Jeong (Republic of Korea)  
Kim, Min Ki (Republic of Korea)  
Kim, Min Kwan (Republic of Korea)  
Kim, Min Su (Republic of Korea)  
Kim, Min Su (Republic of Korea)  
Kim, Min Suck (Republic of Korea)  
Kim, Minchan (Republic of Korea)  
Kim, Mingyun (Republic of Korea)  
Kim, Minhyong (Republic of Korea)  
Kim, Minjoo (Republic of Korea)  
Kim, Minkyu (Republic of Korea)  
Kim, Minseo (Republic of Korea)  
Kim, Minseok (Republic of Korea)  
Kim, Min-Seok (Republic of Korea)  
Kim, Min-Soo (Republic of Korea)  
Kim, Minsung (Republic of Korea)  
Kim, Minwoo (Republic of Korea)  
Kim, Moran (Republic of Korea)  
Kim, Myeon Hu (Republic of Korea)  
Kim, Myeong Sik (Republic of Korea)  
Kim, Myeonghyeok (Republic of Korea)  
Kim, Myeongjun (Republic of Korea)  
Kim, Myeung Soo (Republic of Korea)  
Kim, Myoungnyoun (Republic of Korea)  
Kim, Myunggyu (Republic of Korea)  
Kim, Myungho (Republic of Korea)  
Kim, Myung-Hwan (Republic of Korea)  
Kim, Nam Kyun (Republic of Korea)  
Kim, Namhoon (Republic of Korea)  
Kim, Namkwon (Republic of Korea)  
Kim, Panki (Republic of Korea)  
Kim, Rae Young (Republic of Korea)  
Kim, Samrang (Republic of Korea)  
Kim, San (Republic of Korea)  
Kim, Sang Jin (Republic of Korea)  
Kim, Sang Yoon (Republic of Korea)  
Kim, Sang Yup (Republic of Korea)  
Kim, Sang-Hyun (Republic of Korea)  
Kim, Sangmin (Republic of Korea)  
Kim, Sang-Mok (Republic of Korea)  
Kim, Sangwoo (Republic of Korea)  
Kim, Sangwook (Republic of Korea)  
Kim, Sarah (Republic of Korea)  
Kim, Se Ho (Republic of Korea)  
Kim, Se-Goo (Republic of Korea)  
Kim, Segyong (Republic of Korea)  
Kim, Seick (Republic of Korea)  
Kim, Sejoon (Republic of Korea)  
Kim, Seo Eun (Republic of Korea)  
Kim, Seo Young (Republic of Korea)  
Kim, Seog-Jin (Republic of Korea)  
Kim, Seokchan (Republic of Korea)  
Kim, Seoloh (Republic of Korea)

Kim, Seol-Oh (Republic of Korea)  
 Kim, Seon Jeong (Republic of Korea)  
 Kim, Seong Hyeon (Republic of Korea)  
 Kim, Seong Rae (Republic of Korea)  
 Kim, Seong-A (Republic of Korea)  
 Kim, Seonggwang (Republic of Korea)  
 Kim, Seongjeong (Republic of Korea)  
 Kim, Seongwoo (Republic of Korea)  
 Kim, Seonhwa (Republic of Korea)  
 Kim, Seonja (Republic of Korea)  
 Kim, Seonkuk (Republic of Korea)  
 Kim, Seonkyu (Republic of Korea)  
 Kim, Seonwoo (Republic of Korea)  
 Kim, Seoyoung (Republic of Korea)  
 Kim, Seul-Gi (Republic of Korea)  
 Kim, Seung Chan (Republic of Korea)  
 Kim, Seung Chan (Republic of Korea)  
 Kim, Seung Hyun (Republic of Korea)  
 Kim, Seung Won (Republic of Korea)  
 Kim, Seunghui (Republic of Korea)  
 Kim, Seunghyun (Republic of Korea)  
 Kim, Seungil (Republic of Korea)  
 Kim, Shinuk (Republic of Korea)  
 Kim, Shin-Young (Republic of Korea)  
 Kim, Si Myung (Republic of Korea)  
 Kim, Sijun (Republic of Korea)  
 Kim, Sin (Republic of Korea)  
 Kim, Siwon (Republic of Korea)  
 Kim, So Hyun (Republic of Korea)  
 Kim, So Young (Republic of Korea)  
 Kim, So Young (Republic of Korea)  
 Kim, Sojung (Republic of Korea)  
 Kim, Soo Hyun (Republic of Korea)  
 Kim, Soo Young (Republic of Korea)  
 Kim, Soohan (Republic of Korea)  
 Kim, Soohwan (Republic of Korea)  
 Kim, Soojung (Republic of Korea)  
 Kim, Soonrae (Republic of Korea)  
 Kim, Soonyoung (Republic of Korea)  
 Kim, Sooyoung (Republic of Korea)  
 Kim, Soyeon (Republic of Korea)  
 Kim, Soyeon (Republic of Korea)  
 Kim, Soyeon (Republic of Korea)  
 Kim, Sue Na (Republic of Korea)  
 Kim, Suh-Ryung (Republic of Korea)  
 Kim, Sun Ah (Republic of Korea)  
 Kim, Sun Ho (Republic of Korea)  
 Kim, Sun Hyung (Republic of Korea)  
 Kim, Sunah (Republic of Korea)  
 Kim, Sun-Chul (Republic of Korea)  
 Kim, Sung Ho (Republic of Korea)  
 Kim, Sung Hoon (Republic of Korea)  
 Kim, Sung Ki (Republic of Korea)  
 Kim, Sung Sook (Republic of Korea)  
 Kim, Sung Yeon (Republic of Korea)  
 Kim, Sung Yoon (Republic of Korea)  
 Kim, Sung Youn (Republic of Korea)  
 Kim, Sunghan (Republic of Korea)  
 Kim, Sungja (Republic of Korea)  
 Kim, Sungmin (Republic of Korea)  
 Kim, Sungook (Republic of Korea)  
 Kim, Sungwoon (Republic of Korea)  
 Kim, Sunyoung (Republic of Korea)  
 Kim, Suyeong (Republic of Korea)  
 Kim, Tae Hyoung (Republic of Korea)  
 Kim, Tae Hyung (Republic of Korea)  
 Kim, Tae In (Republic of Korea)  
 Kim, Tae Kyeom (Republic of Korea)  
 Kim, Tae Seo (Republic of Korea)  
 Kim, Tae Soo (Republic of Korea)  
 Kim, Tae Yeong (Republic of Korea)  
 Kim, Taehee (Republic of Korea)  
 Kim, Taehee (Republic of Korea)  
 Kim, Taeheon (Republic of Korea)  
 Kim, Taehyeon (Republic of Korea)  
 Kim, Taejeong (Republic of Korea)  
 Kim, Tae-Jin (Republic of Korea)  
 Kim, Taekyung (Republic of Korea)  
 Kim, Taemin (Republic of Korea)  
 Kim, Taewan (Republic of Korea)  
 Kim, Taeyon (Republic of Korea)  
 Kim, Tak Won (Republic of Korea)  
 Kim, Tea Hwan (Republic of Korea)  
 Kim, Tea Su (Republic of Korea)  
 Kim, Wan Su (Republic of Korea)  
 Kim, Wansoon (Republic of Korea)  
 Kim, Wol Sun (Republic of Korea)  
 Kim, Won Kyu (Republic of Korea)  
 Kim, Won-Sook (Republic of Korea)  
 Kim, Woo Chan (Republic of Korea)  
 Kim, Woo Seong (Republic of Korea)  
 Kim, Woo Tae (Republic of Korea)  
 Kim, Woochan (Republic of Korea)  
 Kim, Woojae (Republic of Korea)  
 Kim, Woojeong (Republic of Korea)  
 Kim, Wook (Republic of Korea)  
 Kim, Woonyeon (Republic of Korea)  
 Kim, Wootae (Republic of Korea)  
 Kim, Wooyeol (Republic of Korea)  
 Kim, Ye Geun (Republic of Korea)  
 Kim, Yeon Jeung (Republic of Korea)  
 Kim, Yeon-Eung (Republic of Korea)  
 Kim, Yeong Jong (Republic of Korea)  
 Kim, Yeonghyeon (Republic of Korea)

Kim, Yeongrak (Republic of Korea)  
Kim, Yeonjun (Republic of Korea)  
Kim, Yeseon (Republic of Korea)  
Kim, Yesule (Republic of Korea)  
Kim, Yi Jun (Republic of Korea)  
Kim, Yi Keon (Republic of Korea)  
Kim, Yisak (Republic of Korea)  
Kim, Yoenha (Republic of Korea)  
Kim, Yong Hyeon (Republic of Korea)  
Kim, Yong Jun (Republic of Korea)  
Kim, Yongduk (Republic of Korea)  
Kim, Yonghwan (Republic of Korea)  
Kim, Yongsik (Republic of Korea)  
Kim, Yongsun (Republic of Korea)  
Kim, Yoon-Joo (Republic of Korea)  
Kim, Yoosik (Republic of Korea)  
Kim, Yoosuk (Republic of Korea)  
Kim, Youchan (Republic of Korea)  
Kim, Youn Sung (Republic of Korea)  
Kim, Young Deuk (Republic of Korea)  
Kim, Young Hak (Republic of Korea)  
Kim, Young Ho (Republic of Korea)  
Kim, Young Hyoun (Republic of Korea)  
Kim, Young In (Republic of Korea)  
Kim, Young Joon (Republic of Korea)  
Kim, Young Kuk (Republic of Korea)  
Kim, Young Kyun (Republic of Korea)  
Kim, Young Mi (Republic of Korea)  
Kim, Young Rock (Republic of Korea)  
Kim, Young Seo (Republic of Korea)  
Kim, Young Wook (Republic of Korea)  
Kim, Younghee (Republic of Korea)  
Kim, Youngho (Republic of Korea)  
Kim, Youngju (Republic of Korea)  
Kim, Youngkey (Republic of Korea)  
Kim, Younjin (Republic of Korea)  
Kim, Younng-Jin (Republic of Korea)  
Kim, Yu Been (Republic of Korea)  
Kim, Yu Jeong (Republic of Korea)  
Kim, Yu Jin (Republic of Korea)  
Kim, Yujeong (Republic of Korea)  
Kim, Yul (Republic of Korea)  
Kim, Yunbae (Republic of Korea)  
Kim, Yun-Hwan (Republic of Korea)  
Kim, Yurim (Republic of Korea)  
Kim, Sehjeong (UAE)  
Kim, Kyoung-Hee Arlene (UK)  
Kim, Brian (USA)  
Kim, Chan-Ho (USA)  
Kim, Chiheon (USA)  
Kim, Dongkwan (USA)  
Kim, Genn Ia (USA)  
Kim, Inyoung (USA)  
Kim, Jae Kyoung (USA)  
Kim, Jisu (USA)  
Kim, Katherine (USA)  
Kim, Kunwoo (USA)  
Kim, Saeja (USA)  
Kim, Seonghak (USA)  
Kim, Yeansu (USA)  
Kim, Yuree (USA)  
Kimm, Ha Jine (Republic of Korea)  
Kimura, Shunichi (Japan)  
Kimura, Takashi (USA)  
Kiratu, Beth (Kenya)  
Kisaka, Masashi (Japan)  
Kiselman, Christer (Sweden)  
Kishi, Yasuhiro (Japan)  
Kitano, Teruaki (Japan)  
Kittipassorn, Teeradej (Thailand)  
Kiwook, Kim (Republic of Korea)  
Kiyohara, Hisae (Japan)  
Kiyohara, Kazuyoshi (Japan)  
Klaus, Friedrich (Germany)  
Klaus, Stephan (Germany)  
Klavik, Pavel (Czech Republic)  
Kleshchev, Alexander (USA)  
Knezevic-Miljanovic, Julka (Serbia)  
Knopova, Viktoriya (Ukraine)  
Ko, Dae Hyeon (Republic of Korea)  
Ko, Dongnam (Republic of Korea)  
Ko, Dongsoo (Republic of Korea)  
Ko, Dongsu (Republic of Korea)  
Ko, Eungil (Republic of Korea)  
Ko, Gyeonghoon (Republic of Korea)  
Ko, Ha Ram (Republic of Korea)  
Ko, Ho Kyoung (Republic of Korea)  
Ko, Hyerim (Republic of Korea)  
Ko, Hyoung June (Republic of Korea)  
Ko, Hyukjin (Republic of Korea)  
Ko, Il Seog (Republic of Korea)  
Ko, Jihoon (Republic of Korea)  
Ko, Jin-Young (Republic of Korea)  
Ko, Ki Hyoung (Republic of Korea)  
Ko, Minsu (Republic of Korea)  
Ko, Minwoo (Republic of Korea)  
Ko, Sangmin (Republic of Korea)  
Ko, Seungchan (Republic of Korea)  
Ko, Won Woo (Republic of Korea)  
Ko, Young Jun (Republic of Korea)  
Ko, Youngjin (Republic of Korea)  
Ko, Hankyung (USA)  
Ko, Young Kun (USA)  
Kobayashi, Masanori (Japan)

- Koch, Lena (Germany)  
 Koch, Thorsten (Germany)  
 Kochubei, Anatoly (Ukraine)  
 Kodama, Hiroki (Japan)  
 Koga, Hirotaka (Japan)  
 Koga, Jiro (Japan)  
 Koga, Jun-Ichi (Japan)  
 Koh, Doowon (Republic of Korea)  
 Koh, Eunjin (Republic of Korea)  
 Koh, Eunseo (Republic of Korea)  
 Koh, Geon Ho (Republic of Korea)  
 Koh, Hayeong (Republic of Korea)  
 Koh, Sung-Eun (Republic of Korea)  
 Koh, Wook (Republic of Korea)  
 Koh, Youngmee (Republic of Korea)  
 Koh, Youngwoo (Republic of Korea)  
 Kohayakawa, Yoshiharu (Brazil)  
 Koiso, Miyuki (Japan)  
 Koiso, Norihito (Japan)  
 Kokilashvili, Vakhtang (Georgia)  
 Kokubu, Hiroshi (Japan)  
 Kolesnikov, Alexei (USA)  
 Kollár, János (USA)  
 Kondo, Takefumi (Japan)  
 Kong, Il Hwan (Republic of Korea)  
 Kong, Jeong Hwan (Republic of Korea)  
 Kong, Meeyong (Republic of Korea)  
 Kong, Su Ryeoun (Republic of Korea)  
 Konjik, Sanja (Serbia)  
 Konno, Hiroshi (Japan)  
 Kono, Tsugio (Japan)  
 Koo, Bon Gyeong (Republic of Korea)  
 Koo, Eunkyung (Republic of Korea)  
 Koo, Hanbin (Republic of Korea)  
 Koo, Ja Hyun (Republic of Korea)  
 Koo, Ja Kyung (Republic of Korea)  
 Koo, Min Woo (Republic of Korea)  
 Koo, Namhun (Republic of Korea)  
 Koo, Namjip (Republic of Korea)  
 Koo, Yunyeong (Republic of Korea)  
 Kook, Woong (Republic of Korea)  
 Kopamu, Samuel (Papua New Guinea)  
 Koshiba, Yoichi (Japan)  
 Kostenko, Aleksey (Austria)  
 Kotani, Motoko (Japan)  
 Kotorii, Yuka (Japan)  
 Kovacs, Sandor (USA)  
 Kown, Minhó (Republic of Korea)  
 Koyama, Shin-Ya (Japan)  
 Kozlov, Vladimir (Sweden)  
 Kozono, Hideo (Japan)  
 Kozuka, Kazuhito (Japan)  
 Kozuma, Rintaro (Japan)  
 Kpata, Akon Abokon Berenger Patrick (Ivory Coast)  
 Kraljevic, Hrvoje (Croatia)  
 Kramer, Jürg (Germany)  
 Krauskopf, Bernd (New Zealand)  
 Krieger, Louise (Germany)  
 Krieger, Wolfgang (Germany)  
 Krivelevich, Michael (Israel)  
 Krumm, David (USA)  
 Krupinska, Katarzyna (Poland)  
 Krupinski, Krzysztof (Poland)  
 Krzywkowski, Marcin (Poland)  
 Ku, Se-Hyun (Republic of Korea)  
 Kubayi, David (South Africa)  
 Kudaibergenov, Sabit (Kazakhstan)  
 Kuelshammer, Julian (Germany)  
 Kuessner, Thilo (Germany)  
 Kühn, Daniela (UK)  
 Kuk, Seungwoo (Republic of Korea)  
 Kuku, Aderemi (USA)  
 Kum, Sangho (Republic of Korea)  
 Kumagai, Takashi (Japan)  
 Kumar, Ajay (India)  
 Kumar, Ajit (India)  
 Kumar, Sanjeev (India)  
 Kumar, Shiv Datt (India)  
 Kumar, Vineet (India)  
 Kumarasamy, Sakthivel (India)  
 Kumlin, Peter (Sweden)  
 Kun, Rattana (Cambodia)  
 Kun, Gabor (Hungary)  
 Kuncham, Syam Prasad (India)  
 Kuo, Kun-Lin (Taiwan)  
 Kupeli Erken, Irem (Turkey)  
 Kurina, Galina (Russia)  
 Kurmanova, Sovetkan (Kyrgyzstan)  
 Kusniyanti, Elvira (Indonesia)  
 Kutzschebauch, Werner Frank (Switzerland)  
 Kuusi, Tuomo (Finland)  
 Kuwada, Kazumasa (Japan)  
 Kuwae, Hiroki (Japan)  
 Kuwae, Kazuhiro (Japan)  
 Kwa, Kiam Heong (Malaysia)  
 Kwack, Yeonghoo (Republic of Korea)  
 Kwak, Do Young (Republic of Korea)  
 Kwak, Dooyoung (Republic of Korea)  
 Kwak, Insung (Republic of Korea)  
 Kwak, Ji Hun (Republic of Korea)  
 Kwak, Minkyu (Republic of Korea)  
 Kwak, Sijong (Republic of Korea)  
 Kwan, Junghee (Republic of Korea)

Kwietniak, Anna (Poland)  
Kwietniak, Dominik (Poland)  
Kwon, Beom Seok (Republic of Korea)  
Kwon, Dae Young (Republic of Korea)  
Kwon, Daehwi (Republic of Korea)  
Kwon, Dohyun (Republic of Korea)  
Kwon, Doyong (Republic of Korea)  
Kwon, Eui Joon (Republic of Korea)  
Kwon, Gukwon (Republic of Korea)  
Kwon, Guwan (Republic of Korea)  
Kwon, Hakbong (Republic of Korea)  
Kwon, Hee-Dae (Republic of Korea)  
Kwon, Heocyk Jun (Republic of Korea)  
Kwon, Hyeok Kyu (Republic of Korea)  
Kwon, Hyeok-Jun (Republic of Korea)  
Kwon, Hyeuknam (Republic of Korea)  
Kwon, Hyokchon (Republic of Korea)  
Kwon, Hyoungjin (Republic of Korea)  
Kwon, Hyukmin (Republic of Korea)  
Kwon, Jae Yong (Republic of Korea)  
Kwon, Jae-Hoon (Republic of Korea)  
Kwon, Jaihee (Republic of Korea)  
Kwon, Kiwoon (Republic of Korea)  
Kwon, Min Jae (Republic of Korea)  
Kwon, Min Jeong (Republic of Korea)  
Kwon, Min-Hyuk (Republic of Korea)  
Kwon, Minseong (Republic of Korea)  
Kwon, Minwoo (Republic of Korea)  
Kwon, Myeonggi (Republic of Korea)  
Kwon, Nam Ho (Republic of Korea)  
Kwon, Oh Nam (Republic of Korea)  
Kwon, Ohsung (Republic of Korea)  
Kwon, O-Joung (Republic of Korea)  
Kwon, Okyu (Republic of Korea)  
Kwon, Sanghoon (Republic of Korea)  
Kwon, Soon Hyun (Republic of Korea)  
Kwon, Soon-Geol (Republic of Korea)  
Kwon, Soonsik (Republic of Korea)  
Kwon, Soun-Hi (Republic of Korea)  
Kwon, Ui-Jin (Republic of Korea)  
Kwon, Yong Jae (Republic of Korea)  
Kwon, Yong Jun (Republic of Korea)  
Kwon, Young Soo (Republic of Korea)  
Kwon, Youngjun (Republic of Korea)  
Kwon, Young-Sam (Republic of Korea)  
Kye, Seung-Hyeok (Republic of Korea)  
Kye, Young Hee (Republic of Korea)  
Kyeong, Daehyeon (Republic of Korea)  
Kyeong, Sunghyon (Republic of Korea)  
Kyoung, Jeong Gyu (Republic of Korea)  
Kytola, Kalle (Finland)  
Kyu Han, Cho (Republic of Korea)  
La, Joonhyun (Republic of Korea)  
Laba, Izabella (Canada)  
Labadin, Jane (Malaysia)  
Laborde, Colette (France)  
Laborde, Jean-Marie (France)  
Labuda, Iwo (USA)  
Lafuerza-Guillen, Bernardo (Spain)  
Lahiri, Ananya (India)  
Lai, Hsin-Hao (Taiwan)  
Lam, Kee Yuen (Canada)  
Lange, Kenneth (USA)  
Lao, Angelyn (Philippines)  
Lap, James T (USA)  
Larpenteur, Veronique (France)  
Lason, Michal (Poland)  
Lau, Siu-Cheong (USA)  
Lauda, Aaron (USA)  
Laurent, Monique (Netherlands)  
Lauret, Emilio (Argentina)  
Laurière, Mathieu (France)  
Lauter, Kristin (USA)  
Lavor, Carlile (Brazil)  
Lawler, Gregory (USA)  
Laya, Fazlollahi (Iran)  
Lazaroiu, Calin (Romania)  
Le, Pengyu (Switzerland)  
Le, Thi Nhu Bich (Vietnam)  
Le, Thi Thanh Nhan (Vietnam)  
Le, Tuan Hoa (Vietnam)  
Le, Van Thanh (Vietnam)  
Le Bris, Claude (France)  
Le Duc, Thoang (Vietnam)  
Le Gall, Jean-François (France)  
Le Huy, Tien (Vietnam)  
Le Thanh Hoang, Nhat (Vietnam)  
Lebedeva, Elena (Russia)  
Leclerc, Bernard (France)  
Ledoux, Michel (France)  
Ledua, Victor (Fiji)  
Lee, Gye-Seon (Germany)  
Lee, Lai Soon (Malaysia)  
Lee, Sun Rye (Republic of Korea)  
Lee, Ah Jeong (Republic of Korea)  
Lee, Aro (Republic of Korea)  
Lee, Beom Young (Republic of Korea)  
Lee, Bon Woo (Republic of Korea)  
Lee, Byeong Geol (Republic of Korea)  
Lee, Byeongchan (Republic of Korea)  
Lee, Byeong Hun (Republic of Korea)  
Lee, Chaegu (Republic of Korea)  
Lee, Chan Min (Republic of Korea)  
Lee, Chang Hoon (Republic of Korea)

- Lee, Chang Hun (Republic of Korea)  
 Lee, Chang Hyeon (Republic of Korea)  
 Lee, Chang Woo (Republic of Korea)  
 Lee, Changhee (Republic of Korea)  
 Lee, Changhoon (Republic of Korea)  
 Lee, Changjin (Republic of Korea)  
 Lee, Chang-Ock (Republic of Korea)  
 Lee, Chanho (Republic of Korea)  
 Lee, Cheol Hee (Republic of Korea)  
 Lee, Chong Gyu (Republic of Korea)  
 Lee, Choong Hoon (Republic of Korea)  
 Lee, Chul-Hee (Republic of Korea)  
 Lee, Chun Jar (Republic of Korea)  
 Lee, Da Young (Republic of Korea)  
 Lee, Dae Gwan (Republic of Korea)  
 Lee, Dae-Woong (Republic of Korea)  
 Lee, Daeseok (Republic of Korea)  
 Lee, Dahye (Republic of Korea)  
 Lee, Dean (Republic of Korea)  
 Lee, Deok Young (Republic of Korea)  
 Lee, Dong Ha (Republic of Korea)  
 Lee, Dong Heun (Republic of Korea)  
 Lee, Dong Hun (Republic of Korea)  
 Lee, Dong Jun (Republic of Korea)  
 Lee, Dong Uk (Republic of Korea)  
 Lee, Dong Won (Republic of Korea)  
 Lee, Donghi (Republic of Korea)  
 Lee, Dong-Hun (Republic of Korea)  
 Lee, Donghyeok (Republic of Korea)  
 Lee, Donghyun (Republic of Korea)  
 Lee, Dong-Il (Republic of Korea)  
 Lee, Dongkeon (Republic of Korea)  
 Lee, Dongkwan (Republic of Korea)  
 Lee, Dongsoo (Republic of Korea)  
 Lee, Dongsun (Republic of Korea)  
 Lee, Donsung (Republic of Korea)  
 Lee, Doo Seok (Republic of Korea)  
 Lee, Doo Young (Republic of Korea)  
 Lee, Eiu Jung (Republic of Korea)  
 Lee, Eo-Jin (Republic of Korea)  
 Lee, Eon-Kyung (Republic of Korea)  
 Lee, Euiwoo (Republic of Korea)  
 Lee, Eun Hye (Republic of Korea)  
 Lee, Eun Kyoung (Republic of Korea)  
 Lee, Eun Taek (Republic of Korea)  
 Lee, Eunbyul (Republic of Korea)  
 Lee, Eunhye (Republic of Korea)  
 Lee, Eunhyuk (Republic of Korea)  
 Lee, Eunjeong (Republic of Korea)  
 Lee, Eunjeong (Republic of Korea)  
 Lee, Eunjung (Republic of Korea)  
 Lee, Eunkyung (Republic of Korea)  
 Lee, Eunsaem (Republic of Korea)  
 Lee, Eun-Young (Republic of Korea)  
 Lee, Ga Won (Republic of Korea)  
 Lee, Ga-Eun (Republic of Korea)  
 Lee, Ganghun (Republic of Korea)  
 Lee, Gangyong (Republic of Korea)  
 Lee, Geon Woo (Republic of Korea)  
 Lee, Geonho (Republic of Korea)  
 Lee, Gue Myung (Republic of Korea)  
 Lee, Gunho (Republic of Korea)  
 Lee, Gun-Won (Republic of Korea)  
 Lee, Gwangyeon (Republic of Korea)  
 Lee, Gyeong-Hoon (Republic of Korea)  
 Lee, Ha Rim (Republic of Korea)  
 Lee, Haeri (Republic of Korea)  
 Lee, Haesung (Republic of Korea)  
 Lee, Hak June (Republic of Korea)  
 Lee, Hanee (Republic of Korea)  
 Lee, Hanjin (Republic of Korea)  
 Lee, Harkjoon (Republic of Korea)  
 Lee, Hayoon (Republic of Korea)  
 Lee, Heak Jin (Republic of Korea)  
 Lee, Hee Chul (Republic of Korea)  
 Lee, Hee Kwon (Republic of Korea)  
 Lee, Heejin (Republic of Korea)  
 Lee, Heisook (Republic of Korea)  
 Lee, Ho (Republic of Korea)  
 Lee, Ho (Republic of Korea)  
 Lee, Ho Joong (Republic of Korea)  
 Lee, Hojoo (Republic of Korea)  
 Lee, Ho-Jun (Republic of Korea)  
 Lee, Hokeon (Republic of Korea)  
 Lee, Hosoo (Republic of Korea)  
 Lee, Hoyeon (Republic of Korea)  
 Lee, Hun Hee (Republic of Korea)  
 Lee, Hwa Jeong (Republic of Korea)  
 Lee, Hwangrae (Republic of Korea)  
 Lee, Hwayoung (Republic of Korea)  
 Lee, Hwayoung (Republic of Korea)  
 Lee, Hyang-Sook (Republic of Korea)  
 Lee, Hye Yeon (Republic of Korea)  
 Lee, Hyejung (Republic of Korea)  
 Lee, Hyeonggon (Republic of Korea)  
 Lee, Hyeongwoo (Republic of Korea)  
 Lee, Hyeonwoo (Republic of Korea)  
 Lee, Hyo Jung (Republic of Korea)  
 Lee, Hyo Yoon (Republic of Korea)  
 Lee, Hyojeong (Republic of Korea)  
 Lee, Hyojun (Republic of Korea)  
 Lee, Hyomin (Republic of Korea)  
 Lee, Hyowon (Republic of Korea)  
 Lee, Hyun (Republic of Korea)

Lee, Hyun Geun (Republic of Korea)  
Lee, Hyun Ho (Republic of Korea)  
Lee, Hyun Jong (Republic of Korea)  
Lee, Hyun Ju (Republic of Korea)  
Lee, Hyun Seok (Republic of Korea)  
Lee, Hyun Su (Republic of Korea)  
Lee, Hyundong (Republic of Korea)  
Lee, Hyung Chun (Republic of Korea)  
Lee, Hyunhui (Republic of Korea)  
Lee, Hyunjin (Republic of Korea)  
Lee, Hyun-Seok (Republic of Korea)  
Lee, Ilseok (Republic of Korea)  
Lee, In Ha (Republic of Korea)  
Lee, Ingyu (Republic of Korea)  
Lee, Insook (Republic of Korea)  
Lee, Inwoo (Republic of Korea)  
Lee, Jae Gook (Republic of Korea)  
Lee, Jae Hee (Republic of Korea)  
Lee, Jae Hwa (Republic of Korea)  
Lee, Jae Hwang (Republic of Korea)  
Lee, Jae Hyouk (Republic of Korea)  
Lee, Jae Kab (Republic of Korea)  
Lee, Jae Min (Republic of Korea)  
Lee, Jae Woo (Republic of Korea)  
Lee, Jae Woong (Republic of Korea)  
Lee, Jae Yong (Republic of Korea)  
Lee, Jaechang (Republic of Korea)  
Lee, Jaegeun (Republic of Korea)  
Lee, Jae-Hee (Republic of Korea)  
Lee, Jaehyup (Republic of Korea)  
Lee, Jaeseon (Republic of Korea)  
Lee, Jaesuk (Republic of Korea)  
Lee, Jae-Uk (Republic of Korea)  
Lee, Jae-Weon (Republic of Korea)  
Lee, Jaewon (Republic of Korea)  
Lee, Jaeyoung (Republic of Korea)  
Lee, Jangjoo (Republic of Korea)  
Lee, Jeong Gwan (Republic of Korea)  
Lee, Jeong Hwan (Republic of Korea)  
Lee, Jeong Min (Republic of Korea)  
Lee, Jeong Min (Republic of Korea)  
Lee, Jeong Woo (Republic of Korea)  
Lee, Jeongbum (Republic of Korea)  
Lee, Jeongheon (Republic of Korea)  
Lee, Jeongjae (Republic of Korea)  
Lee, Jeong-Yup (Republic of Korea)  
Lee, Ji Eun (Republic of Korea)  
Lee, Ji Oon (Republic of Korea)  
Lee, Ji U (Republic of Korea)  
Lee, Ji Yeon (Republic of Korea)  
Lee, Ji Yoon (Republic of Korea)  
Lee, Jieun (Republic of Korea)  
Lee, Jieun (Republic of Korea)  
Lee, Jihoon (Republic of Korea)  
Lee, Jihoon (Republic of Korea)  
Lee, Jimin (Republic of Korea)  
Lee, Jin Bong (Republic of Korea)  
Lee, Jin Hyung (Republic of Korea)  
Lee, Jin Seong (Republic of Korea)  
Lee, Jin Woo (Republic of Korea)  
Lee, Jinhee (Republic of Korea)  
Lee, Jinho (Republic of Korea)  
Lee, Jin-Ho (Republic of Korea)  
Lee, Jinhoo (Republic of Korea)  
Lee, Jinhyeong (Republic of Korea)  
Lee, Jinwoo (Republic of Korea)  
Lee, Jinyeop (Republic of Korea)  
Lee, Jiye (Republic of Korea)  
Lee, Jiyeon (Republic of Korea)  
Lee, Jiyoung (Republic of Korea)  
Lee, Jiyun (Republic of Korea)  
Lee, Jong Bum (Republic of Korea)  
Lee, Jong Eun (Republic of Korea)  
Lee, Jong Hyeon (Republic of Korea)  
Lee, Jong Mun (Republic of Korea)  
Lee, Jong-Chan (Republic of Korea)  
Lee, Jonggul (Republic of Korea)  
Lee, Jongho (Republic of Korea)  
Lee, Jonghyeon (Republic of Korea)  
Lee, Jonghyun (Republic of Korea)  
Lee, Jong-Ryong (Republic of Korea)  
Lee, Jongwoo (Republic of Korea)  
Lee, Joo Yeon (Republic of Korea)  
Lee, Jooho (Republic of Korea)  
Lee, Joongul (Republic of Korea)  
Lee, Joonkyung (Republic of Korea)  
Lee, Ju A (Republic of Korea)  
Lee, Ju Hyeun (Republic of Korea)  
Lee, Ju Yeon (Republic of Korea)  
Lee, Ju Yeon (Republic of Korea)  
Lee, Ju Yeon (Republic of Korea)  
Lee, Ju Young (Republic of Korea)  
Lee, Juhee (Republic of Korea)  
Lee, Juhyun (Republic of Korea)  
Lee, Jun Ho (Republic of Korea)  
Lee, Jun Hyuk (Republic of Korea)  
Lee, Jun Hyuk (Republic of Korea)  
Lee, Jun Seo (Republic of Korea)  
Lee, Jun Seok (Republic of Korea)  
Lee, Jun Yong (Republic of Korea)  
Lee, June Bok (Republic of Korea)  
Lee, June Hee (Republic of Korea)  
Lee, June Seo (Republic of Korea)  
Lee, June-Yub (Republic of Korea)

Lee, Jung Hee (Republic of Korea)  
 Lee, Jung Hoon (Republic of Korea)  
 Lee, Jung Kyung (Republic of Korea)  
 Lee, Jung Rak (Republic of Korea)  
 Lee, Jung Rye (Republic of Korea)  
 Lee, Jung Soo (Republic of Korea)  
 Lee, Jung Woo (Republic of Korea)  
 Lee, Jung-In (Republic of Korea)  
 Lee, Jungseob (Republic of Korea)  
 Lee, Junguk (Republic of Korea)  
 Lee, Jungwon (Republic of Korea)  
 Lee, Jungwoo (Republic of Korea)  
 Lee, Junseok (Republic of Korea)  
 Lee, Junyeong (Republic of Korea)  
 Lee, Kang Min (Republic of Korea)  
 Lee, Kang-Hyurk (Republic of Korea)  
 Lee, Kang-Ju (Republic of Korea)  
 Lee, Kangwon (Republic of Korea)  
 Lee, Ke Seung (Republic of Korea)  
 Lee, Kee Young (Republic of Korea)  
 Lee, Keonhee (Republic of Korea)  
 Lee, Key-Nyoung (Republic of Korea)  
 Lee, Ki Hoon (Republic of Korea)  
 Lee, Ki-Ahm (Republic of Korea)  
 Lee, Kimyeong (Republic of Korea)  
 Lee, Kisuk (Republic of Korea)  
 Lee, Kury (Republic of Korea)  
 Lee, Kwang Kyu (Republic of Korea)  
 Lee, Kwangsu (Republic of Korea)  
 Lee, Kwangwoo (Republic of Korea)  
 Lee, Kwankyuu (Republic of Korea)  
 Lee, Kyeongwon (Republic of Korea)  
 Lee, Kyoung-Hyun (Republic of Korea)  
 Lee, Kyoung-Seog (Republic of Korea)  
 Lee, Kyung Ryul (Republic of Korea)  
 Lee, Kyunghwan (Republic of Korea)  
 Lee, Kyungseung (Republic of Korea)  
 Lee, Man Keun (Republic of Korea)  
 Lee, Manseob (Republic of Korea)  
 Lee, Mee-Jung (Republic of Korea)  
 Lee, Mikyoung (Republic of Korea)  
 Lee, Min Gi (Republic of Korea)  
 Lee, Min Goo (Republic of Korea)  
 Lee, Min Jung (Republic of Korea)  
 Lee, Min Young (Republic of Korea)  
 Lee, Minjae (Republic of Korea)  
 Lee, Minjung (Republic of Korea)  
 Lee, Minku (Republic of Korea)  
 Lee, Moon Sung (Republic of Korea)  
 Lee, Myoung Hi (Republic of Korea)  
 Lee, Myung Hee (Republic of Korea)  
 Lee, Myung Suk (Republic of Korea)  
 Lee, Na-Hyun (Republic of Korea)  
 Lee, Nam-Hoon (Republic of Korea)  
 Lee, Nany (Republic of Korea)  
 Lee, Nari (Republic of Korea)  
 Lee, Nayeong (Republic of Korea)  
 Lee, Pa Ra (Republic of Korea)  
 Lee, Philku (Republic of Korea)  
 Lee, Pyeongjae (Republic of Korea)  
 Lee, Reeha (Republic of Korea)  
 Lee, Sage (Republic of Korea)  
 Lee, Sang Deok (Republic of Korea)  
 Lee, Sang Hyeon (Republic of Korea)  
 Lee, Sang Hyeon (Republic of Korea)  
 Lee, Sang Hyun (Republic of Korea)  
 Lee, Sang Hyun (Republic of Korea)  
 Lee, Sang Jin (Republic of Korea)  
 Lee, Sang Jun (Republic of Korea)  
 Lee, Sang June (Republic of Korea)  
 Lee, Sang Min (Republic of Korea)  
 Lee, Sang Yeop (Republic of Korea)  
 Lee, Sang Youl (Republic of Korea)  
 Lee, Sang-Gu (Republic of Korea)  
 Lee, Sangheon (Republic of Korea)  
 Lee, Sangheon (Republic of Korea)  
 Lee, Sanghoon (Republic of Korea)  
 Lee, Sanghyeon (Republic of Korea)  
 Lee, Sang-Jin (Republic of Korea)  
 Lee, Sangjo (Republic of Korea)  
 Lee, Sangmin (Republic of Korea)  
 Lee, Sangwoo (Republic of Korea)  
 Lee, Sangwook (Republic of Korea)  
 Lee, Sangyeop (Republic of Korea)  
 Lee, Sanha (Republic of Korea)  
 Lee, Se Hwan (Republic of Korea)  
 Lee, See-Woo (Republic of Korea)  
 Lee, Seok Bin (Republic of Korea)  
 Lee, Seok Hyeong (Republic of Korea)  
 Lee, Seokjin (Republic of Korea)  
 Lee, Seok-Min (Republic of Korea)  
 Lee, Seon Jae (Republic of Korea)  
 Lee, Seongdyuk (Republic of Korea)  
 Lee, Seongok (Republic of Korea)  
 Lee, Seonjeong (Republic of Korea)  
 Lee, Seul-Gi (Republic of Korea)  
 Lee, Seung Chul (Republic of Korea)  
 Lee, Seung Eun (Republic of Korea)  
 Lee, Seung Hun (Republic of Korea)  
 Lee, Seung Hyun (Republic of Korea)  
 Lee, Seung Jae (Republic of Korea)  
 Lee, Seung Jai (Republic of Korea)  
 Lee, Seung Jin (Republic of Korea)  
 Lee, Seung Min (Republic of Korea)

Lee, Seung Mok (Republic of Korea)  
 Lee, Seung Wook (Republic of Korea)  
 Lee, Seungha (Republic of Korea)  
 Lee, Seunghee (Republic of Korea)  
 Lee, Seunghee (Republic of Korea)  
 Lee, Seunghoon (Republic of Korea)  
 Lee, Seunghyun (Republic of Korea)  
 Lee, Seunghyun (Republic of Korea)  
 Lee, Seungick (Republic of Korea)  
 Lee, Seungik (Republic of Korea)  
 Lee, Seungjae (Republic of Korea)  
 Lee, Seung-Jun (Republic of Korea)  
 Lee, Seung-On (Republic of Korea)  
 Lee, Seungtae (Republic of Korea)  
 Lee, Seungwoo (Republic of Korea)  
 Lee, Seungyeon (Republic of Korea)  
 Lee, Seung-Yun (Republic of Korea)  
 Lee, Seyeon (Republic of Korea)  
 Lee, Shin-Myung (Republic of Korea)  
 Lee, Sik (Republic of Korea)  
 Lee, Siyong (Republic of Korea)  
 Lee, So Hyeon (Republic of Korea)  
 Lee, So Min (Republic of Korea)  
 Lee, So Yeon (Republic of Korea)  
 Lee, Sojin (Republic of Korea)  
 Lee, Soo Gon (Republic of Korea)  
 Lee, Soo Hong (Republic of Korea)  
 Lee, Soo Jeong (Republic of Korea)  
 Lee, Soo Jung (Republic of Korea)  
 Lee, Soohyeon (Republic of Korea)  
 Lee, Soojung (Republic of Korea)  
 Lee, Soo-Kyeong (Republic of Korea)  
 Lee, Soon-Phil (Republic of Korea)  
 Lee, Sori (Republic of Korea)  
 Lee, Soungha (Republic of Korea)  
 Lee, Su Bin (Republic of Korea)  
 Lee, Su Yeon (Republic of Korea)  
 Lee, Su Yeon (Republic of Korea)  
 Lee, Sue (Republic of Korea)  
 Lee, Su-Jin (Republic of Korea)  
 Lee, Sun (Republic of Korea)  
 Lee, Sun Ju (Republic of Korea)  
 Lee, Sun Jung (Republic of Korea)  
 Lee, Sung Jun (Republic of Korea)  
 Lee, Sungjin (Republic of Korea)  
 Lee, Sungjoon (Republic of Korea)  
 Lee, Sungmin (Republic of Korea)  
 Lee, Sungwoo (Republic of Korea)  
 Lee, Sung-Wook (Republic of Korea)  
 Lee, Sunho (Republic of Korea)  
 Lee, Suyoung (Republic of Korea)  
 Lee, Tae Ho (Republic of Korea)  
 Lee, Tae Young (Republic of Korea)  
 Lee, Taehee (Republic of Korea)  
 Lee, Taeho (Republic of Korea)  
 Lee, Taehun (Republic of Korea)  
 Lee, Uiryel (Republic of Korea)  
 Lee, Wan Luyr (Republic of Korea)  
 Lee, Wanho (Republic of Korea)  
 Lee, Wanki (Republic of Korea)  
 Lee, Wanseok (Republic of Korea)  
 Lee, Wonwoong (Republic of Korea)  
 Lee, Yang-Hi (Republic of Korea)  
 Lee, Yebin (Republic of Korea)  
 Lee, Yeihaing (Republic of Korea)  
 Lee, Yeon Jun (Republic of Korea)  
 Lee, Yeongmi (Republic of Korea)  
 Lee, Yeonhee (Republic of Korea)  
 Lee, Yong Gyu (Republic of Korea)  
 Lee, Yong Hoon (Republic of Korea)  
 Lee, Yong Jin (Republic of Korea)  
 Lee, Yongnam (Republic of Korea)  
 Lee, Yoo Sung (Republic of Korea)  
 Lee, Yoon Gu (Republic of Korea)  
 Lee, Yoonbok (Republic of Korea)  
 Lee, Yoonjin (Republic of Korea)  
 Lee, Yoonkyeong (Republic of Korea)  
 Lee, Yoonweon (Republic of Korea)  
 Lee, Young Hee (Republic of Korea)  
 Lee, Young Jae (Republic of Korea)  
 Lee, Young Min (Republic of Korea)  
 Lee, Young Soon (Republic of Korea)  
 Lee, Youngae (Republic of Korea)  
 Lee, Youngchan (Republic of Korea)  
 Lee, Young-Chan (Republic of Korea)  
 Lee, Youngjae (Republic of Korea)  
 Lee, Youngmee (Republic of Korea)  
 Lee, Youngwoong (Republic of Korea)  
 Lee, Yun Jeong (Republic of Korea)  
 Lee, Yun Min (Republic of Korea)  
 Lee, Choongbum (USA)  
 Lee, Ik Jae (USA)  
 Lee, Junho (USA)  
 Lee, Kyung-Bai (USA)  
 Leem, Su Min (Republic of Korea)  
 Leenawong, Chartchai (Thailand)  
 Lemence, Richard (Philippines)  
 Lempert, Laszlo (USA)  
 Lenstra, Arjen (Switzerland)  
 Lephodisa, Benjamin Masego (Botswana)  
 Lesmono, Dharma (Indonesia)  
 Letzter, Shoham (Israel)  
 Levesque, Claude (Canada)  
 Leviatan, Talma (Israel)

Levine, Adam (USA)  
 Lewintan, Peter (Germany)  
 Lewis, Adrian Stephen (USA)  
 Lewkeeratiyutkul, Wicharn (Thailand)  
 Leyson, Dennis (Philippines)  
 Lho, Hyen Ho (Republic of Korea)  
 Li, Bing (P.R. China)  
 Li, Di (P.R. China)  
 Li, Guanghan (P.R. China)  
 Li, Haifang (P.R. China)  
 Li, Hui (P.R. China)  
 Li, Jiyou (P.R. China)  
 Li, Linsong (P.R. China)  
 Li, Nan (P.R. China)  
 Li, Ni (P.R. China)  
 Li, Ping (P.R. China)  
 Li, Qifeng (P.R. China)  
 Li, Qingzhong (P.R. China)  
 Li, Xiaoyu (P.R. China)  
 Li, Xu (P.R. China)  
 Li, Yinghong (P.R. China)  
 Li, Zhuchun (P.R. China)  
 Li, Guang-Liang (Hong Kong)  
 Li, Victor (Hong Kong)  
 Li, Changzheng (Japan)  
 Li, Chunlan (Republic of Korea)  
 Li, Xiaofei (Republic of Korea)  
 Li, Kai (Sweden)  
 Li, Xue-Mei (UK)  
 Li, Aihua (USA)  
 Li, Tao (USA)  
 Liang, Chao (P.R. China)  
 Liberti, Leo (USA)  
 Liendo, Alvaro (Chile)  
 Lifyand, Elijah (Israel)  
 Lifshits, Mikhail (Russia)  
 Lih, Ko-Wei (Taiwan)  
 Lim, Byung Chan (Republic of Korea)  
 Lim, Dong Hee (Republic of Korea)  
 Lim, Gwang-Bun (Republic of Korea)  
 Lim, Gyuseok (Republic of Korea)  
 Lim, Han Jun (Republic of Korea)  
 Lim, Han Kyul (Republic of Korea)  
 Lim, Heejin (Republic of Korea)  
 Lim, Heeye (Republic of Korea)  
 Lim, Ho Jin (Republic of Korea)  
 Lim, Hyoung Rae (Republic of Korea)  
 Lim, Hyunn Su (Republic of Korea)  
 Lim, Jin Wook (Republic of Korea)  
 Lim, Jongryul (Republic of Korea)  
 Lim, Jun Yeong (Republic of Korea)  
 Lim, Ki Hyuk (Republic of Korea)  
 Lim, Kyung Ju (Republic of Korea)  
 Lim, Mikyoung (Republic of Korea)  
 Lim, O Seon (Republic of Korea)  
 Lim, Sangho (Republic of Korea)  
 Lim, Seongan (Republic of Korea)  
 Lim, Seong-Jea (Republic of Korea)  
 Lim, Seonhee (Republic of Korea)  
 Lim, Seung Won (Republic of Korea)  
 Lim, Subong (Republic of Korea)  
 Lim, Sung Hyun (Republic of Korea)  
 Lim, Yeon-Hwa (Republic of Korea)  
 Lim, Yong Han (Republic of Korea)  
 Lim, Yongdo (Republic of Korea)  
 Lim, Young Sa (Republic of Korea)  
 Lim, Chang Mou (Singapore)  
 Lim, Junghwan (UK)  
 Limbupasiriporn, Jirapha (Thailand)  
 Lin, Mongkolsery (Cambodia)  
 Lin, Peishan (P.R. China)  
 Lin, Shi-Ying (P.R. China)  
 Lin, Shen (France)  
 Lin, Yanping (Hong Kong)  
 Lin, Chang-Shou (Taiwan)  
 Lin, Hai (USA)  
 Lin, Longzhi (USA)  
 Linares, Felipe (Brazil)  
 Lind, Tony (Sweden)  
 Ling, San (Singapore)  
 Lingeswaran, Shangerganesh (India)  
 Linowitz, Benjamin (USA)  
 Linton, Fred (USA)  
 Lipikorn, Rajalida (Thailand)  
 Little, Bert (USA)  
 Liu, Dangzheng (P.R. China)  
 Liu, Gongxiang (P.R. China)  
 Liu, Le Ping (P.R. China)  
 Liu, Pin (P.R. China)  
 Liu, Shu-Jun (P.R. China)  
 Liu, Wei (P.R. China)  
 Liu, Xin (P.R. China)  
 Liu, Yongxia (P.R. China)  
 Liu, Po-Hung (Taiwan)  
 Liu, Tai-Ping (Taiwan)  
 Liu, Zhengwei (USA)  
 Liu, Zhuangyi (USA)  
 Liubov, Shenderova (France)  
 Lkhanga, Oyuntsetseg (Mongolia)  
 Loeser, François (France)  
 Loewy, Raphael (Israel)  
 Loh, Hyunbin (Republic of Korea)  
 Loh, Din-Sui (USA)  
 Loh, Po-Shen (USA)

Lombe, Mubanga (Zambia)  
Long, Yiming (P.R. China)  
Longhi, Ignazio (P.R. China)  
Lope, Jose Ernie (Philippines)  
Lopez, Hiram H. (Mexico)  
Lopez, Rafael (Spain)  
Lorch, David (Germany)  
Lottermann, Annina (Germany)  
Lovász, László (Hungary)  
Lu, Shannian (P.R. China)  
Lubberts, Zachary (USA)  
Luca, Florian (Mexico)  
Ludwig, Ursula (France)  
Luk, Jonathan (USA)  
Lukarevski, Martin (Macedonia)  
Luli, Garving (USA)  
Lungu, Edward (Botswana)  
Luo, Jun (P.R. China)  
Luo, Qiu-Ming (P.R. China)  
Luo, Zhen (P.R. China)  
Luo, Tie (USA)  
Luu, Hoang Duc (Vietnam)  
Luu Quoc, Dat (Vietnam)  
Luzon, Ana (Spain)  
Lyons, Terry (UK)  
Lyons, Russell (USA)  
Lyons, Timothy (USA)  
Lytkina, Daria (Russia)  
Lyubich, Mikhail (USA)  
Ma, Junjie (P.R. China)  
Ma, Letian (P.R. China)  
Ma, Zhi-Ming (P.R. China)  
Mabuchi, Toshiki (Japan)  
Macintyre, Angus (UK)  
Macpherson, Andrew (UK)  
Madahar, Keerti Vardhan (India)  
Madeti, Prabhakar (India)  
Maeng, Chae Jung (Republic of Korea)  
Maeng, Da Hae (Republic of Korea)  
Maeng, Hoyoung (Republic of Korea)  
Mafi, Amir (Iran)  
Mahamane, Amadou (Mali)  
Maitournam, Aboubakar (Niger)  
Makhatova, Saule (Kazakhstan)  
Makinde, Oluwole Daniel (South Africa)  
Malaga Sabogal, Alba Marina (France)  
Malaspina, Uldarico (Peru)  
Malchiodi, Andrea (Italy)  
Male, Camille (France)  
Malik, Shabnam (Pakistan)  
Malinin, Dmitry (Jamaica)  
Mallahi Karai, Keivan (Germany)  
Malyutin, Konstantin (Ukraine)  
Mam, Mareth (Cambodia)  
Manchanda, Pammy (India)  
Manderscheid, David (USA)  
Mandrescu, Eugen (Israel)  
Mango Magero, John (Uganda)  
Mani, Arun (Australia)  
Manit, Tauch (Cambodia)  
Manjunath, Madhusudan (USA)  
Manning, Jason (USA)  
Manojlovic, Vesna (Serbia)  
Mantilla, Irla (Peru)  
Mapes-Szekelyhidi, Sonja (USA)  
Mar, Ohn (Myanmar)  
Mara, Muhlasah (Indonesia)  
March, Peter (USA)  
Marcos, Aboubacar (Benin)  
Marini, Luisa Donatella (Italy)  
Markarian, Roberto (Uruguay)  
Markina, Irina (Norway)  
Marklof, Jens (Germany)  
Markovic, Vladimir (USA)  
Markowich, Peter (Austria)  
Markowsky, Gregory (Australia)  
Markwardt, Sylwia (Germany)  
Maronne, Sebastien (France)  
Martha, Subash Chandra (India)  
Martinez, Antonio (Spain)  
Martinez, Aurea (Spain)  
Martinez Lopez, Consuelo (Spain)  
Martinez-Avendano, Ruben (Mexico)  
Martirosyan, Mher (Armenia)  
Maruyama, Kyoko (Japan)  
Maruyama, Naomasa (Japan)  
Maryati, Tita (Indonesia)  
Mase, Makiko (Japan)  
Mason, Darren (USA)  
Masuda, Shigeru (Japan)  
Masutova, Kamilyam (Uzbekistan)  
Mataga, Etsuko (Japan)  
Mataga, Yoshiharu (Japan)  
Mathis, Hélène (France)  
Matsuda, Osamu (Japan)  
Matsumoto, Shigenori (Japan)  
Matsumura, Kumiko (Japan)  
Matsumura, Takeshi (Japan)  
Matsuyama, Hiromi (Japan)  
Matsuyama, Yoshio (Japan)  
Mattila, Pertti (Finland)  
Mattingly, Jonathan (USA)  
Matveev, Mikhail (Russia)  
Matveev, Sergey (Russia)

Matveev, Konstantin (USA)  
 Matveeva, Inessa (Russia)  
 Matveeva, Liubov (Russia)  
 Mauduit, Christian (France)  
 Maulik, Daves (USA)  
 Maumary, Serge (Switzerland)  
 Maya, Daniel (Mexico)  
 Maya, Joaquin (Mexico)  
 Mazanti, Guilherme (France)  
 Mbang, Joseph (Cameroon)  
 Mbarawa, Eunice (United Republic of Tanzania)  
 Mccallum, Rupert (Germany)  
 Mccann, Robert J. (Canada)  
 Mcclain, Christopher (USA)  
 Mcclure, Donald (USA)  
 Mcclure, Mary (USA)  
 Mcduff, Dusa (USA)  
 Mcferon, Donovan (USA)  
 Mcguinness, Michelle (USA)  
 Mckay, Brendan (Australia)  
 Mckubre-Jordens, Maarten (New Zealand)  
 Mclarty, Colin (USA)  
 McMullen, Curtis (USA)  
 Meakin, John (USA)  
 Medina, Luis (Chile)  
 Medvedeva, Yulia (Russia)  
 Meher, Jaban (India)  
 Mei, Shu-Yuan (P.R. China)  
 Mekheimer, Kh (Egypt)  
 Melikian, Elena (USA)  
 Mena, Hermann (Austria)  
 Mendez-Hernandez, Pedro (Costa Rica)  
 Mendoza, Renier (Austria)  
 Menezes, Ana (France)  
 Meng, Li (P.R. China)  
 Meng, Jie (Republic of Korea)  
 Merino, Pedro (Ecuador)  
 Merle, Frank (France)  
 Mermri, El Bekkaye (Morocco)  
 Merriman, Barry (USA)  
 Meskhi, Alexander (Georgia)  
 Meyer, Johannes (South Africa)  
 Mezard, Ariane (France)  
 Mghazli, Zoubida (Morocco)  
 Miasnikov, Alexei (USA)  
 Miatello, Roberto (Argentina)  
 Mielke, Alexander (Germany)  
 Miguel, Francisca (Spain)  
 Mikayelyan, Hayk (P.R. China)  
 Mikolajewska, Barbara (USA)  
 Mileti, Joseph (USA)  
 Milnor, John (USA)  
 Milovanovic, Gradimir (Serbia)  
 Mimura, Masato (Japan)  
 Min, Byongjae (Republic of Korea)  
 Min, Chohong (Republic of Korea)  
 Min, Geon Ho (Republic of Korea)  
 Min, Jae Won (Republic of Korea)  
 Min, Jeongwon (Republic of Korea)  
 Min, Jin Won (Republic of Korea)  
 Min, Kyung Bun (Republic of Korea)  
 Min, Kyungchan (Republic of Korea)  
 Min, Sung-Hong (Republic of Korea)  
 Min, Taywon (Republic of Korea)  
 Min, Youngmi (Republic of Korea)  
 Min Gi, Choi (Republic of Korea)  
 Mingarelli, Angelo (Canada)  
 Mingo, James (Canada)  
 Minh, Jeon (Republic of Korea)  
 Minjeong, Kim (Republic of Korea)  
 Minkeviciene, Danguole (Lithuania)  
 Minkevicius, Saulius (Lithuania)  
 Mirasol, Lowilton (Philippines)  
 Mirzakhani, Maryam (USA)  
 Mishchenko, Alexander (Russia)  
 Mishchenko, Tatiana (Russia)  
 Misra, Kailash (USA)  
 Misra, Suprava (USA)  
 Mitsui, Kentaro (Japan)  
 Mittal, Hari Vansh Rai (India)  
 Miura, Keiji (Japan)  
 Miyazaki, Rinko (Japan)  
 Mizumach, Tetsu (Japan)  
 Mj, Mahan (India)  
 Mm, Radhika (India)  
 Mo, Chae Young (Republic of Korea)  
 Moalosi, Kebareng Ibeni (Botswana)  
 Moche, Gugu (South Africa)  
 Mochizuki, Takuro (Japan)  
 Mocz, Lucia (USA)  
 Mofidi, Alireza (Iran)  
 Mogilski, Jerzy (USA)  
 Mohammad, Rhudaina (Philippines)  
 Mohammed, Ramy Ramadan Mahmoud (Egypt)  
 Mohanta, Charulata (India)  
 Mohd Kasim, Abdul Rahman (Malaysia)  
 Molati, Motlatsi (Lesotho)  
 Molev, Alexander (Australia)  
 Moleva, Anna (Australia)  
 Mond, David (UK)  
 Monina, Mariia (Russia)  
 Monita, Chanroath (Cambodia)  
 Monnesland, Irene (Norway)  
 Monnier, Samuel (Switzerland)

Montalbán, Antonio (USA)  
Montans, Fernando (Uruguay)  
Montibeller, Celine (France)  
Moon, Gi Tak (Republic of Korea)  
Moon, Gunho (Republic of Korea)  
Moon, Hyunsuk (Republic of Korea)  
Moon, Hyunsuk (Republic of Korea)  
Moon, Ji Eun (Republic of Korea)  
Moon, Kitak (Republic of Korea)  
Moon, Kwang Hee (Republic of Korea)  
Moon, Kyoung-Sook (Republic of Korea)  
Moon, Kyunghwan (Republic of Korea)  
Moon, Myounggho (Republic of Korea)  
Moon, Sangho (Republic of Korea)  
Moon, Sang-Hyeok (Republic of Korea)  
Moon, Seung Hwan (Republic of Korea)  
Moon, Soyoung (Republic of Korea)  
Moon, Sunghwan (Republic of Korea)  
Moon, Sunyo (Republic of Korea)  
Moon, Woo Hyeon (Republic of Korea)  
Moon, Yeonjun (Republic of Korea)  
Moon, Youngtae (Republic of Korea)  
Morales, Carlos (Brazil)  
Moran, Gadi (Israel)  
Moreira, Carlos Gustavo (Brazil)  
Morel, Jean-Michel (France)  
Moreno, Agustin (Colombia)  
Morgan, Kerri (Australia)  
Morgan, Frank (USA)  
Morgan, John (USA)  
Morgenthaler, Stephan (Switzerland)  
Mori, Reiko (Japan)  
Mori, Shigefumi (Japan)  
Mori, Yoshiyuki (Japan)  
Morimoto, Mitsuo (Japan)  
Moriya, Bhavinkumar Kishor Sinh (India)  
Mostafid, Mohammad Hadi (Iran)  
Motegi, Kohei (Japan)  
Mouayn, Zouhair (Morocco)  
Moussa, Seydou (Niger)  
Mu, Lihua (P.R. China)  
Muchtadi-Alamsyah, Intan (Indonesia)  
Mueller, Carl (USA)  
Mugisha, Joseph Tindimubona (Uganda)  
Mugochi, Martin M. (Namibia)  
Muhammad, Nazeer (Pakistan)  
Muirhead, Stephen (UK)  
Mukhaiyar, Utriweni (Indonesia)  
Mukhamedov, Farrukh (Malaysia)  
Mukhamedova, Shirin (Malaysia)  
Mukmin, Zaini (Malaysia)  
Muminov, Zahridin (Malaysia)  
Mun, Juhyeok (Republic of Korea)  
Mun, Sun Bin (Republic of Korea)  
Munemasa, Akihiro (Japan)  
Munembe, Joao Sebastiao Paulo (Mozambique)  
Munoz, Cladio (France)  
Munteanu, Marian Ioan (Romania)  
Murathan, Cengizhan (Turkey)  
Musat, Magdalena (Denmark)  
Muslu, Gulcin Mihriye (Turkey)  
Mustafa, Ghulam (Pakistan)  
Mustapha, Norzieha (Malaysia)  
Mustață, Mircea (USA)  
Muthuvalu, Mohana (Malaysia)  
Myeong Suk, Choi (Republic of Korea)  
Myeongjoon, Park (Republic of Korea)  
Myo Aye, Khin (Myanmar)  
Myong, Jinhong (Republic of Korea)  
Myung, Dae Ho (Republic of Korea)  
Myung, Hyun Ki (Republic of Korea)  
Myung, Ji Yoon (Republic of Korea)  
Myung, Noh Hyun (Republic of Korea)  
Myung, Sunghyun (Republic of Korea)  
Na, Eun Young (Republic of Korea)  
Na, Geonho (Republic of Korea)  
Na, Heui-Gyeong (Republic of Korea)  
Na, In Hyuk (Republic of Korea)  
Na, Joohan (Republic of Korea)  
Na, Sanghoon (Republic of Korea)  
Na, Yeongkwan (Republic of Korea)  
Na, Young Hoon (Republic of Korea)  
Nabati, Hossein (Iran)  
Naber, Aaron (USA)  
Nadeem, Sohail (Pakistan)  
Nagaev, Sergey (Russia)  
Nagao, Taro (Japan)  
Nagata, Kayo (Japan)  
Nagura, Makoto (Japan)  
Nahm, Ghee Hyun (Republic of Korea)  
Naidoo, Dreyeshlin (South Africa)  
Naidoo, Inderasan (South Africa)  
Naito, Toshiki (Japan)  
Najafi, Zahra (Iran)  
Nakahara, Toru (Pakistan)  
Nakajima, Yukiyoshi (Japan)  
Nakamura, Inasa (Japan)  
Nakamura, Gen (Republic of Korea)  
Nakandakari, Masatomo (Japan)  
Nakane, Michiyo (Japan)  
Nakane, Shizuo (Japan)  
Nakano, Masatoshi (Japan)  
Nakano, Daniel (USA)  
Nakata, Toshio (Japan)

Nakatsuka, Harunori (Japan)  
 Nakayama, Muneyasu (Japan)  
 Näkki, Raimo (Finland)  
 Nam, Chang Wan (Republic of Korea)  
 Nam, Chi Hyun (Republic of Korea)  
 Nam, Giung (Republic of Korea)  
 Nam, Hae Won (Republic of Korea)  
 Nam, Hakho (Republic of Korea)  
 Nam, Hayan (Republic of Korea)  
 Nam, Hyunsoo (Republic of Korea)  
 Nam, Seo Yeon (Republic of Korea)  
 Nam, Sunjoo (Republic of Korea)  
 Nam, Sun-Young (Republic of Korea)  
 Nam, Yun-Woo (Republic of Korea)  
 Namgoong, Jeongil (Republic of Korea)  
 Namm, Robert (Russia)  
 Nang, Philibert (Gabon)  
 Narayanaswami, Padma (Canada)  
 Narayanaswami, Pallasena (Canada)  
 Narita, Makoto (Japan)  
 Nastasescu, Laura Elena (Romania)  
 Nastasescu, Maria (USA)  
 Nath, Gorakh (India)  
 Natocho Mango, Lovisa (Uganda)  
 Natroshvili, David (Georgia)  
 Navarro, Alberto (Spain)  
 Navarro, Jose (Spain)  
 Nawa, Victor Mooto (Zambia)  
 Ndiaye, Babacar Mbaye (Senegal)  
 Nebres, Bienvenido (Philippines)  
 Negro, Giuseppe (Spain)  
 Negut, Andrei (Romania)  
 Nemenzo, Fidel (Philippines)  
 Nešetřil, Jaroslav (Czech Republic)  
 Nesetřilova, Helena (Czech Republic)  
 Netay, Elena (Russia)  
 Netay, Igor (Russia)  
 Neumann, Frank (UK)  
 Nevanlinna, Olavi (Finland)  
 Neves, André (UK)  
 Newelski, Ludomir (Poland)  
 Ng, Lenhard (USA)  
 Ngendakumana, Ancille (Burundi)  
 Ngo, Bao Chau (USA)  
 Ngo, Trung (Vietnam)  
 Ngo Thi Thanh, Huong (Vietnam)  
 Ngonn, Seam (Cambodia)  
 Ngounda, Edgard (South Africa)  
 Nguyen, Thien Binh (Republic of Korea)  
 Nguyen, Ha Thu (UK)  
 Nguyen, Yen (USA)  
 Nguyen, Hoang Son (Vietnam)  
 Nguyen, Huy Chieu (Vietnam)  
 Nguyen, Quan (Vietnam)  
 Nguyen, Quoc Thang (Vietnam)  
 Nguyen, Thi Hong Van (Vietnam)  
 Nguyen, Thi Ngoc Diep (Vietnam)  
 Nguyen, Thi Thu Hang (Vietnam)  
 Nguyen, Thi Thu Thuy (Vietnam)  
 Nguyen, Thi Thu Van (Vietnam)  
 Nguyen, Thi Thuy Quynh (Vietnam)  
 Nguyen, Thinh (Vietnam)  
 Nguyen, Tu Cuong (Vietnam)  
 Nguyen, Xuan Tan (Vietnam)  
 Nguyen Khoa, Son (Vietnam)  
 Nguyen Ngoc, Hai (Vietnam)  
 Nguyen Thi, Nga (Vietnam)  
 Nho, Yoonjae (Republic of Korea)  
 Ni, David (Taiwan)  
 Nielsen, Pace (USA)  
 Niethammer, Barbara (Germany)  
 Nikandish, Reza (Iran)  
 Nikolayevsky, Yuri (Australia)  
 Nilov, Fedor (Russia)  
 Nisse, Mounir (Algeria)  
 Nistor, Ana Irina (Romania)  
 Niu, Yanyan (P.R. China)  
 Niyomploy, Akarat (Thailand)  
 Nkemzi, Boniface (Cameroon)  
 No, Sungjong (Republic of Korea)  
 Noh, Heawon (Republic of Korea)  
 Noh, Hee Sang (Republic of Korea)  
 Noh, Hyeonho (Republic of Korea)  
 Noh, Pu Reum (Republic of Korea)  
 Noh, Sangyoon (Republic of Korea)  
 Noh, Si Ung (Republic of Korea)  
 Noh, Sunsook (Republic of Korea)  
 Noh, Tae Hyun (Republic of Korea)  
 Nokoe, Kaku Sagary (Ghana)  
 Nolin, Pierre (Switzerland)  
 Nongxa, Buhle (South Africa)  
 Nongxa, Loyiso (South Africa)  
 Nopendri, Nopendri (Indonesia)  
 Normand, Raoul (Taiwan)  
 Notsu, Hirofumi (Japan)  
 Noy, Marc (Spain)  
 Nuñez, Inés (Spain)  
 Núñez-Betancourt, Luis (USA)  
 Nurtazina, Karlygash (Kazakhstan)  
 Nyayate, Shubhada (India)  
 O, Jeongsik (Republic of Korea)  
 O, Sangrok (Republic of Korea)  
 O, Suil (USA)  
 Oak, Joonsung (Republic of Korea)

Obidjon, Abdullayev (Uzbekistan)  
 Obitsu, Kunio (Japan)  
 Oboudi, Mohammad Reza (Iran)  
 Ocampo Uribe, Oscar Eduardo (Brazil)  
 Octavia, Gael (France)  
 Odell, Matthew (USA)  
 O'Donnell, Ryan (USA)  
 Oeding, Luke (USA)  
 Ogana, Wandera (Kenya)  
 Oguiso, Keiji (Japan)  
 Oh, Byeong-Kweon (Republic of Korea)  
 Oh, Byung-Geun (Republic of Korea)  
 Oh, Chunyoung (Republic of Korea)  
 Oh, Do-Hyun (Republic of Korea)  
 Oh, Dong Yeol (Republic of Korea)  
 Oh, Duk-Soon (Republic of Korea)  
 Oh, Eun Bi (Republic of Korea)  
 Oh, Gyeong Won (Republic of Korea)  
 Oh, Gyujin (Republic of Korea)  
 Oh, Han Young (Republic of Korea)  
 Oh, Hu Taek (Republic of Korea)  
 Oh, Hung-Kuk (Republic of Korea)  
 Oh, Hwanhee (Republic of Korea)  
 Oh, Hwa-Pyoung (Republic of Korea)  
 Oh, Hwapyung (Republic of Korea)  
 Oh, Hyun Ju (Republic of Korea)  
 Oh, Hyungseok (Republic of Korea)  
 Oh, Hyunjeong (Republic of Korea)  
 Oh, Jae-Pill (Republic of Korea)  
 Oh, Jangheon (Republic of Korea)  
 Oh, Jehan (Republic of Korea)  
 Oh, Jeongbin (Republic of Korea)  
 Oh, Jeongseok (Republic of Korea)  
 Oh, Jin-Woo (Republic of Korea)  
 Oh, Ju Young (Republic of Korea)  
 Oh, Jumi (Republic of Korea)  
 Oh, Jun Seok (Republic of Korea)  
 Oh, Jungtaek (Republic of Korea)  
 Oh, Jung-Woo (Republic of Korea)  
 Oh, Sei-Qwon (Republic of Korea)  
 Oh, Se-Jin (Republic of Korea)  
 Oh, Semin (Republic of Korea)  
 Oh, Se-Min (Republic of Korea)  
 Oh, Seungtaik (Republic of Korea)  
 Oh, Seyeon (Republic of Korea)  
 Oh, Sukyung (Republic of Korea)  
 Oh, Sunul (Republic of Korea)  
 Oh, Taek Keun (Republic of Korea)  
 Oh, Wontae (Republic of Korea)  
 Oh, Yong-Geun (Republic of Korea)  
 Oh, Young-Tak (Republic of Korea)  
 Oh, Yousang (Republic of Korea)  
 Oh, Yuhyun (Republic of Korea)  
 Oh, Hee (USA)  
 Oh, Sung-Jin (USA)  
 Ohm, Mi-Ray (Republic of Korea)  
 Ohno, Masahiro (Japan)  
 Oinarov, Ryskul (Kazakhstan)  
 Ok, Jihoon (Republic of Korea)  
 Ok, Seongmin (Republic of Korea)  
 Okada, Tatsuya (Japan)  
 Okamoto, Hisashi (Japan)  
 Okano, Keiji (Japan)  
 Okazaki, Ryota (Japan)  
 Okiyoshi, Mami (Japan)  
 Oksendal, Bernt (Norway)  
 Øksendal, Eva (Norway)  
 Okyay, Mahmut Sait (Republic of Korea)  
 Olela Otafudu, Olivier (South Africa)  
 Oleynik, Oxana (Russia)  
 Oliverio, Paolo Antonio (Italy)  
 Olshanski, Grigori (Russia)  
 Olver, Peter (USA)  
 Omirov, Bakhrom (Uzbekistan)  
 Omoleye Adewumi, Christy (Nigeria)  
 Ondo Melang Ondo, Grace Marlene Peggy  
 (Gabon)  
 Onozuka, Tomokazu (Japan)  
 Onshuus, Alf (Colombia)  
 Orive Illera, Rafael (Spain)  
 Orlowsky, Anita (Germany)  
 Orsted, Bent (Denmark)  
 Osada, Hirofumi (Japan)  
 O'Shea, Donal (USA)  
 Osher, Kathryn (USA)  
 Osher, Stanley (USA)  
 Osinga, Hinke (New Zealand)  
 Osinovskaya, Anna (Belarus)  
 Ospanov, Kordan (Kazakhstan)  
 Osthus, Deryk (UK)  
 Ostrik, Victor (USA)  
 Ostrover, Yaron (Israel)  
 Otani, Shin-Ichi (Japan)  
 Ou, Phichhang (Cambodia)  
 Ouaro, Stanislas (Burkina Faso)  
 Ouedraogo, Marie Françoise (Burkina Faso)  
 Oum, Sang-Il (Republic of Korea)  
 Ouyang, Geng (P.R. China)  
 Oyono, Roger (Polynesia)  
 Ozawa, Tohru (Japan)  
 Ozisik, Sevtap (USA)  
 Pach, János (Hungary)  
 Pacheeripadikkal, Jidesh (India)  
 Pacifico, Maria Jose (Brazil)

Pae, Jun Il (Republic of Korea)  
 Paik, Seunghoon (Republic of Korea)  
 Pak, Eunmi (Republic of Korea)  
 Pakharev, Grigorie (Kyrgyzstan)  
 Pakharev, Alexey (Russia)  
 Pakovich, Fedor (Israel)  
 Palis, Jacob (Brazil)  
 Pamuk, Mehmetcik (Turkey)  
 Pamuk, Semra (Turkey)  
 Pan, Xiaodong (P.R. China)  
 Panackal, Harikrishnan (India)  
 Pani, Amiya Kumar (India)  
 Panthee, Mahendra (Brazil)  
 Pardo, Juan Carlos (Mexico)  
 Parhusip, Hanna Arini (Indonesia)  
 Park, Amen (Republic of Korea)  
 Park, Boram (Republic of Korea)  
 Park, Byeong U. (Republic of Korea)  
 Park, Chan Ho (Republic of Korea)  
 Park, Chan Woo (Republic of Korea)  
 Park, Chang-Kyu (Republic of Korea)  
 Park, Changsoon (Republic of Korea)  
 Park, Chanjae (Republic of Korea)  
 Park, Chansu (Republic of Korea)  
 Park, Chanwoo (Republic of Korea)  
 Park, Cheolmin (Republic of Korea)  
 Park, Cinna (Republic of Korea)  
 Park, Dae Won (Republic of Korea)  
 Park, Donghui (Republic of Korea)  
 Park, Dongjae (Republic of Korea)  
 Park, Dongsei (Republic of Korea)  
 Park, Elena (Republic of Korea)  
 Park, Euisung (Republic of Korea)  
 Park, Euiyong (Republic of Korea)  
 Park, Eun Ik (Republic of Korea)  
 Park, Eun Ji (Republic of Korea)  
 Park, Eun-Hee (Republic of Korea)  
 Park, Eun-Jae (Republic of Korea)  
 Park, Eunku (Republic of Korea)  
 Park, Geonwoo (Republic of Korea)  
 Park, Gibeom (Republic of Korea)  
 Park, Gihyun (Republic of Korea)  
 Park, Ha Yong (Republic of Korea)  
 Park, Haemin (Republic of Korea)  
 Park, Hanchul (Republic of Korea)  
 Park, Hansol (Republic of Korea)  
 Park, Hayan (Republic of Korea)  
 Park, Heesang (Republic of Korea)  
 Park, Ho (Republic of Korea)  
 Park, Hyangdong (Republic of Korea)  
 Park, Hye Sook (Republic of Korea)  
 Park, Hyejin (Republic of Korea)  
 Park, Hyeon Ji (Republic of Korea)  
 Park, Hyeonjun (Republic of Korea)  
 Park, Hyeonwoo (Republic of Korea)  
 Park, Hyo Jin (Republic of Korea)  
 Park, Hyoung Suk (Republic of Korea)  
 Park, Hyoung-Won (Republic of Korea)  
 Park, Hyowon (Republic of Korea)  
 Park, Hyung Ju (Republic of Korea)  
 Park, Hyung Ju (Republic of Korea)  
 Park, Hyungju (Republic of Korea)  
 Park, Il Ah (Republic of Korea)  
 Park, Ilah (Republic of Korea)  
 Park, Inchul (Republic of Korea)  
 Park, Jae Hyeong (Republic of Korea)  
 Park, Jae Kyun (Republic of Korea)  
 Park, Jae Suk (Republic of Korea)  
 Park, Jaehee (Republic of Korea)  
 Park, Jaehoon (Republic of Korea)  
 Park, Jaehwan (Republic of Korea)  
 Park, Jae-Suk (Republic of Korea)  
 Park, Jai Hyun (Republic of Korea)  
 Park, Jay (Republic of Korea)  
 Park, Je Yong (Republic of Korea)  
 Park, Jeehoon (Republic of Korea)  
 Park, Jeong Hyo (Republic of Korea)  
 Park, Jeong Min (Republic of Korea)  
 Park, Jeong Rye (Republic of Korea)  
 Park, Jeong Soo (Republic of Korea)  
 Park, Jeonghoon (Republic of Korea)  
 Park, Jeonghyeong (Republic of Korea)  
 Park, Jeonguk (Republic of Korea)  
 Park, Jeung Eun (Republic of Korea)  
 Park, Ji Hyun (Republic of Korea)  
 Park, Ji Hyun (Republic of Korea)  
 Park, Jiewon (Republic of Korea)  
 Park, Ji-Ho (Republic of Korea)  
 Park, Jihun (Republic of Korea)  
 Park, Jihun (Republic of Korea)  
 Park, Jihyang (Republic of Korea)  
 Park, Jihye (Republic of Korea)  
 Park, Jimi (Republic of Korea)  
 Park, Jin Wan (Republic of Korea)  
 Park, Jinyung (Republic of Korea)  
 Park, Jinman (Republic of Korea)  
 Park, Jinoh (Republic of Korea)  
 Park, Jinsung (Republic of Korea)  
 Park, Jin-Woo (Republic of Korea)  
 Park, Jinyeong (Republic of Korea)  
 Park, Jinyoung (Republic of Korea)  
 Park, Jisu (Republic of Korea)  
 Park, Jiwon (Republic of Korea)  
 Park, Jiwoong (Republic of Korea)

Park, Ji-Yoon (Republic of Korea)  
Park, Jong An (Republic of Korea)  
Park, Jong Hyuck (Republic of Korea)  
Park, Jong Youll (Republic of Korea)  
Park, Jong-Do (Republic of Korea)  
Park, Jongil (Republic of Korea)  
Park, Joonhyun (Republic of Korea)  
Park, Joonsang (Republic of Korea)  
Park, Joowon (Republic of Korea)  
Park, Ju Dong (Republic of Korea)  
Park, Ju Sang (Republic of Korea)  
Park, Jucheol (Republic of Korea)  
Park, Jun Hyeok (Republic of Korea)  
Park, Jun Oh (Republic of Korea)  
Park, Jun Young (Republic of Korea)  
Park, Jun Young (Republic of Korea)  
Park, Jung-Tae (Republic of Korea)  
Park, Jung-Youl (Republic of Korea)  
Park, Junhee (Republic of Korea)  
Park, Junhyung (Republic of Korea)  
Park, Junmi (Republic of Korea)  
Park, Junyeong (Republic of Korea)  
Park, Juseong (Republic of Korea)  
Park, Ki Hong (Republic of Korea)  
Park, Koung Pyo (Republic of Korea)  
Park, Kwang-Soon (Republic of Korea)  
Park, Kyeong-Dong (Republic of Korea)  
Park, Kyewon Koh (Republic of Korea)  
Park, Kyoo-Hong (Republic of Korea)  
Park, Kyoung Il (Republic of Korea)  
Park, Kyoungsuk (Republic of Korea)  
Park, Kyu Tae (Republic of Korea)  
Park, Kyungmee (Republic of Korea)  
Park, Mi Hee (Republic of Korea)  
Park, Mincheol (Republic of Korea)  
Park, Minjae (Republic of Korea)  
Park, Minjoo (Republic of Korea)  
Park, Minsu (Republic of Korea)  
Park, Miyoung (Republic of Korea)  
Park, Moojin (Republic of Korea)  
Park, Peungja (Republic of Korea)  
Park, Poo-Sung (Republic of Korea)  
Park, Puegun (Republic of Korea)  
Park, Saint (Republic of Korea)  
Park, Sang Hu (Republic of Korea)  
Park, Sang Hyun (Republic of Korea)  
Park, Sang Hyun (Republic of Korea)  
Park, Sang Uk (Republic of Korea)  
Park, Sanghoon (Republic of Korea)  
Park, Sang-Hyeon (Republic of Korea)  
Park, Sangjun (Republic of Korea)  
Park, Sehawn (Republic of Korea)  
Park, Sehie (Republic of Korea)  
Park, Seho (Republic of Korea)  
Park, Seongbae (Republic of Korea)  
Park, Seonjeong (Republic of Korea)  
Park, Seoree (Republic of Korea)  
Park, Seung Kyun (Republic of Korea)  
Park, Seung Seol (Republic of Korea)  
Park, Seung Won (Republic of Korea)  
Park, Seungjin (Republic of Korea)  
Park, Seungkook (Republic of Korea)  
Park, Seungwan (Republic of Korea)  
Park, Shin Hae (Republic of Korea)  
Park, So Hyun (Republic of Korea)  
Park, So-Hee (Republic of Korea)  
Park, Soohyun (Republic of Korea)  
Park, Su Min (Republic of Korea)  
Park, Su Yong (Republic of Korea)  
Park, Sun Hee (Republic of Korea)  
Park, Sun Hoo (Republic of Korea)  
Park, Sun Woo (Republic of Korea)  
Park, Sung Gi (Republic of Korea)  
Park, Sung Jae (Republic of Korea)  
Park, Sung Jin (Republic of Korea)  
Park, Sung Jun (Republic of Korea)  
Park, Sung Woo (Republic of Korea)  
Park, Sungguk (Republic of Korea)  
Park, Sungjin (Republic of Korea)  
Park, Sunwoo (Republic of Korea)  
Park, Tae Hwan (Republic of Korea)  
Park, Tae Young (Republic of Korea)  
Park, Wi Gon (Republic of Korea)  
Park, Won Seok (Republic of Korea)  
Park, Wonjin (Republic of Korea)  
Park, Won-Kyu (Republic of Korea)  
Park, Yong Jin (Republic of Korea)  
Park, Yong Moon (Republic of Korea)  
Park, Yoon Jae (Republic of Korea)  
Park, Yoon Kyung (Republic of Korea)  
Park, Young Ho (Republic of Korea)  
Park, Young Min (Republic of Korea)  
Park, Young Woong (Republic of Korea)  
Park, Younghee (Republic of Korea)  
Park, Young-Hwan (Republic of Korea)  
Park, Yunbeom (Republic of Korea)  
Park, Yun-Ha (Republic of Korea)  
Park, Yun-Jung (Republic of Korea)  
Park, Junhyung (UK)  
Park, Alice (USA)  
Park, Haesun (USA)  
Park, Hyungbin (USA)  
Park, Jane (USA)  
Park, Jun Yong (USA)

Park, Kyungbae (USA)  
 Park, Peter (USA)  
 Park, Seongshim (USA)  
 Parsa, S. Reza (Iran)  
 Pascasio, Arlene (Philippines)  
 Paseman, Gerhard (USA)  
 Pastor Ferreira, Ademir (Brazil)  
 Patel, Ajit (India)  
 Patel, Sanjaykumar (India)  
 Patel, Shital (India)  
 Patel, Jemini (USA)  
 Pathak, Vinod (India)  
 Patidar, Vinod (India)  
 Patidar, Kailash C. (South Africa)  
 Paulhus, Jennifer (USA)  
 Paun, Mihai (France)  
 Pauna, Matti (Finland)  
 Péché, Sandrine (France)  
 Pechen, Alexander (Russia)  
 Peche-Semadeni, Paloma (France)  
 Pedroza, Andres (Mexico)  
 Pei, Yufeng (P.R. China)  
 Pekonen, Osmo (Finland)  
 Pelletier, Arya Devi (USA)  
 Pelletier, Daniel Phillip (USA)  
 Pelletier, Kalyan Bhargava (USA)  
 Peltola, Eveliina (Finland)  
 Peltonen, Kirsi (Finland)  
 Peng, Shige (P.R. China)  
 Peng, Danping (USA)  
 Perepelkina, Yulianna (Russia)  
 Perez, David (Spain)  
 Perez, Juan De Dios (Spain)  
 Perez-Chavela, Ernesto (Mexico)  
 Persson, Tomas (Sweden)  
 Perthame, Benoit (France)  
 Petrache, Mircea (France)  
 Petrov, Leonid (USA)  
 Peyghami, Mohammad Reza (Iran)  
 Peypouquet, Juan (Chile)  
 Phalavonk, Utomporn (Thailand)  
 Pham Dinh, Tung (Vietnam)  
 Pham Huu Anh, Ngoc (Vietnam)  
 Pham Minh, Hien (Vietnam)  
 Phan, Quoc Khanh (Vietnam)  
 Phan, Thi Ha Duong (Vietnam)  
 Phauk, Sokkhey (Cambodia)  
 Phu, Hoang Xuan (Vietnam)  
 Phung, Ho Hai (Vietnam)  
 Phuong, Sokchann (Cambodia)  
 Piao, Daxiong (P.R. China)  
 Piao, Guangri (P.R. China)  
 Piao, Yongjie (P.R. China)  
 Piccione, Paolo (Brazil)  
 Pichika, Srinivasu (India)  
 Piene, Ragni (Norway)  
 Pieroni, Andrea (Italy)  
 Pietraho, Jennifer (USA)  
 Pietraho, Thomas (USA)  
 Pila, Jonathan (UK)  
 Pillai, Nastesh (USA)  
 Pilyugin, Sergey (Russia)  
 Pineda, Angel (USA)  
 Pinheiro, Vilton (Brazil)  
 Pinto, Alberto (Portugal)  
 Pintz, János (Hungary)  
 Pinzari, Gabriella (Italy)  
 Pipher, Jill (USA)  
 Piryatinska, Alexandra (USA)  
 Pirzada, Shariefuddin (India)  
 Pisier, Gilles (France)  
 Planchon, Fabrice (France)  
 Plaut, Conrad (USA)  
 Pochai, Nopparat (Thailand)  
 Podesta, Ricardo Alberto (Argentina)  
 Pokela, Heikki (Finland)  
 Pollanen, Marco (Canada)  
 Pollicott, Mark (UK)  
 Polthier, Konrad (Germany)  
 Polyakova, Lyudmila (Russia)  
 Pomareda, Rolando (Chile)  
 Ponge, Raphael (Republic of Korea)  
 Pontim, Carolina (Brazil)  
 Poon, Yat Sun (USA)  
 Pop, Christine (Germany)  
 Potapova, Aiyyna (Russia)  
 Potapova, Sargylana (Russia)  
 Pradhan, Debasish (India)  
 Praeger, Cheryl (Australia)  
 Prakash, Om (India)  
 Prasattong, Santipong (Thailand)  
 Prause, Istvan (Finland)  
 Pravda-Starov, Karel (France)  
 Prins, Abraham (South Africa)  
 Promislow, David (Canada)  
 Promislow, Shirley (Canada)  
 Proske, Frank (Norway)  
 Provido, Eden Delight (Germany)  
 Przytycki, Feliks (Poland)  
 Ptak, Marek (Poland)  
 Purin, Marju (USA)  
 Purnama, Anton (Oman)  
 Purohit, Sunil Dutt (India)  
 Pyo, Gina (Republic of Korea)

Pyo, Ginwoo (Republic of Korea)  
 Pyo, Jae-Hong (Republic of Korea)  
 Pyo, Jaewoo (Republic of Korea)  
 Pyo, Ji Soo (Republic of Korea)  
 Pyo, Juncheol (Republic of Korea)  
 Pyun, Dobyung (Republic of Korea)  
 Qazaqzeh, Khaled (Jordan)  
 Qi, Feng (P.R. China)  
 Qi, Jiayue (P.R. China)  
 Qiao, Xiurang (P.R. China)  
 Qu, Anjing (P.R. China)  
 Qu, Jingjing (P.R. China)  
 Quach, Tri (Finland)  
 Quarteroni, Alfio (Switzerland)  
 Quehenberger, Renate (Austria)  
 Quintero, Jose (Colombia)  
 Qureshi, Muhammad Imran (Pakistan)  
 Qureshi, Rabia (Pakistan)  
 Rabajante, Jomar (Philippines)  
 Rabarison, Fanomezantsoa Patrick (Madagascar)  
 Radu, Remus (USA)  
 Radzhabova, Lutfiya (Tadjikistan)  
 Raghunathan, Madabusi (India)  
 Rahmoeller, Margaret (USA)  
 Rajchgot, Jenna (USA)  
 Rajendran, Venkatesh (India)  
 Raka, Madhu (India)  
 Rakhimov, Isamiddin (Uzbekistan)  
 Rakhimova, Elena (Uzbekistan)  
 Rakic, Zoran (Serbia)  
 Rakotondrajao, Fanja (Madagascar)  
 Ramachandran, Balasubramanian (India)  
 Ramachandran, Raja (India)  
 Ramakrishnan, Jothilakshmi (India)  
 Raman, Parimala (USA)  
 Ramanan, Kavita (USA)  
 Ramirez Ospina, Hector Fabian (Spain)  
 Ramirez-Solano, Maria (Denmark)  
 Ramos, Daniel (Spain)  
 Ranasinghe, Ranasinghe (Sri Lanka)  
 Randriamanirisoa, Saha Hasina (Madagascar)  
 Rangel Quintino, Karina Aparecida (Brazil)  
 Rannen, Amal (Tunisia)  
 Rao, Sheng (P.R. China)  
 Raphael, Pierre (France)  
 Rapinchuk, Andrei S. (USA)  
 Rapinchuk, Igor (USA)  
 Rapinchuk, Tatiana (USA)  
 Rappoport, Juri (Russia)  
 Rasila, Antti (Finland)  
 Ratan, Karam (India)  
 Rathee, Saloni (India)  
 Ratnaparkhi, Gayatri (India)  
 Ratnaparkhi, Sunanda (India)  
 Ravelonirina, Hanitriniaina Sammy Gregoire (Madagascar)  
 Ray, Nigel (UK)  
 Razani, Abdolrahman (Iran)  
 Rebiai, Salah Eddine (Algeria)  
 Reddy, Batmanathan Dayanand (South Africa)  
 Redhu, Poonam (India)  
 Redondo Buitrago, Antonia (Spain)  
 Ree, Sang Yub (Republic of Korea)  
 Ree, Sangwook (Republic of Korea)  
 Refiei Demneh, Rafat (Iran)  
 Reggiani, Silvio (Argentina)  
 Reid, Miles (UK)  
 Reinfelds, Andrejs (Latvia)  
 Reinova, Liudmila (Russia)  
 Rémy, Bertrand (France)  
 Renchin-Ochir, Mijiddorj (Mongolia)  
 Rennemo, Jorgen (UK)  
 Rentsen, Enkhbat (Mongolia)  
 Resende, Maria Joao (Brazil)  
 Ressayre, Nicolas (France)  
 Reuter, Andreas (Germany)  
 Reyes Ahumada, Graciela Astrid (Mexico)  
 Reynov, Oleg (Russia)  
 Rezakhah, Saeid (Iran)  
 Rezunencko, Oleksandr (Ukraine)  
 Rhee, Eunjai (Republic of Korea)  
 Rhee, Hyewon (Republic of Korea)  
 Rhee, Ki Hun (Republic of Korea)  
 Rho, Yoomi (Republic of Korea)  
 Richert, Lucy (USA)  
 Richert, Norman (USA)  
 Ries-Bossemeyer, Renate Anna (Germany)  
 Rim, Jeongdae (Republic of Korea)  
 Rim, Kyung Soo (Republic of Korea)  
 Ringström, Hans (Sweden)  
 Rivas, Cristobal (Chile)  
 Rivera, Marissa (Brazil)  
 Rivero, Victor (Mexico)  
 Roath, Chan (Cambodia)  
 Robbiano, Luc (France)  
 Robert, Damien (France)  
 Robertson, Neil (USA)  
 Roche-Newton, Oliver (UK)  
 Rodkina, Alexandra (Jamaica)  
 Rödl, Vojtech (Czech Republic)  
 Rodriguez, Rubi (Chile)  
 Rognes, John (Norway)  
 Roh, Gil Ho (Republic of Korea)  
 Roh, Jaiok (Republic of Korea)

- Roh, Se Hyeong (Republic of Korea)  
 Rojas, Anita (Chile)  
 Romaskevich, Olga (Russia)  
 Roney-Dougal, Colva (UK)  
 Rong, Feng (P.R. China)  
 Roque, Marian (Philippines)  
 Rordam, Mikael (Denmark)  
 Rosadi, Dedi (Indonesia)  
 Rosales, Leobardo (USA)  
 Rossman, Benjamin (Japan)  
 Rouchon, Pierre (France)  
 Rousseau, Brian (Canada)  
 Rousseau, Christiane (Canada)  
 Rovenski, Vladimir (Israel)  
 Roy, Marie-Francoise (France)  
 Roy, Prosenjit (India)  
 Rozikov, Utkir (Uzbekistan)  
 Ruan, Zhuoping (P.R. China)  
 Rubinstein-Salzedo, Simon (USA)  
 Ruchjana, Budi Nurani (Indonesia)  
 Rudnick, Zeev (Israel)  
 Rufin Ghys, Martine (France)  
 Ruivivar, Leonor (Philippines)  
 Ruiz, Angel (Costa Rica)  
 Russell, Heather (USA)  
 Rusu, Galina (Moldova)  
 Ryoo, Jung Hyun (Republic of Korea)  
 Ryoo, Sang Woo (Republic of Korea)  
 Ryou, Ho Joon (Republic of Korea)  
 Ryu, Bo-Hwa (Republic of Korea)  
 Ryu, Chunmi (Republic of Korea)  
 Ryu, Eojin (Republic of Korea)  
 Ryu, Homoon (Republic of Korea)  
 Ryu, Hong (Republic of Korea)  
 Ryu, Inyoung (Republic of Korea)  
 Ryu, Jeongseog (Republic of Korea)  
 Ryu, Ji Hyang (Republic of Korea)  
 Ryu, Jiyong (Republic of Korea)  
 Ryu, Jong Sook (Republic of Korea)  
 Ryu, Jun Hwan (Republic of Korea)  
 Ryu, Jun Seung (Republic of Korea)  
 Ryu, Juyoung (Republic of Korea)  
 Ryu, Seungjin (Republic of Korea)  
 Ryu, Sunyoung (Republic of Korea)  
 Ryu, Youngpyo (Republic of Korea)  
 S, Kumaresan (India)  
 S Sastry, Challa (India)  
 S.R. Srinivasa Rao, Arni (USA)  
 Saadetoglu, Muge (Cyprus)  
 Sabatti, Chiara (USA)  
 Sabau, Sorin V. (Japan)  
 Sabzrou, Hossein (Iran)  
 Sacawa, Paul (Canada)  
 Saddi, Daryl Allen (Philippines)  
 Sadhu, Vivek (India)  
 Sadirbajevs, Felikss (Latvia)  
 Sadullaev, Azimbay (Uzbekistan)  
 Sadullaeva, Nilufar (Uzbekistan)  
 Sadullaeva, Shakhlo (Uzbekistan)  
 Sae-Jie, Wichuta (Thailand)  
 Saez, Mariel (Chile)  
 Safarov, Utkir (Uzbekistan)  
 Sahadevan, Ramajayam (India)  
 Sahin, Mesut (Turkey)  
 Sahoo, Pradyumn Kumar (India)  
 Sahraoui, Fatiha (Algeria)  
 Saidou, Adamou (Niger)  
 Saifullah, Khalid (Pakistan)  
 Sain, Debmalya (India)  
 Saint Raymond, Laure (France)  
 Saint-Donat, Bernard (USA)  
 Saito, Shingo (Japan)  
 Saito, Yasuhisa (Japan)  
 Sakai, Keiichi (Japan)  
 Sakata, Mika (Japan)  
 Sakue, Kazuhiro (Japan)  
 Sakulrang, Sasikarn (Thailand)  
 Sal Moslehian, Mohammad (Iran)  
 Salavati, Erfan (Iran)  
 Saleh, Khaerudin (Indonesia)  
 Salimath, Chandrashekarayya (India)  
 Salimath, Rajeshwari (India)  
 Salur, Sema (USA)  
 Sanders, Tom (UK)  
 Sane, Sharad (India)  
 Sang Rae, Kim (Republic of Korea)  
 Sangare, Daouda (Ivory Coast)  
 Sankaran, Abhisekh (India)  
 Sanogo, Moumine (Mali)  
 Sanz-Sole, Marta (Spain)  
 Sarbadhikari, Haimanti (India)  
 Sargsyan, Alla (Armenia)  
 Sarich, Marco (Germany)  
 Sarkar, Jaydeb (India)  
 Sarmiento, Jumela (Philippines)  
 Sasamoto, Akira (Japan)  
 Sato, Kumi (Japan)  
 Sato, Masahisa (Japan)  
 Sato, Takako (Japan)  
 Sato, Wataru (Japan)  
 Savin, Gordan (USA)  
 Sawae, Naoko (Japan)  
 Sawae, Ryuichi (Japan)  
 Sawon, Justin (USA)

Sayfy, Ali (UAE)  
 Sbierski, Jan (UK)  
 Schacht, Mathias (Germany)  
 Schaeffer, George (USA)  
 Schaffhauser, Florent (Colombia)  
 Schaper, Judith (USA)  
 Scheidler, Renate (Canada)  
 Scheven, Christoph (Germany)  
 Schick, Thomas (Germany)  
 Schlag, Wilhelm (USA)  
 Schlue, Volker (Canada)  
 Schmitt, Alexander (Germany)  
 Schmitt, Daniela (Germany)  
 Schmuland, Byron (Canada)  
 Scholze, Peter (Germany)  
 Schreiber, Bertram (USA)  
 Schreiber, Rita (USA)  
 Schreyer, Frank-Olaf (Germany)  
 Schroers, Bernd (Germany)  
 Schulze-Pillot, Rainer (Germany)  
 Scrimshaw, Travis (USA)  
 Sebastian, Elizabeth (India)  
 Seeliger, Birgit (Germany)  
 Seet, Jonathan (Australia)  
 Seggev, Itai (USA)  
 Seguin, Nicolas (France)  
 Sehatkhah, Mehdi (Iran)  
 Selinger, Nikita (USA)  
 Selmane, Schehrazad (Algeria)  
 Semenov, Vladimir I. (Russia)  
 Semenov, Mikhail (UK)  
 Semmler, Angelika (Germany)  
 Semmler, K.-D. (Switzerland)  
 Senapathi, Chaitanya (India)  
 Sene, Abdou (Senegal)  
 Seo, Aeryeong (Republic of Korea)  
 Seo, Been (Republic of Korea)  
 Seo, Bongun (Republic of Korea)  
 Seo, Boyoon (Republic of Korea)  
 Seo, Changwon (Republic of Korea)  
 Seo, Dana (Republic of Korea)  
 Seo, Dong Woo (Republic of Korea)  
 Seo, Donggyun (Republic of Korea)  
 Seo, Donghwi (Republic of Korea)  
 Seo, Dongjin (Republic of Korea)  
 Seo, Gyeong-Sig (Republic of Korea)  
 Seo, Haesong (Republic of Korea)  
 Seo, Hoseob (Republic of Korea)  
 Seo, Hyowon (Republic of Korea)  
 Seo, Hyung Joo (Republic of Korea)  
 Seo, Ihyeok (Republic of Korea)  
 Seo, Insuk (Republic of Korea)  
 Seo, Jae Hyeon (Republic of Korea)  
 Seo, Jeong Wan (Republic of Korea)  
 Seo, Jeonghyeon (Republic of Korea)  
 Seo, Jeong-Woo (Republic of Korea)  
 Seo, Ji Hwan (Republic of Korea)  
 Seo, Jihwan (Republic of Korea)  
 Seo, Jihyuk (Republic of Korea)  
 Seo, Jin Keun (Republic of Korea)  
 Seo, Jinwoo (Republic of Korea)  
 Seo, Joo Hyoung (Republic of Korea)  
 Seo, Ju Young (Republic of Korea)  
 Seo, Junseok (Republic of Korea)  
 Seo, Keomkyo (Republic of Korea)  
 Seo, Kwang Jin (Republic of Korea)  
 Seo, Kyung Duck (Republic of Korea)  
 Seo, Myoungsoo (Republic of Korea)  
 Seo, Seong Mi (Republic of Korea)  
 Seo, Seunghyun (Republic of Korea)  
 Seo, Seungsuk (Republic of Korea)  
 Seo, Soogil (Republic of Korea)  
 Seo, Hoseong (UK)  
 Seok, Hye Young (Republic of Korea)  
 Seok, Jeongjoo (Republic of Korea)  
 Seok, Jinmyoung (Republic of Korea)  
 Seok Won, Choi (Republic of Korea)  
 Seol, Han-Guk (Republic of Korea)  
 Seol, Hongjin (Republic of Korea)  
 Seol, Ji Yoon (Republic of Korea)  
 Seol, Jin Seok (Republic of Korea)  
 Seol, Seouk Bong (Republic of Korea)  
 Seon-Hong, Ahn (Republic of Korea)  
 Seong, Juno (Republic of Korea)  
 Seong, See-Hak (Republic of Korea)  
 Seong, Yeongho (Republic of Korea)  
 Seongtae, Wi (Republic of Korea)  
 Seongu, Kim (Republic of Korea)  
 Sepanski, Mark (USA)  
 Seppäläinen, Timo (USA)  
 Seretlo, Thekiso (South Africa)  
 Serganova, Vera (USA)  
 Sergeev, Armen (Russia)  
 Sergeichuk, Vladimir (Ukraine)  
 Sesum, Natasa (USA)  
 Seung Jun, Lee (Republic of Korea)  
 Seung Kyu, Lee (Republic of Korea)  
 Seunghye, Kim (Republic of Korea)  
 Shabbir, Ayesha (Pakistan)  
 Shafie, Sharidan (Malaysia)  
 Shakiban, Cheri (USA)  
 Shalaiko, Taras (Germany)  
 Shamsi, Zahid (Republic of Korea)  
 Shankar, Arul (India)

Shanmugam, Saravanan (India)  
 Shao, Hongliang (P.R. China)  
 Shao, Zhiqiang (P.R. China)  
 Sharafullina, Albina (USA)  
 Sharipov, Olimjon (Uzbekistan)  
 Sharma, Bibhya (Fiji)  
 Sharma, Vikram (India)  
 Sharma, Vishnu (India)  
 Shatashvili, Samson (Ireland)  
 Sheen, Dongwoo (Republic of Korea)  
 Sheikh, Neyaz (India)  
 Shen, Weixiao (Singapore)  
 Sheng, Linxue (P.R. China)  
 Sheng, Yuqiu (P.R. China)  
 Shi, Yu-Ying (P.R. China)  
 Shim, Eun Hwa (Republic of Korea)  
 Shim, Jae Seon (Republic of Korea)  
 Shim, Jae Ung (Republic of Korea)  
 Shim, Kyung-Ah (Republic of Korea)  
 Shim, Soho (Republic of Korea)  
 Shim, Woo-Joo (Republic of Korea)  
 Shim, Youngin (Republic of Korea)  
 Shim, Yugeun (Republic of Korea)  
 Shim, Yun (Republic of Korea)  
 Shim, Eunha (USA)  
 Shimokoshi, Hiroko (Japan)  
 Shimura, Hajime (Japan)  
 Shimura, Wataru (Japan)  
 Shin, Jongson (Japan)  
 Shin, Yeong Lin (Republic of Korea)  
 Shin, An Sook (Republic of Korea)  
 Shin, Bomi (Republic of Korea)  
 Shin, Bum Geun (Republic of Korea)  
 Shin, Bumsoo (Republic of Korea)  
 Shin, Byeong-Chun (Republic of Korea)  
 Shin, Daniel (Republic of Korea)  
 Shin, Dong Uy (Republic of Korea)  
 Shin, Dong Yun (Republic of Korea)  
 Shin, Donghyeok (Republic of Korea)  
 Shin, Dongsoo (Republic of Korea)  
 Shin, Dong-Wook (Republic of Korea)  
 Shin, Eun Joo (Republic of Korea)  
 Shin, Gicheol (Republic of Korea)  
 Shin, Giyeon (Republic of Korea)  
 Shin, Haeng Beom (Republic of Korea)  
 Shin, Heayong (Republic of Korea)  
 Shin, Heesung (Republic of Korea)  
 Shin, Ho Sook (Republic of Korea)  
 Shin, Hwanyong (Republic of Korea)  
 Shin, Hyangkeun (Republic of Korea)  
 Shin, Hyeok Gyo (Republic of Korea)  
 Shin, Hyung-Seok (Republic of Korea)  
 Shin, Hyunkyung (Republic of Korea)  
 Shin, Hyunyong (Republic of Korea)  
 Shin, Inchul (Republic of Korea)  
 Shin, Jae Moon (Republic of Korea)  
 Shin, Jae Woong (Republic of Korea)  
 Shin, Jaeho (Republic of Korea)  
 Shin, Jaemin (Republic of Korea)  
 Shin, Jaemin (Republic of Korea)  
 Shin, Jaemin (Republic of Korea)  
 Shin, Jaesun (Republic of Korea)  
 Shin, Jaeyong (Republic of Korea)  
 Shin, Jang Wan (Republic of Korea)  
 Shin, Jeongsu (Republic of Korea)  
 Shin, Jeonho (Republic of Korea)  
 Shin, Jinwoo (Republic of Korea)  
 Shin, Joon Woo (Republic of Korea)  
 Shin, Joonhyung (Republic of Korea)  
 Shin, Joonkook (Republic of Korea)  
 Shin, Jun Yong (Republic of Korea)  
 Shin, Minsub (Republic of Korea)  
 Shin, Sang Peak (Republic of Korea)  
 Shin, Seong Jae (Republic of Korea)  
 Shin, Seung Hun (Republic of Korea)  
 Shin, Sungchan (Republic of Korea)  
 Shin, Suyeon (Republic of Korea)  
 Shin, Wonsik (Republic of Korea)  
 Shin, Woo Ho (Republic of Korea)  
 Shin, Woosung (Republic of Korea)  
 Shin, Yeong Jun (Republic of Korea)  
 Shin, Yeongtae (Republic of Korea)  
 Shin, Yong Min (Republic of Korea)  
 Shin, Yongjoo (Republic of Korea)  
 Shin, Yoon Joo (Republic of Korea)  
 Shin, Youjin (Republic of Korea)  
 Shin, Young Jun (Republic of Korea)  
 Shin, James (USA)  
 Shin, Sug Woo (USA)  
 Shindyaoin, Andrey (Mozambique)  
 Shinya, Kadota (Japan)  
 Shirai, Tomoyuki (Japan)  
 Shishikura, Mitsuhiro (Japan)  
 Shishkov, Andrey (Ukraine)  
 Shlyk, Valiantsina (Belarus)  
 Shlyk, Vladimir (Belarus)  
 Shtilmark, Maria (Russia)  
 Shu, Lin (P.R. China)  
 Shu, Chi-Wang (USA)  
 Shurayeva, Damegul (Kazakhstan)  
 Shvai, Nadiya (Ukraine)  
 Si, Duc Quang (Vietnam)  
 Sidana, Swati (India)  
 Siddiqi, Abul Hasan (India)

Sidoravicius, Vldas (Brazil)  
Siebert, Bernd (Germany)  
Siegel, Charles (Japan)  
Siegmond-Schultze, Reinhard (Norway)  
Sigala, Nikoleta (Greece)  
Siggers, Mark (Republic of Korea)  
Sihwaningrum, Idha (Indonesia)  
Sikhov, Mirbulat (Kazakhstan)  
Silaban, Denny Riama (Indonesia)  
Siljander, Juhana (Finland)  
Silvestre, Luis (USA)  
Sim, Doyi (Republic of Korea)  
Sim, Goen-Hee (Republic of Korea)  
Sim, Han Na (Republic of Korea)  
Sim, Imbo (Republic of Korea)  
Sim, Inbo (Republic of Korea)  
Sim, Young Jae (Republic of Korea)  
Simanjuntak, Rinovia (Indonesia)  
Simic, Slavko (Serbia)  
Simons, James (USA)  
Simsir, Fatma Muazzez (Turkey)  
Sin, Donkha (Russia)  
Sin Yin, Teh (Malaysia)  
Singh, Mansa (Canada)  
Singh, Meera (Canada)  
Singh, Mahender (India)  
Singh, Vineet Kumar (India)  
Singh, Ajaya (Nepal)  
Singhun, Sirirat (Thailand)  
Sinha, Kayan B. (India)  
Sirisack, Sackmone (Laos)  
Sison, Virgilio (Philippines)  
Siu, Man Keung (P.R. China)  
Siu Chan, Fung Kit (P.R. China)  
Siwach, Vikash (India)  
Skalski, Adam (Poland)  
Skopenkov, Mikhail (Russia)  
Skopina, Maria (Russia)  
Skowera, Jonathan (Switzerland)  
Skulkhu, Ruth J. (Thailand)  
Skvortsov, Valentin (Russia)  
Slamin, Slamin (Indonesia)  
Slavova, Angela (Bulgaria)  
Slominska, Jolanta (Poland)  
Smania, Daniel (Brazil)  
Smith, Karen E (USA)  
Snipes, Marie (USA)  
Snopche, Ilir (Brazil)  
So, Hoseop (Republic of Korea)  
So, Hyoungsuk (Republic of Korea)  
So, Jae Won (Republic of Korea)  
So, Jae Young (Republic of Korea)  
Sodam, Yi (Republic of Korea)  
Sodin, Alexander (USA)  
Sohn, Jaebum (Republic of Korea)  
Sohn, Ji Hoon (Republic of Korea)  
Sohn, Sung-Ik (Republic of Korea)  
Sohn, Woon Ha (Republic of Korea)  
Sohn, Wuhyun (Republic of Korea)  
Sola Conde, Luis Eduardo (Spain)  
Solano Albornoz, Omar Javier (Brazil)  
Solecki, Slawomir (USA)  
Solotar, Andrea (Argentina)  
Sommerfield, Karla (USA)  
Son, Bumsuk (Republic of Korea)  
Son, Byeong Uk (Republic of Korea)  
Son, Dae Won (Republic of Korea)  
Son, Dong Hawan (Republic of Korea)  
Son, Eun A (Republic of Korea)  
Son, Hyeon-Min (Republic of Korea)  
Son, Kwanhong (Republic of Korea)  
Son, Min June (Republic of Korea)  
Son, Sang Jun (Republic of Korea)  
Son, Woo Hyung (Republic of Korea)  
Son, Woo-Sik (Republic of Korea)  
Son, Youngjun (Republic of Korea)  
Son, Youngmin (Republic of Korea)  
Song, Chong (P.R. China)  
Song, He (P.R. China)  
Song, Shu (P.R. China)  
Song, Arim (Republic of Korea)  
Song, Chae Lin (Republic of Korea)  
Song, Chahwan (Republic of Korea)  
Song, Chan Woo (Republic of Korea)  
Song, Chang Hun (Republic of Korea)  
Song, Daeuk (Republic of Korea)  
Song, Dong Hun (Republic of Korea)  
Song, Donghoon (Republic of Korea)  
Song, Eunhee (Republic of Korea)  
Song, Heejung (Republic of Korea)  
Song, Hwanwoong (Republic of Korea)  
Song, Jaesub (Republic of Korea)  
Song, Ji Sue (Republic of Korea)  
Song, Jinsub (Republic of Korea)  
Song, Jong Won (Republic of Korea)  
Song, Jongbaek (Republic of Korea)  
Song, Joong Hyun (Republic of Korea)  
Song, Jounghmin (Republic of Korea)  
Song, Junghyun (Republic of Korea)  
Song, Kyeong (Republic of Korea)  
Song, Minjeong (Republic of Korea)  
Song, Minju (Republic of Korea)  
Song, Samuel (Republic of Korea)  
Song, Sang Yeop (Republic of Korea)

- Song, Seohyeon (Republic of Korea)  
 Song, Seok-Zun (Republic of Korea)  
 Song, Seo-Young (Republic of Korea)  
 Song, Shin-Eui (Republic of Korea)  
 Song, Taeseung (Republic of Korea)  
 Song, Wonyeong (Republic of Korea)  
 Song, Yeaeun (Republic of Korea)  
 Song, Yeong-Ho (Republic of Korea)  
 Song, Yongjin (Republic of Korea)  
 Song, Yongsoo (Republic of Korea)  
 Song, Yoon Ae (Republic of Korea)  
 Song, Younghun (Republic of Korea)  
 Song, Yun Min (Republic of Korea)  
 Song, Eunseo (Republic of Korea)  
 Sood, Garima (India)  
 Soparman, Basuki Widodo (Indonesia)  
 Sorger, Christoph (France)  
 Sostaks, Aleksandrs (Latvia)  
 Soto-Andrade, Jorge (Chile)  
 Sottinen, Tommi (Finland)  
 Soundararajan, Gnanavel (India)  
 Souza Brandao, Paulo Rogerio (Brazil)  
 Soyeon, Ju (Republic of Korea)  
 Speicher, Moritz (Germany)  
 Speicher, Roland (Germany)  
 Spicer, Chris (USA)  
 Spitkovsky, Ilya (USA)  
 Srinivas, Vasudevan (India)  
 Srinivasan, Kesavan (India)  
 Srinivasan, Raman Paranj (USA)  
 Srivastava, Nikhil (India)  
 Srivastava, Shashi Mohan (India)  
 Ssebuliba, Joseph (Uganda)  
 Staib, Erich (USA)  
 Stamatovic, Biljana (Montenegro)  
 Steger, Angelika (Switzerland)  
 Stein, Maya (Chile)  
 Steinhorn, Charles (USA)  
 Stenseth, Nils Chr (Norway)  
 Sternheimer, Daniel (France)  
 Stillwell, John (Australia)  
 Stillwell, Elaine (USA)  
 Stipsicz, Andras (Hungary)  
 Stoica, Ioana (USA)  
 Stoimenow, Alexander (Republic of Korea)  
 Strickland, Elisabetta (Italy)  
 Strien, Sebastian Van (UK)  
 Stuart, Andrew (UK)  
 Stuhl, Izabella (Brazil)  
 Sturmfels, Bernd (USA)  
 Stussak, Christian (Germany)  
 Su, Xifeng (P.R. China)  
 Su, Yiming (P.R. China)  
 Su, Zhanjun (P.R. China)  
 Suarez Alvarez, Mariano (Argentina)  
 Sugeng, Kiki Ariyanti (Indonesia)  
 Sugiyama, Toshi (Japan)  
 Suh, Dong Youp (Republic of Korea)  
 Suh, Geewon (Republic of Korea)  
 Suh, Kiseok (Republic of Korea)  
 Suh, Kisoo (Republic of Korea)  
 Suh, Pyongwon (Republic of Korea)  
 Suh, Uhi Rinn (Republic of Korea)  
 Suh, Young Jin (Republic of Korea)  
 Suh(Ju), Eunice Eunok (Republic of Korea)  
 Suhyeon, Kim (Republic of Korea)  
 Sukhotin, Alexander (Russia)  
 Sullivan, John (Germany)  
 Sum, Sze Wan Emily (Hong Kong)  
 Sumarti, Novriana (Indonesia)  
 Sumetkijakan, Songkiat (Thailand)  
 Sumin, Lim (Republic of Korea)  
 Sun, Cong (P.R. China)  
 Sun, Cui Xia (P.R. China)  
 Sun, Leping (P.R. China)  
 Sun, Shenghao (P.R. China)  
 Sun, Yun (P.R. China)  
 Sun, Yeneng (Singapore)  
 Sun, Michael (USA)  
 Sundararaman, Ramanan (India)  
 Sung, Byoungchan (Republic of Korea)  
 Sung, Chanyoung (Republic of Korea)  
 Sung, Hyoshin (Republic of Korea)  
 Sung, Jihyun (Republic of Korea)  
 Sungnul, Surattana (Thailand)  
 Suragan, Durvudkhan (UK)  
 Surapholchai, Chotiros (Thailand)  
 Suriajaya, Ade Irma (Indonesia)  
 Sury, Balasubramanian (India)  
 Sushchenko, Andrei (Russia)  
 Suzuki, Fumika (Canada)  
 Suzuki, Kaori (Japan)  
 Suzuki, Kohei (Japan)  
 Suzuki, Masaaki (Japan)  
 Svrtan, Dragutin (Croatia)  
 Sweatman, Winston (New Zealand)  
 Sy, Polly W. (Philippines)  
 Sy, Mamadou (Senegal)  
 Syrbu, Parascovia (Moldova)  
 Szeftel, Jeremie (France)  
 Szekelyhidi, Laszlo (Hungary)  
 Székelyhidi, László (Germany)  
 Székelyhidi, Gábor (USA)  
 Szemeridine Kepes, Anna Viktoria (Hungary)

Szomolay, Barbara (UK)  
Ta, Viet Ton (Japan)  
Ta, Thi Hoai An (Vietnam)  
Ta Cong, Son (Vietnam)  
Taam, Alexander (USA)  
Tabata, Masahisa (Japan)  
Tabata, Ryo (Japan)  
Taehee, Kim (Republic of Korea)  
Taehyeon, Son (Republic of Korea)  
Taffin, Johan (France)  
Tag, Hyungjoon (Republic of Korea)  
Taguchi, Yuichiro (Japan)  
Tak, Byungjoo (Republic of Korea)  
Tak, Hee-Joon (Republic of Korea)  
Tak, Ji Heon (Republic of Korea)  
Tak, Seong Won (Republic of Korea)  
Takaesu, Toshimitsu (Japan)  
Takamura, Hiroyuki (Japan)  
Takemura, Tomoko (Japan)  
Takesaki, Kyoko (Japan)  
Takesaki, Masamichi (Japan)  
Taki, Shingo (Japan)  
Talaue, Cheryl (Philippines)  
Talay, Denis (France)  
Talukder, Abdur Raafi (Bangladesh)  
Talukder, Mohammad Rashed (Bangladesh)  
Tamrakar, Nirula (Nepal)  
Tamura, Hiroshi (Japan)  
Tamura, Manami (Japan)  
Tanahashi, Kotaro (Japan)  
Tanaka, Ryo (Japan)  
Tanaka, Shohei (Japan)  
Tanase, Raluca (USA)  
Tanbay, Betul (Turkey)  
Tang, Lianjie (P.R. China)  
Tang, Lixin (P.R. China)  
Tang, Tai Man (P.R. China)  
Tann, Chantara (Cambodia)  
Tao, Yuanhong (P.R. China)  
Tao, Thi Huyen (Vietnam)  
Tasaka, Koji (Japan)  
Tatimakula, Vasanthi (India)  
Tayfeh Rezaie, Behruz (Iran)  
Tchapnda, Sophonie Blaise (Cameroon)  
Tchoundja, Edgar Landry (Cameroon)  
Teh, Wen Chean (Malaysia)  
Teicher, Mina (Israel)  
Teixeira, Eduardo (Brazil)  
Teleman, Constantin (USA)  
Telschow, Gerhard (Germany)  
Temirgaliyev, Nurlan (Kazakhstan)  
Tennstaedt, Tobias (Germany)  
Teo, Lee Peng (Malaysia)  
Terakawa, Hiroyuki (Japan)  
Terasawa, Yutaka (Japan)  
Teschke, Olaf (Germany)  
Teschner, Jörg (Germany)  
Testerman, Donna (Switzerland)  
Tetunashvili, Shakro (Georgia)  
Thangavelu, Geetha (India)  
Tiba, Dan (Romania)  
Tichy, Robert (Austria)  
Tikhonov, Sergey (Belarus)  
Tillmann, Ulrike (UK)  
Timimoun, Chahnaz Zakia (Algeria)  
Timotin, Dan Grigore (Romania)  
Tinoco, David (Mexico)  
Tjhin, Ferry Jaya Permana (Indonesia)  
Toan, Phan Thanh (Republic of Korea)  
Toda, Reiko (Japan)  
Toda, Yukinobu (Japan)  
Todjihounde, Leonard (Benin)  
Todorov, Dmitry (Russia)  
Toën, Bertrand (France)  
Tokunaga, Hiroo (Japan)  
Toland, John (UK)  
Tomisaki, Matsuyo (Japan)  
Tomiyama, Jun (Japan)  
Tondeur, Philippe (USA)  
Tonita, Valentin (Japan)  
Topping, Peter (UK)  
Torres-Carvajal, Luis Miguel (Ecuador)  
Tossa, Joel (Benin)  
Toure, Saliou (Ivory Coast)  
Tournes, Dominique (France)  
Tovar, Luis Manuel (Mexico)  
Tran, Thanh (Australia)  
Tran, Mai Lan (Republic of Korea)  
Tran, Thi-Thu-Huong (Vietnam)  
Tran, Tuan Nam (Vietnam)  
Tran, Van Tan (Vietnam)  
Tran Do Minh, Chau (Vietnam)  
Tran Giang, Nam (Vietnam)  
Trevino, Rodrigo (USA)  
Trillo, Juan Carlos (Spain)  
Trivedi, Saurabh (Poland)  
Trott, Michael (USA)  
Truong, Thu Huong (Vietnam)  
Truong Ha, Hai (Vietnam)  
Tsai, Yen-Lung (Taiwan)  
Tsanava, Tsira (Georgia)  
Tsandzana, Afonso Fernando (Mozambique)  
Tschinkel, Yuri (USA)  
Tsend-Ayush, Selenge (Mongolia)

Tseng, Jui-Pin (Taiwan)  
 Tsou, Sheung Tsun (UK)  
 Tsou, Judy (USA)  
 Tsuboi, Takashi (Japan)  
 Tsujii, Masato (Japan)  
 Tsukamoto, Masaki (Japan)  
 Tsushima, Ryuji (Japan)  
 Tsybakov, Alexandre (France)  
 Tugyonov, Zohid (Uzbekistan)  
 Tuisku, Petri (Finland)  
 Tumuluri, Suman Kumar (India)  
 Tungatarov, Aliaskar (Kazakhstan)  
 Tupan, Alexandru (USA)  
 Turesson, Bengt Ove (Sweden)  
 Turova, Tatyana (Sweden)  
 Turunen, Ville (Finland)  
 Uddin, M. Ashraf (Bangladesh)  
 Ueno, Kohei (Japan)  
 Uguz, Selman (Turkey)  
 Ulecia, Teresa (Spain)  
 Ullmo, Emmanuel (France)  
 Um, So Yeon (Republic of Korea)  
 Umehara, Morimichi (Japan)  
 Upadhyay, Shyamashree (India)  
 Uppal, Surindar Mohan (Kenya)  
 Urazboev, Gayrat (Uzbekistan)  
 Urbina-Romero, Wilfredo (USA)  
 Uribe, Bernardo (Colombia)  
 Urzua, Giancarlo (Chile)  
 Ushakova, Elena (Russia)  
 Ushakova, Kristina (Russia)  
 Ustinov, Alexey (Russia)  
 Utsumi, Kazuki (Japan)  
 Vaananen, Jouko (Finland)  
 Vaccon, Tristan (France)  
 Vaderlind, Paul (Sweden)  
 Vaisova, Moxira (Uzbekistan)  
 Vakil, Ravi (USA)  
 Valdez, Ferran (Mexico)  
 Valdez, Lilibeth (Philippines)  
 Valette, Alain (Switzerland)  
 Valle, Cristina (Italy)  
 Vallette, Bruno (France)  
 Van Garrel, Michel (Republic of Korea)  
 Van Koert, Otto (Netherlands)  
 Van Leer, Laurel (USA)  
 Van Wyk, Leon (South Africa)  
 Vannucci, Manila (Italy)  
 Vanualailai, Jito (Fiji)  
 Varagnolo, Michela (France)  
 Vargas, Edson (Brazil)  
 Varpanen, Harri (Finland)  
 Vashakmadze, Tamaz S. (Georgia)  
 Vasiliev, Alexander (Norway)  
 Vasilyev, Vladimir (Russia)  
 Vasserot, Eric (France)  
 Vasy, Andras (USA)  
 Vatutin, Vladimir (Russia)  
 Vatutina, Elena (Russia)  
 Vavilov, Nikolai (Russia)  
 Vavilova, Olga (Russia)  
 Vazirani, Monica (USA)  
 Vazquez, Juan-Luis (Spain)  
 Vazquez Noguera, Jose Luis (Paraguay)  
 Velasquez, Oswaldo (Peru)  
 Velich, Ilja (Slovakia)  
 Velichova, Daniela (Slovakia)  
 Velichova, Roberta (Slovakia)  
 Venkadachalam, Ramesh (India)  
 Ventura, Jade (Philippines)  
 Verbitsky, Misha (Russia)  
 Verchinine, Vladimir (France)  
 Verma, Jugal (India)  
 Verrette, Jean (USA)  
 Vezzoni, Luigi (Italy)  
 Viana, Marcelo (Brazil)  
 Vidyasagar, Mathukumalli (USA)  
 Viitasaari, Lauri (Finland)  
 Vilela, Jocelyn (Philippines)  
 Villamizar-Roa, Elder J (Colombia)  
 Villarroja Alvarez, Francisco (Sweden)  
 Vinayaka Prasad, Kerehalli (India)  
 Vinet, Luc (Canada)  
 Vinson-Rouchon, Blandine (France)  
 Violet, Bianca (Germany)  
 Virag, Balint (Canada)  
 Vittone, Francisco (Argentina)  
 Vogtmann, Karen (USA)  
 Voiculescu, Dan-Virgil (USA)  
 Voiculescu, Ioana (USA)  
 Voight, John (USA)  
 Volpert, Klaus (USA)  
 Vondrak, Anahita (USA)  
 Vondrak, Jan (USA)  
 Vonk, Jan (UK)  
 Vu, Van (USA)  
 Vulpe, Nicolae (Moldova)  
 Wadsley, Simon (UK)  
 Wahab, Abdul (Pakistan)  
 Wahyuni, Sri (Indonesia)  
 Wainwright, Martin (USA)  
 Waldschmidt, Michel (France)  
 Waldspurger, Irene (France)  
 Walker, Debbie (USA)

Walker, James (USA)  
Wallbridge, James (Japan)  
Wang, Xiaoheng (Canada)  
Wang, Baihua (P.R. China)  
Wang, Chengbo (P.R. China)  
Wang, Fang (P.R. China)  
Wang, Gehao (P.R. China)  
Wang, Jing (P.R. China)  
Wang, Kai-Rui Or Rui (P.R. China)  
Wang, Liping (P.R. China)  
Wang, Longmin (P.R. China)  
Wang, Lusheng (P.R. China)  
Wang, Menglin (P.R. China)  
Wang, Xiaojie (P.R. China)  
Wang, Xiaokun (P.R. China)  
Wang, Xiaoxia (P.R. China)  
Wang, Yi (P.R. China)  
Wang, Yuan (P.R. China)  
Wang, Joe-Sung Ho (Republic of Korea)  
Wang, Siyang (Sweden)  
Wang, Shin-Hwa (Taiwan)  
Wang, Shitao (UK)  
Wang, Lidan (USA)  
Wanka, Gert (Germany)  
Ward, Kenneth (P.R. China)  
Ward, Casey (USA)  
Wardhana, I.G.A.W. (Indonesia)  
Wee, In-Suk (Republic of Korea)  
Wei, Qingmeng (P.R. China)  
Wei, Juncheng (Hong Kong)  
Weibel, Charles (USA)  
Weisstein, Eric (USA)  
Wenger, Stefan (Switzerland)  
Werner, Wendelin (Switzerland)  
Whang, Jun Ho (USA)  
Wiegand, Sylvia (USA)  
Wigman, Igor (UK)  
Wilkinson, Amie (USA)  
Williams, Gareth (UK)  
Williams, Ryan (USA)  
Willson, Benjamin (Republic of Korea)  
Wilson, Robert (UK)  
Wilton, Henry (UK)  
Winklmeier, Monika (Colombia)  
Wise, Daniel (Canada)  
Witt, Emily (USA)  
Wolfram, Stephen (USA)  
Won, Dae Yeon (Republic of Korea)  
Won, Joonyeong (Republic of Korea)  
Won, Jun-Hee (Republic of Korea)  
Won, Junho (Republic of Korea)  
Won, Tae Kyong (Republic of Korea)  
Won, Taegyung (Republic of Korea)  
Wong, Adeline (USA)  
Woo, Changhwa (Republic of Korea)  
Woo, Chong Min (Republic of Korea)  
Woo, Gyungsoo (Republic of Korea)  
Woo, Hye Young (Republic of Korea)  
Woo, Hyenkyun (Republic of Korea)  
Woo, Moo Ha (Republic of Korea)  
Woo, Youngho (Republic of Korea)  
Woo, Younsun (Republic of Korea)  
Woo, Jae Oh (USA)  
Wood, Carol (USA)  
Wooley, Trevor (UK)  
Wornnarith, Chhit (Cambodia)  
Wrzesien, Andrzej (Poland)  
Wu, Congmin (P.R. China)  
Wu, Dan (P.R. China)  
Wu, Donglun (P.R. China)  
Wu, Faen (P.R. China)  
Wu, Fan (P.R. China)  
Wu, Qiuyi (P.R. China)  
Wu, Xiang (P.R. China)  
Wu, Zhongtao (Hong Kong)  
Wu, Pei Yuan (Taiwan)  
Xiang, Shuhuang (P.R. China)  
Xiao, Qinghua (P.R. China)  
Xie, Chunjing (P.R. China)  
Xie, Yanna (P.R. China)  
Xie, Yunli (P.R. China)  
Xie, Zhifu (USA)  
Xing, Yu (P.R. China)  
Xu, Minqi (P.R. China)  
Xu, Qingxiang (P.R. China)  
Xu, Runzhang (P.R. China)  
Xu, Shicheng (P.R. China)  
Xu, Wenxue (P.R. China)  
Xu, Xianghui (P.R. China)  
Xuan Duc Ha, Truong (Vietnam)  
Yadav, Raj Narayan (Nepal)  
Yagasaki, Tatsuhiko (Japan)  
Yamada, Hiromichi (Japan)  
Yamagishi, Manabu (Japan)  
Yamagishi, Marika (Japan)  
Yamagishi, Masakazu (Japan)  
Yamaguchi, Yoshikazu (Japan)  
Yamane, Hiroyuki (Japan)  
Yamasaki, Aiichi (Japan)  
Yan, Tao (P.R. China)  
Yanagawa, Makoto (Japan)  
Yanchevskii, Vyacheslav (Belarus)  
Yang, Jihyeon Jessie (Canada)  
Yang, Weiling (P.R. China)

Yang, Xiaoping (P.R. China)  
 Yang, Xinhua (P.R. China)  
 Yang, Xuxin (P.R. China)  
 Yang, Yanbing (P.R. China)  
 Yang, Zhihua (P.R. China)  
 Yang, Hyejin (Republic of Korea)  
 Yang, Jaewoon (Republic of Korea)  
 Yang, Jeha (Republic of Korea)  
 Yang, Jongho (Republic of Korea)  
 Yang, Juwon (Republic of Korea)  
 Yang, Min Su (Republic of Korea)  
 Yang, Minsuk (Republic of Korea)  
 Yang, Seol A (Republic of Korea)  
 Yang, Seong-Deog (Republic of Korea)  
 Yang, Seungwook (Republic of Korea)  
 Yang, Sujin (Republic of Korea)  
 Yang, Sung Jin (Republic of Korea)  
 Yang, Yoonjeong (Republic of Korea)  
 Yang, Jingxuan (UK)  
 Yang, Annie (USA)  
 Yang, David (USA)  
 Yang, Lan (USA)  
 Yao, Guowu (P.R. China)  
 Yaparova, Natalia (Russia)  
 Yatagawa, Yuri (Japan)  
 Yau, Mei-Lin (Taiwan)  
 Ye, Xiangdong (P.R. China)  
 Yechan, Kim (Republic of Korea)  
 Yeh, Li-Ming (Taiwan)  
 Yeji, Lee (Republic of Korea)  
 Yekhanin, Sergey (USA)  
 Yen, Chih-Hung (Taiwan)  
 Yendamuri, Lakshmi Naidu (India)  
 Yeo, Gwan Goo (Republic of Korea)  
 Yeo, Joo Hyun (Republic of Korea)  
 Yeo, Joohyun (Republic of Korea)  
 Yeo, Sunmin (Republic of Korea)  
 Yeojin, Jung (Republic of Korea)  
 Yeon, Mijeong (Republic of Korea)  
 Yeong Seok, Song (Republic of Korea)  
 Yeonghoon, Shin (Republic of Korea)  
 Yerim, Nam (Republic of Korea)  
 Yeshkeyev, Aibat (Kazakhstan)  
 Yhee, Donggeon (Republic of Korea)  
 Yhim, Wonbeen (Republic of Korea)  
 Yi, Zhan (P.R. China)  
 Yi, Jung Won (Republic of Korea)  
 Yi, Kang San (Republic of Korea)  
 Yi, Ki Youn (Republic of Korea)  
 Yi, Seunghun (Republic of Korea)  
 Yi, Sodam (Republic of Korea)  
 Yi, Soheun (Republic of Korea)  
 Yi, Su-Cheol (Republic of Korea)  
 Yi, Taeil (USA)  
 Yie, Ikkwon (Republic of Korea)  
 Yildirim, Cem Yalcin (Turkey)  
 Yim, Eun Jeong (Republic of Korea)  
 Yim, Joon-Hyeok (Republic of Korea)  
 Yin, Yongxue (P.R. China)  
 Yin, Mei (USA)  
 Yokoyama, Tomoo (Japan)  
 Yon, Hongyun (Republic of Korea)  
 Yoneyama, Taisuke (Japan)  
 Yong, Jiongmin (USA)  
 Yongchan, Hong (Republic of Korea)  
 Yongliang, Huo (P.R. China)  
 Yoo, Dooyoul (Republic of Korea)  
 Yoo, Eon Ok (Republic of Korea)  
 Yoo, Erlin (Republic of Korea)  
 Yoo, Hongbeom (Republic of Korea)  
 Yoo, Hwajong (Republic of Korea)  
 Yoo, Hwanchul (Republic of Korea)  
 Yoo, Hyun Jae (Republic of Korea)  
 Yoo, Hyung-Ki (Republic of Korea)  
 Yoo, Jaeseong (Republic of Korea)  
 Yoo, Jeong Young (Republic of Korea)  
 Yoo, Jinjoo (Republic of Korea)  
 Yoo, Jisang (Republic of Korea)  
 Yoo, Jooyeon (Republic of Korea)  
 Yoo, Kijo (Republic of Korea)  
 Yoo, Kyeongsik (Republic of Korea)  
 Yoo, Minha (Republic of Korea)  
 Yoo, Philsang (Republic of Korea)  
 Yoo, Sang-Bum (Republic of Korea)  
 Yoo, Se Min (Republic of Korea)  
 Yoo, Seong Hoon (Republic of Korea)  
 Yoo, Seong-Moon (Republic of Korea)  
 Yoo, Seonguk (Republic of Korea)  
 Yoo, Seungwoo (Republic of Korea)  
 Yoo, Sung-Jae (Republic of Korea)  
 Yoo, Tae-Yeon (Republic of Korea)  
 Yoo, Younggeun (Republic of Korea)  
 Yoo, Youngjun (Republic of Korea)  
 Yoo Jeong, Kwon (Republic of Korea)  
 Yoon, Byeonghoon (Republic of Korea)  
 Yoon, Changwook (Republic of Korea)  
 Yoon, Chi Won (Republic of Korea)  
 Yoon, Dae Won (Republic of Korea)  
 Yoon, Dong Hyun (Republic of Korea)  
 Yoon, Dong Sung (Republic of Korea)  
 Yoon, Doyeon (Republic of Korea)  
 Yoon, Gangjoon (Republic of Korea)  
 Yoon, Haewon (Republic of Korea)  
 Yoon, Hee Rhang (Republic of Korea)

Yoon, Hyeong Seo (Republic of Korea)  
 Yoon, Hyunsuk (Republic of Korea)  
 Yoon, Jae Min (Republic of Korea)  
 Yoon, Jae Ung (Republic of Korea)  
 Yoon, Jaehoon (Republic of Korea)  
 Yoon, Jeong Ho (Republic of Korea)  
 Yoon, Ji Hee (Republic of Korea)  
 Yoon, Jihee (Republic of Korea)  
 Yoon, Jihun (Republic of Korea)  
 Yoon, Jin Hee (Republic of Korea)  
 Yoon, Jiyoung (Republic of Korea)  
 Yoon, Jonghun (Republic of Korea)  
 Yoon, Joon Ho (Republic of Korea)  
 Yoon, Minhyeok (Republic of Korea)  
 Yoon, Miseon (Republic of Korea)  
 Yoon, Myounggho (Republic of Korea)  
 Yoon, Ryeongkyung (Republic of Korea)  
 Yoon, Seong Hyun (Republic of Korea)  
 Yoon, Seongjun (Republic of Korea)  
 Yoon, Seonhee (Republic of Korea)  
 Yoon, Soyoung (Republic of Korea)  
 Yoon, Sukyoung (Republic of Korea)  
 Yoon, Sung Hyun (Republic of Korea)  
 Yoon, Sung Min (Republic of Korea)  
 Yoon, Sunjoo (Republic of Korea)  
 Yoon, Woo Sung (Republic of Korea)  
 Yoon, Yong Ho (Republic of Korea)  
 Yoon, Youngho (Republic of Korea)  
 Yoon, Seok Ho (UK)  
 Yoon, Jeong-Mi (USA)  
 You, Geon Hei (Republic of Korea)  
 You, Je Min (Republic of Korea)  
 You, Ki Hyun (Republic of Korea)  
 You, Min Sang (Republic of Korea)  
 You, Young (USA)  
 Youn, Hyo Jung (Republic of Korea)  
 Youn, Jinju (Republic of Korea)  
 Youn, Sanggyoun (Republic of Korea)  
 Youn, Younji (Republic of Korea)  
 Young Ju, Choi (Republic of Korea)  
 Youssef, Pierre (Canada)  
 Yu, Pin (P.R. China)  
 Yu, Byeongsu (Republic of Korea)  
 Yu, Chung Hyun (Republic of Korea)  
 Yu, Heekyung (Republic of Korea)  
 Yu, Hyun (Republic of Korea)  
 Yu, Jihwan (Republic of Korea)  
 Yu, Mi-Gyoung (Republic of Korea)  
 Yu, Min-Jae (Republic of Korea)  
 Yu, Myeonghun (Republic of Korea)  
 Yu, Sanghyeon (Republic of Korea)  
 Yu, Seung Hyeon (Republic of Korea)  
 Yu, Shih-Hsien (Singapore)  
 Yu, Myungjun (USA)  
 Yuan, Liping (P.R. China)  
 Yuan, Wenjun (P.R. China)  
 Yuan, Ya-Xiang (P.R. China)  
 Yuk, Hyung Bin (Republic of Korea)  
 Yulia, Fock (France)  
 Yun, Beong In (Republic of Korea)  
 Yun, Gabjin (Republic of Korea)  
 Yun, Goang Gyun (Republic of Korea)  
 Yun, Hansung (Republic of Korea)  
 Yun, Hera (Republic of Korea)  
 Yun, Jae Heon (Republic of Korea)  
 Yun, Jaehoon (Republic of Korea)  
 Yun, Ji-Yong (Republic of Korea)  
 Yun, Juho (Republic of Korea)  
 Yun, Ki-Heon (Republic of Korea)  
 Yun, Kihyun (Republic of Korea)  
 Yun, Na Yeon (Republic of Korea)  
 Yun, Sangwoon (Republic of Korea)  
 Yun, Seok-Bae (Republic of Korea)  
 Yun, Yong Sik (Republic of Korea)  
 Yune, Seokhun (Republic of Korea)  
 Zagrebnoy, Valentin (France)  
 Zakharevich, Ilya (USA)  
 Zaki, Rachad (UAE)  
 Zakrzewski, Michał (Poland)  
 Zannier, Umberto (Italy)  
 Zarea, Sana'A A (Saudi Arabia)  
 Zariphopoulou, Thaleia (USA)  
 Zeitouni, Naomi (Israel)  
 Zeitouni, Ofer (Israel)  
 Zelenyuk, Yuliya (South Africa)  
 Zelmanov, Efim (USA)  
 Zeltser, Maria (Estonia)  
 Zeng, Liuchuan (P.R. China)  
 Zeng, Wei (P.R. China)  
 Zhainibekova, Mekhribanu (Kazakhstan)  
 Zhang, Hai (P.R. China)  
 Zhang, Hengrui (P.R. China)  
 Zhang, Jie (P.R. China)  
 Zhang, June (P.R. China)  
 Zhang, Kun (P.R. China)  
 Zhang, Lei (P.R. China)  
 Zhang, Lin (P.R. China)  
 Zhang, Mengping (P.R. China)  
 Zhang, Qin Hai (P.R. China)  
 Zhang, Ruiming (P.R. China)  
 Zhang, Ruixiang (P.R. China)  
 Zhang, Runxuan (P.R. China)  
 Zhang, Shijin (P.R. China)  
 Zhang, Ting (P.R. China)

Zhang, Tingting (P.R. China)  
Zhang, Wei (P.R. China)  
Zhang, Xiongtao (P.R. China)  
Zhang, Zhenlei (P.R. China)  
Zhang, Zhenning (P.R. China)  
Zhang, Genkai (Sweden)  
Zhang, Jun (USA)  
Zhang, Yitang (USA)  
Zhao, Mo (P.R. China)  
Zhao, Peibiao (P.R. China)  
Zhao, Quanting (P.R. China)  
Zhao, Xu (P.R. China)  
Zhao, Hongmei (Sweden)  
Zhao, Hongkai (USA)  
Zhao, Yufei (USA)  
Zhe, Hyoungbeom (Republic of Korea)  
Zheng, Xinye (P.R. China)  
Zhou, Sanming (Australia)  
Zhou, Chun Qin (P.R. China)  
Zhou, Haigang (P.R. China)  
Zhou, Jiazuo (P.R. China)  
Zhou, Jie (P.R. China)  
Zhou, Liangdong (P.R. China)  
Zhou, Min (P.R. China)  
Zhu, Baocheng (P.R. China)  
Zhu, Huilin (P.R. China)  
Zhu, Minxian (P.R. China)  
Zhu, Yuanguo (P.R. China)  
Zhu, Chengbo (Singapore)  
Zhubanysheva, Axaule (Kazakhstan)  
Zhunussova, Zhanat (Kazakhstan)  
Zhuravlev, Sergey (Russia)  
Ziegler, Günter M. (Germany)  
Ziegler, Tamar (Israel)  
Zieve, Michael (USA)  
Zlotnik, Alexander (Russia)  
Zoo, Ho Sung (Republic of Korea)  
Zuddas, Daniele (Italy)  
Zviagin, Andrei (Russia)  
Zvyagin, Victor (Russia)

## Participants by Country

Albania	3	India	164	Pakistan	17
Algeria	11	Indonesia	32	Papua New Guinea	1
Argentina	21	Iran	37	Paraguay	1
Armenia	6	Iraq	3	Peru	6
Australia	25	Ireland	1	Philippines	52
Austria	10	Israel	18	Poland	27
Azerbaijan	1	Italy	23	Polynesia	2
Bangladesh	5	Ivory Coast	4	Portugal	2
Belarus	9	Jamaica	2	Puerto Rico	3
Belgium	2	Japan	205	Republic of Korea	2654
Benin	5	Jordan	1	Romania	11
Botswana	4	Kazakhstan	29	Russian Federation	77
Brazil	54	Kenya	6	Saudi Arabia	9
Brunei Darussalam	1	Kuwait	1	Senegal	7
Bulgaria	2	Kyrgyzstan	6	Serbia	7
Burkina Faso	4	Laos	1	Singapore	10
Burundi	1	Latvia	4	Slovakia	3
Cambodia	15	Lesotho	1	South Africa	21
Cameroon	7	Lithuania	2	Spain	49
Canada	54	Macedonia	1	Sri Lanka	1
Chile	22	Madagascar	5	Sudan	2
Colombia	11	Malawi	1	Sweden	26
Congo	1	Malaysia	23	Switzerland	24
Costa Rica	2	Mali	2	Tadjikistan	3
Croatia	3	Malta	1	Taiwan	31
Cyprus	1	Mexico	24	Thailand	24
Czech Republic	8	Moldova	5	Tunisia	2
Denmark	7	Mongolia	8	Turkey	29
Ecuador	3	Montenegro	2	Uganda	4
Egypt	8	Morocco	7	Ukraine	10
Estonia	1	Mozambique	7	United Arab Emirates	9
Ethiopia	2	Myanmar	2	United Kingdom	82
Fiji	3	Namibia	3	United Republic	
Finland	24	Nepal	10	of Tanzania	1
France	121	Netherlands	6	United States	
Gabon	4	New Zealand	10	of America	430
Georgia	9	Niger	3	Uruguay	3
Germany	81	Nigeria	4	Uzbekistan	25
Ghana	2	Norway	14	Venezuela	1
Greece	3	Oman	4	Vietnam	66
Hong Kong	8	P.R. China	227	Zambia	2
Hungary	10	Pakistan	17		



## Author Index

### A

Abgrall, Rémi Vol IV, 699  
Abouzaid, Mohammed Vol II, 815  
Agol, Ian Vol I, 141  
Alekseev, Anton Vol III, 983  
Andruskiewitsch, Nicolás Vol II, 119  
Ardakov, Konstantin Vol III, 1  
Arora, Sanjeev Vol I, 81  
Arthur, James Vol I, 171  
Ayoub, Joseph Vol II, 1103

### B

Bader, Uri Vol III, 71  
Baladi, Viviane Vol III, 525  
Ball, Deborah Vol I, 739  
Bao, Weizhu Vol IV, 971  
Barak, Boaz Vol IV, 509  
Barton, Bill Vol I, 739  
Behrend, Kai Vol II, 593  
Belolipetsky, Mikhail Vol II, 839  
Benkart, Georgia Vol I, 633  
Benoist, Yves Vol III, 11  
Bhargava, Manjul Vol I, 657  
Biquard, Olivier Vol II, 855  
Bodineau, Thierry Vol III, 721  
Borodin, Alexei Vol I, 199  
Bourguignon, Jean-Pierre Vol I, 787  
Braides, Andrea Vol IV, 997  
Braverman, Mark Vol IV, 535  
Breuillard, Emmanuel Vol III, 27  
Brezzi, Franco Vol I, 217  
Brown, Francis Vol II, 297  
Brundan, Jonathan Vol III, 51  
Buffa, Annalisa Vol IV, 727  
Bulatov, Andrei A. Vol IV, 561

### C

Cancès, Eric Vol IV, 1017  
Candes, Emmanuel Vol I, 235  
Cederbaum, Carla Vol I, 755  
Chatterjee, Sourav Vol IV, 1  
Chatzidakis, Zoé Vol II, 3  
Chierchia, Luigi Vol III, 547  
Christodoulou, Demetrios Vol I, 259  
Chudnovsky, Maria Vol IV, 291  
Chuzhoy, Julia Vol IV, 585  
Ciocan-Fontanine, Ionuț Vol II, 617  
Codá Marques, Fernando Vol I, 283  
Colom, Miguel Vol IV, 1061  
Colli, Edurado Vol I, 799  
Conlon, David Vol IV, 303  
Cortiñas, Guillermo Vol II, 145  
Corwin, Ivan Vol III, 1007  
Crovisier, Sylvain Vol III, 571

### D

Dafermos, Mihalis Vol III, 747  
Daubechies, Ingrid Vol I, 787  
Davenport, James H. Vol I, 743  
Daskalopoulos, Panagiota Vol III, 773  
Dickenstein, Alicia Vol I, 755  
Duplantier, Bertrand Vol III, 1035

### E

Efendiev, Yalchin Vol IV, 749  
Eisenbrand, Friedrich Vol IV, 829  
Emerton, Matthew Vol II, 321  
Entov, Michael Vol II, 1149  
Erdős, László Vol III, 213  
Eynard, Bertrand Vol III, 1063

### F

Facciolo, Gabriele Vol IV, 1061

- |                         |                         |                        |               |
|-------------------------|-------------------------|------------------------|---------------|
| Fang, Fuquan            | Vol II, 869             | <b>J</b>               |               |
| Farah, Ilijas           | Vol II, 17              | Jerrard, Robert L.     | Vol III, 789  |
| Farb, Benson            | Vol II, 1175            |                        |               |
| Fathi, Albert           | Vol III, 597            | <b>K</b>               |               |
| Faure, Frédéric         | Vol III, 683            | Kahn, Jeremy           | Vol II, 899   |
| Fedkiw, Ron             | Vol I, 90               | Kang, Seok-Jin         | Vol II, 181   |
| Figalli, Alessio        | Vol III, 237            | Kassabov, Martin       | Vol II, 205   |
| Fock, Vladimir V.       | Vol III, 1087           | Katz, Nets Hawk        | Vol III, 303  |
| Fox, Jacob              | Vol IV, 329             | Kedem, Rinat           | Vol III, 1141 |
| Frieze, Alan            | Vol I, 311              | Keys, Kevin L.         | Vol IV, 95    |
| Furman, Alex            | Vol III, 71             | Kharlampovich, Olga    | Vol II, 225   |
|                         |                         | Khot, Subhash          | Vol I, 711    |
| <b>G</b>                |                         | Kim, Bumsig            | Vol II, 617   |
| Galatius, Søren         | Vol II, 1199            | Kim, Byunghan          | Vol II, 43    |
| Gallagher, Isabelle     | Vol III, 721            | Kim, Myung-Hwan        | Vol I, 787    |
| Gan, Wee Teck           | Vol II, 345             | Klainerman, Sergiu     | Vol III, 895  |
| Gentry, Craig           | Vol IV, 609             | Kleshchev, Alexander   | Vol III, 97   |
| Gerasimov, Anton A.     | Vol III, 1097           | Kolesnikov, Alexei     | Vol II, 43    |
| Ghys, Étienne           | Vol I, 47, Vol IV, 1187 | Kollár, János          | Vol I, 395    |
| Gilbert, Anna C.        | Vol IV, 1043            | Krivelevich, Michael   | Vol IV, 355   |
| Goldston, D. A.         | Vol II, 421             | Kumagai, Takashi       | Vol IV, 75    |
| Goodrick, John          | Vol II, 43              | Kuznetsov, Alexander   | Vol II, 637   |
| Grünberg, David         | Vol I, 755              | Kühn, Daniela          | Vol IV, 381   |
| Green, Ben              | Vol I, 341              |                        |               |
| Green, Mark L.          | Vol I, 114              | <b>L</b>               |               |
| Greuel, Gert-Martin     | Vol I, 755              | Laba, Izabella         | Vol III, 315  |
| Grimmett, Geoffrey R.   | Vol IV, 25              | Laborde, Jean-Marie    | Vol I, 739    |
| Gross, Benedict H.      | Vol I, 56               | Lange, Kenneth         | Vol IV, 95    |
| Gross, Mark             | Vol II, 725             | Laurent, Monique       | Vol IV, 843   |
| Guralnick, Robert       | Vol II, 165             | Lebrun, Marc           | Vol IV, 1061  |
|                         |                         | Ledoux, Michel         | Vol IV, 117   |
| <b>H</b>                |                         | Lee, Ki-Ahm            | Vol III, 811  |
| Ha, Seung-Yeal          | Vol III, 1123           | Le Gall, Jean-François | Vol I, 421    |
| Hairer, Martin          | Vol I, 685, Vol IV, 49  | Lewis, Adrian S.       | Vol IV, 871   |
| Han, Qi                 | Vol IV, 1217            | Li, Tao                | Vol II, 1247  |
| Harris, Michael         | Vol II, 369             | Lin, Chang-Shou        | Vol III, 331  |
| Helfgott, Harald Andrés | Vol II, 393             | Loeser, François       | Vol II, 61    |
| Hill, Michael A.        | Vol II, 1221            | Loos, Andreas          | Vol IV, 1203  |
| Hingston, Nancy         | Vol II, 883             | Lyons, Russell         | Vol IV, 137   |
| Hirachi, Kengo          | Vol III, 257            | Lyons, Terry           | Vol IV, 163   |
| Hopkins, Michael J.     | Vol II, 1221            | Lyubich, Mikhail       | Vol I, 443    |
| Hwang, Jun-Muk          | Vol I, 369              |                        |               |
| Hytönen, Tuomas         | Vol III, 279            | <b>M</b>               |               |
|                         |                         | Malchiodi, Andrea      | Vol III, 345  |

- Marcus, Adam W. Vol III, 363  
 Marklof, Jens Vol III, 623  
 Markovic, Vladimir Vol II, 899  
 Maulik, Davesh Vol II, 663  
 McCann, Robert J. Vol III, 835  
 McMullen, Curtis T. Vol I, 73  
 Merle, Frank Vol I, 475  
 Mochizuki, Takuro Vol I, 499  
 Montalbán, Antonio Vol II, 81  
 Moreira, Carlos Gustavo T. de A. Vol III, 647  
 Morel, Jean-Michel Vol I, 90, Vol IV, 1061  
 Mustăță, Mircea Vol II, 675  
 Myasnikov, Alexei Vol II, 225
- N**
- Naber, Aaron Vol II, 913  
 Nemenzo, Fidel R. Vol I, 799  
 Neves, André Vol II, 941  
 Niethammer, Barbara Vol IV, 1087  
 Noy, Marc Vol IV, 407
- O**
- O'Donnell, Ryan Vol IV, 633  
 Oguiso, Keiji Vol II, 697  
 Olshanski, Grigori Vol IV, 431  
 Olver, Peter J. Vol I, 773  
 Osinga, Hinke M. Vol IV, 1101  
 Osthus, Deryk Vol IV, 381  
 Ostrik, Victor Vol III, 121  
 Ostrover, Yaron Vol II, 961
- P**
- Péché, Sandrine Vol III, 1159  
 Pach, János Vol IV, 455  
 Paenza, Adrián Vol I, 729  
 Park, Hyungju Vol I, 755  
 Park, Youngah Vol I, 787  
 Perthame, Benoît Vol I, 529  
 Pierazzo, Nicola Vol IV, 1061  
 Pintz, J. Vol II, 421  
 Pinzari, Gabriella Vol III, 547  
 Pipher, Jill Vol III, 387  
 Pila, Jonathan Vol I, 547  
 Pollicott, Mark Vol III, 661
- Polthier, Konrad Vol I, 799
- R**
- Rais, Martin Vol IV, 1061  
 Raphaël, Pierre Vol III, 849  
 Rapinchuk, Andrei S. Vol II, 249  
 Ravenel, Douglas C. Vol II, 1221  
 Reddy, B. Daya Vol IV, 1125  
 Ressayre, Nicolas Vol III, 165  
 Rezk, Charles Vol II, 1127  
 Ringström, Hans Vol II, 985  
 Rödl, Vojtěch Vol I, 573  
 Robbiano, Luc Vol IV, 897  
 Rodnianski, Igor Vol III, 895  
 Rognes, John Vol II, 1261  
 Rouchon, Pierre Vol IV, 921  
 Rousseau, Christiane Vol I, 799  
 Rudnick, Zeev Vol II, 445  
 Rémy, Bertrand Vol III, 143
- S**
- Saint-Raymond, Laure Vol III, 721  
 Sanders, Tom Vol III, 401  
 Sapiro, Guillermo Vol I, 90  
 Schick, Thomas Vol II, 1287  
 Schlag, Wilhelm Vol III, 425  
 Scholze, Peter Vol II, 463  
 Seiringer, Robert Vol III, 1175  
 Seppäläinen, Timo Vol IV, 185  
 Serganova, Vera Vol I, 603  
 Sesum, Natasa Vol II, 1003  
 Shatashvili, Samson L. Vol III, 1195  
 Shen, Weixiao Vol III, 699  
 Shu, Chi-Wang Vol I, 90, Vol IV, 767  
 Sidoravicius, Vladas Vol IV, 199  
 Siebert, Bernd Vol II, 725  
 Siegmund-Schultze, Reinhard Vol IV, 1231  
 Silvestre, Luis Vol III, 873  
 Siu, Man Keung Vol I, 739  
 Smith, Karen E. Vol II, 273  
 Sodin, Sasha Vol III, 451  
 Solecki, Sławomir Vol II, 105  
 Speicher, Roland Vol III, 477

- |                         |               |                        |              |
|-------------------------|---------------|------------------------|--------------|
| Spielman, Daniel A.     | Vol III, 363  | Virág, Bálint          | Vol IV, 247  |
| Srivastava, Nikhil      | Vol III, 363  | Vu, Van H.             | Vol IV, 489  |
| Steger, Angelika        | Vol IV, 475   |                        |              |
| Steurer, David          | Vol IV, 509   | <b>W</b>               |              |
| Strien, Sebastian van   | Vol III, 699  | Wainwright, Martin J.  | Vol IV, 273  |
| Stuart, Andrew M.       | Vol IV, 1145  | Waldspurger, J.-L.     | Vol II, 489  |
| Székelyhidi Jr., László | Vol III, 503  | Wang, Yi-Qing          | Vol IV, 1061 |
| Szeftel, Jérémie        | Vol III, 895  | Wei, Juncheng          | Vol III, 941 |
| Székelyhidi, Gábor      | Vol II, 1019  | Wenger, Stefan         | Vol II, 1051 |
|                         |               | Williams, Ryan         | Vol IV, 659  |
| <b>T</b>                |               | Wise, Daniel T.        | Vol II, 1077 |
| Talay, Denis            | Vol IV, 787   | Wooley, Trevor D.      | Vol II, 507  |
| Teleman, Constantin     | Vol II, 1311  |                        |              |
| Teschner, Jörg          | Vol III, 1223 | <b>Y</b>               |              |
| Toda, Yukinobu          | Vol II, 747   | Yekhanin, Sergey       | Vol IV, 683  |
| Topping, Peter M.       | Vol II, 1035  | Yong, Jiongmin         | Vol IV, 947  |
| Tournès, Dominique      | Vol IV, 1255  | Yu, Shih-Hsien         | Vol III, 965 |
| Toën, Bertrand          | Vol II, 771   | Yuan, Ya-xiang         | Vol IV, 807  |
| Tsujii, Masato          | Vol III, 683  | Yıldırım, C. Y.        | Vol II, 421  |
| Tsybakov, Alexandre B.  | Vol IV, 225   | Yin, Wotao             | Vol I, 90    |
|                         |               |                        |              |
| <b>V</b>                |               | <b>Z</b>               |              |
| Varagnolo, Michela      | Vol III, 191  | Zannier, Umberto       | Vol II, 533  |
| Vasserot, Eric          | Vol III, 191  | Zariphopoulou, Thaleia | Vol IV, 1163 |
| Vasy, András            | Vol III, 915  | Zeitouni, Ofer         | Vol I, 65    |
| Verbitsky, Misha        | Vol II, 795   | Zhang, Yitang          | Vol II, 559  |
| Villani, Cédric         | Vol I, 755    | Ziegler, Günter M.     | Vol IV, 1203 |
|                         |               | Ziegler, Tamar         | Vol II, 571  |